1 Simulation and Basics of Simple Random Variate Generation

Stochastic simulation methods are central tools in modern science and technology. Simulation methods for random variate generation and numerical approximation via Monte Carlo integration are, at a real practical level, simply approaches to exploring (usually mathematically complicated and multivariate) distributions. We use the term (stochastic) simulation interchangeably with Monte Carlo.

1.1 Uniform Pseudo-Random Generators

The uniform distribution underlies all practical simulation methods. See Gamerman & Lopes (§1.1) for discussion of algorithms for generating high-quality approximations to uniform U(0,1) variates. These are all deterministic algorithms, including chaotic process models, and generate pseudo-random numbers - sequences that are, these days and for most practical purposes, independent and uniformly distributed in the sense that a finite sequence of values so generated cannot be distinguished from a theoretically exact uniform random sample using statistical tests. Random quantities so generated are always in fact based on integer sequences taking values in (0, 1, ..., M) where M is a very large integer. Most effective are so-called congruential methods have periods of recurrence (return to starting value, or seed, and then repeat) that are very large. Scientific software and packages such as Matlab and R/Splus implement very high-quality and high-period algorithms, and these should be used as standard in most statistical and scientific work. The current Matlab (v6) \texttt{rand} generator, for example, claims to be able to generate all the floating point numbers in the closed interval $[2^{-53}, 1 - 2^{-53}]$ and with a period in excess of $10^{400}$. For some studies it is of interest to explore the results of a simulation analysis based on repeat evaluation using the same sequence of uniform random variates; in such cases, if a long or complex simulation is involved, the underlying pseudo-random number generator can be reset to begin at the same initial value - the seed - for repeat analyses.

1.2 Univariate Distributions via the Inverse CDF Transform

For a c.d.f. $P(\cdot)$, generate $x \sim P(x)$ via

- $x = P^-(u)$ with $u \sim U(0,1)$ where
  \[ P^-(u) = \inf\{ x : P(x) \geq u \}. \]

If $P(x)$ is continuous, $P^-(u) = P^{-1}(u)$ is the standard inverse function, or quantile function of the distribution. The generalized inverse $P^-$ applies also for discrete distributions, and for more interesting cases of mixed discrete and continuous distributions, examples in which $P(x)$ is piecewise continuous but exhibits jumps, for example. $P^-$ defines the generalized quantile function of the distribution.

Some key examples and specific results are of note.

- **Location and/or scale transformations.**
  
  If $P(x) = P_0((x-m)/s)$ for some location $m$ and scale $s > 0$, and where $P_0$ is a specified standard distribution, then $x = m + sz$ where $z \sim P_0$. See this by simply noting that $P^-(u) = m + sP_0^{-1}(u)$.

  Key examples of the normal, Cauchy and other T distributions, but also exponential, gamma and other distributions that arise commonly.

- $x$ is exponential with rate $1/t$, or mean $t$, if $P(x) = 1 - \exp(-x/t)$ so that $x = -t\log(1 - u)$. Equivalently, $x = -t\log(u)$. This is also an example of scaling from the standard (unit) exponential when $t = 1$.

- $x \sim \Gamma(0,1)$ has $p(x) = 1/\{\pi(1+x^2)\}$ so that $P(x) = 1/2 + \arctan(x)/\pi$; thus $x = \tan(\pi(u-1/2))$.

  Location/scale extensions have $x = m + \tan(\pi(u - 1/2))$. 

1.3 Transformation Methods

In many examples, direct transformations create samples from distributions of interest based on samples from more standard distributions.

- $x \sim \text{Ga}(a, b)$ is a scale transform from the standard gamma with shape $a$, i.e., $x = y/b$ where $y \sim \text{Ga}(a, 1)$.
- $x$ is standard lognormal if $x = \exp(z)$ where $z \sim N(0, 1)$.
- Beta and Dirichlet (generalised, multivariate versions of the beta) are transformations of gamma variables.
- Consider the AR(1) model and suppose we know the innovations variance and assume the conditional $y$ vector may be a complicated function of $y$ for some $p$.

