Monte Carlo Inference

Part III Course, Lent 2010

Revision Notes

Daniel Guetta

guetta@cantab.net

Based on lectures by R. Gramacy, Lent 2010

Generally Useful R Code

Distributions

- Normal
  - `rnorm(n,m,s)` generates `n` data points from $N(\mathtt{m}, \mathtt{s}^2)$
  - `pnorm(x,m,s)`= $\mathbb{P}(N(m,s^2) \leq x)$
  - `dnorm(x,m,s)`= $f_X(x)$ if $X \sim N(m,s^2)$. If the argument `log = TRUE` is specified, $\log\left(f_X(x)\right)$ is returned.
- Gamma
  - `rgamma(n,a, rate=b)` generates `n` data points from $\Gamma(a,b)$
  - Others as above
- Uniform
  - `runif(n,a,b)` generates `n` data points from $U(a, b)$
  - Others as above
- Students' t
  - `rt(n,df=1)` generates `n` data points from a t disrtibution with 1 degree of freedom (ie: a Cauchy distribution).
  - Others as above
- Chi-squared
  - `rchisq(n,df=5)` generates `n` data points from a chi-squared distribution with 5 degrees of freedom.
  - Others as above

Graphics

- `par()` is used to modify settings in the current graphics environment. Possible settings:
  - `mfrow = c(rows, columns)` sets the numbers of rows and columns of graphs
  - `cex.main=s, cex.lab=s, cex.axis=s` respectively set the title, axes labels and axes font sizes to `s`
- `plot(x, y)` creates a new plot area and plots two vectors against each other. Possible settings
  - `type = "l"` plots a line (by default, individual points are plotted)

  - `lwd = 2` makes the line (or points) thicker
  - `col = "blue"` makes the line blue
  - `ylim = c(miny,maxy)` sets the limits on the axes [could also use `range` function below]
  - `lty = 2` or `3` give different kinds of dashing.
  - `main = "title"` sets the title to "main"
- `lines(x, y)` adds a line to an existing plot. Possible settings as for `plot`.
- `points(x, y)` adds points to an existing plot. Possible settings as for `plot`.
- `abline(v=1)` adds a vertical line at $x = 1$, and `abline(h=1)` adds a horizontal line at $y = 1$. Possible settings as for `plot`.
- `hist(data)` plots a histogram of the input data. Possible settings as for `plot`, as well as
  - `freq = FALSE` normalises the area of the histogram to 1. The default, `freq = TRUE`, simply plots counts.
- `text(x, y,"text")` adds text to a plot. `col` and `cex` are possible options for the colour and font size. Note that `y = 0` puts the text straight on the x-axis.
- `d <- dev.cur` sets a handle for the current graphics window. A new graphics window can then be created using `x11()`, and the old graphics window can be returned to using `dev.set(d)`.

Data handling

- `plot(density(x))` plots the empirical density of `x`. Very useful if trying to pictorially show the efficiency of various estimators; simulate the value of interest a number of times, and plot the density each time to see what it looks like.
- `data.frame(a=aData, b=bData)` creates a data frame with variable `a` taking values in `aData`, etc... The data can consist of vectors; R will add suffices accordingly.
- `summary(dataFrame)` gives all kinds of useful statistics on the items in the data frame.

- `boxplot(dataFrame)` plots a box-and-whisker plot showing the spread of data in each of the variables in the data frame.
- `quantiles(x, v)` finds the `v` quantile of the data `x`. For example, if `v = 0.5`, it finds the median. Extremely useful in approximating confidence intervals.

Vectors, matrices, etc…

- Vector numbering starts from 1, not 0.
- `x <- seq(a,b,length=c)` fills `x` with `c` values between `a` and `b`. Another possible argument is `by=c`, which increments by `c` each time.
- `rep(a,b)` produces a vector containing `b` instances of `a`. `a` can be set to `NA` to create an empty array.
- `range(vector)` gives a column containing the lowest and highest value of a vector. Useful when setting the limits of an axis.
- `rev(x)` reverses the input vector.
- `length(x)` gives the number of items in a vector.
- `x>2` returns a vector of the same length as `x` containing `TRUE` wherever the condition is met and `FALSE` otherwise. Can be used in a number of ways
  - `x[x > 2]` returns a vector containing all items in `x` greater than `2`.
  - `mean[x > 2]` assigns `1` to each component of `x` greater than `2`, and `0` to others, and finds the mean of these numbers.
- `apply(matrix, index, <operation>)` applies the operation `<operation>` to rows (if `index = 1`) or columns (if `index = 2`) of `matrix`.
- `matrix(data, nrow, ncol, byrow = FALSE)` creates a matrix and fills it with `data`, column by column (unless `byrow = TRUE`).
- `t(m)` transposes the matrix `m`.
- `x %*% y` performs the matrix multiplication `xy`.
- `prod(x, na.rm=FALSE)` finds the product of the elements in the vector `x`. If `na.rm = TRUE`, missing values are removed.
- `Numeric(0)` creates an empty vector.

Control blocks

- `for (i in 1:5) { }` loops the item in the braces 5 times.

Other

- `x <- readline("promt")` prompts the user and inserts their response as a string into `x`.
- `paste(string1,string2,sep=".")` concatenates the strings in the argument, adding . between each string.
- `help(functionName)` gives help.

Random Number Generation

Assume we have an infinite supply of random numbers U that are distributed randomly over [0,1]. We discuss general methods for using these numbers to generate random numbers from other, more sophisticated distributions.

Method of Inversion

If X has a continuous CDF F, then $U = F(X) \sim U(0,1)$[1], and so $X = F^{-1}(U)$. This is the basis of the method of inversion:

> **Method of inversion**: To generate a sample x from a distribution with CDF F
>
> 1. Simulate $u \sim U(0,1)$
> 2. Set $x = F^{-1}(u)$

Proof: Consider that

$$\mathbb{P}\left(X \le x\right) = \mathbb{P}\left(F^{-1}(U) \le x\right)$$

since F is a strictly increasing[2], monotonic and continuous function of x, we can write

$$= \mathbb{P}(U \le F(x))$$

since $U \sim U(0, 1)$, $P(U \le ?) = ?$, and so

$$= F(x)$$

as desired. ■

Notes:

- The inverse CDF relationship exists between any two continuous random variables; $X = F_X^{-1}\left(F_Y(Y)\right)$. This method is therefore often used with $Y =$ normal distribution.
- Even when F^{-1} exists in closed form, it may be more computationally intensive to calculate than some alternative methods for generating random numbers.

[1] To see why, consider that $\mathbb{P}\left(F(P) \le p\right) = \mathbb{P}\left(P \le F^{-1}(p)\right) = F\left(F^{-1}(p)\right) = p$.

[2] For functions that are not *strictly* increasing, use a generalised inverse.

- When F^{-1} does not exist in closed form, the method of inversion can still be used by solving $F_X(x) - u = 0$ numerically.

This method can also be used for discrete distributions

Method of inversion (discrete distributions): Consider a distribution with levels m_j and $\mathbb{P}(m_j) = p_j$. The CDF is

$$F_j = \sum_{k=1}^{j} p_k$$

so that $p_j = F_j - F_{j-1}$. To generate a sample x from this distribution

1. Simulate $u \sim U(0,1)$
2. Set $x = m_j$ if $F_{j-1} < u < F_j$

Proof: Consider that

$$\mathbb{P}\left(X = j\right) = \mathbb{P}\left(F_{j-1} < U \leq F_j\right)$$

since $U \sim U(0, 1)$, $P(a < U \leq b) = b - a$, and so

$$= P_j - P_{j-1}$$
$$= p_j$$

as desired. ■

In practice, the algorithm above can be applied by first simulating u and then calculating each $\mathbb{P}(m_j)$ for each j, until the sum of all probabilities calculated is greater or equal to u.

This method can also be used for mixture densities

Method of inversion (mixture distributions) Consider a distribution with mixture density

$$f = \sum_{i=1}^{k} w_i f_i$$

where the f_i are PDFs, $w_i > 0$ and $\sum w_i = 1$. To sample from this distribution:

1. Choose $I = i$ with probability w_i
2. Sample from f_I

Rejection Sampling

Rejection sampling: Suppose it is difficult to sample directly from a density f, but that we have a **majorising** or **envelope** density g from which it is easy to sample and a constant $M \in [1,\infty]$ such that

$$f(x) \leq Mg(x) \qquad \forall x \in \mathbb{R}$$

Or alternatively

$$\sup_{x\in\mathbb{R}} \left(\frac{f(x)}{g(x)} \right) < \infty$$

To generate a sample x from the distribution with desntiy g:

1. Generate $y \sim g$ and independent $u \sim U(0,1)$
2. If $u > \dfrac{f(y)}{Mg(y)}$, return to step 1.
3. Return $x = y$.

Proof: We have

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(Y \leq x \mid Y \text{ is accepted}) \\ &= \frac{\mathbb{P}(Y \leq x \text{ and } Y \text{ is accepted})}{\mathbb{P}(Y \text{ is accepted})}\end{aligned}$$

But we know that

$$\begin{aligned}\mathbb{P}(Y \text{ is accepted}) &= \mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)} \right) \\ &= \int_{-\infty}^{\infty} \mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)} \mid Y = y \right) \\ &\qquad\qquad \mathbb{P}(Y = y)\ \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}\left(U \leq \frac{f(y)}{Mg(y)} \right) g(y)\ \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \int_{0}^{\frac{f(y)}{Mg(y)}} g(y)\ \mathrm{d}u\ \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y)\ \mathrm{d}y \\ &= \frac{1}{M} \int_{-\infty}^{\infty} f(y)\ \mathrm{d}y\end{aligned}$$

Similarly

$$\mathbb{P}\left(\begin{matrix} Y \leq x \text{ and} \\ Y \text{ is accepted} \end{matrix}\right) = \int_{-\infty}^{x} \mathbb{P}\left[U \leq \frac{f(y)}{Mg(y)}\right] g(y)\ \mathrm{d}y$$
$$= \frac{1}{M}\int_{-\infty}^{x} f(y)\ \mathrm{d}y$$

as desired. ■

Note that the denominator in the final expression normalises the function, so we only need to know f up to a multiplicative constant.

Note that the proportion of "accepted" trials is given by

$$\pi = \frac{\text{Area under } f}{\text{Area under } g} = \frac{1}{M}$$

So we want to choose M as small as possible subject to $M \geq f(x)/g(x)$. Since M always needs to be bigger than the RHS, the smallest value M can take is the supremum of the RHS. So the optimal M is given by

$$M^* = \sup_{x \in \mathbb{R}} \left(\frac{f(x)}{g(x)}\right)$$

Sometimes, we may choose a *family* of enveloping functions characterised by a parameter β, say. In that case, minimizing M^* with respect to β provides the best choice of function.

EXAMPLE: Let $\alpha \in (1,\infty)$, and consider sampling from $\Gamma(\alpha,1)$

$$f(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} \qquad x \in (0,\infty)$$

We may be tempted to choose a majoring function $g(x) = x^{\alpha-1}$, but that's no good, because it is not bounded, and therefore is not a distribution. Instead, we choose the family

$$g_\beta(x) = \beta e^{-\beta x}$$

and we note that

$$\sup_{x \in \mathbb{R}^+} \left(\frac{f(x)}{g_\beta(x)}\right) = \frac{\left(\frac{\alpha-1}{1-\beta}\right)^{\alpha-1} e^{-(\alpha-1)}}{\beta\Gamma(\alpha)} < \infty$$

So g does indeed work as a majoring function. We also note that the minimum of this supremum is attained at $\beta^* = \frac{1}{\alpha}$.

EXAMPLE: Consider sampling from the distribution

$$f(x) = e^{-(x+1)} + (e-1)e^{-ex} \qquad x \in (0,\infty)$$

In this case, choosing $g(x) = e^{-x}$ gives

$$M = \sup_{x\in\mathbb{R}^+} \frac{f(x)}{g(x)} = \sup_{x\in\mathbb{R}^+} \left(e^{-1} + (e-1)e^{(1-e)x}\right) = e^{-1} + e - 1$$

If $f(x)$ is difficult to compute, then the test in step 2 can be slow to evaluate. We can use simple **squeeze functions** that bracket f.

- For example, a squeeze function $s(y)$ *below* f has $s(y) \le f(y)$.
- In a given trial, we evaluate $s(y)$ before $f(y)$. If $u \le \frac{s(y)}{Mg(y)} \Rightarrow u \le \frac{f(y)}{Mg(y)}$ and so we can automatically accept the point without computng $f(y)$.
- Clearly, we want the area under s to be as large as possible subject to $s(y) \le f(y)$, so that we have a larger chance of accepting points directly.

Ratio of Uniforms

Theorem: Consider a distribution

$$f_X(x) = \frac{h(x)}{\int_{-\infty}^{\infty} h(x)\,\mathrm{d}x} \qquad \text{and } \int_{-\infty}^{\infty} h(x)\,\mathrm{d}x < \infty$$

Then for random variables $(U,\ V)$ sampled uniformly distributed over the region

$$C_h = \left\{(u,v) : 0 \le u \le \sqrt{h(v\,/\,u)}\right\}$$

The random variable $X = V\,/\,U$ has PDF $f_X(x)$.

Proof: Let $Z = (U,\ V)$ be uniformly distributed on C_h.

$$f_Z(u,v) = \frac{1}{A}\mathbb{I}_{\left\{0\le u\le\sqrt{h(v/u)}\right\}}$$

where A is the area of C_h. Now, let's apply the variable transformation:

$$X = V\,/\,U \qquad\qquad Y = U$$
$$\Rightarrow \qquad U = Y \qquad\qquad V = XY$$

The transformation from $Z = (U,\ V)$ to $W = (X,\ Y)$ therefore has Jacobian

$$J = \begin{vmatrix} \partial u\,/\,\partial x & \partial u\,/\,\partial y \\ \partial v\,/\,\partial x & \partial v\,/\,\partial y \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ y & x \end{vmatrix} = -y$$

The joint density of $W = (X,\ Y)$ is then given by

$$\begin{aligned} f_W(x,y) &= f_Z\Big(u(x,y), v(x,y)\Big)\Big|J\Big| \\ &= f_Z(y, xy)y \\ &= \frac{y}{A}\mathbb{I}_{\left\{0\le y\le\sqrt{h(x)}\right\}} \end{aligned}$$

The marginal density of X is given by

$$f_X(x) = \int_0^{\sqrt{h(x)}} \frac{y}{A}\mathrm{d}y = \left[\frac{y^2}{2A}\right]_0^{\sqrt{h(x)}} = \frac{h(x)}{2A}$$

However, f_X is a density, so

$$\begin{aligned} 1 = \int_{-\infty}^{\infty} f_X(x)\ \mathrm{d}x &= \int_{-\infty}^{\infty} \frac{h(x)}{2A}\ \mathrm{d}x \\ \Rightarrow \int_{-\infty}^{\infty} h(x)\ \mathrm{d}x &= 2A \end{aligned}$$

So the marginal density of $X = V/U$ is given by

$$f_X(x) = \frac{h(x)}{\displaystyle\int_{-\infty}^{\infty} h(x)\ \mathrm{d}x}$$

as desired. ■

This method is most useful if C_h can be contained in a rectangle. We therefore develop the following theorem:

Theorem: If $h(x)$ and $x^2h(x)$ are bounded, then

$$C_h \subseteq \Big[0, a\Big]\times\Big[b_-, b_+\Big]$$

where

$$\begin{aligned} a &= \sup_{x\in\mathbb{R}}\sqrt{h(x)} \\ b_- &= -\sup_{x\le 0}\sqrt{x^2h(x)} \\ b_+ &= \sup_{x\ge 0}\sqrt{x^2h(x)} \end{aligned}$$

Proof: If $(u,v)\in C_h$, then

$$0 < u < \sqrt{h(v\,/\,u)} \le \sqrt{\sup_{x\in\mathbb{R}} h(x)} = \sup_{x\in\mathbb{R}}\sqrt{h(x)} = a$$

so u is indeed bounded as predicted.

For the v coordinate, we check two cases:

- If $\boxed{v \ge 0}$, then we substitute $t = \frac{v}{u} \ge 0$ into $0 \le u^2 \le h(\frac{v}{u})$ and get

$$0 \le v^2 \le t^2 h(t) \le \sup_{t \ge 0} t^2 h(t) = b_+^2$$

Since v is positive, this implies $v < b_+$.

- If $\boxed{v < 0}$, a similar trick gives $v^2 < b_-^2$, and implies that $v > b_-$

We have therefore derived both bounds. ■

In the case were these conditions are met, we have a new method:

Ratio of uniforms: To generate a sample x from a distribution with CDF F

1. Simulate $u_1, u_2 \sim U(0,1)$
2. Let
$$u = a u_1 \qquad v = b_- + (b_+ - b_-) u_2$$
This ensures that $(u,\ v)$ is uniformly chosen from $[0,a] \times [b_-, b_+]$.
3. If $u^2 \le h(\frac{v}{u}) \Rightarrow (u,v) \in C_h$, then return $x = v/u$. Otherwise, return to step 1.

Analogously to rejection sampling, it may be possible to find sets $C_- \subseteq C_h \subseteq C_+$ for which it is easier to determine the membership of $(u,\ v)$. [$C_+ = [0,a] \times [b_-, b_+]$ is an example of such an "upper bounding" set].

EXAMPLE: Consider sampling from the Cauchy distribution

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Set

$$h(x) = \frac{1}{(1+x^2)} \qquad \text{clearly, } \int_{-\infty}^{\infty} h(x)\ \mathrm{d}x = \pi$$

In this case, the exact form of C_h is

$$C_h = \left\{(u,v) : 0 \le u, u^2 + v^2 \le 1\right\}$$

$h(x)$ and $x^2 h(x)$ are both bounded [to see why, consider what happens as $x \to \infty$]. Now, we calculate

$$a = \sup_{x\in\mathbb{R}^+} \sqrt{h(x)} = 1$$
$$b_- = -\sup_{x\le 0} \sqrt{x^2 h(x)} = -1$$
$$b_+ = -\sup_{x\ge 0} \sqrt{x^2 h(x)} = 1$$

As such, our "approximate" region is

$$C_h = \left\{(u,v) : u \in [0,1], v \in [-1,1]\right\}$$

Note that this method can also be used with $f(x) = h(x)$.