The Box-Muller transformation (Gamerman & Lopes, §1.3.2) is a (venerable) example of a bivariate transformation: two $U(0, 1)$ variates $u_1 \perp u_2$ transform to two $N(0, 1)$ variates $x_1 \perp x_2$ via $x_1 = \sqrt{-2\log(u_1)}\cos(2\pi u_2)$ and $x_2 = \sqrt{-2\log(u_1)}\sin(2\pi u_2)$, easily verified by direct transformation of the bivariate p.d.f. of $(u_1, u_2)$ utilizing the Jacobian. (Most standard software does not use this method nowadays, though it still represents a useful example of transformation methods.)

- Location/scale transformations provide other examples, including the multivariate normal: $x = m + L\epsilon$ samples the $p-$dimensional $N(m, LL')$ distribution by transformation from a set of $p$ independent univariate normal elements of $\epsilon$.

It is easy to underestimate the importance of transformation methods - they are central to simulation-based analysis of complex, high-dimensional models. Suppose we have a large sample from a distribution $P(x)$ for some $p-$vector $x$, and are interested in a set of $q$ quantities in a $q-$vector $y$ where each element of $y$ may be a complicated function of $x$. Analytic calculations are impossible, but transforming the Monte Carlo sample is trivial so long as each $y_i$ can be evaluated as a function of $x$. This generates a sample of vectors $y$ and we can just by inspection look at histograms and other aspects to explore the implied marginal distribution of any element of $y$.

1.4 Example of Transformation: Dirichlet distributions

Beta and Dirichlet (generalised, multivariate versions of the beta) are transformations of gamma variables. These are central distributions in many areas of statistics, including categorical data analysis, non- and semi-parametric modelling, and mixture modelling.

- Suppose $\theta = (\theta_1, \ldots, \theta_k)'$ is a (non-degenerate) vector of random variable such that $\theta_j > 0$ for $j = 1, \ldots, k$ and $\sum_{j=1}^k \theta_j = 1$. The values of $\theta$ lie in the $k-$dimensional simplex, and represent a discrete probability distribution on $k$ cells.

- $\theta$ has a Dirichlet distribution with vector parameter $a = (a_1, \ldots a_k)'$, where each $a_j > 0$ and, with $\alpha = \sum_{j=1}^k a_j$, the p.d.f. is

$$p(\theta) = c(a)^{-1} \prod_{j=1}^k \theta_j^{a_j-1}$$
on the simplex, where the normalizing constant is
\[
c(a) = \frac{\prod_{j=1}^{k} \Gamma(a_j)}{\Gamma(\alpha)}.
\]

- We write \( \theta \sim \text{Dir}(a) \), or \( \text{Dir}_k(a) \) to make the dimension explicit.
- The prior mean is \( m = a/\alpha \) with elements \( m_j = E(\theta_j) = a_j/\alpha \). Clearly \( m \) is a probability distribution. The equivalent notation \( \theta \sim \text{Dir}(\alpha m) \) is common.
- \( \alpha \) represents a concentration parameter - the precision parameter of this Dirichlet distribution - showing up, for example, in the variances \( V(\theta_j) = m_j(1 - m_j)/(\alpha + 1) \) and covariances \( \text{Cov}(\theta_j, \theta_h) = -m_j m_h/(\alpha + 1) \).
- Dirichlets aggregate. For example, define \( \mu = (\theta_1, \ldots, \theta_{k-2}, \theta_{k-1} + \theta_k) \), a set of \( k - 1 \) probabilities obtained by aggregating the last two “cells”. Then \( \mu \sim \text{Dir}_{k-1}(b) \) where \( b = (a_1, \ldots, a_{k-2}, a_{k-1} + a_k)' \). More general aggregations are obvious.
- As a special case of the above, the univariate margins for each \( \theta_j \) are beta distributions (See exercises).
- Dirichlet distributed variables are generated by transformation of independent gamma random variables. Suppose that \( \lambda = (\lambda_1, \ldots, \lambda_k)' \) is a set of \( k \) independent gamma random variables with \( \lambda_j \sim \text{Ga}(a_j, 1) \) for some given shape parameters \( a_j > 1 \) (and a common scale parameter of unity). Define \( \Lambda = \sum_{j=1}^{k} \lambda_j \) and the transformation to \( \theta \) where
\[
\theta_j = \frac{\lambda_j}{\Lambda}, \quad j = 1, \ldots, k.
\]
Then \( \theta \sim \text{Dir}(a) \).