EXAMPLE: Consider sampling from the distribution

$$f(x) = e^{-(x+1)} + (e-1)e^{-ex} \qquad x \in (0,\infty)$$

We use $h(x) = f(x)$. Now, the function is monotonously decreasing, and so reaches its maximum at $x = 0$ [**Note**: it is *important* to remember that this can be the case – differentiating and setting to 0 in such a case won't help at all]. So

$$a = \sup_{x\in\mathbb{R}^+} \sqrt{f(x)} = \sqrt{e^{-1} + e - 1}$$

Similarly, the only non-positive value the function can have is 0, and so

$$b_- = -\sup_{x\le 0} \sqrt{x^2 f(x)} = 0$$

Furthermore, by the triangle inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, and so

$$b_+ = \sup_{x\ge 0} \sqrt{x^2 f(x)} \le \sup_{x\ge 0} x\left\{e^{-\frac{x}{2}-\frac{1}{2}} + (e-1)^{\frac{1}{2}} e^{-\frac{ex}{2}}\right\} = 2e^{-\frac{3}{2}} + \tfrac{2}{e^2}(e-1)^{\frac{1}{2}}$$

Composition

Let $\left\{f(x;\theta) : \theta \in \Theta\right\}$ denote a family of densities, and let $p(\theta)$ denote a density (or mass) over Θ. Then the density

$$f(x) = \begin{cases} \int_\Theta f(x;\theta)p(\theta)\,\mathrm{d}\theta & \theta \text{ continuous} \\ \sum_{\theta\in\Theta} f(x;\theta)p(\theta) & \theta \text{ discrete} \end{cases}$$

is a **mixture density**. We can sample x from $f(x)$ as follows

- Generating θ with density/mass function $p(\theta)$
- Generating x with density $f(x;\theta)$

This is called the **method of composition**.

EXAMPLE: Consider sampling from the non-central chi-squared distribution, with $n \in \mathbb{N}$ degrees of freedom and non-centrality parameter $\lambda \in (0,\infty)$, denoted $\chi^2_n(\lambda)$

$$f(x;n,\lambda) = \sum_{r=0}^{\infty} \frac{e^{-\lambda/2}\lambda^r}{2^r r!} \frac{x^{\frac{n}{2}+r-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}+r}\Gamma\left(\frac{n}{2}+r\right)} \qquad x \in (0,\infty)$$

Notice also that the distribution of the Poisson and chi-squared distributions are

$$f_{\mathrm{Po}(\lambda)}\left(n;\lambda\right) = \frac{\lambda^n e^{-\lambda}}{n!} \qquad f_{\chi^2_k}(x;k) = \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}$$

Clearly, therefore, we can generate the above by first generating $R \sim \mathrm{Po}(\frac{\lambda}{2})$, and then generating $X \sim \chi^2_{n+2R}$.

EXAMPLE: Consider sampling from the distribution

$$f(x) = e^{-(x+1)} + (e-1)e^{-ex} \qquad x \in (0,\infty)$$

We note that this can be re-written as

$$f(x) = \frac{1}{e}e^{-x} + \frac{e-1}{e}\cdot e \cdot e^{-ex}$$

This is a mixture of exponentials, with

$$f(x;\theta) = \theta e^{-\theta x} \qquad p(\theta) = \begin{cases} 1/e & \theta = 1 \\ (e-1)/e & \theta = e \end{cases}$$

We therefore generate two random variables, U_1 and U_2, and

- If $U_1 > (1/e)$, return $-\log(U_2)$
- Else, return $-e^{-1}\log(U_2)$

Specific distributions

We now go through a number of commonly used algorithms for standard discrete and continuous distributions

- **Normal distribution** $\boxed{Z \sim N(0,1)}$
  - The **Box-Muller Method** uses $U_1, U_2 \sim U(0,1)$ and generates

    $$Z_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2) \qquad Z_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

    This can be proven as follows
    - The joint distribution of U_1 and U_2 is $f_{U_1,U_2}(u_1,u_2) = \mathbb{I}_{\{(u_1,u_2)\in(0,1)^2\}}$.

- $U_1 = \exp\left\{-\frac{1}{2}(Z_1^2 + Z_2^2)\right\}$ and $U_2 = \frac{1}{2\pi}\tan^{-1}(Z_2 / Z_1)$.
- We can then find the Jacobian

$$J = \begin{vmatrix} \partial U_1 / \partial Z_1 & \partial U_1 / \partial Z_2 \\ \partial U_2 / \partial Z_1 & \partial U_2 / \partial Z_2 \end{vmatrix}$$

- This allows us to write

$$\begin{aligned} f_{Z_1,Z_2}(z_1,z_2) &= f_{U_1,U_2}\left(u_1(z_1,z_2), u_2(z_1,z_2)\right)\left|J\right| \\ &= \frac{1}{2\pi}\exp\left\{-(z_1^2 + z_2^2)/2\right\} \end{aligned}$$

Which is indeed the distribution of $N(0, 1)$.

- Evaluating trigonometric functions is slow. A faster method uses rejection sampling. Generate $V_1, V_2 \sim U(-1,1)$ and set $r^2 = V_1^2 + V_2^2$. If $r^2 \geq 1$, reject. Otherwise, deliver

$$Z_1 = V_1\sqrt{\frac{-2\log(r^2)}{r^2}} \qquad Z_2 = V_2\sqrt{\frac{-2\log(r^2)}{r^2}}$$

This can be proven as follows

- The random variables V_1 and V_2 that pass the rejection step are distributed uniformly on the unit disk, so $f_{V_1,V_2}(v_1, v_2) = \frac{1}{\pi}\mathbb{I}_{\{v_1^2+v_2^2\leq 1\}}$.
- We then write

$$V_1 = R\cos\Theta = \sqrt{R^2}\cos\Theta \quad V_2 = R\sin\Theta = \sqrt{R^2}\sin\Theta$$

- We find the Jacobian for the transformation involving R^2

$$J = \begin{vmatrix} \partial V_1 / \partial R^2 & \partial V_1 / \partial \Theta \\ \partial V_2 / \partial R^2 & \partial V_2 / \partial \Theta \end{vmatrix} = \begin{vmatrix} \frac{1}{2\sqrt{R^2}}\cos\Theta & -\sqrt{R^2}\sin\Theta \\ \frac{1}{2\sqrt{R^2}}\sin\theta & \sqrt{R^2}\cos\Theta \end{vmatrix} = \frac{1}{2}$$

- This immediately gives

$$f_{R^2,\Theta}(r^2,\theta) = \frac{1}{2\pi}\mathbb{I}_{\{0<R^2\leq 1, 0\leq\theta<\pi\}}$$

So R^2 and $\tilde{\Theta} = \frac{\Theta}{2\pi}$ are uniformly distributed on (0, 1].

- We can therefore use the Box-Muller method with $U_1 = R^2$ and $U_2 = \tilde{\Theta}$. Since $\cos\left(2\pi\tilde{\Theta}\right) = \cos\left(\Theta\right) = V_1 / R$ and $\cos\left(2\pi\tilde{\Theta}\right) = V_2 / R$, the Box-Muller equations become exactly as above, in terms of R^2, V_1 and V_2.

- **Exponential distribution** $\boxed{X \sim \exp(\mu)}$: the inverse CDF method is easy to implement and considered satisfactory. Note that the inverse CDF is $F^{-1}(U) = -\mu^{-1}\log(U)$.

- **Gamma distribution** $\boxed{X \sim \Gamma(\alpha,\beta)}$: Note that if $Z \sim \Gamma(\alpha,1)$, then $X = Z / \beta$, so sampling $Z \sim \Gamma(\alpha,1)$ is enough. A number of algorithms based on rejection methods exist depending on whether $\alpha > 1$ or $\alpha < 1$. (If $\alpha = 1$, we have an exponential distribution).

  For an inverse gamma, note that $X \sim \Gamma^{-1}(\alpha,\beta) \Rightarrow X^{-1} \sim \Gamma(\alpha,\beta)$, so simply take the reciprocal of gamma draws.

- **Chi-squared distribution** $\boxed{X \sim \chi^2_\nu}$: This is a special case of the gamma; $X \sim \Gamma(\frac{\nu}{2}, \frac{1}{2})$.

  For small ν , we can use the Box-Muller method to generate ν normally distributed variables, square them, and add them together. Note that if Z_1 and Z_2 are two Box-Muller generated variables, $Z_1^2 + Z_2^2 = -2\log(U_1)$, and so when ν is even, the sum is simply given by $-2\log\left(\prod_{i=1}^{\nu/2} U_i\right)$.

- **Poisson distribution** $\boxed{X \sim \text{Po}(\lambda)}$: A slow but clever way of sampling X is to realise that if the number of arrivals in an interval $[0, t]$ is Poisson distributed with mean λt, then the time between each Poisson arrival is distributed as $\exp(\lambda^{-1})$, for which $F^{-1}(U) = -\frac{1}{\lambda}\log(U)$.

  Thus, we simply generate such exponential variables and continuously sum them until the sum is greater than 1 – let N be the number of random variables required for this to happen. Then N is a realisation of X.

Non-parameteric Inference

Given a sample $X_1,\cdots,X_n$ of independent variables with distribution F, we are often interested in estimating some parameter $\theta = \theta(F)$. Common examples:

- $\theta = \mathbb{E}_f\left\{\phi(x)\right\}$
- $\theta = P_f\left(X_1 \in A\right)$ – this is a special case of the above with $\phi(x) = \mathbb{I}_{\{x\in A\}}$.
- $\theta = F^{-1}(1/2)$, the median.

The **plugin principle** is often used in estimating such values:

> **Plugin estimator**: The plugin estimator for θ is
>
> $$\hat{\theta} = \theta(\hat{F})$$
>
> Where $\hat{F}$ is an estimator of F. Often, the **empirical distribution function** (ECDF) $\hat{F}_n$ is used:
>
> $$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{x_i\le x\}}$$
>
> in which case we denote the plugin estimate $\hat{\theta}_n$.

R-CODE: The ECDF of a set of points in R can be found using

```
ecdf(dataPoints)
```

So for example, to plot find the ECDF of the normal distribution, based on 10 points, use

```
plot(ecdf(rnorm(10,0,1)),verticals=TRUE)
```

(The last statement ensures vertical lines are drawn to connect the "steps" in the graph).

To then add a line representing the *real* density, use

```
x <- seq(-2.5,2.5,length=1000)
```

(this generates 1000 x values between –2.5 and 2.5), and then

```
lines(x,pnorm(x,0,1),col=2,lty=2)
```

(the last two statements make the line red and dotted).

Using the ECDF, the plugin estimators of the above examples are

- $\hat{\theta}_n = \frac{1}{n}\sum\phi(x_i)$
- $\hat{\theta}_n = \frac{1}{n}\sum\mathbb{I}_{\{x_i\in A\}}$
- $\hat{\theta}_n = F_n^{-1}(\frac{1}{2}) = x_{\left(\lceil n/2\rceil\right)}$, where $X_{(1)} < \cdots < X_{(n)}$ are the ordered X.

These estimators are unbiased, and it can be shown[3] that

$$\mathbb{V}\mathrm{ar}(\hat{\theta}_n) = \frac{1}{n}\mathbb{V}\mathrm{ar}\left\{\phi(X)\right\}$$

This variance, however, may be very large. In **variance reduction**, we try to reduce the variance of estimators while maintaining other good qualities.

Importance Sampling (IS)

Importance sampling reduces the variance of our estimator, and also allows us to sample from another, simpler distribution. We define the **support** of a function g, $\mathcal{Y}_g$ as $\mathcal{Y}_g = \left\{y : g(y) > 0\right\}$.

Importance sampling: Suppose we are trying to estimate $\theta = \mathbb{E}_f\left\{\phi(x)\right\}$. The density of interest is f, and let g denote another density that is easily sampled from and such that

$$f(x)\left|\phi(x)\right| > 0 \Rightarrow g(x) > 0$$

(Or in other words, $\mathcal{Y}_{f(x)|\phi(x)|} \subseteq \mathcal{Y}_g$).

Sample $y_1, \cdots, y_n$ independently from g and consider estimators of the form

$$\hat{\theta}_g = \frac{1}{n}\sum_{i=1}^{n} w_i \phi(y_i) \qquad\qquad w_i = w(y_i) = \frac{f(y_i)}{g(y_i)}$$

where the w_i are called **importance weights**.

Then the estimator $\hat{\theta}_g$ is also unbiased, and its variance is minimized when

$$g(x) = g_0(x) = \frac{\left|\phi(x)\right| f(x)}{\int_{\mathcal{Y}_{f|\phi|}} \left|\phi(y)\right| f(y)\ \mathrm{d}y}$$

[3] We prove this result as follows

$$\mathbb{V}\mathrm{ar}\left(\hat{\theta}_n\right) = \mathbb{V}\mathrm{ar}\left(\tfrac{1}{n}\textstyle\sum \phi(x_i)\right) = \tfrac{1}{n^2}\mathbb{V}\mathrm{ar}\left(\textstyle\sum \phi(x_i)\right) = \tfrac{1}{n}\mathbb{V}\mathrm{ar}\left\{\phi(X)\right\}$$

(note that in the last step, we only extra a factor of n from the variance – not n^2 – because the different items in the sum are independent).

Intuitively – the last line shows that we are sampling from a region where both f and ϕ are large; in other words, very informative regions.

Proof: We first show that our estimator is unbiased

$$\begin{aligned}
\mathbb{E}_g\left\{\hat{\theta}_g\right\} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{w(Y)\phi(Y)\right\} \\
&= \mathbb{E}_g\left\{w(Y)\phi(Y)\right\} \\
&= \int_{\mathcal{Y}_g}\left(\frac{f(y)}{g(y)}\phi(y)\right)g(y)\ \mathrm{d}y \\
&= \int_{\mathcal{Y}_g}\phi(y)f(y)\ \mathrm{d}y \\
&= \int_{\mathcal{Y}_{f|\phi|}}\phi(y)f(y)\ \mathrm{d}y \\
&= \mathbb{E}_f\left\{\phi(x)\right\} \\
&= \theta
\end{aligned}$$

(We were able to change the range of integration in the second-to-last step, because in all those members of $\mathcal{Y}_g$ that are not members of $\mathcal{Y}_{f|\phi|}$, the integral would have been 0 anyway).

The variance of our estimator is given by

$$\begin{aligned}
\mathrm{var}_g\left\{\hat{\theta}_g\right\} &= \mathrm{var}_g\left\{\frac{1}{n}\sum_{i=1}^{n}w(Y)\phi(Y)\right\} \\
&= \frac{1}{n}\mathrm{var}_g\left\{w(Y)\phi(Y)\right\} \\
&= \frac{1}{n}\mathrm{var}_g\left\{\frac{f(Y)}{g(Y)}\phi(Y)\right\}
\end{aligned}$$

So

$$\begin{aligned}
n\,\mathrm{var}_g\left\{\hat{\theta}_g\right\} &= \mathbb{E}_g\left\{\left(\frac{f(Y)}{g(Y)}\phi(Y)\right)^2\right\} - \mathbb{E}_g\left\{\frac{f(Y)}{g(Y)}\phi(Y)\right\}^2 \\
&= \int_{\mathcal{Y}_g}\frac{f(y)^2\phi(y)^2}{g(y)^2}g(y)\ \mathrm{d}y - \theta^2 \\
&= \int_{\mathcal{Y}_{f|\phi|}}\frac{f(y)^2\phi(y)^2}{g(y)}\ \mathrm{d}y - \theta^2
\end{aligned}$$

(We were able to change the range of integration in the last step because at any point in $\mathcal{Y}_g$ which is not in $\mathcal{Y}_{f|\phi|}$, the integral is 0 anyway).

Clearly, the variance is minimized when the integral in the last line above is minimized. By Jensen's Inequality, however, we can find this lower bound:

$$\mathbb{E}_g\left\{\left(\frac{f(Y)}{g(Y)}\phi(Y)\right)^2\right\} \geq \mathbb{E}_g\left\{\frac{f(Y)\left|\phi(Y)\right|}{g(Y)}\right\}^2$$
$$= \left(\int_{\mathcal{Y}_{f|\phi|}} \left|\phi(Y)\right| f(y)\ \mathrm{d}y\right)^2$$

This bound is obviously achieved when g is as given in the theorem. ■

The only issue is that if we know the integral in the expression for g_0, we probably know the integral of interest anyway! It is furthermore unclear how we might sample from g_0. The theorem is useful, however, in suggesting that we should seek a g "close" to g_0 from which it is easy to sample.

Another interesting note: if $\phi(x) = \mathbb{I}_{\{x\in A\}}$, then g_0 is simply the conditional density of X given $x \in A$ … This makes sense; this is where the data is most informative (we don't care about points outside A).

A final point is that if f or g are only known up to a normalising constant, then the method can still be used with the estimator

$$\hat{\theta}_g = \frac{\sum_{i=1}^n w_i\phi(x_i)}{\sum_{i=1}^n w_i}$$

is asymptotically unbiased, because

$$\mathbb{E}_g(W) = \mathbb{E}_g\left\{f(Y)\ /\ g(Y)\right\} = 1$$

so as $n \to \infty$, $\frac{1}{n}\sum w_i \to 1$. Short of such an asymptotic situation, the estimator above will always exhibit some (small) bias. However, choosing g correctly will still give us a small variance estimator. This is the **bias-variance tradeoff**.

This method can be used to estimate the CDF F by allowing ϕ to take two arguments.

$$\phi(y_i, x) = \mathbb{I}_{\{y_i \le x\}}$$
$$\hat{\theta}_g(x) = \frac{\sum_{i=1}^n w_i \phi(y_i, x)}{\sum_{i=1}^n w_i} = \sum_{i=1}^n \frac{w_i}{\sum_{i=1}^n w_i} \mathbb{I}_{\{y_i \le x\}} = \hat{F}_{g,n}(x)$$

This can easily be worked out as follows

- Sample y_i from g [a uniform distribution, for example]. Let $y_{(i)}$ be the i^{th} ordered part of the sample.
- Work out the weights, normalise them – call the normalised weights $\tilde{w}$.
- Let $\hat{F}_{g,n}\left(y_{(1)}\right) = \tilde{w}\left(y_{(1)}\right)$
- Thence, $\hat{F}_{g,n}\left(y_{(i+1)}\right) = \hat{F}_{g,n}\left(y_{(i)}\right) + \tilde{w}\left(y_{(i+1)}\right)$

Sometimes, it is useful to retain a sample from f rather than an estimate of a parameter. This can be achieved by **sampling with replacement** from $y_1, \cdots, y_n$ with a discrete distribution proportional to $w_1, \cdots, w_n$. This is called **sampling importance re-sampling** (SIR).

R-CODE: The following function

```
sample(y, num, replace=TRUE, prob=w)
```

Samples `num` items from the vector `y`, *with* replacement, and places a probability `w` on each item.

EXAMPLE: Suppose that X has Cauchy distribution, with density

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad x \in \mathbb{R}$$

and cumulative density

$$F(x) = \tfrac{1}{\pi}\arctan(x) + \tfrac{1}{2}$$

and we want to estimate $\theta = \mathbb{P}(X > 2)$. In other words, we have $\phi(x) = \mathbb{I}_{\{x>2\}}$.

We want g to be as close as possible to

$$\begin{aligned} f(X=x \mid X>2) &= \frac{f(x>2 \mid X=x) f(X=x)}{\mathbb{P}\Big(X>2\Big)} \\ &= \frac{\mathbb{I}_{\{x>2\}} \frac{1}{\pi(1+x^2)}}{1-\frac{1}{\pi}\arctan(2)+\frac{1}{2}} \\ &= \frac{1}{\Big(\frac{\pi}{2}-\arctan 2\Big)\Big(1+x^2\Big)} \mathbb{I}_{\{x>2\}} \end{aligned}$$

A sensible choice seems to be $g(x)=\frac{2}{x^2}\mathbb{I}_{\{x>2\}}$, which is easy to sample from by inversion [the factor of 2 is to ensure the distribution is normalised].