- The proof is simple: the joint density \( p(\lambda, \Lambda) \) is just the product of the gamma densities of the \( \lambda_j \) on the support \( \lambda_j > 0 \) for each \( j \) subject to \( \Lambda = \sum_{j=1}^{k} \lambda_j \). The Jacobian of the transformation to \( (\theta, \Lambda) \) is \( \Lambda^k \); integrating the resulting joint p.d.f. \( p(\theta, \Lambda) \) yields the Dirichlet density above.
- The reverse is true: If \( \theta \sim \text{Dir}(a) \) as above and \( \Lambda \sim \text{Ga}(\alpha, 1) \) independently of \( \theta \), then \( \lambda_j = \theta_j \Lambda \) are independent \( \text{Ga}(a_j, 1) \).
- The Dirichlet is the natural conjugate prior for the multinomial sampling distribution.

1. Multinomial sampling on \( k \) unordered categories is the canonical example. A random sample of size \( n \) is drawn from a (conceptually infinite or at least very large) population with one characteristic of interest represented \( k \) times with unknown population probabilities \( \theta_j \) of type/cell \( j \). The data is the set of counts \( n_j \) of cases of type/cell \( j \), with \( n = \sum_{j=1}^{k} n_j \). Survey sampling is a key example. The sampling model gives the likelihood function
\[
(n!/\prod_{j=1}^{k} n_j!) \prod_{j=1}^{k} \theta_j^{n_j}
\]
so that a \( \text{Dir}(\alpha m) \) prior is updated to a \( \text{Dir}(\alpha m + nf) \) posterior where \( f \) is the vector of cell frequencies \( f_j = n_j/n \). Evidently \( m \) plays the role of “prior cell frequencies” and \( \alpha \) the “prior sample size”.

2. Notation: \((n_1, \ldots, n_k)’ \sim Mn(n, \theta) \). The (fixed in advance) total sample size and the population cell probabilities are the two parameters.
3. A very important class of practical examples arises when the \( \theta_j \) are probabilities under a distribution function on some partition of a sample space. The simplest case is as follows. Suppose \( y \in [0, 1] \) is a scalar random quantity with a c.d.f. \( F(y) \) on \([0, 1] \). Define any partition of the unit interval into \( k \) cells,

\[
0 = x_0 < x_1 < \ldots < x_{k-1} < x_k = 1
\]

and name the corresponding partition cells \( I_j = (x_{j-1}, x_j) \) for \( j = 1, \ldots, k \). Write \( \theta_j = F(x_j) - F(x_{j-1}) = Pr(y \in I_j) \) for the \( k \) cell probabilities. The \( \theta \) is a discretized version of \( F(y) \).

- Suppose that we have a problem in which \( n \) observations are independently generated from \( F(y) \) but we observe only the partition cells counts \( n_j \). We then have a multinomial data outcome with the difference that the cells are ordered.
- If we specify a prior \( \theta \sim \text{Dir}_k(a) \) for some \( a \), the the Dirichlet posterior for \( \theta \) provides inferences on the discretised version of \( F(y) \). This leads into Bayesian histograms and non-(or semi-) parametric density estimation generally (later on).

### 1.5 Simulation via Convolutions: Mixtures of Distributions

We have already met this in several key example.

- **T distributions**: A scalar quantity \( x \) has a standard (Student) T distribution with \( \nu > 0 \) degrees of freedom if the p.d.f. is proportional to \( (1 + x^2/\nu)^{-(\nu+1)/2} \). The p.d.f. has the form

\[
p(x) \propto \int_0^\infty \lambda^{1/2} \exp(-x^2\lambda/2)p(\lambda)d\lambda
\]

where \( p(\lambda) \) is the p.d.f. of the gamma distribution \( \lambda \sim \text{Ga}(\nu/2, 1/2) \). Location and scale transformations of \( x \) generate the full class of univariate T distributions.

- **Multivariate T distributions**: The above is a special case of \( p = 1 \) in the multivariate T distribution where \( p(x) \propto \{1 + (x - m)'V^{-1}(x - m)/\nu\}^{-(\nu+p)/2} \). This is a scale mixture of normals too: with conditional variance matrix \( V \), we have \( (x|\lambda) \sim N(m, V/\lambda) \) and \( \lambda \sim \text{Ga}(\nu/2, 1/2) \). Simulate the T distribution via the mixture (i.e., via convolution) as \( x = m + \lambda^{-1/2}L \epsilon \) where \( L \) is the Cholesky lower-triangular component of \( V \), \( \epsilon \) is a vector of \( p \) independent standard univariate normals, and \( \lambda \) is a draw from the gamma distribution.

- Consider a stochastic process - such as the AR(1) model - for a series \( x_1, x_t, \ldots \) with a model that defines the joint distribution in terms of compositional form forward in time:

\[
p(x_{1:n}) = p(x_1) \prod_{t=2}^n p(x_t|x_{1:(t-1)}).
\]

If we sequentially simulate draws from \( p(x_1) \), then from \( p(x_2|x_1) \) conditional on the simulated value of \( x_1 \), and so forth, we generate a sequence \( x_{1:n} \) such that, by construction, each \( x_t \) is a draw from the implied marginal distribution \( p(x_t) \). Hence, within this compositional analysis we have a whole series of embedded convolutional simulations.

### 1.6 Compositional Sampling

Sampling mixtures is an example of compositional sampling, though the latter is usually used to denote sampling of a joint distribution of interest in its own right, whereas - often - mixture/convolutional sampling
quite often simply uses the mathematical form of a density as a mixture, if such is available, as a technical device.

The joint density of any set of $p$ random quantities can be written in many compositional forms, and sometimes one is preferred over others for its technical structure in connection with simulation. Generally, $x \sim p(x)$ can be simulated by sequencing through

- simulate $x_1$ from the univariate margin for $x_1$, and given that specific value,
- simulate $x_2$ from the conditional $p(x_2|x_1)$; given that value,
- simulate $x_3 \sim p(x_3|x_2, x_1)$, and so on.

Any subset of elements of $x$ so sampled represents a draw from the corresponding marginal distribution.

We have seen several examples already: the example of the normal/gamma structure of the T distribution - where $x$ and $\lambda$ are drawn from the joint distribution via composition, and also as in the example of simulation of the AR(1), and other stochastic processes where the $x_i$ are naturally ordered in time.

### 1.7 Posterior Simulation in Bayesian Analysis

Simulation of posterior distributions provides approaches to evaluation of inferences about complicated functions of model parameters that would, otherwise, be difficult to evaluate or even estimate.

- **AR(1) Example**

  In the reference Bayesian analysis of the AR(1) model we can simulate the posterior for the AR(1) model parameters, (conditioning on the initial value $x_1$) from the posterior directly, as it has a conditional normal-gamma form. Sample the gamma (scaled chi-squared) distribution. In the earlier notation, we simulate $p(\phi, v|x_{1:n})$ via composition, drawing $(v^{-1}|x_{1:n}) \sim Ga((n - 2)/2, Q(b)/2)$ and then $(\phi|x_{1:n}, v) \sim N(b, vB^{-1})$.

  One aspect of interest in this example is stationarity: the posterior is not constrained to a stationary model, for which $|\phi| < 1$. If we generate a large random sample as above, then any draw $(\phi^i, v^i)$ such that $|\phi^i| \geq 1$ indicates a nonstationary model, so that we can assess, by Monte Carlo, just how well supported the stationarity assumption is by looking at the proportion of such draws. This line of thinking is formalized below in connection with Monte Carlo integration.

Very often the posterior distribution is much more complicated, mathematically, and direct simulation via composition or convolutional methods is just not possible. In such cases, accept/reject methods, approximations using analytic approximations and importance (weighted) sampling, and especially MCMC are more relevant.