We then sample y_i from g. In this case, every value we sample is greater than 2, and so $\phi(x)=1$. As such, the IS estimator is simply

$$\hat{\theta}_g = \frac{1}{n}\sum_{i=1}^{n} w_i \qquad\qquad w_i = \frac{f(y_i)}{g(y_i)} = \frac{y_i^2}{2\pi\Big(1+y_i^2\Big)}$$

Consider the variance of this estimator

$$\begin{aligned} \operatorname{var}_g\Big\{\hat{\theta}_g\Big\} &= \operatorname{var}_g\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{2\pi\Big(1+y_i^2\Big)}\right\} \\ &= \frac{1}{4\pi^2 n}\operatorname{var}_g\left\{\frac{y_i^2}{1+y_i^2}\right\} \\ &= \frac{1}{4\pi^2 n}\left\{\int_{-\infty}^{\infty}\frac{y_i^4}{(1+y_i^2)^2}\frac{2}{y_i^2}\mathbb{I}_{\{x>2\}}\,\mathrm{d}x-\theta^2\right\} \\ &= \frac{1}{2\pi^2 n}\left\{\int_{2}^{\infty}\frac{y_i^2}{(1+y_i^2)^2}\,\mathrm{d}x-\theta^2\right\} \\ &= \frac{1}{n}\left(\frac{1}{10\pi^2}+\frac{\theta}{4\pi}-\theta^2\right) \end{aligned}$$

Furthermore,

$$\begin{aligned} \operatorname{var}_f\Big\{\hat{\theta}_n\Big\} &= \operatorname{var}_f\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{x_i>2\}}\right\} \\ &= \frac{1}{n}\operatorname{var}_f\Big\{\mathbb{I}_{\{X>2\}}\Big\} \\ &= \frac{1}{n}\Big\{\Big[0\cdot\mathbb{P}(X<2)+1\cdot\mathbb{P}(X>2)\Big]-\theta^2\Big\} \\ &= \frac{1}{n}\theta(1-\theta) \end{aligned}$$

Feeding in the true value of $\theta=\frac{1}{2}-\frac{1}{\pi}\arctan 2$, we find that the independence sampler is a large improvement.

Control variates

Definition (Control variates): Suppose Y is an unbiased estimator of θ. C is a *control variate* for Y if it is correlated with Y and its mean μ_C is known.

Control variates: Suppose Y is an unbiased estimator of θ, and C is a *control variate* for Y. Then

$$Y_\beta = Y - \beta(C - \mu_C)$$

is also unbiased for θ, and its variance is minimized when

$$\beta = \beta^* = \frac{\operatorname{cov}\{Y, C\}}{\operatorname{var}\{C\}}$$

at which point

$$\operatorname{var}\{Y_{\beta^*}\} = (1 - \rho^2)\operatorname{var}\{Y\} \leq \operatorname{var}\{Y\}$$

where $\rho = \operatorname{corr}\{Y, C\}$.

Proof: It is pretty obvious that Y_β is unbiased. We then have

$$\begin{aligned}
\operatorname{var}\{Y_\beta\} &= \operatorname{var}\{Y - \beta C + \beta\mu_C\} \\
&= \operatorname{var}\{Y - \beta C\} \\
&= \mathbb{E}\left\{\left(\{Y - \theta\} + \{-\beta C + \beta\mu_C\}\right)^2\right\} \\
&= \mathbb{E}\left\{(Y - \theta)^2\right\} + \mathbb{E}\left\{\beta^2 (C - \mu_C)^2\right\} \\
&\qquad -2\beta\mathbb{E}\left\{\{Y - \theta\}\{C - \mu_C\}\right\} \\
&= \operatorname{var}\{Y\} + \beta^2 \operatorname{var}\{C\} - 2\beta \operatorname{cov}\{Y, C\}
\end{aligned}$$

Differentiating with respect to β and setting to 0, we obtain the results above. ■

Note that even if $\operatorname{cov}(Y, C)$ is not known, there is an optimal β that relies only on the size of the covariance. Furthermore, the ideas can be extended to more than one control variate as follows:

$$\tilde{Y} = Y - \beta_1(C_1 - \mu_{C_1}) - \cdots - \beta_k(C_k - \mu_{C_k})$$

EXAMPLE: We return to the Cauchy example of the last section, in which we were estimating $\theta = \mathbb{P}(X > 2)$. A different approach would be to use the estimator

$$\theta = \frac{1}{2} - \mathbb{P}\Big(0 < X < 2\Big) = \frac{1}{2} - \int_0^2 f(x)\ \mathrm{d}x$$

We can estimate the integral using Monte-Carlo integration. In this case, we use $\phi(x) = \mathbb{I}_{\{x\in(0,2)\}}$, and

$$g \sim U(0,2) \Rightarrow g(x) = \tfrac{1}{2}\mathbb{I}_{\{x\in(0,2)\}}$$

We then generate $y_1, \cdots, y_n$ independently of g. Every one of these values will be between 0 and 2, and so $\phi(x) = 1$. The IS estimator of the integral based on g is then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i)}{\frac{1}{2}} = \frac{2}{n}\sum_{i=1}^{n} f(y_i)$$

and so our estimator of θ will be

$$\tilde{\theta} = \frac{1}{2} - \frac{2}{n}\sum_{i=1}^{n} f(y_i)$$

A Taylor Expansion of f suggests the following improved estimator:

$$\tilde{\theta} = \frac{1}{2} - \frac{2}{n}\sum_{i=1}^{n} f(y_i) - \beta\Big(y_i^2 - \mathbb{E}[y_i^2]\Big)$$

This is clearly of control variate form, with

$$C = \frac{2}{n}\sum_{i=1}^{n} y_i^2$$

The variance of this estimator is minimised at

$$\beta = \beta^* = \frac{\mathrm{cov}\Big\{Y, C\Big\}}{\mathrm{var}\Big\{C\Big\}}$$

Now remember that the different y_i are independent. Thus

$$\begin{aligned}
\mathrm{cov}\Big\{Y, C\Big\} &= \tfrac{4}{n^2}\mathrm{cov}\Big\{\textstyle\sum f(y_i), \sum y_i^2\Big\} \\
&= \frac{4}{n^2} n\,\mathrm{cov}\left\{\frac{1}{\pi(1+Y^2)}, Y^2\right\} \\
&= \frac{4}{n}\left\{\mathbb{E}\left(\frac{Y^2}{\pi(1+Y^2)}\right) - \mathbb{E}\left(\frac{1}{\pi(1+Y^2)}\right)\mathbb{E}(Y^2)\right\}
\end{aligned}$$

Remembering Y is uniformly distributed, the above are relatively simple to calculate. Similarly

$$\mathrm{var}\left\{\tfrac{2}{n}\sum_{i=1}^{n} y_i^2\right\} = \tfrac{4}{n}\mathrm{var}\left\{Y^2\right\}$$

We find

$$\frac{\mathrm{cov}\left\{Y,C\right\}}{\mathrm{var}\left\{C\right\}} = \frac{\mathrm{cov}\left\{\frac{1}{\pi(1+Y^2)},Y^2\right\}}{\mathrm{var}\left\{Y^2\right\}} = \frac{45}{64\pi}\left(1-\frac{7}{6}\arctan 2\right)$$

Finding the actual variance of the estimator involves similar steps.

Antithetic variables

Antithetic variables: Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of θ with the same expectation. We say they are *antithetic variables* if they are negatively correlated.

We then consider estimators formed as a convex combination of $\hat{\theta}_1$ and $\hat{\theta}_2$

$$\hat{\theta}_\lambda = \lambda\hat{\theta}_1 + \left(1-\lambda\right)\hat{\theta}_2 \qquad \lambda \in (0,1)$$

We have $\mathbb{E}(\hat{\theta}_\lambda) = \mathbb{E}(\hat{\theta}_1) = \mathbb{E}(\hat{\theta}_2)$. The variance of the estimator is minimized by

$$\lambda^* = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

at which point it takes value

$$\mathrm{var}\left(\hat{\theta}_{\lambda^*}\right) = \frac{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

Where $\sigma_1^2 = \mathrm{var}(\hat{\theta}_1)$, $\sigma_2^2 = \mathrm{var}(\hat{\theta}_2)$ and $\rho = \mathrm{corr}(\hat{\theta}_1,\hat{\theta}_2)$.

Proof: We have

$$\begin{aligned}
\mathrm{var}(\hat{\theta}_\lambda) &= \mathrm{var}\left\{\lambda\hat{\theta}_1 + \left(1-\lambda\right)\hat{\theta}_2\right\} \\
&= \mathbb{E}\left\{\left[\lambda\left(\hat{\theta}_1-\mu_1\right) + (1-\lambda)\left(\hat{\theta}_2-\mu_2\right)\right]^2\right\} \\
&= \lambda^2\mathbb{E}\left\{\left(\hat{\theta}_1-\mu_1\right)^2\right\} + (1-\lambda)^2\mathbb{E}\left\{\left(\hat{\theta}_2-\mu_2\right)^2\right\} \\
&\qquad +2\lambda(1-\lambda)\mathbb{E}\left\{\left(\hat{\theta}_1-\mu_1\right)\left(\hat{\theta}_2-\mu_2\right)\right\} \\
&= \lambda^2\sigma_1^2 + (1-\lambda)^2\sigma_2^2 + 2\lambda(1-\lambda)\,\mathrm{cov}(\hat{\theta}_1,\hat{\theta}_2) \\
&= \lambda^2\sigma_1^2 + (1-\lambda)^2\sigma_2^2 + 2\lambda(1-\lambda)\rho\sigma_1\sigma_2
\end{aligned}$$

Differentiating with respect to λ and setting to 0, we obtain the results above. ■

Note: the computational cost of computing $\hat{\theta}_{\lambda^*}$ is twice that of computing each of the individual estimators, but we get a significant reduction in variance.

The method is especially useful when used with the following theorem. (Note: a **non-degenerate distribution** is one for which there is no $a \in \mathbb{R}$ such that $F(x) = \mathbb{I}_{\{x \geq a\}}$)

Theorem: Let F be a non-degenerate distribution function and $U \sim U(0,1]$, then

$$\operatorname{cov}\left\{F^{-1}(U), F^{-1}(1-U)\right\} < 0$$

Proof: Let

$$\theta = \mathbb{E}\left\{F^{-1}(U)\right\} = \mathbb{E}\left\{F^{-1}(1-U)\right\}$$

Observe that

$$F^{-1}(1-u) > \theta \Leftrightarrow u < 1 - F(\theta)$$

We have

$$\begin{aligned}
&\operatorname{cov}\left\{F^{-1}(U), F^{-1}(1-U)\right\} \\
&\quad = \mathbb{E}\left\{\left(F^{-1}(U) - \theta\right)\left(F^{-1}(1-U) - \theta\right)\right\} \\
&\quad = \mathbb{E}\left\{F^{-1}(U)\left(F^{-1}(1-U) - \theta\right)\right\} \\
&\quad = \int_0^{1-F(\theta)} F^{-1}(u)\left[F^{-1}(1-u) - \theta\right]\,\mathrm{d}u \\
&\qquad\qquad + \int_{1-F(\theta)}^{1} F^{-1}(u)\left[F^{-1}(1-u) - \theta\right]\,\mathrm{d}u
\end{aligned}$$

We want to try and get an upper bound for this quantity; ie: find the **largest** it could ever be. To do that, we note that

- The quantity in the square brackets is positive in the first integral.
- Thus, if we replace $F^{-1}(u)$ by a constant equal to the largest value it can take in that range of integration, we get an upper bound.

- Since $F^{-1}(u)$ is a non-decreasing function, the largest value it can take in the range of the first integral is $F^{-1}\big(1-F(\theta)\big)$.
- So the first integral is smaller than

$$F^{-1}\big(1-F(\theta)\big)\int_0^{1-F(\theta)} F^{-1}(1-u)-\theta \ \mathrm{d}u$$

The quantity in the square brackets is negative in the second integral, and a similar argument applies. Together, these imply that

$$\begin{aligned}
&\operatorname{cov}\Big\{F^{-1}(U),F^{-1}(1-U)\Big\}\\
&\quad< F^{-1}\big(1-F(\theta)\big)\Big\{\int_0^{1-F(\theta)} F^{-1}(1-u)-\theta \ \mathrm{d}u\\
&\qquad\qquad +\int_{1-F(\theta)}^{1} F^{-1}(1-u)-\theta \ \mathrm{d}u\Big\}\\
&\quad= F^{-1}\big(1-F(\theta)\big)\int_0^{1} F^{-1}(1-u)-\theta \ \mathrm{d}u\\
&\quad= F^{-1}\big(1-F(\theta)\big)\Big\{\int_0^{1} F^{-1}(1-u) \ \mathrm{d}u-\theta\Big\}\\
&\quad= 0
\end{aligned}$$

Our covariance is therefore negative. ■

Effectively, the method of inversion provides two correlated samples; one based on U and one based on $1 - U$. This method is useful when we have an estimator of the form $\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n} g(y_i)$, and the y_i can be simulated by inversion.

EXAMPLE: Back once again to the Cauchy distribution, we return to our estimator of the form

$$\tilde{\theta}=\frac{1}{2}-\frac{2}{n}\sum_{i=1}^{n}\frac{1}{\pi(1+y_i^2)}$$

where $y_1,\cdots,y_n \sim U(0,2),\text{IID}$. An antithetic variable estimator based on this is simply

$$\tilde{\theta}=\frac{1}{2}-\frac{2}{n}\sum_{i=1}^{n}\frac{1}{2}\frac{1}{\pi(1+y_i^2)}+\frac{1}{2}\frac{1}{\pi(1+(2-y_i)^2)}$$

EXAMPLE: Consider that

$$\int_0^1 \sqrt{1-u^2} \ \mathrm{d}u=\frac{\pi}{4}$$

This means that a Monte-Carlo estimator of π is

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} 4\sqrt{1-U_i^2}$$

Where U_i are independent $U(0,1)$ random variables. An estimator based on the method of antithetic variables is

$$\hat{\pi}_{\text{AV}} = \frac{1}{n}\sum_{i=1}^{n} 4\left\{\tfrac{1}{2}\sqrt{1-U_i^2} + \tfrac{1}{2}\sqrt{1-(1-U_i)^2}\right\}$$

Now, note that

$$\begin{aligned}\mathbb{V}\text{ar}(\hat{\pi}) &= \frac{1}{n^2}\mathbb{V}\text{ar}\left(\sum_{i=1}^{n} 4\sqrt{1-U_i^2}\right)\\ &= \frac{16}{n}\mathbb{V}\text{ar}\left(\sqrt{1-U_i^2}\right)\end{aligned}$$

And

$$\begin{aligned}\mathbb{V}\text{ar}\left(\hat{\pi}_{\text{AV}}\right) &= \frac{1}{n^2}\mathbb{V}\text{ar}\left[\sum_{i=1}^{n} 4\left\{\tfrac{1}{2}\sqrt{1-U_i^2} + \tfrac{1}{2}\sqrt{1-(1-U_i)^2}\right\}\right]\\ &= \frac{4}{n}\Bigg[\mathbb{V}\text{ar}\left(\sqrt{1-U_i^2}\right) + \mathbb{V}\text{ar}\left(\sqrt{1-(1-U_i)^2}\right)\\ &\qquad +2\mathbb{C}\text{ov}\left(\sqrt{1-U_i^2},\sqrt{1-(1-U_i)^2}\right)\Bigg]\\ &= \frac{4}{n}\left[2\mathbb{V}\text{ar}\left(\sqrt{1-U_i^2}\right) + 2\mathbb{C}\text{ov}\left(\sqrt{1-U_i^2},\sqrt{1-(1-U_i)^2}\right)\right]\end{aligned}$$

Note that there is no factor of n in front of the covariance because the U_i are independent and therefore uncorrelated with each other. Thus, cross terms vanish from the covariance.

As such

$$\begin{aligned}\frac{\mathbb{V}\text{ar}(\hat{\pi}_{\text{AV}})}{\mathbb{V}\text{ar}(\hat{\pi})} &= \frac{1}{4}\frac{\left[2\mathbb{V}\text{ar}\left(\sqrt{1-U_i^2}\right) + 2\mathbb{C}\text{ov}\left(\sqrt{1-U_i^2},\sqrt{1-(1-U_i)^2}\right)\right]}{\mathbb{V}\text{ar}\left(\sqrt{1-U_i^2}\right)}\\ &= \frac{1}{2}\left\{1 + \frac{\mathbb{C}\text{ov}\left(\sqrt{1-U_i^2},\sqrt{1-(1-U_i)^2}\right)}{\mathbb{V}\text{ar}\left(\sqrt{1-U_i^2}\right)}\right\}\\ &= \frac{1}{2}\left\{1 + \frac{\int_0^1 \sqrt{1-u^2}\sqrt{1-(1-u)^2}\ \mathrm{d}u - \frac{\pi^2}{16}}{(1-\frac{1}{3}) - \frac{\pi^2}{16}}\right\}\\ &= 0.140\end{aligned}$$

Non-parameteric Inference

Monte-Carlo Tests

Let $X_1,\cdots,X_n$ be independent with distribution function F, and suppose we want to use a statistic $T = T(X_1,\cdots,X_n)$ to test

$$H_0 : F = F_0 \qquad \text{against}$$
$$H_1 : F \neq F_0$$

If **small values of T** represent **departure from H_0**, then a test of size $\alpha \in (0,1)$ would **reject H_0** if $T < c_\alpha$, where c_α is the α^{th} quantile of T.

If the null distribution of T is unkown, however, we way not be able to compute c_α. In a **Monte Carlo test** of approximate size α, we estimate c_α as follows:

1. Choose a large $B \in \mathbb{N}$.
2. For each $k \in \{1,\cdots,B\}$, associate an α, namely $\alpha = k/(B+1)$, and restrict the choice of α to those values.
3. For each $k \in \{1,\cdots,B\}$, assume H_0 is true, and simulate a random sample of N variables
$$X_{k1}^*,\cdots,X_{kn}^* \qquad \text{for } k = 1,\cdots,B$$
4. For each $k \in \{1,\cdots,B\}$, compute $T_k^* = T\left(X_{k1}^*,\cdots,X_{kn}^*\right)$
5. Let $c_\alpha \equiv T_{(k)}^*$, where $T_{(k)}^*$ is the k^{th} item when the T_k^* are ordered.

Since the critical point is random, the critical region is "blurred"

> **Theorem:** Assume that under H_0, T has density f_0 supported on an interval. Then the MC test has *exact* size α.

Proof: The size of the test is (by definition) $\mathbb{P}\left(T < T_{(k)}^* \mid H_0 \text{ true}\right)$. We can condition on the value of T:

$$\begin{aligned}\mathbb{P}\left(T < T_{(k)}^*\right) &= \int_{-\infty}^{\infty} \mathbb{P}\left(T < T_{(k)}^* \mid T = t\right) f_0(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \mathbb{P}\left(T_{(k)}^* > t\right) f_0(t)\,\mathrm{d}t\end{aligned}$$

Note, however, that $\mathbb{P}\left(T^*_{(k)} > t\right)$ is just the probability that *less than* k of the T^* are $\leq t$, so

$$\begin{aligned}\mathbb{P}\left(T^*_{(k)} > t\right) &= \mathbb{P}\left(\text{bin}\left(B, \mathbb{P}(T^* \leq t)\right) < k\right) \\ &= \sum_{r=0}^{k-1} {}^{B}C_r F_0(t)^r \left\{1 - F_0(t)\right\}^{B-r}\end{aligned}$$

Note that this last line assumes the T^* have the same distribution as T. It is therefore *not* acceptable to obtain the X^* using the jackknife or bootstrap.