### 1.8 Mixtures, Compositional Sampling & Missing Data: Some Key Examples

- Mixtures arise natural in certain models - the Student T as a mixture of normals is a nice example. They also arise from parameter uncertainty. The predictive distribution in a Bayesian analysis is a mixture over the (prior or posterior) distribution representing parameter uncertainty: $p(x) = \int p(x|\theta)p(\theta)d\theta$ is the prior predictive distribution in a model for data $x$ and with parameters $\theta$. In predicting new data $y$ from the corresponding model component $p(y|x, \theta)$ - to forecast, or to generate insights into model fit and adequacy - we often use simulation of the implied posterior predictive distribution with p.d.f.

  $$p(y|x) = \int p(y|x, \theta)p(\theta|x)d\theta.$$
We have already seen an example in the AR(1) model where \( x = x_{1:n} \) and \( y = x_{(n+1):(n+m)} \) is the future \( m \) values of a time series process. Note that random sampling models are very special cases in which \( y \independent x|\theta \) so that \( p(y|x, \theta) = p(y|\theta) \).

- Discrete mixtures of parametric distributions - such as a discrete mixture (weighted average) of normal distributions - are used in many statistical studies, including kernel density estimation, to represent non-standard distributional forms. For example, a discrete mixture of \( k \) normal distributions \( N(m_j, v_j), (j = 1, \ldots, k) \), is written as

\[
x \sim \sum_{j=1}^{k} w_j N(m_j, v_j),
\]

where each \( w_j > 0 \) and \( \sum_{j=1}^{k} w_j = 1 \). The p.d.f. is simply

\[
p(x) = \sum_{j=1}^{k} w_j (2\pi v_j)^{-1/2} \exp(- (x - m_j)^2/(2v_j)).
\]

As \( k \) and the \((w_j, m_j, v_j)\) are varied, this can generate densities that are multimodal and skewed, and in fact provide direct analytic approximation to essentially any practically plausible continuous distribution function.

Here it is of interest to note that it is easy to sample such discrete mixture via convolution: the mixture can be interpreted in terms of a latent (missing, hidden) underlying variable \( z \) that "chooses" a mixture component: \( z \) is a discrete random variable taking a value from \( \{1, 2, \ldots, k\} \) and with \( Pr(z = j) = w_j, (j = 1, \ldots, k) \). That is, \( z \) is a multinomial random quantity with sample size 1 and cell probabilities \( w = (w_1, \ldots, w_n)' \), \( Z \sim Mn(1, w) \). Then we have the discrete convolutional representation of \( p(x) \), i.e.,

\[
p(x) = \sum_{z=1}^{k} p(x|z)p(z)
\]

where \( p(x|z) = N(m_z, v_z) \). Simulate a draw from \( p(x) \) via (a) draw a value of \( z \sim p(z) \), and then (b) sample the implied normal conditional for \( (x|z) \).

As a complete aside, this mixture framework also provides examples of multivariate distributions that are not multivariate normal but whose margins are: a mixture of bivariate normals, for example, may be such that the marginal normal for, say, \( x_1 \) in each mixture component is \( N(0,1) \), so that the overall margin \( p(x_1) \) is also standard normal, whereas of course the full joint distribution is far from normal.

- Mixtures are often induced in statistical analyses by missing data. The above mixture model can be viewed that way, with the implicit latent variable \( z \) interpreted as missing data. A simple example in arising with uncertain outcomes in classification testing provides a nice illustration, as well as contacting a number of additional distributional forms.

A gene variant created during RNA splicing often underlies much of what matters in terms of the function of a gene. A molecular test assess the presence or absence of one of two types of a variant of a specific gene: variant type A or type B. The test reports an outcome \( x = 0 \) (no variant - the gene is the common or "wild type"), \( x = 1 \) (type A variant) or \( x = 2 \) (type B variant). Write \( \theta_i = Pr(x = i) \) so that the parameter is \( \theta = (\theta_0, \theta_1, \theta_2)' \) subject to \( \sum_{i=0}^{2} \theta_i = 1 \). Prior information about base rates for gene variants suggests a prior distribution \( p(\theta) = Dir(a) \) where \( a = (a_0, a_1, a_2)' \) and \( a_j = \alpha m_j \) with \( E(\theta_j) = m_j \) for each \( j \). The p.d.f. is

\[
p(\theta) = c(a)^{-1} \prod_{i=0}^{2} \theta_i^{a_i-1} \quad \text{on} \quad \sum_{i=0}^{2} \theta_i = 1,
\]
with normalizing constant
\[ c(a) = \left\{ \prod_{i=0}^{2} \Gamma(a_i) \right\}/\Gamma(\alpha). \]

The prior mean is \( \mu \) and \( \alpha \) is the precision (or concentration) parameter. As we know, the Dirichlet is the natural conjugate prior for the multinomial sampling distribution so that if, for example, if a measurement is made on a new patient and the test outcome \( x \) observed, then
\[ p(\theta|x) \propto p(\theta)p(x|\theta) = p(\theta) \prod_{i=0}^{2} \theta^{e_i} \]
where \( e_i = I(x = i) \). Thus \( (\theta|x) \sim Dir(a + e) \) with \( e = (e_0, e_1, e_2)' \); we simply “update” the parameter for the cell that the patient falls in. A “Type A” variant observation updates \( a \) to \( a + 1 \), for example; the resulting Dirichlet has updated parameter \( (a_0, a_1 + 1, a_2)' = (\alpha m_0, \alpha m_1 + 1, \alpha m_2)' \) – the mean of this appropriately reflects the new information and the precision (also called “total mass”) increases to \( \alpha + 1 \).

Now for a missing data twist: suppose that, on this specific individual, the test is unable to distinguish gene variant \( A \) from \( B \), generating the observation \( y = \{x \in (1,2)\} \); the true nature of the variant is hidden, though we learn that it is certainly not the wild type. In this case the data probability is \( p(y|\theta) = \theta_1 + \theta_2 \) and so the actual posterior is now
\[ p(\theta|y) \propto p(\theta)p(y|\theta) = C \left\{ \prod_{i=0}^{2} \theta_i^{a_i-1} \right\} (\theta_1 + \theta_2) \]
for some (to be computed) normalization constant \( C \). This can be shown (homework exercise) to be a mixture of two Dirichlet distributions,
\[ (\theta|y) = w_1 Dir(a + \epsilon_1) + w_2 Dir(a + \epsilon_2) \]
where \( \epsilon_1 = (0,1,0)' \) and \( \epsilon_2 = (0,0,1)' \), and with \( w_j = c(a + \epsilon_j)/(c(a + \epsilon_1) + c(a + \epsilon_2)) \) for \( j = 1, 2 \). The mixing reflects the uncertainty about the type of gene variant on the posterior, and learning process, for the underlying mutation rates. Posterior simulation that generates a mixture indicator according to the probabilities \( w_j \) then mirrors this uncertainty in simulating relevant \( \theta \) values.

### 1.9 Monte Carlo Integration

Much of what underlies the use of simulation is the theory underlying Monte Carlo integration. See Gamerman & Lopes (§3.4). Since a primary use is in simulation-based Bayesian computation, we use \( \theta \) to denote the random quantity of interest; \( \theta \) will typically represent a vector of parameters, or latent variables, in a statistical model, and the distribution \( P(\theta) \) is the posterior distribution based on a model analysis and fit to observed data. The development of Monte Carlo integration is of course quite general and applies to any distribution \( P(\theta) \). We work in terms of density functions \( p(\theta) \) throughout.

- Interest lies in expectations, typically for many different functions \( h(\cdot) \),
\[ H = \int h(\theta)p(\theta)d\theta. \]

Generally includes vector functions, though scalars suffice for most practical purposes (the mean of a vector is the vector of means).
• Random sample $\theta_{1:m}$ where $\theta_i \perp \perp p(\theta)$ for $i = 1, \ldots, m$.
• Monte Carlo approximation:

\[
\tilde{h} = m^{-1} \sum_{i=1}^{m} h(\theta_i).
\]

• **Convergence theories:** Each (transformed random quantity) $h(\theta_i)$ has mean $H$ and, assuming that $E(h^2(\theta)) < \infty$ so that the common variance $\int (h(\theta) - H)^2 p(\theta) d\theta$ of each $h(\theta_i)$ is finite, we have:

- Law of Large Numbers: ensures almost sure converge of $\tilde{h}$ to $H$;
- Central Limiting Theorem: $\tilde{h} - H$ is asymptotically normal, with asymptotic variance

\[
m^{-1} \int (h(\theta) - H)^2 p(\theta) d\theta
\]

that can be approximated by

\[
v_m = m^{-2} \sum_{i=1}^{m} (h(\theta_i) - \tilde{h})^2.
\]

Hence, taking large Monte Carlo sample sizes $m$ (in the thousands or tens of thousands) can yield very precise, and cheaply computed, numerical approximations to mathematically difficult integrals. The Central Limit Theorem provides effective rule-of-thumb guidelines about precision: for example, an approximate 95% interval estimate associated with the Monte Carlo estimate $\tilde{h}$ is $\pm 1.96\sqrt{v_m}$. We refer to $\sqrt{v_m}$ as the (numerical or) Monte Carlo standard error.