Feeding this back into our integral

$$\begin{aligned}\mathbb{P}\left(T < T^*_{(k)}\right) &= \int_{-\infty}^{\infty} \sum_{r=0}^{k-1} {}^{B}C_r F_0(t)^r \left\{1 - F_0(t)\right\}^{B-r} f_0(t)\, \mathrm{d}t \\ &= \int_{-\infty}^{\infty} \sum_{r=0}^{k-1} {}^{B}C_r u^r \left\{1 - u\right\}^{B-r}\, \mathrm{d}u\end{aligned}$$

(The last line uses the existence of a density). Since the sum is finite:

$$\begin{aligned}&= \sum_{r=0}^{k-1} \int_{-\infty}^{\infty} {}^{B}C_r u^r \left\{1 - u\right\}^{B-r}\, \mathrm{d}u \\ &= \sum_{r=0}^{k-1} r^{-1} \int_{-\infty}^{\infty} ur\, {}^{B}C_r u^{r-1} \left\{1 - u\right\}^{B-r}\, \mathrm{d}u\end{aligned}$$

Replacing the factorials in the binomial coefficients with gamma functions, we see that this equivalent to taking $\mathbb{E}\left\{\beta\left(r, B - r + 1\right)\right\} = r \,/\, (B+1)$, which gives

$$\begin{aligned}&= \sum_{r=0}^{k-1} r^{-1} \frac{r}{B+1} \\ &= \frac{k}{B+1} \\ &= \alpha\end{aligned}$$

As required. ■

The hard part of the process above is step 3; simulating random samples under H_0. This can be done using **resampling methods**, which involve the use of samples taken from a single observed sample, and can be used when very little is known about the underlying distribution. We study two such methods…

The Jackknife

The jackknife involves subsampling without replacement from an observed sample. It involves the use of leave-one-out data sets and is most often used to estimate such quantities as the variance or bias of an estimator.

We first develop some notation. Let $X_1,\cdots,X_n$ be a sample of independent variables with distribution F, and let $\theta = \theta(F)$ be a parameter of interest. Let $\hat{\theta}^n = \hat{\theta}^n(X_1,\cdots,X_n)$ be an estimator of θ with variance v, and for $i = 1,\cdots,n$ let $\hat{\theta}^{n-1}_{(-i)} = \hat{\theta}^{n-1}(X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_n)$ be an estimator based on all the variables in the sample save one.

Jackknife estimator of variance: The *jackknife estimator of variance* is given by

$$\hat{v}_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}^{n-1}_{(-i)} - \hat{\theta}^{n-1}_{(\text{av})}\right)^2$$

where

$$\hat{\theta}^{n-1}_{(\text{av})} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^{n-1}_{(-i)}$$

Motivation: We can motivate this definition by considering a **linear statistic** of the form

$$\hat{\theta} = \mu + \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i)$$

for some function α.

In this case

$$\hat{\theta}^{n-1}_{(-i)} = \mu + \frac{1}{n-1}\sum_{j\neq i}\alpha(X_j)$$

and so

$$\begin{aligned}\hat{\theta}^{n-1}_{(\text{av})} &= \mu + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i}\alpha(X_j)\\ &= \mu + \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i}\alpha(X_j)\end{aligned}$$

note that the sum involves every $\alpha(X_i)$ exactly $n-1$ times, because we sum from $i = 1 \to n$, leaving out one item each time. We can therefore write it as

$$= \mu + \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i)$$
$$\equiv \mu + \bar{\alpha}$$

Where $\bar{\alpha}$ denotes the average of $\alpha(X_i)$.

We can now write:

$$\hat{v}_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(-1)}^{n-1} - \mu - \bar{\alpha}\right)^2$$
$$= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i}\alpha(X_j) - \bar{\alpha}\right)^2$$
$$= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{(n-1)^2}\sum_{j,k\neq i}\alpha(X_j)\alpha(X_k) - \frac{2\bar{\alpha}}{n-1}\sum_{j\neq i}\alpha(X_j) + \bar{\alpha}^2\right)^2$$

We now combine the outer and inner sums as follows:

- The first sum is hard! We can decompose it into two parts:
  - First, we'll get terms of the form $\alpha(X_i)^2$. We'll get $n-1$ of each of those, because we sum from $i = 1 \rightarrow n$, leaving out one item each time.
  - We'll then gets terms of the form $\alpha(X_i)\alpha(X_j)$. We'll get $n-2$ of each one of those, because we sum from $i = 1 \rightarrow n$, and we miss out <u>two</u> items each time ($\alpha(X_i)\alpha(X_j)$ and $\alpha(X_j)\alpha(X_i)$).
- The second sum is exactly as we saw above – once we combine the two sums, each $\alpha(X_i)$ appears $n-1$ times.
- $\bar{\alpha}$ is a constant, and so simply needs to be multiplied by n.

Overall, we now get:

$$
\begin{aligned}
&= \tfrac{n-1}{n}\Bigg\{\tfrac{n-2}{(n-1)^2}\sum_{i\neq j}\alpha(X_i)\alpha(X_j) + \tfrac{n-1}{(n-1)^2}\sum_{i=1}^{n}\alpha(X_i)^2 \\
&\qquad\qquad - \tfrac{2\bar{\alpha}(n-1)}{n-1}\sum_{i=1}^{n}\alpha(X_i) + n\bar{\alpha}^2\Bigg\} \\
&= \tfrac{n-1}{n}\Bigg\{\tfrac{n-2}{(n-1)^2}\sum_{i\neq j}\alpha(X_i)\alpha(X_j) + \tfrac{1}{n-1}\sum_{i=1}^{n}\alpha(X_i)^2 \\
&\qquad\qquad - 2n\bar{\alpha}^2 + n\bar{\alpha}^2\Bigg\} \\
&= \tfrac{n-1}{n}\Bigg\{\tfrac{n-2}{(n-1)^2}\sum_{i\neq j}\alpha(X_i)\alpha(X_j) + \tfrac{1}{n-1}\sum_{i=1}^{n}\alpha(X_i)^2 - n\bar{\alpha}^2\Bigg\} \\
&= \tfrac{n-1}{n}\Bigg\{\tfrac{n-2}{(n-1)^2}\left(\sum_{i=1}^{n}\alpha(X_i)\right)^2 + \tfrac{1}{(n-1)^2}\sum_{i=1}^{n}\alpha(X_i)^2 - n\bar{\alpha}^2\Bigg\}
\end{aligned}
$$

In the last step, consider the fact that the first term contributes $\frac{n-2}{(n-1)^2}$ lots of $\sum\alpha(x_i)^2$, whereas the second term contributes $\frac{1}{(n-1)^2}$ lots of it. Together, these make

$$\frac{n-2}{(n-1)^2} + \frac{1}{(n-1)^2} = \frac{n-1}{(n-1)^2} = \frac{1}{n-1}$$

Which is indeed what we need, from the second-to-last line. Finally, we write

$$\hat{v}_{\text{jack}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big(\alpha(X_i) - \bar{\alpha}\Big)^2$$

Now, we also have

$$
\begin{aligned}
\operatorname{var}\Big(\hat{\theta}\Big) &= \operatorname{var}\left(\mu + \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i)\right) \\
&= \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n}\alpha(X_i)\right)
\end{aligned}
$$

Since the X_i are independent, we can write

$$
\begin{aligned}
\operatorname{var}\Big(\hat{\theta}\Big) &= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}\Big(\alpha(X_i)\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\Big\{\alpha(X_i) - \bar{\alpha}\Big\}^2\right)
\end{aligned}
$$

Taking the expectation of $\hat{v}_{\text{jack}}$ above, we see that it is indeed an unbiased estimator of v in this case. This results only holds true for a linear statistics, but

many smooth statistics can be well approximated by linear statistics. ■

Jackknife estimator of bias: The *jackknife estimator of bias* is given by

$$\widehat{\text{bias}}_{\text{jack}} = \left(n-1\right)\left(\hat{\theta}^{n-1}_{(\text{av})} - \hat{\theta}^{n}\right)$$

Motivation: We cannot use the same linear statistic as we did in the last proof, because the bias of that statistic is in fact 0. Instead, we need to the *quadratic statistic*

$$\hat{\theta} = \mu + \frac{1}{n}\sum_{i=1}^{n}\alpha(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta\left(X_i, X_j\right)$$

as an estimator of

$$\theta = \mu + \mathbb{E}\{\alpha(X_1)\} + \mathbb{E}\{\beta(X_1, X_2)\} = \mu + a + b$$

We first note that

$$\hat{\theta}^{n-1}_{(-i)} = \mu + \frac{1}{n-1}\sum_{i\neq j}\alpha(X_i) + \frac{1}{(n-1)^2}\sum_{j,k\neq i}\beta(X_i, X_j)$$

Now consider finding the average of each of the terms above

- $\left\langle \frac{1}{n-1}\Sigma_{i\neq j}\alpha(X_i)\right\rangle = \frac{1}{n}\Sigma_{i=1}^{n}\frac{1}{n-1}\Sigma_{i\neq j}\alpha(X_i)$. Looking at the sums, we find that each $\alpha(X_i)$ is summed over n – 1 times. So we can re-write this simply as $\frac{1}{n}\sum_{i=1}^{N}\alpha(X_i)$.
- $\left\langle \frac{1}{(n-1)^2}\Sigma_{j,k\neq i}\beta(X_i, X_j)\right\rangle = \frac{1}{n}\Sigma_{i=1}^{n}\frac{1}{(n-1)^2}\Sigma_{j,k\neq i}\beta(X_i, X_j)$
  We can split the second sum into two parts:
  - Those terms where $i = j$. There are $(n-1)$ such terms, which gives $\frac{1}{n(n-1)}\Sigma_{i=j}\beta(X_i, X_j)$.
  - Those terms where $i \neq j$. We can simply write those $\Sigma_{i\neq j}\beta(X_i, X_j)$. The sum above involves n – 2 lots of this sum (n from the outer sum, –2 to

exclude each of the subscripts being equal to i).

Thus, all in all, we can write (we omit the arguments of α and β to save space):

$$\hat{\theta}^{n-1}_{(\text{av})} = \mu + \tfrac{1}{n}\sum_{i\neq j}\alpha + \tfrac{n-2}{n(n-1)^2}\sum_{i\neq j}\beta + \tfrac{1}{n(n-1)}\sum_{i=j}\beta$$

It follows that

$$\begin{aligned}\widehat{\text{bias}}_{\text{jack}} &= (n-1)(\hat{\theta}^{n-1}_{(\text{av})} - \hat{\theta}^{n}) \\ &= (n-1)\left(\tfrac{n-2}{n(n-1)^2}\sum_{i\neq j}\beta + \tfrac{1}{n(n-1)}\sum_{i=j}\beta - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta\right)\end{aligned}$$

The last term, however, can be split as follows

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta = \tfrac{1}{n^2}\sum_{i=j}\beta + \tfrac{1}{n^2}\sum_{i\neq j}\beta$$

And so

$$\begin{aligned}\widehat{\text{bias}}_{\text{jack}} &= (n-1)(\hat{\theta}^{n-1}_{(\text{av})} - \hat{\theta}^{n}) \\ &= (n-1)\left(\left[\tfrac{n-2}{n(n-1)^2} - \tfrac{1}{n^2}\right]\sum_{i\neq j}\beta + \left[\tfrac{1}{n(n-1)} - \tfrac{1}{n^2}\right]\sum_{i=j}\beta\right) \\ &= (n-1)\left(\left[\tfrac{n(n-2)-(n-1)^2}{n^2(n-1)^2}\right]\sum_{i\neq j}\beta + \left[\tfrac{n-(n-1)}{n^2(n-1)}\right]\sum_{i=j}\beta\right) \\ &= -\tfrac{1}{n^2(n-1)}\sum_{i\neq j}\beta + \tfrac{1}{n^2}\sum_{i=j}\beta\end{aligned}$$

Taking expectations

$$\begin{aligned}\mathbb{E}\left(\widehat{\text{bias}}_{\text{jack}}\right) &= -\tfrac{1}{n^2(n-1)}n(n-1)\mathbb{E}\left\{\beta(X_1, X_2)\right\} \\ &\qquad + \tfrac{1}{n^2}n\mathbb{E}\left\{\beta(X_1, X_1)\right\} \\ &= \tfrac{1}{n}\left[\mathbb{E}\left\{\beta(X_1, X_1)\right\} - \mathbb{E}\left\{\beta(X_1, X_2)\right\}\right] \\ &= \tfrac{1}{n}\left[\mathbb{E}\left\{\beta(X_1, X_1)\right\} - b\right] \\ &= \text{bias}\left(\hat{\theta}_n\right)\end{aligned}$$

As expected. ■

The jackknife can also be used to generate $B = n$ "samples" from $X_1, \cdots, X_n$ by missing out a single element from each sample. However, if n is small, B will be small, and in the context of a Monte Carlo test, this will severely limit the possible choices of significance level α.

The Bootstrap

The non-parametric bootstrap is another form of resampling which involves forming samples *with* replacement from observed samples.

Consider a set of independent random variables $X = \left(X_1, \cdots, X_n\right)$, with distribution function F, of which we have a realisation $x_1, \cdots, x_n$. We are interested in the distribution of a root or pivot $R_n\left(X; F\right)$, which we denote $K_n(F)$. Such a "root" could include any statistic $T_n(X)$, but also quantities like $\frac{\sqrt{n}}{\hat{\sigma}}\left\{\overline{X} - \mathbb{E}_F(X_1)\right\}$.

Our strategy will be to use a different distribution function $\hat{F}$ that approximates F, and then estimate $K_n(F)$ by $K_n(\hat{F})$, as follows

Bootstrap estimator: We estimate $K_n(F)$ by $K_n(\hat{F})$, which we find as follows

1. Draw B independent *bootstrap samples*, each of size n
$$X_b^* = \left(X_{b1}^*, \cdots, X_{bn}^*\right) \quad b = 1, \cdots, B$$
where each of the X^* are independently drawn from the distribution $\hat{F}$.
2. Approximate $K_n(\hat{F})$ by the ECDF of
$$\left\{R_n\left(X_b^*; \hat{F}\right) : b = 1, \cdots, b\right\}$$

In general, $B = 100$ or $B = 200$ is often sufficient to estimate a variance or quantile, but $B = 1000$ is recommended to estimate the entire distribution.

Notes: The question of how to choose $\hat{F}$ remains.

- In a parametric model $\mathcal{F} = \left\{F_\theta : \theta \in \Theta\right\}$, the *parametric bootstrap* uses $\hat{F} = F_{\hat{\theta}}$, where $\hat{\theta}$ is the MLE of θ.
- The *non-parametric bootstrap* simply uses the ECDF $\hat{F}_n$ of $X = \left(X_1, \cdots, X_n\right)$. In that case, step 1 above simply involves taking a re-sample of size n with replacement from

$x_1,\cdots,x_n$, with equal probability placed on each expression.

Notes: If N_i denotes the number of times x_i appears in the bootstrap sample, then the vector $\left(N_1,\cdots,N_n\right)$ (note that $N_+ = 1$) has a one-to-one mapping with the $(2n - 1)$ binary tuple that contains N_1 ones, followed by a zero, followed by N_2 ones, etc... There are ${}^{2n-1}C_n$ such tuples, and this is therefore the number of bootstrap samples.

Furthermore, $N_i \sim \text{Multi}\left(n;\frac{1}{n},\cdots,\frac{1}{n}\right)$, and so

$$\mathbb{P}\left(N_1 = n_1,\cdots,N_n = n_n\right) = \frac{n!}{n_1!\cdots n_n!}\left(\frac{1}{n}\right)^n \leq \frac{n!}{n^n}$$

So the most likely sample is the original data, with probability $n!/n^n$.

R-Code: The bootstrap samples themselves can be taken using

```
X.star <- matrix(NA, nrow=B, ncol=n)
For(b in 1:B) {
    X.star[b,] <- sample(x, n, replace=TRUE)
}
```

Where `x` contains our data, `n` is the number of data items, and `B` is the number of bootstrap samples we want to take.

We cover a few examples to clarify this concept:

1. **Finding a confidence interval for** $\mu = \mathbb{E}_F\{X_1\}$

   An analytic approach would be to define the root

   $$R_n(X;F) = \sqrt{n}\left(\overline{X} - \mu\right) \to^d N\left(0, \mathbb{V}\text{ar}_F(X_1)\right)$$

   and then to estimate[4]

[4] To prove this result, consider that the ECDF of X_1^*, and associated density, are

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{x_i \leq x\}} \qquad \hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{x_i = x\}}$$

and so

$$\mathbb{E}\left(X_1^*\right) = \int x\hat{f}_n(x)\,\mathrm{d}x = \int \frac{1}{n}\sum_{i=1}^n x\mathbb{I}_{\{x_i=x\}}\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}$$

and similarly

$$\mathbb{V}\mathrm{ar}_F(X_1) = \mathbb{V}\mathrm{ar}_{\hat{F}}(X_1^*) = \tfrac{1}{n}\sum(x_i - \overline{x})^2$$

and finally estimating the $(1-\alpha)$ confidence interval as

$$\left[\overline{x} - \frac{z_{\sigma^2,\frac{1}{2}\alpha}}{\sqrt{n}}, \overline{x} + \frac{z_{\sigma^2,\frac{1}{2}\alpha}}{\sqrt{n}}\right]$$

Where $z_{\sigma^2,\alpha}$ **denotes the upper** α point of a $N(0,\sigma^2)$ distribution.

In reality, however, we do *not* know the distribution of the root. We therefore use the following bootstrap method:

- Fix a large B such that $\frac{1}{2}\alpha(B+1)$ is an integer.
- For each, b, generate independent bootstrap samples X_b^* by re-sampling, and compute
$$R_b^* = R_n\left(X_b^*; \hat{F}_n\right) = \sqrt{n}\left(\overline{X}_b^* - \overline{X}\right)$$
- Approximate the $(1-\alpha)$ confidence interval as
$$\left[\overline{x} - R^*_{\left([1-\frac{\alpha}{2}][B+1]\right)}, \overline{x} + R^*_{\left(\frac{\alpha}{2}[B+1]\right)}\right]$$

More generally, the bootstrap confidence intervals for a parameter $N(0,\sigma^2)$ are often based on the root $R_n\left(X;F\right) = \sqrt{n}\left(\tilde{\theta} - \theta\right)$, where $\tilde{\theta}$ is an estimator of θ. These are called *percentile intervals*.