1.9.1 A Key Example: Estimating the c.d.f. at a Point

One example takes $h(\theta) = I(\theta \leq x)$ for any $x$, so that $H$ is the value of the c.d.f. $P(x)$. Then

- $\tilde{h}$ is the proportion of the sample values $\theta_{1:m}$ that are below $x$.
- Each event $\{\theta_i < x\}$ has probability $H$, and independence implies that $m\tilde{h} \sim Bin(m, H)$.
- In this case, $V(\tilde{h}) = H(1 - H)/m \approx \tilde{h}(1 - \tilde{h})/m$. (It is easily checked that this standard binomial variance approximation coincides with the general sample variance formula for $v_m$ above.) Under the normal central limit, a 95% interval estimate of $H$ is then $\tilde{h} \pm 1.96\sqrt{\tilde{h}(1 - \tilde{h})/m}$.

For example, in the AR(1) model we are interested in $Pr(|\phi| < 1) = P(1) - P(-1)$ where $P(\cdot)$ is the posterior c.d.f. for $\phi$. Here, then, we want to use Monte Carlo integration to estimate $P(1)$ and $P(-1)$ so as to approximately evaluate the posterior probability that the AR process is stationary.

1.9.2 Histograms of Monte Carlo Draws

Histograms of $\theta_{1:m}$ represent Monte Carlo approximations to $p(\theta)$. The above theory provides access to uncertainty assessments. It is very common to use complex simulation methods to generate large Monte Carlo samples and to then use those samples in creating histograms as well as computing approximate expectations for various functions of interest - Monte Carlo posterior means $(h(\theta) = \theta)$, various posterior probabilities $(h(\theta)$ is an indicator function on some set or range of values of $\theta$,) and more complex transformations of what may be a high-dimensional parameter $\theta$. 

8
1.10 Importance Sampling

Importance sampling is one of the first steps into Monte Carlo analysis in which simulated variates from one distribution are used to explore another - simulation from the “wrong distribution” can be extraordinarily useful. Rejection sampling is another such method, and Metropolis MCMC methods define the encompassing framework. Currently, importance sampling is still of practical interest in

- fairly small problems, in terms of dimension,
- in which the density of the distribution of interest can be easily evaluated, but when it is difficult to sample from directly, and
- when it is relatively easy to identify and simulate from distributions that approximate the distribution of interest.

The core example is Bayesian inference on a parameter $\theta$ and, as earlier, the distribution $P(\theta)$ is the posterior distribution based on a model analysis and fit to observed data. We work in terms of density functions $p(\theta)$ throughout. Here, then, we must be able to evaluate $p(\theta)$ as a function at any point, and we assume that we have available an importance sampling distribution with p.d.f. $g(\theta)$. Two key requirements are that (a) $g(\cdot)$ is easy to sample from, and (b) the p.d.f. $g(\theta)$ is easy to evaluate at any point, as for $p(\theta)$. Often, the context is one in which $g(\theta)$ has been derived as an analytic approximation to $p(\theta)$, and the closer the approximation, the more accurate the resulting importance sampling MC analysis will be. However, sometimes $g(\theta)$ is just a convenient density that is particularly easy to simulate and evaluate, and in such cases we may easily generate very large Monte Carlo samples to increase the accuracy even though $g(\theta)$ may be a rather crude approximation to $p(\theta)$.

The underlying idea is simple:

- Interest lies in expectations of the form

$$H = \int h(\theta)p(\theta)d\theta.$$ 

- We can write

$$H = \int h(\theta)w(\theta)g(\theta)d\theta \quad \text{with} \quad w(\theta) = p(\theta)/g(\theta).$$

This shows that the expectation of $h(\theta)$ under $p(\theta)$ is just that of $h(\theta)w(\theta)$ under $g(\theta)$.