2. Estimating $\mathbb{V}\mathrm{ar}_F\left\{\tilde{\theta}(X)\right\}$, where $\tilde{\theta}(X)$ is an estimator of $\theta = \theta(F)$

Our step should be to try and calculate the non-parametric bootstrap estimator $\mathbb{V}\mathrm{ar}_{\hat{F}_n}\left\{\tilde{\theta}(X^*)\right\}$ analytically, then we apply the following algorithm

- Generate B independent bootstrap samples X_b^* (by re-sampling)
- For each bootstrap sample, calculate $\tilde{\theta}_b^* = \tilde{\theta}(X_b^*)$
- Approximate $\mathbb{V}\mathrm{ar}_{\hat{F}_n}\left\{\tilde{\theta}(X^*)\right\}$ by
$$\frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{\theta}_b^* - \overline{\tilde{\theta}^*}\right)^2 \qquad \text{where } \overline{\tilde{\theta}^*} = \frac{1}{B}\sum_{i=1}^{B}\tilde{\theta}_b^*$$

$$\mathbb{V}\mathrm{ar}\left(X_1^*\right) = \int (x-\overline{x})^2 \hat{f}_n(x)\,\mathrm{d}x = \int \frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^2 \mathbb{I}_{\{x_i=x\}}\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

3. Estimating the bias of $\tilde{\theta}(X)$, an estimator of $\theta = \theta(F)$

The bias of $\tilde{\theta}$ is

$$\text{bias}_F\left\{\tilde{\theta}(X)\right\} = \mathbb{E}_F\left\{\tilde{\theta}(X)\right\} - \theta$$

the non-parametric bootstrap estimator of the bias is

$$\text{bias}_{\hat{F}_n}\left\{\tilde{\theta}(X^*)\right\} = \mathbb{E}_{\hat{F}_n}\left\{\tilde{\theta}(X^*)\right\} - \hat{\theta} \quad \text{where } \hat{\theta} = \theta(\hat{F}_n)$$

If we cannot compute this expression directly, we can apply the following algorithm instead:

- Generate B independent bootstrap samples X_b^* (by re-sampling)
- For each bootstrap sample, calculate $\tilde{\theta}_b^* = \tilde{\theta}(X_b^*)$
- Approximate $\text{bias}_{\hat{F}_n}(\tilde{\theta})$ by

$$\frac{1}{B}\sum_{b=1}^{B}\left(\tilde{\theta}_b^* - \hat{\theta}\right)$$

4. Estimating the distribution of the sample median, $F^{-1}\left(\frac{1}{2}\right)$

In this case, the distribution can be calculated analytically. Let $x_{(j)}$ be the j^{th} ordered statistic of $x_1, \cdots, x_n$, the *original* sample. Let also $\hat{F}_n^*$ be the empirical distribution function of a given bootstrap sample. Then

$$\begin{aligned}\mathbb{P}\left(\hat{F}_n^{*-1}(\tfrac{1}{2}) \le x_{(j)}\right) &= \mathbb{P}\left(\sum_{i=1}^{n}\mathbb{I}_{\{X_i^* \le x_{(j)}\}} \ge \left\lceil \tfrac{1}{2}n \right\rceil\right) \\ &= \sum_{k=\left\lceil \frac{1}{2}n \right\rceil}^{n}\binom{n}{k}\left(\frac{j}{n}\right)^k\left(\frac{n-j}{n}\right)^{n-k}\end{aligned}$$

Effectively, this simply the probability of our bootstrap sample containing more than $n/2$ items which are $\le x_{(j)}$ [each with probability j/n]. The sum above is over all the possible number of "more than $n/2$" items. From this, we have, for $j = 2, \ldots, n$

$$\mathbb{P}\left(\hat{F}_n^{*-1}(\tfrac{1}{2}) = x_{(j)}\right) = \mathbb{P}\left(\hat{F}_n^{*-1}(\tfrac{1}{2}) \le x_{(j)}\right) - \mathbb{P}\left(\hat{F}_n^{*-1}(\tfrac{1}{2}) \le x_{(j-1)}\right)$$

and

$$\mathbb{P}\left(\hat{F}_n^{*-1}(\tfrac{1}{2}) = x_{(1)}\right) = \sum_{k=\left\lceil \frac{1}{2}n \right\rceil}^{n}\binom{n}{k}\left(\frac{1}{n}\right)^k\left(\frac{n-1}{n}\right)^{n-k}$$

We now consider a final interesting example in which the nonparametric bootstrap *fails*. Consider $X_1,\cdots,X_n \sim U(0,\theta]$, and define the root

$$R_n\left(X;F_\theta\right)=\frac{n\left(\theta-X_{(n)}\right)}{\theta}$$

where $X_n=\max\left(X_1,\cdots,X_n\right)$.

We first find the distribution of $R_n\left(X;F\right)$

$$\mathbb{P}_F\left(R_n>x\right)=\mathbb{P}_F\left(X_{(n)}<\frac{(n-x)\theta}{n}\right)=\left(1-\frac{x}{n}\right)^n\to e^{-x}$$

So $R_n \sim \text{Exp}(1)$. Note, however, that if we use $\hat{F}_n$ as our distribution, then our best estimator of $\theta=x_{(n)}$, and so

$$R_n\left(X^*,\hat{F}_n\right)=\frac{n\left(x_{(n)}-X^*_{(n)}\right)}{x_{(n)}}$$

And so

$$\begin{aligned}\mathbb{P}_{\hat{F}_n}\left(R_n=0\right)&=\mathbb{P}\left(X^*_{(n)}=x_{(n)}\right)\\&=\mathbb{P}\left(x_{(n)}\in\text{bootstrap sample}\right)\\&=1-\mathbb{P}\left(x_{(n)}\notin\text{bootstrap sample}\right)\\&=1-\left[\mathbb{P}\left(x_{(n)}\neq i^{\text{th}}\text{ element in sample}\right)\right]^n\\&=1-\left(1-\frac{1}{n}\right)^n\\&\to 1-e^{-1}\end{aligned}$$

Clearly, therefore, the asymptotic distribution of the nonparametric bootstrap here does not tend to that of the root.

Let us consider, instead, the parametric bootstrap, which uses the MLE of $\hat{\theta}=x_{(n)}$:

$$R_n\left(X^*,F_{\hat{\theta}}\right)=\frac{n\left(\hat{\theta}-X^*_{(n)}\right)}{\hat{\theta}}$$

Now

$$\mathbb{P}_{F_{\hat{\theta}}}\left(R_n>x\right)=\mathbb{P}_{F_{\hat{\theta}}}\left(X^*_{(n)}<\frac{(n-x)\hat{\theta}}{n}\right)=\left(1-\frac{x}{n}\right)^n\to e^{-x}$$

Which does indeed tend to the distribution of the root.

Note the key difference between $\hat{F}_n$, which picks each value with equal probability, and $F_{\hat{\theta}}$, in which the probability of picking values in an interval depends on the value itself.

This behaviour is due to the non-standard asymptotics of the uniform distribution.

Bayesian Inference

Bayes' Theorem states that

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{L(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int L(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}$$

Where

- $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ is the *posterior distribution* of $\boldsymbol{\theta}$
- $L(\boldsymbol{x} \mid \boldsymbol{\theta})$ is the *likelihood*
- $p(\boldsymbol{\theta})$ is the *prior* on $\boldsymbol{\theta}$.

EXAMPLE: Consider an auto-regressive (times-series) model of order k, in which we observe data $\boldsymbol{x} = \left(x_1, \cdots, x_N\right)$, which, for $t > k + 1$ is generated by the process

$$x_t = \sum_{r=1}^{k} a_r x_{t-r} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

We can express this as

$$\boldsymbol{x}_k = X_k \boldsymbol{a} + \boldsymbol{\varepsilon}$$

Where

- $\boldsymbol{x}_k = \left(x_{k+1}, \cdots, x_N\right)^T$
- X_k is an $(N - k) \times k$ matrix with $(X_k)_{ij} = x_{k+i-j}$.
- $\boldsymbol{a} = \left(a_1, \cdots, x_k\right)$
- $\boldsymbol{\varepsilon} = \left(\varepsilon_{k+1}, \cdots, \varepsilon_N\right)^T$

We then have

$$L\left(\boldsymbol{x}_k \mid \boldsymbol{a}, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}(n-k)}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_k - X_k \boldsymbol{a}\right)^T \left(\sigma^2 I\right)^{-1} \left(\boldsymbol{x}_k - X_k \boldsymbol{a}\right)\right)$$

We might place the following priors on $\boldsymbol{a}$ and σ^2

$\boldsymbol{a} \sim N_k\left(\boldsymbol{\mu}_a, \Sigma_a\right)$

$$\Rightarrow p(\boldsymbol{a}) = \frac{1}{(2\pi)^{k/2}\sqrt{\left|\Sigma_a\right|}} \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}_a)^T \Sigma_a^{-1} (\boldsymbol{a} - \boldsymbol{\mu}_a)\right)$$

$\sigma^2 \sim \text{InverseGamma}(\alpha, \beta)$

$$\Rightarrow p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2}$$

We are interested here in calculating marginal summary statistics for $\boldsymbol{\theta}$. For example, the posterior expectation

$$\mathbb{E}_{\pi}\left\{\phi(\boldsymbol{\theta})\right\} = \int \phi(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{x})\ \mathrm{d}\boldsymbol{\theta}$$

Recall that, using Monte-Carlo integration, we can approximate

$$\mathbb{E}_{\pi}\left\{\phi(\boldsymbol{\theta})\right\} \approx \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{\theta}_i) \qquad \boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n \sim \pi(\boldsymbol{\theta} \mid \boldsymbol{x})$$

Unfortunately π may be difficult to sample from, since it is often high dimensional and of unfamiliar form. We resolve this problem as follows:

- We construct a Markov chain $\boldsymbol{\theta}^{(0)},\boldsymbol{\theta}^{(1)},\cdots$ which has π as its stationary distribution. In other words, such that
$$\left\|\mathcal{K}^t\left(\boldsymbol{\theta}^{(0)},?\right) - \pi(? \mid \boldsymbol{x})\right\| \to 0 \qquad \text{as } t \to \infty$$
Where $\mathcal{K}^t\left(\boldsymbol{\theta}^{(0)},?\right) = \mathbb{P}\left(\boldsymbol{\theta}^{(t)} = ? \mid \boldsymbol{\theta}^{(0)}\right)$.
- We run the chain until it reaches equilibrium.
- Further realisations can be regarded as a dependent sample from π.

This method is called **Markov Chain Monte Carlo**. We consider various methods to construct such a chain.

The Gibbs Sampler

Gibbs Sampling: Imagine the random vector we are sampling has p dimensions, and distribution $\pi(\boldsymbol{\theta})$. The Gibbs sampler samples from this distribution as follows:

1. Begin in some arbitrary state $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)},\cdots,\theta_p^{(0)}\right)$.
2. At time t, in state $\boldsymbol{\theta}^{(t)}$, update the state vector one components at a time:
$$\theta_1^{(t+1)} \sim \pi(\theta_1 \mid \cancel{\theta_1^{(t)}},\theta_2^{(t)},\cdots,\theta_p^{(t)})$$
$$\theta_2^{(t+1)} \sim \pi(\theta_2 \mid \theta_1^{(t+1)},\cancel{\theta_2^{(t)}},\theta_3^{(t)},\cdots,\theta_p^{(t)})$$
$$\theta_p^{(t+1)} \sim \pi(\theta_p \mid \theta_1^{(t+1)},\cdots,\theta_{p-1}^{(t+1)},\cancel{\theta_p^{(t)}})$$
3. Collect a total of T samples.
4. Discard the first b samples as "*burn-in*".
5. Treat $\left\{\boldsymbol{\theta}^{(t)}\right\}_{t=b+1}^{T}$ as a dependent sample from π.

The distribution of one component conditional on another is called the **full joint conditional distribution**.

EXAMPLE: Consider the AR model from above. The full joint conditions for $\boldsymbol{a}$ and σ^2 are

$$\begin{aligned}
\pi(\sigma^2 \mid \boldsymbol{x}_k, \boldsymbol{a}) &\propto L(\boldsymbol{x}_k \mid \boldsymbol{a}, \sigma^2)\pi(\sigma^2) \\
&\propto \frac{1}{(\sigma^2)^{\frac{1}{2}(N-k)}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_k - X_k\boldsymbol{a}\right)^T\left(\boldsymbol{x}_k - X_k\boldsymbol{a}\right)\right) \\
&\qquad (\sigma^2)^{-\alpha-1} e^{-\beta/\sigma^2} \\
&\propto \frac{1}{(\sigma^2)^{\frac{1}{2}(N-k)}} (\sigma^2)^{-\alpha-\frac{1}{2}(N-k)-1} \\
&\qquad \exp\left(-\frac{1}{\sigma^2}\left\{\frac{1}{2}\left(\boldsymbol{x}_k - X_k\boldsymbol{a}\right)^T\left(\boldsymbol{x}_k - X_k\boldsymbol{a}\right) + \beta\right\}\right) \\
&\sim \text{InverseGamma}\begin{pmatrix}\alpha + \frac{1}{2}(N-k), \\ \beta + \frac{1}{2}(\boldsymbol{x}_k - X_k\boldsymbol{a})^T(\boldsymbol{x}_k - X_k\boldsymbol{a})\end{pmatrix}
\end{aligned}$$

And

$$\begin{aligned}
\pi(\boldsymbol{a} \mid \boldsymbol{x}_k, \sigma^2) &= L(\boldsymbol{x}_k \mid \boldsymbol{a}, \sigma^2)\pi(\boldsymbol{a}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_k - X_k\boldsymbol{a}\right)^T\left(\boldsymbol{x}_k - X_k\boldsymbol{a}\right)\right) \\
&\qquad \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}_a)^T\Sigma_a^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_a)\right)
\end{aligned}$$

Since we want a quadratic in $\boldsymbol{a}$, we write

$$\begin{aligned}
\pi(\boldsymbol{a} \mid \boldsymbol{x}_k, \sigma^2) &\propto \exp\left(-\frac{1}{2}\begin{Bmatrix}\boldsymbol{a}^T\left(-\sigma^{-2}X_k^TX_k + \Sigma_a^{-1}\right)\boldsymbol{a} \\ -2\boldsymbol{a}^T\left(\sigma^{-2}X_k^T\boldsymbol{x}_k + \Sigma_a^{-1}\boldsymbol{\mu}_a\right)\end{Bmatrix}\right) \\
&\propto \exp\left(-\frac{1}{2}\left\{\boldsymbol{a}^T\Sigma^{-1}\boldsymbol{a} - 2\boldsymbol{a}^T\Sigma^{-1}\boldsymbol{\mu}\right\}\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{a} - \boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\boldsymbol{a} - \boldsymbol{\mu}\right)\right) \\
&\sim N_k\left(\boldsymbol{\mu}, \Sigma\right)
\end{aligned}$$

With

$$\Sigma^{-1} = \sigma^{-2}X_k^TX_k + \Sigma_a^{-1} \qquad \boldsymbol{\mu} = \Sigma(\sigma^{-2}X_k^T\boldsymbol{x}_k + \Sigma_a^{-1}\boldsymbol{\mu}_a)$$

The Gibbs sampler can be used to great effect when

- There is some missing data

- The likelihood is difficult to use in its native form, but easy to use when conditional on some unobserved data (which we can treat as "missing" data).

In both the cases above, we can explicitly calculate $L\left(y_{\text{obs}}, y_{\text{miss}} \mid \theta\right)$. We then simulate θ and y_{miss} together using a Gibbs sampler, with

$$\pi(\theta, y_{\text{miss}} \mid y_{\text{obs}}) \propto L\left(y_{\text{obs}}, y_{\text{miss}} \mid \theta\right) p(\theta)$$

Of course, in reality, the distribution we are interested in is

$$\pi(\theta \mid y_{\text{obs}}) = \int \pi(\theta, y_{\text{miss}} \mid y_{\text{obs}}) \, \mathrm{d}y_{\text{miss}}$$

The values of θ returned by the Gibbs sampler, however, have precisely that distribution, because by ignoring the values of y_{miss} that the Gibbs sampler returns, we are effectively "integrating over" y_{miss}.

EXAMPLE: Consider a group of N animals, each assigned to one of four categories

$$\left(y_1, y_2, y_3, y_4\right)$$

with probabilities

$$\left(\tfrac{2+\theta}{4}, \tfrac{1-\theta}{4}, \tfrac{1-\theta}{4}, \tfrac{\theta}{4}\right)$$

The likelihood is then multinomial

$$L\left(\boldsymbol{y} \mid \theta\right) \propto \left(\tfrac{2+\theta}{4}\right)^{y_1} \left(\tfrac{1-\theta}{4}\right)^{y_2+y_3} \left(\tfrac{\theta}{4}\right)^{y_4}$$

Suppose we place a $\text{Beta}(\alpha, \beta)$ prior on θ. Then

$$\begin{aligned} \pi(\theta \mid \boldsymbol{y}) &\propto L\left(\boldsymbol{y} \mid \theta\right) p(\theta) \\ &\propto \left(\tfrac{2+\theta}{4}\right)^{y_1} \left(\tfrac{1-\theta}{4}\right)^{y_2+y_3} \left(\tfrac{\theta}{4}\right)^{y_4} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \left(2+\theta\right)^{y_1} \left(1-\theta\right)^{y_2+y_3+\beta-1} \theta^{y_4+\alpha-1} \end{aligned}$$

This is hard to sample from.

Consider, instead, splitting our data into the following *five* groups

$$\left(y_1 - z, z, y_2, y_3, y_4\right)$$

With probabilities

$$\left(\tfrac{1}{2}, \tfrac{\theta}{4}, \tfrac{1-\theta}{4}, \tfrac{1-\theta}{4}, \tfrac{\theta}{4}\right)$$

We then have

$$\begin{aligned}\pi(\theta, z \mid \boldsymbol{y}) &\propto L\Big(\boldsymbol{y} \mid \theta, z\Big) p(\theta) \\ &\propto \tfrac{N!}{z!(y_1-z)!\cdots}\Big(\tfrac{1}{2}\Big)^{y_1-z}\Big(\tfrac{\theta}{4}\Big)^{z}\Big(\tfrac{1-\theta}{4}\Big)^{y_2+y_3}\Big(\tfrac{\theta}{4}\Big)^{y_4}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto {}^{y_1}C_{z}\Big(\tfrac{1}{2}\Big)^{y_1-z}\Big(\tfrac{\theta}{4}\Big)^{z}\theta^{y_4+\alpha-1}\Big(1-\theta\Big)^{y_2+y_3+\beta-1}\end{aligned}$$

[Note – it is important, in this case, to keep the factor of $\theta\,/\,4$ intact, because getting rid of the ¼ would imply getting rid of a factor of $1/4^z$. Since, however, we will be needing the distribution of z, this is not legitimate].

And so

$$\theta \mid z, \boldsymbol{y} \sim \text{Beta}\Big(z + y_4 + \alpha, y_2 + y_3 + \beta\Big)$$

Finding the joint conditional for z is slightly harder. Here are two ways to do it

1. Note that

   $$\pi(\theta, z \mid \boldsymbol{y}) \propto {}^{y_1}C_{z}\Big(\tfrac{1}{2}\Big)^{y_1-z}\Big(\tfrac{\theta}{4}\Big)^{z}$$

   This looks like a binomial, but ½ and $\theta\,/4$ don't sum to 1. We'll therefore try and find something we can multiply both of them by to get something that *does* sum to 1.

   $$\tfrac{1}{2}?+\tfrac{\theta}{4}? = 1$$

   $$? = \frac{1}{\frac{\theta}{4}+\frac{1}{2}} = \frac{4}{2+\theta}$$

   And so

   $$\pi(\theta, z \mid \boldsymbol{y}) \propto {}^{y_1}C_{z}\Big(\tfrac{2}{2+\theta}\Big)^{y_1-z}\Big(\tfrac{\theta}{2+\theta}\Big)^{z}$$

   As such

   $$z \mid \theta, \boldsymbol{y} \sim \text{Bin}\Big(y_1, \tfrac{\theta}{2+\theta}\Big)$$

2. *Or*, we can simply note that once θ is known, the conditional probability of an observation being in z <u>given</u> it's in y_1 is

   $$\frac{\mathbb{P}\Big(\in z\Big)}{\mathbb{P}\Big(\in y_1\Big)} = \frac{\frac{\theta}{4}}{\frac{1}{2}+\frac{\theta}{4}} = \frac{\theta}{2+\theta}$$

   Which yields the same result.

The Metropolis-Hastings Algorithm

The Gibbs sampler is a special case of the Metropolis-Hastings (MH) algorithm. The algorithm basically samples values from an "approximate" distribution and then "corrects" these so that they asymptotically behave as if they came from the stationary distribution.

The Metropolis Hastings (MH) algorithm:

1. Begin with some arbitrary state $\boldsymbol{\theta}^{(0)}$.
2. Simulate $\boldsymbol{\phi} \sim q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi})$; q is our "approximate transition probability" from $\boldsymbol{\theta}^{(0)}$ to $\boldsymbol{\phi}$.
3. Let
$$\alpha\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}\right) = \min\left\{1, \frac{\pi(\boldsymbol{\phi} \mid \boldsymbol{x}) q(\boldsymbol{\phi}, \boldsymbol{\theta}^{(t)})}{\pi(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x}) q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi})}\right\}$$
4. Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\phi}$ with probability $\alpha\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}\right)$, or else reject it and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

So the Markov transition Kernel for the chain is given by

$$\mathcal{P}_{H}\left(\boldsymbol{\theta}, B\right) = \underbrace{\int_{B} \mathcal{K}_{H}(\boldsymbol{\theta}, \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi}}_{\substack{\text{Moving forward to} \\ \text{some } \boldsymbol{\phi} \in B}} + \underbrace{r(\boldsymbol{\theta}) I_{B}(\boldsymbol{\theta})}_{\substack{\text{Staying, if it} \\ \text{so happened that} \\ \boldsymbol{\theta} \text{ was already } \in B}} + \underbrace{0}_{\substack{\text{We neither move} \\ \text{to any state in } B, \\ \text{nor do we remain} \\ \text{there}}}$$

Where

$$\mathcal{K}_{H}(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}, \boldsymbol{\phi}) \alpha(\boldsymbol{\theta}, \boldsymbol{\phi})$$
$$r_{B}(\boldsymbol{\theta}) = 1 - \int_{\forall \boldsymbol{x}} \mathcal{K}_{H}(\boldsymbol{\theta}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

This leads to values that behave as if they come from $\pi(\boldsymbol{\theta})$.

Proof: First, note that

$$\begin{aligned}
\pi(x)\mathcal{K}_{H}(x, y) &= \pi(x) q(x, y) \min\left\{1, \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\right\} \\
&= \min\left(\pi(x) q(x, y), \pi(y) q(y, x)\right) \\
&= \min\left\{\frac{\pi(x) q(x, y)}{\pi(y) q(y, x)}, 1\right\} \pi(y) q(y, x) \\
&= \pi(y) \mathcal{K}_{H}(y, x)
\end{aligned}$$

This implies that

$$\begin{aligned}
\int_{\forall x} \pi(x) \mathcal{K}_{H}(x, y) \, \mathrm{d}x &= \pi(y) \int_{\forall x} \mathcal{K}_{H}(y, x) \, \mathrm{d}x \\
&= \pi(y)\left[1 - r(y)\right]
\end{aligned}$$

With this key result in hand, we are now ready to prove our theorem. Effectively, we want to show that if we sample $\boldsymbol{x}$ from the stationary distribution, the probability of the Markov chain moving us from $\boldsymbol{x}$ to any element in B is the same as the probability of being in B under the posterior distribution. In other words, we want to show that

$$\int_{\forall x} \pi(x)\mathcal{P}_{H}(x,B)\ \mathrm{d}x = \int_{B} \pi(y)\ \mathrm{d}y$$

To do that, consider

$$\begin{aligned}
&\int_{\forall x} \pi(x)\mathcal{P}_{H}(x,B)\ \mathrm{d}x \\
&\quad = \int_{\forall x} \pi(x)\Big[\int_{B} \mathcal{K}_{H}(x,y)\ \mathrm{d}y + r(x)I_{B}(x)\Big]\ \mathrm{d}x \\
&\quad = \int_{\forall x}\int_{B} \pi(x)\mathcal{K}_{H}(x,y)\ \mathrm{d}y\ \mathrm{d}x + \int_{\forall x} \pi(x)r(x)I_{B}(x)\ \mathrm{d}x \\
&\quad = \int_{B}\int_{\forall x} \pi(x)\mathcal{K}_{H}(x,y)\ \mathrm{d}x\ \mathrm{d}y + \int_{B} \pi(x)r(x)\ \mathrm{d}x
\end{aligned}$$

We can now use our previously derived result for $\int_{\forall x} \pi(x)\mathcal{K}_{H}(x,y)$, which gives

$$\begin{aligned}
&\int_{\forall x} \pi(x)\mathcal{P}_{H}(x,B)\ \mathrm{d}x \\
&\quad = \int_{B} \pi(y)\Big[1 - r(y)\Big]\ \mathrm{d}y + \int_{B} \pi(x)r(x)\ \mathrm{d}x \\
&\quad = \int_{B} \pi(y)\ \mathrm{d}y
\end{aligned}$$

As required. ■

Note: A great advantage of the MH algorithm over Gibbs sampling is that we do not need to know the normalisation constant of π. Also, we do not need to know any of the joint conditionals.

The acceptance function might be easier to understand when written as follows

$$\alpha\Big(\boldsymbol{\theta}^{(t)},\boldsymbol{\phi}\Big) = \min\left(1, \frac{\pi(\boldsymbol{\phi} \mid \boldsymbol{x})}{\pi(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x})} \div \frac{q(\boldsymbol{\theta}^{(t)},\boldsymbol{\phi})}{q(\boldsymbol{\phi},\boldsymbol{\theta}^{(t)})}\right)$$

This makes it clearer that it is indeed in the form (actual density / proposal density) that we saw in importance sampling.

Note that when the proposal distribution q is symmetric, it is called a *symmetric* (or *Metropolis*) proposal. In that case, $q(\boldsymbol{\phi},\boldsymbol{\theta}) = q(\boldsymbol{\theta},\boldsymbol{\phi})$ and so the acceptance probability reduces to

$$\alpha(\boldsymbol{\theta},\boldsymbol{\phi}) = \min\left\{1, \frac{\pi(\boldsymbol{\phi} \mid \boldsymbol{x})}{\pi(\boldsymbol{\theta} \mid \boldsymbol{x})}\right\}$$

There are a number of possible choices for the proposal distribution q. Here are two common examples:

- **Random-Walk (RW) Metropolis**

  Here, we specify

  $$\boldsymbol{\phi} = \boldsymbol{\theta} + \boldsymbol{z} \qquad\qquad \boldsymbol{z} \sim f$$

  Common choices for f may include a uniform distribution or a multivariate normal. The distribution is often – but not always – chosen to be symmetric. In RW metropolis, we are effectively choosing our new value to be close to the current one, the reasoning being that when the stationary distribution has been reached, we're more likely to spend longer at points of high density.

- **Independence Sampler**

  Here, the candidate observation is drawn independently of the current state, so that $q\left(\boldsymbol{\theta},\boldsymbol{\phi}\right) = g(\boldsymbol{\phi})$. The corresponding acceptance probability can be written

  $$\alpha(\boldsymbol{\theta},\boldsymbol{\phi}) = \min\left\{1, \frac{\omega(\boldsymbol{\phi})}{\omega(\boldsymbol{\theta})}\right\} \qquad\qquad \omega(\cdot) = \frac{\pi(\cdot \mid \boldsymbol{x})}{g(\cdot)}$$

  This is precisely the importance weight function that would be used in importance sampling, given observations from g being used to sample from π.

Theorem: The Gibbs sampler is a special case of the Metropolis Hastings algorithm.

Proof: Suppose we have a current estimate of our parameters $\boldsymbol{\theta} = \left(\theta_1, \cdots, \theta_p\right)$. We break each iteration of the MH algorithm into steps, each of which update a single value of $\boldsymbol{\theta}$.

1. Start by setting $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}$. Set $t = 1$.

2. Propose a new value ϕ for a single component θ_t, so produce a new vector
$$\boldsymbol{\theta}^{(t)} = \left(\theta_1^{(t-1)},\cdots,\theta_{t-1}^{(t-1)},\phi,\theta_{t+1}^{(t-1)}\cdots,\theta_p^{(t-1)}\right)$$
with proposal density
$$q_t(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\theta}^{(t)}) = \pi(\phi \mid \boldsymbol{\theta}_{(-t)}^{(t-1)},\boldsymbol{x}) = \pi(\theta_t^{(t)} \mid \boldsymbol{\theta}_{(-t)}^{(t-1)},\boldsymbol{x})$$
Where $\boldsymbol{\theta}_{(-t)}^{(t-1)} = \left(\theta_1^{(t-1)},\cdots,\cancel{\theta_t^{(t-1)}},\cdots,\theta_p^{(t-1)}\right)$. Denote
3. Accept the proposal with acceptance probability $\alpha(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\theta}^{(t)})$.
    a. If it is accepted, keep $\boldsymbol{\theta}^{(t)}$.
    b. Otherwise, keep the previous iteration $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$.

All we need to do is to show that the acceptance probability in this case is always 1. In that case, the algorithm above is equivalent to the Gibbs sampler. Now, $\alpha(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\theta}^{(t)}) = \min\left\{1, A_t\right\}$, where

$$\begin{aligned} A_t &= \frac{\pi(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x}) q_t(\boldsymbol{\theta}^{(t)},\boldsymbol{\theta}^{(t-1)})}{\pi(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{x}) q_t(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\theta}^{(t)})} \\ &= \frac{\pi(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x}) \pi(\theta_t^{(t-1)} \mid \boldsymbol{\theta}_{(-t)}^{(t)},\boldsymbol{x})}{\pi(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{x}) \pi(\theta_t^{(t)} \mid \boldsymbol{\theta}_{(-t)}^{(t-1)},\boldsymbol{x})} \\ &= \frac{\pi(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x}) \,/\, \pi(\theta_t^{(t)} \mid \boldsymbol{\theta}_{(-t)}^{(t-1)},\boldsymbol{x})}{\pi(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{x}) \,/\, \pi(\theta_t^{(t-1)} \mid \boldsymbol{\theta}_{(-t)}^{(t)},\boldsymbol{x})} \end{aligned}$$

Note, however, that $\boldsymbol{\theta}_{(-t)}^{(t)} = \boldsymbol{\theta}_{(-t)}^{(t-1)}$, so

$$\begin{aligned} A_t &= \frac{\pi(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x}) \,/\, \pi(\theta_t^{(t)} \mid \boldsymbol{\theta}_{(-t)}^{(t)},\boldsymbol{x})}{\pi(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{x}) \,/\, \pi(\theta_t^{(t-1)} \mid \boldsymbol{\theta}_{(-t)}^{(t-1)},\boldsymbol{x})} \\ &= \frac{\pi\left(\boldsymbol{\theta}_{(-t)}^{(t)}\right)}{\pi\left(\boldsymbol{\theta}_{(-t)}^{(t-1)}\right)} \\ &= 1 \end{aligned}$$

(To go from the first line to the second line, use Bayes' Theorem on conditional π probabilities). As required. ■

It is generally accepted that if one wants T independent samples from π, but must instead resort to obtaining T depend samples (eg: via MCMC), then

somehow, the resulting effective sample size (relative to independent sampling) is less than T, due to the autocorrelation of the chain.

Radford Neal defined

Effective sample size due to autocorrelation: (for marginally scalar samples)

$$\mathrm{ESS}(\boldsymbol{\theta}) = \frac{T}{1 + 2\Sigma_{\ell=1}^{T-1}\hat{\rho}(\ell)}$$

Where $\hat{\rho}(\ell)$ is the sample autocorrelation at lag ℓ:

$$\hat{\rho}(\ell) = \frac{\hat{\gamma}(\ell)}{\hat{\gamma}(0)}$$

where

$$\hat{\gamma}(\ell) = \frac{1}{T-\ell}\sum_{t=1}^{T-\ell}\left(\theta^{(t)} - \bar{\theta}\right)\left(\theta^{(t+\ell)} - \bar{\theta}\right)$$

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$$

Model uncertainty

So far, our algorithms have considered *fixed* parameter spaces. What if, however, there is some uncertainty as to which model is the "correct" one from a set $\{\mathcal{M}_1, \cdots, \mathcal{M}_k\}$? We can include the model as an additional parameter to be estimated

$$\pi(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M} \mid \boldsymbol{x}) \propto L(\boldsymbol{x} \mid \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}) p(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}) p(\mathcal{M})$$

The idea is then to construct a Markov chain that is able to move between different models, and has stationary distribution π , to generate samples $\{\mathcal{M}^{(t)}, \boldsymbol{\theta}^{(t)}_{\mathcal{M}^{(t)}}\}_{t=1}^{T}$.

Using these samples, here are two examples of statistics we might be interested in calculating:

1. The posterior probability of each model

$$\pi\left(\mathcal{M}_i \mid \boldsymbol{x}\right) = \frac{L(\boldsymbol{x} \mid \mathcal{M}_i) p(\mathcal{M}_i)}{\Sigma_{i=1}^{k} L(\boldsymbol{x} \mid \mathcal{M}_i) p(\mathcal{M}_i)}$$

where

$$L(\boldsymbol{x} \mid \mathcal{M}_i) = \int L(\boldsymbol{x} \mid \mathcal{M}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) p(\boldsymbol{\theta}_{\mathcal{M}_i} \mid \mathcal{M}_i) \, \mathrm{d}\boldsymbol{\theta}_{\mathcal{M}_i}$$

This can be calculated, using $\{\mathcal{M}^{(t)}, \boldsymbol{\theta}^{(t)}_{\mathcal{M}^{(t)}}\}_{t=1}^{T}$, as

$$\hat{\pi}\left(\mathcal{M}_i \mid \boldsymbol{x}\right) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\left\{\mathcal{M}^{(t)}=\mathcal{M}_i\right\}}$$

2. Estimates concerning the parameters themselves, based on the distribution

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{k} \pi(\boldsymbol{\theta} \mid \boldsymbol{x}, \mathcal{M}_i)\pi(\mathcal{M}_i \mid \boldsymbol{x})$$

We can calculate such statistics by simply ignoring the model part of our Markov chain. For example

$$\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{\theta}^{(t)}$$

$$\mathrm{var}\left(\boldsymbol{\theta}\right) = \frac{1}{T}\sum_{i=1}^{T}\left(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}\right)^2$$

Of course, this only works if $\boldsymbol{\theta}$ is a statistic that is common to all the models (for example, the mean – see the example below).

The algorithm we use to generate our desired Markov Chain is called the **reversible jump algorithm** (some written as RJ-MCMC).

Reversible-Jump Monte Carlo:

1. Given a model $\mathcal{M}$, update the parameters $\boldsymbol{\theta}_{\mathcal{M}}$ using MH or GS.
2. Reversible jump (RJ)-step: with probability $P\left(\mathcal{M} \to \mathcal{M}'\right)$ [which must be chosen], propose to replace the model $\mathcal{M}$ with a new model $\mathcal{M}'$. This proposal can be accepted or rejected.

The RJ step is complicated because it involves proposing new parameters $\boldsymbol{\theta}_{\mathcal{M}'}$ for the new model. Imagine the move we are proposing is

$$\left\{\mathcal{M}, \boldsymbol{\theta}_{\mathcal{M}}\right\} \to \left\{\mathcal{M}', \boldsymbol{\theta}_{\mathcal{M}'}\right\}$$

Our approach will be to choose

$$\left(\boldsymbol{\theta}_{\mathcal{M}'}, \boldsymbol{u}'\right) = g_{\mathcal{M},\mathcal{M}'}\left(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u}\right)$$

Where

- g is a known, deterministic, bijective function[5] between $\left(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u}\right)$ and $\left(\boldsymbol{\theta}_{\mathcal{M}'}, \boldsymbol{u}'\right)$.
- $\boldsymbol{u}$ and $\boldsymbol{u}'$ are vectors chosen such that

$$\dim \boldsymbol{u} + \dim\left(\boldsymbol{\theta}_{\mathcal{M}}\right) = \dim \boldsymbol{u}' + \dim\left(\boldsymbol{\theta}_{\mathcal{M}'}\right)$$

  In other words, it "matches" the dimensions of the parameters. Either one or both of $\boldsymbol{u}$ and $\boldsymbol{u}'$ will be 0, depending on the size of the models.

  If $\dim(\boldsymbol{\theta}_{\mathcal{M}'}) > \dim(\boldsymbol{\theta}_{\mathcal{M}})$, $\boldsymbol{u}$ is nonzero. We choose it by sampling from a proposal density $q_{\mathcal{M},\mathcal{M}'}(\boldsymbol{u})$

We then accept the move with probability $\alpha\left(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{\theta}_{\mathcal{M}'}\right) = \min\left\{1, A\right\}$

$$A = \underbrace{\frac{\pi(\mathcal{M}', \boldsymbol{\theta}_{\mathcal{M}'} \mid \boldsymbol{x})}{\pi(\mathcal{M}, \boldsymbol{\theta}_{\mathcal{M}} \mid \boldsymbol{x})}}_{\substack{\text{Model ratio} \\ \left\{=\frac{L(\boldsymbol{x}\mid\mathcal{M}', \boldsymbol{\theta}_{\mathcal{M}'})p(\boldsymbol{\theta}_{\mathcal{M}'}\mid\mathcal{M}')p(\mathcal{M}')}{L(\boldsymbol{x}\mid\mathcal{M}, \boldsymbol{\theta}_{\mathcal{M}})p(\boldsymbol{\theta}_{\mathcal{M}}\mid\mathcal{M})p(\mathcal{M})}\right\}}} \cdot \underbrace{\frac{P\left(\mathcal{M}' \to \mathcal{M}\right) q\left(\boldsymbol{u}'\right)}{P\left(\mathcal{M} \to \mathcal{M}'\right) q\left(\boldsymbol{u}\right)} \left| \frac{\partial g(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u})}{\partial\left(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u}\right)} \right|}_{\text{Proposal ratio}}$$

Note that

$$\left(\frac{\partial g(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u})}{\partial\left(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u}\right)}\right)_{ij} = \left(\frac{\partial(\boldsymbol{\theta}_{\mathcal{M}'}, \boldsymbol{u}')}{\partial(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u})}\right)_{ij} = \frac{\partial\left((\boldsymbol{\theta}_{\mathcal{M}'}, \boldsymbol{u}')_i\right)}{\partial\left((\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u})_j\right)}$$

Where $\left(\boldsymbol{\theta}_{\mathcal{M}}, \boldsymbol{u}\right)$ is the vector formed by combining $\boldsymbol{\theta}_{\mathcal{M}}$ and $\boldsymbol{u}$.

Further note that the reverse move from $\left\{\mathcal{M}', \boldsymbol{\theta}_{\mathcal{M}'}\right\} \to \left\{\mathcal{M}, \boldsymbol{\theta}_{\mathcal{M}}\right\}$ is fully defined by g^{-1} and $\alpha\left(\boldsymbol{\theta}_{\mathcal{M}'}, \boldsymbol{\theta}_{\mathcal{M}}\right) = \min\left\{1, A^{-1}\right\}$.

EXAMPLE: Suppose that we observe data $\boldsymbol{x}$ which are IID, but where the distribution is unknown; either Exp(λ) or Gamma(α, β), with all parameters unknown[6]. We let $\mathbb{P}\left(\text{Exp}\right)$ and $\mathbb{P}\left(\text{G}\right)$ be the prior probability on each model, and let the priors on the parameters be

$$\lambda \sim \text{Gamma}(a_1, b_1) \qquad \alpha \sim \text{Gamma}(a_2, b_2) \qquad \beta \sim \text{Gamma}(a_3, b_3)$$

Let's find posterior distributions

Exponential model

[5] A bijective function f from a set X to a set Y has the property that for every y in Y, there is exactly one x in X such that $f(x) = y$ and no unmapped element exists in either X or Y.

[6] This example is really quite silly, because Gamma$(1, \lambda) \sim$ Exp(λ), but it's useful for demonstration.

[Reminder: we need all the constants in the original distribution, because λ is something we'll want the distribution of!]

$$L\left(\boldsymbol{x} \mid \lambda, \text{Exp}\right) = \prod_{i=1}^{n} \lambda \exp\left(-\lambda x_i\right) = \lambda^n \exp\left(-\lambda x_+\right)$$

$$\Downarrow$$

$$\pi(\lambda \mid \boldsymbol{x}) \propto L(\boldsymbol{x} \mid \lambda, \text{Exp}) p(\lambda) \propto \lambda^{n+a_1-1} e^{-\lambda x_+ - \lambda b_1}$$

$$\boxed{\lambda \mid \boldsymbol{x} \sim \text{Gamma}\left(n + a_1, x_+ + b_1\right)}$$

Gamma Model

[Reminder: we need all the constants in the original distribution, because α and β are things we'll want the distribution of later! The Bayes' factors, however, integrate out the relevant variables by definition, and so can be ignored]

$$L\left(\boldsymbol{x} \mid \alpha, \beta, \text{G}\right) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \exp\left(-\beta x_i\right) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n x_i^{n(\alpha-1)} \exp\left(-\beta x_+\right)$$

$$\Downarrow$$

$$\pi(\beta \mid \boldsymbol{x}, \alpha) \propto L\left(\boldsymbol{x} \mid \alpha, \beta, \text{G}\right) p(\beta) \propto \beta^{\alpha n + a_3 - 1} \exp\left(-\beta x_+ - b_3 \beta\right)$$

$$\boxed{\beta \mid \boldsymbol{x}, \alpha \sim \text{Gamma}\left(\alpha n + a_3, x_+ + b_3\right)}$$

$$\Downarrow$$

$$\pi(\alpha \mid \boldsymbol{x}, \beta) \propto L\left(\boldsymbol{x} \mid \alpha, \beta, \text{G}\right) p(\alpha)$$

$$\boxed{\pi(\alpha \mid \boldsymbol{x}, \beta) \propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n x_i^{n(\alpha-1)} \alpha^{a_2 - 1} \exp\left(-b_2 \alpha\right)}$$

At each step, the first two distributions can be sampled from using a GS or the MH algorithm. We given an example for α and β. The latter can be sampled from using a simple Gibbs sampler. The former needs MH:

- Start with a value of $\alpha^{(0)}$
- Sample $\beta^{(t+1)}$ from $\text{Gamma}\left(\alpha^{(t)} n + a_3, x_+ + b_3\right)$
- Simulate ϕ from $U[\frac{3}{4}\alpha, \frac{4}{3}\alpha]$ (this is our proposal density).
- Calculate the acceptance probability

$$A\left(\alpha^{(t)}, \phi\right) = \min\left(1, \frac{\pi(\phi \mid \beta^{(t)}, \boldsymbol{x}) \frac{1}{\left(\frac{4}{3} - \frac{3}{4}\right)\phi} \mathbb{I}_{\alpha^{(t)} \in \left[\frac{3}{4}\phi, \frac{4}{3}\phi\right]}}{\pi(\alpha^{(t)} \mid \beta^{(t)}, \boldsymbol{x}) \frac{1}{\left(\frac{4}{3} - \frac{3}{4}\right)\alpha^{(t)}} \mathbb{I}_{\phi \in \left[\frac{3}{4}\alpha^{(t)}, \frac{4}{3}\alpha^{(t)}\right]}}\right)$$

With probability A, set $\alpha^{(t+1)} = \phi$, and with probability $1 - A$, set $\alpha^{(t+1)} = \alpha^{(t)}$.

We can now consider the RJ step. In each case, we will "propose" to move with probability 1 (so $P(\text{Exp} \to \text{Gamma}) = P(\text{Gamma} \to \text{Exp}) = 1$).

Let us now consider the move from Exp to Gamma. We must deal with the augmentation of the parameter space, from $(\lambda) \to (\alpha, \beta)$.

- The target parameter space is larger, so we set $u' = 0$, and we sample $u \sim \text{Gamma}(\gamma, \delta)$.
- We define our function g so that the mean of the original Exp distribution and of the resulting Gamma distribution is the same. This gives
$$g(\lambda, u) \to (\alpha = u, \beta = \lambda u)$$
- We then have
$$\left|\frac{\partial(\alpha,\beta)}{\partial(\lambda,u)}\right| = \begin{vmatrix} \partial\alpha / \partial\lambda & \partial\alpha / \partial u \\ \partial\beta / \partial\lambda & \partial\beta / \partial u \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ u & \lambda \end{vmatrix} = u$$

The probability of accepting the model move is then given by $\min\{1, A\}$, where

$$\begin{aligned} A &= \frac{\pi(\text{G}, \alpha, \beta \mid \boldsymbol{x})}{\pi(\text{Exp}, \lambda \mid \boldsymbol{x})} \frac{P\big(\text{G} \to \text{Exp}\big)}{P\big(\text{Exp} \to \text{G}\big) q\big(u\big)} \left|\frac{\partial(\alpha,\beta)}{\partial\big(\lambda,u\big)}\right| \\ &= \frac{L(\boldsymbol{x} \mid \text{G}, \alpha, \beta) p(\alpha) p(\beta) \mathbb{P}(\text{G})}{L(\boldsymbol{x} \mid \lambda, \text{Exp}) p(\lambda) \mathbb{P}(\text{Exp})} \frac{u}{\frac{1}{\Gamma(\gamma)} \delta^{\gamma} u^{\gamma-1} e^{-\delta u}} \end{aligned}$$

The reverse step is defined by $(\lambda, u) = g^{-1}(\alpha, \beta) \Rightarrow u = \alpha, \lambda = \beta / \alpha$ and probability A^{-1}.

EXAMPLE: Suppose that we observe data $\boldsymbol{x}$ which are IID, but where the distribution is unknown; either Exp(λ) or Pareto(α, β), with all parameters unknown. We let $\mathbb{P}\big(\text{Exp}\big)$ and $\mathbb{P}\big(\text{Par}\big)$ be the prior probability on each model, and let the priors on the parameters be

$$\lambda \sim \text{Gamma}(a, b) \qquad \alpha \sim \text{Exp}(\mu) \quad \beta \sim \text{Exp}(\nu)$$

Let's find posterior distributions

Exponential model

$$L\left(\boldsymbol{x} \mid \lambda, \text{Exp}\right) = \prod_{i=1}^{n} \lambda \exp\left(-\lambda x_i\right) = \lambda^n \exp\left(-\lambda x_+\right)$$

$$\Downarrow$$

$$\pi(\lambda \mid \boldsymbol{x}) \propto L(\boldsymbol{x} \mid \lambda, \text{Exp}) p(\lambda) \propto \lambda^{n+a-1} e^{-\lambda x_+ - \lambda b}$$

$$\boxed{\lambda \mid \boldsymbol{x} \sim \text{Gamma}\left(n + a, x_+ + b\right)}$$

Pareto Model

The Pareto distribution comes with an additional complication, in that it is imperative that $0 < \alpha \leq x$. Thus

$$L\left(\boldsymbol{x} \mid \alpha, \beta, \text{Par}\right) = \prod_{i=1}^{n} \beta \alpha^{\beta} x_i^{-(\beta+1)} = \beta^n \alpha^{n\beta} \prod_{i=1}^{n} x_i^{-(\beta+1)}$$

$$\Downarrow$$

$$\pi(\alpha \mid \beta, \boldsymbol{x}) \propto \beta^n \alpha^{n\beta} \left(\prod_{i=1}^{n} x_i^{-(\beta+1)}\right) \mu e^{-\mu\alpha} \mathbb{I}_{\{\alpha \in (0, x_{\min}]\}}$$

$$\boxed{\pi(\alpha \mid \beta, \boldsymbol{x}) \propto \alpha^{n\beta} e^{-\mu\alpha} \mathbb{I}_{\{\alpha \in (0, x_{\min}]\}}}$$

$$\Downarrow$$

$$\pi(\beta \mid \alpha, \boldsymbol{x}) \propto \beta^n \alpha^{n\beta} \left(\prod_{i=1}^{n} x_i^{-(\beta+1)}\right) \nu e^{-\nu\beta}$$

$$\pi(\beta \mid \alpha, \boldsymbol{x}) \propto \beta^n \exp\left(-\nu\beta + n\beta \log \alpha - \beta \Sigma_{i=1}^{n} \log x_i\right)$$

$$\boxed{\beta \mid \alpha, \boldsymbol{x} \sim \text{Gamma}\left(n + 1, \nu - n \log \alpha + \Sigma_{i=1}^{n} \log x_i\right)}$$

Once again, the first distribution is not in standard form, and therefore requires an MH sampler. We can use a symmetric random walk MH algorithm as follows:

- Given an α, generate $\alpha' = \alpha + u$, where $u \sim U[0, \sigma^2]$
- Automatically reject α' if $\alpha' \notin (0, x_{\min}]$
- Otherwise, accept it with probability

$$\min\left(1, \frac{\pi(\alpha' \mid \beta, \boldsymbol{x})}{\pi(\alpha \mid \beta, \boldsymbol{x})}\right) = \min\left(1, \frac{\alpha'^{n\beta} \exp\left(-\mu\alpha'\right)}{\alpha^{n\beta} \exp\left(-\mu\alpha\right)}\right)$$

We can now consider the RJ step. In each case, we will "propose" to move with probability 1 (so $P(\text{Exp} \to \text{Par}) = P(\text{Par} \to \text{Exp}) = 1$).

Let us now consider the move from Exp to Pareto. We must deal with the augmentation of the parameter space, from $(\lambda) \rightarrow (\alpha, \beta)$.

- The target parameter space is larger, so we set $u' = 0$, and we sample $u \sim \text{Exp}(\gamma)$.
- We define our function g so that the mean of the original Exp distribution and of the resulting Parto distribution is the same. This gives
$$g(\lambda, u) \rightarrow \left(\alpha = \tfrac{u}{\lambda(u+1)}, \beta = u+1\right)$$
- We then have
$$\begin{aligned}\left|\frac{\partial(\alpha,\beta)}{\partial(\lambda,u)}\right| &= \begin{vmatrix}\partial\alpha / \partial\lambda & \partial\alpha / \partial u \\ \partial\beta / \partial\lambda & \partial\beta / \partial u\end{vmatrix} \\ &= \begin{vmatrix}-\frac{u}{\lambda^2(u+1)} & \frac{1}{\lambda(u+1)} - \frac{u}{\lambda(u+1)^2} \\ 0 & 1\end{vmatrix} \\ &= \frac{u}{\lambda^2(u+1)}\end{aligned}$$

The probability of accepting the model change is then given by $\min\{1, A\}$, where

$$\begin{aligned}A &= \frac{\pi(\text{Par}, \alpha, \beta \mid \boldsymbol{x})}{\pi(\text{Exp}, \lambda \mid \boldsymbol{x})} \frac{P\left(\text{Par} \rightarrow \text{Exp}\right)}{P\left(\text{Exp} \rightarrow \text{Par}\right) q\left(u\right)} \left|\frac{\partial(\alpha,\beta)}{\partial\left(\lambda,u\right)}\right| \\ &= \frac{\pi(\alpha, \beta \mid \boldsymbol{x}, \text{Par}) p(\text{Par})}{\pi(\lambda \mid \boldsymbol{x}, \text{Exp}) p(\text{Exp})} \frac{1}{\phi\left(u \middle/ \sqrt{\sigma^2}\right)} \frac{u}{\lambda^2(u+1)} \\ &= \frac{L(\boldsymbol{x} \mid \alpha, \beta, \text{Par}) p(\alpha) p(\beta) p(\text{Par})}{L(\boldsymbol{x} \mid \lambda, \text{Exp}) p(\lambda) p(\text{Exp})} \frac{1}{\phi\left(u \middle/ \sqrt{\sigma^2}\right)} \frac{u}{\lambda^2(u+1)}\end{aligned}$$

Sequential Importance Sampling

Sequential importance sampling is an alternative to MCMC for sampling from high-dimensional distributions. It helps with the problem of estimating good proposal/importance distributions by building them sequentially.

To motivate the concept, first consider a target distribution $\pi(\boldsymbol{\theta})$ and a proposal distribution $g(\boldsymbol{\theta})$. Denoting $\boldsymbol{\theta}_{(-j)} = \left(\theta_1, \cdots, \not{\theta}_j, \cdots, \theta_p\right)$ We can write these as

$$g(\boldsymbol{\theta}) = g_1(\theta_1) g_2(\theta_2 \mid \theta_1) \cdots g_p(\theta_p \mid \boldsymbol{\theta}_{(-p)})$$
$$\pi(\boldsymbol{\theta}) = \pi(\theta_1) \pi(\theta_2 \mid \theta_1) \cdots \pi(\theta_p \mid \boldsymbol{\theta}_{(-p)})$$

And we can write the importance weights as

$$w(\boldsymbol{\theta}) = \frac{\pi(\theta_1) \pi(\theta_2 \mid \theta_1) \cdots \pi(\theta_p \mid \boldsymbol{\theta}_{(-p)})}{g_1(\theta_1) g_2(\theta_2 \mid \theta_1) \cdots g_p(\theta_p \mid \boldsymbol{\theta}_{(-p)})}$$

Writing $\boldsymbol{\theta}_k = \left(\theta_1, \cdots, \theta_k\right)$, we can define a *partial weight*

$$w_k(\boldsymbol{\theta}_k) = w_{k-1}(\boldsymbol{\theta}_{k-1}) \frac{\pi(\theta_k \mid \boldsymbol{\theta}_{k-1})}{g_k(\theta_k \mid \boldsymbol{\theta}_{k-1})}$$

When we include the entire vector, $w_p(\boldsymbol{\theta}_p) = w(\boldsymbol{\theta})$. Using this method is advantageous for two reasons

- We can stop generating further components if the partial weight gets too small (this will, however, lead to bias).
- We can use the marginal distribution $\pi(\theta_k \mid \boldsymbol{\theta}_{k-1})$ to help us in designing $g_k(\theta_k \mid \boldsymbol{\theta}_{k-1})$.

The only problem is that the decomposition of π is very difficult; it requires

$$\pi(\boldsymbol{\theta}_k) = \int \pi(\boldsymbol{\theta}) \, \mathrm{d}\theta_{k+1} \cdots \mathrm{d}\theta_p$$

Which is as difficult (or harder than) the initial problem.

Sequential importance sampling: Suppose we can find a sequence of "auxiliary distributions", which need not be normalised, such that

$$\pi_k(\boldsymbol{\theta}_k) \approx \pi(\boldsymbol{\theta})_k \qquad k = 1, \cdots, p-1$$
$$\pi_p(\boldsymbol{\theta}_p) = \pi(\boldsymbol{\theta})_p$$

The SIS method is then defined as the following recursive procedure

1. Draw $\theta_k \sim g_k\left(\theta_k \mid \boldsymbol{\theta}_{k-1}\right)$, and let $\boldsymbol{\theta}_k = \left(\boldsymbol{\theta}_{k-1}, \theta_k\right)$.
2. Compute the *incremental weight*

$$u_k = \frac{\pi_k(\boldsymbol{\theta}_k)}{\pi_{k-1}(\boldsymbol{\theta}_{k-1})} \frac{1}{g_k(\theta_k \mid \boldsymbol{\theta}_{k-1})}$$

3. Let $w_k = w_{k-1} u_k$

Note: An important special case arises when we can build g_k using π_k; ie:

$$g_k\left(\theta_k \mid \boldsymbol{\theta}_{k-1}\right) = \pi_k\left(\theta_k \mid \boldsymbol{\theta}_{k-1}\right)$$

in that case,

$$u_k = \frac{\pi_k(\boldsymbol{\theta}_{k-1})}{\pi_{k-1}(\boldsymbol{\theta}_{k-1})}$$

We often obtain several samples $\boldsymbol{\theta}^{(1)},\cdots,\boldsymbol{\theta}^{(T)}$ in parallel. Each of the partial samples $\mathcal{S}_k = \{\boldsymbol{\theta}_k^{(t)}\}_{t=1}^T$ are called *particles* or *streams* when properly weighed by the collection of weights $\mathcal{W}_k = \{w_k^{(t)}\}_{t=1}^T$.

As k increases, the variance of the importance weights also increases, which decreases the overall effective sample size. To fix this problem, we can use *re-sampling*

Resampling: Periodically, randomly, or dynamically (for example, when the ESS) is low, perform the following two steps:

1. Sample a new set of steams $\mathcal{S}_k'$ from $\mathcal{S}_k$, *with* replacement and with weights $\mathcal{W}_k$.
2. Assign equal weights $\Sigma_{t=1}^T w_k^{(t)} / T$ to each of the steams in $\mathcal{S}_k'$. (It is also common to set the weights to $1/T$).

Notes: Resampling can result in few unique; this is known as *particle depletion*. *Enrichment* or *diversification* methods involving kernel smoothing and MCMC have been developed as a remedy.

SIS is part of a broader class of Sequential Monte Carlo (SMC) methods, which are popular for inference in state space models, and are commonly encountered when dealing with time series data.

EXAMPLE: Suppose the auxiliary distributions are only known up to a normalising constant. Denote the normalised distributions by π_k and the un-normalised distributions by

$$q_k(\boldsymbol{\theta}_k) = Z_k \pi_k(\boldsymbol{\theta}_k) \qquad k = 1, \cdots, p$$

The incremental weights take the form

$$u_k = \frac{q_k(\boldsymbol{\theta}_k)}{q_{k-1}(\boldsymbol{\theta}_{k-1}) g_k(\theta_k \mid \boldsymbol{\theta}_{k-1})} = \frac{Z_k \pi_k(\boldsymbol{\theta}_k)}{Z_{k-1} \pi_{k-1}(\boldsymbol{\theta}_{k-1}) g_k(\theta_k \mid \boldsymbol{\theta}_{k-1})}$$

and final weight takes the form

$$w_p = \prod_{i=2}^{p} u_p = \frac{Z_p}{Z_1} \frac{\pi_p(\boldsymbol{\theta}_p)}{g_1(\theta_1) \cdots g_p(\theta_p \mid \boldsymbol{\theta}_{p-1})}$$

Thus, the sample average of $\{w_p^{(t)}\}_{t=1}^{T}$ gives an unbiased estimate of $\mathbb{E}(w_p) = Z_p / Z_1$.

Classical Inference

Classical inference involves maximising a likelihood $L(\boldsymbol{x} \mid \boldsymbol{\theta})$ to obtain estimates $\hat{\boldsymbol{\theta}}$ for the parameters. It will often be more convenient to consider *minimising* $-L(\boldsymbol{x} \mid \boldsymbol{\theta})$.

Simulated Annealing

Sometimes, this can be done analytically, but often, numerical algorithms are required. Some, called *absolute descent* algorithms try to move downhill until they can get no lower. Unfortunately, these algorithms are susceptible to getting stuck at subsidiary maxima. Consider, instead, an algorithm that proposes a move from $\boldsymbol{\theta} \to \boldsymbol{\phi}$ such that

- If $f(\boldsymbol{\phi}) \leq f(\boldsymbol{\theta})$, the move is accepted (absolute descent)
- If $f(\boldsymbol{\phi}) > f(\boldsymbol{\theta})$, the move is accepted with probability α

The problem with this method is that it does not present any obvious place to "stop". One approach would be the run the algorithm for a fixed time (but then it's unclear when to stop).

Another approach, which is adopted by *simulated annealing*, is to decrease α at each step. Eventually, the system freezes, to what we would hope is the global optimum.

More specifically , suppose we want to minimise the function $f(\boldsymbol{\theta})$, and let

$$b_T(\boldsymbol{\theta}) = \frac{\exp\left\{-f(\boldsymbol{\theta}) / T\right\}}{\int \exp\left\{-f(\boldsymbol{\theta}) / T\right\} \, \mathrm{d}\boldsymbol{\theta}} \propto \exp\left\{-\frac{f(\boldsymbol{\theta})}{T}\right\}$$

The parameter T is called the *temperature*. Clearly, b_T is a distribution that favours smaller values of $f(\boldsymbol{\theta})$. As $T \to 0$, smaller values of $f(\boldsymbol{\theta})$ are increasingly preferred.

Ideally, we would like to sample from b_0, but this is a point mass of the unknown minimising value $\hat{\boldsymbol{\theta}}$. Instead, simulated annealing works by simulating from a series of b_T for a decreasing sequence of temperatures:

Simulated Annealing (SA):

1. Take an initial temperature T_0 and a starting value $\boldsymbol{\theta}_0$.

2. Propose a new state $\boldsymbol{\phi}$ with density $q(\boldsymbol{\theta},\boldsymbol{\phi})$ [a typical choice is random-walk Metropolis, with $\boldsymbol{\phi} = \boldsymbol{\theta} + \boldsymbol{z}$].
3. Accept the move with probability $\alpha(\boldsymbol{\theta},\boldsymbol{\phi}) = \min\left\{1, A\right\}$, with
$$A = \frac{b_T(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta})}{b_T(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})}$$
4. Repeat steps 2 and 3 until the chain reaches equilibrium.
5. Check for a stopping criterion. A common criterion is to stop if no moves were accepted at this temperature.
6. Lower the temperature T and return to step 2.

Notes: When maximising a likelihood, we typically set $f(\boldsymbol{\theta}) = -L(\boldsymbol{x} \mid \boldsymbol{\theta})$ or $f(\boldsymbol{\theta}) = -\log L(\boldsymbol{x} \mid \boldsymbol{\theta})$. In the latter case, $b_T(\boldsymbol{\theta}) \propto \left[L(\boldsymbol{x} \mid \boldsymbol{\theta})\right]^{1/T}$.

The overall Markov chain is *inhomogeneous* because the target distribution changes over time.

The ideal proposal distribution q for the MH sampler still depend on the temperature T – the lower T, the closer we'd expect to be to the minimum, and the lower the variance of q. For this reason, Gibbs Sampling is preferred to standard distributions, since no proposal distribution is needed.

EXAMPLE: Consider trying to find the mode of a distribution comprised of a mixture of normals:

$$\begin{aligned} f(m) &= -\log\Big(L(m \mid \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)\Big) \\ &= -\log\Big(0.6N(m \mid \mu_1 = -8, \sigma_1^2 = 0.5^2) \\ &\qquad +0.4N(m \mid \mu_2 = 8, \sigma_2^2 = 0.9^2)\Big) \end{aligned}$$

And so

$$b_T(m) = \left[0.6\phi\left(\frac{m-\mu_1}{\sigma_1}\right) + 0.4\phi\left(\frac{m-\mu_2}{\sigma_2}\right)\right]^{1/T}$$

For fixed T, sampling from b_T will require MH.

EXAMPLE: Consider, now, the AR(k) example. If, once again, we let $f(\boldsymbol{\theta}) = -\log L(\boldsymbol{x} \mid \boldsymbol{\theta})$, we get:

$$\begin{aligned} b_T(\boldsymbol{a}, \sigma^2) &\propto \left(\sigma^{-2}\right)^{\frac{N-k}{2T}} \exp\left(-\frac{1}{2T\sigma^2}(\boldsymbol{x} - X_k\boldsymbol{a})^2\right) \\ &= \left(\sigma^{-2}\right)^{\frac{N-k}{2T}} \exp\left(-\frac{1}{2T\sigma^2}\sum_{t=k+1}^{N}\left(x_t - \Sigma_{r=1}^{k} a_r x_{t-r}\right)^2\right) \end{aligned}$$

Finding the full conditionals

- **For $\boldsymbol{\sigma}^2$**

$$b_T(\boldsymbol{a}, \sigma^2) \propto \left(\sigma^{-2}\right)^{\frac{N-k}{2T}} \exp\left(-\frac{1}{\sigma^2}\frac{1}{2T}\sum_{t=k+1}^{N}\left(x_t - \Sigma_{r=1}^{k} a_r x_{t-r}\right)^2\right)$$

And so

$$b_T(\boldsymbol{a}, \sigma^2) \sim \text{InvGamma}\left(\frac{N-k}{2T} + 1, \frac{1}{2T}\sum_{t=k+1}^{N}\left(x_t - \Sigma_{r=1}^{k} a_r x_{t-r}\right)^2\right)$$

- **And for $\boldsymbol{a}$**

$$\begin{aligned} b_T(\boldsymbol{a} \mid \sigma^2) &\propto \exp\left(-\frac{1}{2T\sigma^2}(\boldsymbol{x} - X_k\boldsymbol{a})^T(\boldsymbol{x} - X_k\boldsymbol{a})\right) \\ &\propto \exp\left(-\frac{1}{2}\left\{\boldsymbol{a}^T\frac{X_k^T X_k}{T\sigma^2}\boldsymbol{a} - 2\frac{\boldsymbol{a}^T X_k^T}{T\sigma^2}\boldsymbol{x}\right\}\right) \\ &\propto \exp\left(-\frac{1}{2}\left\{\boldsymbol{a}^T\Sigma^{-1}\boldsymbol{a} - 2\boldsymbol{a}^T\Sigma^{-1}\boldsymbol{\mu}\right\}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{a} - \boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\boldsymbol{a} - \boldsymbol{\mu}\right)\right) \\ &\sim N\left(\boldsymbol{\mu}, \frac{\Sigma^{-1}}{T\sigma^2}\right) \end{aligned}$$

Where

$$\Sigma^{-1} = \frac{X_k^T X_k}{T\sigma^2} \qquad \boldsymbol{\mu} = \frac{\Sigma X_k^T \boldsymbol{x}}{T\sigma^2}$$

Note that as $T \to 0$, the variance tends to 0, but the mean is unchanged.

Both can be sampled using a Gibbs sampler.

Model Selection Using Simulated Annealing

We can also use simulated annealing to do model selection. For example, to use Akaike's information criterion, we set

$$f(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}) = -2\log L_{\mathcal{M}}(\boldsymbol{x} \mid \boldsymbol{\theta}_{\mathcal{M}}) + 2\dim\left(\mathcal{M}\right)$$
$$g_T\left(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}\right) = \exp\left(-f\left(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}\right)/T\right)$$

Of course, we now require inter-dimensional jumps, and so RJ-MCMC must be used.

EXAMPLE: Consider, now, the AR(k) example. Use

$$f(\boldsymbol{a}_k, \sigma^2, k) = -\log L_k(\boldsymbol{x} \mid \boldsymbol{a}_k, \sigma^2) + k$$
$$b_T(\boldsymbol{a}, \sigma^2, k) \propto \left(\sigma^{-2}\right)^{\frac{N-k}{T}} \exp\left\{-\frac{1}{2\sigma^2 T}\left(\boldsymbol{x} - X_k\boldsymbol{a}\right)^2 - \frac{k}{T}\right\}$$

Now suppose we wish to jump from $k \to k+1$. In terms of parameters, we're going from

$$g\left\{(\boldsymbol{a}_k, u), \sigma^2\right\} = \left\{\boldsymbol{a}_{k+1}, \sigma^2\right\}$$

And we generate $u \sim q$. The Jacobian is the identity, since the mapping g is the identity.

The proposed move is then accepted with probability $\alpha = \min\left\{1, A\right\}$, where

$$A = \frac{b_T(\boldsymbol{a}_{k+1}, \sigma^2, k+1)}{b_T(\boldsymbol{a}_k, \sigma^2, k)} \frac{P(k+1 \to k)}{P\left(k \to k+1\right) q(u)}$$

Now, we saw in the previous example how to work out $b_T\left(\boldsymbol{a}_{k+1} \mid \boldsymbol{a}_k, \sigma^2, k+1\right)$; it's a normal distribution, given by

$$b_T\left(\boldsymbol{a}_{k+1} \mid \boldsymbol{a}_k, \sigma^2, k+1\right) = \frac{b_T\left(\boldsymbol{a}_{k+1}, \sigma^2, k+1\right)}{\int b_T(\boldsymbol{a}_{k+1}, \sigma^2, k+1)\ \mathrm{d}a_{k+1}}$$

It seems sensible to use this as our proposal distribution q. Using that and $P(k+1 \to k) = P\left(k \to k+1\right)$, the acceptance probability reduces to

$$A = \frac{\int b_T(\boldsymbol{a}_{k+1}, \sigma^2, k+1)\ \mathrm{d}a_{k+1}}{b_T(\boldsymbol{a}_k, \sigma^2, k)}$$

ie: the ratio of the marginal distributions of the unchanged parameters under each model.

Expectation Maximisation

Expectation minimisation is a method that is used when the likelihood is not defined explicitly, but is known in the form

$$f(\boldsymbol{\theta}) = \int g(\boldsymbol{\theta}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} \qquad \text{or} \qquad f(\boldsymbol{\theta}) = \int g_1(\boldsymbol{\theta} \mid \boldsymbol{z}) g_2(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}$$

An obvious situation in which this occurs is where we have known data, contained in the vector $\boldsymbol{x}$, and missing data, contained in the vector $\boldsymbol{z}$. In that case, the likelihood we want to maximise is

$$L(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int L(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\theta}) f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}$$

Expectation Maximisation (EM):

1. *E-Step* – given $\boldsymbol{x}$, calculate the expectation of the complete data log-likelihood as a function of the current estimate of $\boldsymbol{\theta}$

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= \mathbb{E}\left\{ \ell(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)} \right\} \\ &= \mathbb{E}_{\boldsymbol{x}, \boldsymbol{\theta}^{(t)}} \left\{ \ell(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \right\} \\ &= \int \ell(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}) \, \mathrm{d}\boldsymbol{z} \end{aligned}$$

   if the log-likelihood is linear in the join sufficient statistics of $[\boldsymbol{x}, \boldsymbol{z}]$, then this step simply involves finding the expectation of $\boldsymbol{z}$ given $\boldsymbol{x}$ and $\boldsymbol{\theta}^{(t)}$ and feeding it into ℓ.

2. *M-Step* – find $\boldsymbol{\theta}^{(t+1)}$ with maximises $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$

Theorem: Every step of the EM algorithm increases the log likelihood. That is

$$\ell(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t+1)}) \geq \ell(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})$$

with equality if and only if

$$Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$$

Proof: The likelihood of the complete data can be factorised as

$$L(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) = L(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})$$

(To understand why this expression makes sense, we can view L as a sort of "prior on $\boldsymbol{x}$"). And so

$$\ell(\boldsymbol{x},\boldsymbol{z} \mid \boldsymbol{\theta}) = \ell(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log f(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{\theta})$$
$$\ell(\boldsymbol{x} \mid \boldsymbol{\theta}) = \ell(\boldsymbol{x},\boldsymbol{z} \mid \boldsymbol{\theta}) - \log f(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{\theta})$$

Taking expectations of both sizes over the distribution of $\boldsymbol{z} \mid \boldsymbol{\theta}^{(t)}$ gives

$$\ell(\boldsymbol{x} \mid \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$$

where

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{x},\boldsymbol{z} \mid \boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{\theta}^{(t)}) \, \mathrm{d}\boldsymbol{z}$$
$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \int \log f(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{\theta}^{(t)}) \, \mathrm{d}\boldsymbol{z}$$

Now, the difference between $\ell(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)})$ and $\ell(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t+1)})$ is given by

$$\begin{aligned}
&\ell(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{x} \mid \boldsymbol{\theta}^{(t)}) \\
&\quad = Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \\
&\qquad\quad - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \\
&\quad = \left\{ Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \right\} \\
&\qquad\quad + \left\{ H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \right\}
\end{aligned}$$

Now:

- The EM algorithm maximises Q, and so $Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) > 0$.
- Jensen's Inequality gives that $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \leq H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$.

Overall, therefore, the likelihood increases. ■

Note: The EM algorithm is therefore a "hill-climbing" technique. It is guaranteed to find local maxima. Global maxima can be sought by running the algorithm with many different starting values.

EXAMPLE: Suppose we have a series of data $x_1, \cdots, x_n$ from some mixture distribution

$$f(x) = \alpha f_1(x) + (1-\alpha) f_2(x)$$

For example, UK heights were f_1 corresponds to men, f_2 corresponds to women, and $f_i(x) = N(\mu_i, \sigma^2)$.

The full likelihood function for these data is

$$L(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \alpha f_1(x_i) + (1-\alpha) f_2(x_i) \right\}$$

Maximising this, however, is extremely nasty.

Instead, let

$$z_i = \begin{cases} 1 & \text{if } x_i \sim f_1(x) \\ 0 & \text{otherwise} \end{cases}$$

The likelihood conditional on this variable is then

$$L(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z}) = \prod_{i=1}^{n} \left\{ \alpha f_1(x_i) \right\}^{z_i} \left\{ (1-\alpha) f_2(x_i) \right\}^{1-z_i}$$

This is much simpler to maximise.

Let's first consider the *E*-step. z_i can only be equal to 0 or 1, so

$$\begin{aligned} \mathbb{E}(z_i \mid \boldsymbol{x}, \boldsymbol{\theta}) &= 1 \times \mathbb{P}(z_i = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) + 0 \times \mathbb{P}(z_i = 0 \mid \boldsymbol{x}, \boldsymbol{\theta}) \\ &= \mathbb{P}(z_i = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) \\ &= \frac{\alpha f_1(x_i)}{\alpha f_1(x_i) + (1-\alpha) f_2(x_i)} = \hat{z}_i \end{aligned}$$

(This is evident from the form of the likelihood, which makes it clear that z is a Bernoulli random variable.)

Now the *M*-step. We have that $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{x}, \hat{\boldsymbol{z}} \mid \boldsymbol{\theta})$

$$\ell(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z}) = \sum_{i=1}^{n} \Big[z_1 \log \left\{ \alpha f_1(x_i) \right\} + (1 - z_1) \log \left\{ (1-\alpha) f_2(x_i) \right\} \Big]$$

and

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \left[\frac{z_1}{\alpha} + \frac{-(1-z_1)}{(1-\alpha)} \right] = \sum_{i=1}^{n} \left[\frac{z_1 - \alpha}{\alpha(1-\alpha)} \right] = 0$$

$$\Rightarrow \boxed{\hat{\alpha} = \frac{\sum_{i=1}^{n} z_i}{n}}$$

and

$$\frac{\partial \ell}{\partial \mu_1} = -\frac{\partial}{\partial \mu_1} \sum_{i=1}^{n} z_1 \frac{1}{2} \left(\frac{x_i - \mu_1}{\sigma} \right)^2 = \sum_{i=1}^{n} z_1 \frac{x_i - \mu_1}{\sigma^2} = 0$$

$$\Rightarrow \boxed{\hat{\mu}_1 = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}}, \boxed{\hat{\mu}_2 = \frac{\sum_{i=1}^{n} (1-z_i) x_i}{\sum_{i=1}^{n} (1-z_i)}}$$

and finally

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^2} &= \sum_{i=1}^{n}\left[z_1\left\{-\frac{1}{2\sigma^2}+\frac{(x-\mu_1)^2}{2(\sigma^2)^2}\right\}+(1-z_1)\left\{-\frac{1}{2\sigma^2}+\frac{(x-\mu_2)^2}{2(\sigma^2)^2}\right\}\right] \\ &= \sum_{i=1}^{n}\left[\frac{z_1(x-\mu_1)^2}{2(\sigma^2)^2}-\frac{1}{2\sigma^2}+\frac{(x-\mu_2)^2}{2(\sigma^2)^2}-\frac{z_1(x-\mu_2)^2}{2(\sigma^2)^2}\right]=0\end{aligned}$$

and so

$$\begin{aligned}\sum_{i=1}^{n}\left[\frac{z_1(x-\mu_1)^2}{2(\hat{\sigma}^2)^2}-\frac{1}{2\hat{\sigma}^2}+\frac{(x-\mu_2)^2}{2(\hat{\sigma}^2)^2}-\frac{z_1(x-\mu_2)^2}{2(\hat{\sigma}^2)^2}\right]&=0 \\ \sum_{i=1}^{n}\left[\frac{z_1(x-\mu_1)^2}{2(\hat{\sigma}^2)^2}+\frac{(x-\mu_2)^2}{2(\hat{\sigma}^2)^2}-\frac{z_1(x-\mu_2)^2}{2(\hat{\sigma}^2)^2}\right]&=\frac{n}{2\hat{\sigma}^2} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big[z_1(x-\mu_1)^2+\big(1-z_1\big)(x-\mu_2)^2\Big]\end{aligned}$$

Thus, our algorithm simply involves finding z in each case (at the *E*-step) and then finding values for the other parameters (at the *M*-step).