- Using direct Monte Carlo integration,

$$\bar{h} = m^{-1} \sum_{i=1}^{m} w(\theta_i)h(\theta_i),$$

where we draw the random sample $\theta_i \perp \perp g(\theta)$ for $i = 1, \ldots, m$. We are sampling from the “wrong” distribution.

- The measure “how wrong” at each simulated $\theta_i$ value is the importance weight

$$w(\theta_i) = p(\theta_i)/g(\theta_i).$$

These ratios “weight” the sample estimates $h(\theta_i)$ to “correct” for the fact that we sampled the wrong distribution.
See Lopes & Gamerman (§3.4) for discussion of convergence of Monte Carlo approximations, as well as comments on choice of importance sampling densities and related questions of optimal (with respect to specific, squared error loss functions) Monte Carlo estimation of a specified expectation (see also Robert and Casella §3.3 for much more lengthy and detailed discussion of optimality questions and convergence (of \( \bar{h} \) to \( H \)) and the association limit theorems.) The focus on computing specific expectations means that choices of \( g \) will depend on \( h \) to deliver best results. However, in much statistical work, and especially this context of a posterior analysis, we are interested in many possible expectations, and generally view the Monte Carlo analysis as one in which \( g(\theta) \) will be chosen to ensure good MC accuracy for various posterior characteristics.

Key considerations:

- MC estimate \( \bar{h} \) has the expectation \( H \), and is generally almost surely convergent to \( H \) under conditions below.
- \( \text{Var}(\bar{h}) = m^{-1} \int h^2(\theta)p^2(\theta)/g(\theta)d\theta - H^2 \).
- For ranges of functions \( h \), these variances are going to be finite for cases in which, generally, \( w(\theta) = p(\theta)/g(\theta) \) is bounded and decays rapidly in the tails of the target \( p(\theta) \). Smaller variance, and hence superior MC approximations, are achieved for densities whose tails dominate those of the target.
- The last feature means that importance sampling distributions should be chosen to have tails at least as fat as the target - common then to consider distributions such as T distributions, that decay like reciprocal powers of \( \theta \), when the target has exponential decay such as is exhibited in normal and other exponential family models. Theoretical investigation of the tail decay of \( p(\theta) \) will help guide thinking about importance sampling distributions.
- Obviously require the support of \( g(\theta) \) to be the same as, or contain, that of \( p(\theta) \).

Problems in which \( w(\theta) \) can be computed are rare in statistics. Quite commonly, we know \( p(\theta) \) only up to a constant of normalization - especially true in Bayesian analysis using Bayes’ Theorem. Then we renormalise the importance weights:

\[
\hat{h} = \frac{1}{m} \sum_{i=1}^{m} w_i h(\theta_i) \quad \text{where} \quad w_i = \frac{w(\theta_i)}{\sum_{j=1}^{m} w(\theta_i)}.
\]

- This is “as if” we are using a discrete distribution to approximate \( p(\theta) \): that distribution with point masses (=probabilities) \( w_i \) at the points \( \theta_i \).
- Good discussion of the convergence of \( \hat{h} \), and associated asymptotic (in \( m \)) theory, is given in Geweke (1989, *Econometrica*, 97, pp1317-1339). In particular, \( \hat{h} - H \) is asymptotically normal under broad conditions, basically those noted above on the relative tail weight matter. The asymptotic numerical or Monte Carlo standard error for guiding thinking about the accuracy of specific approximations is \( \sigma_m \) where

\[
\sigma_m^2 = (mW_m)^{-1} \sum_{i=1}^{m} (h(\theta_i) - \hat{h})^2 w(\theta_i)^2 \quad \text{with} \quad W_m = \sum_{i=1}^{m} w(\theta_i)^2.
\]

Exploring the empirical distribution of the sampled weights \( w_i \) guides understanding of how well, or how poorly, a particular importance sampling approximation may be. Weights that are close to uniform are desirable, and very unevenly distributed weights are not; it is easy to generate examples in which one or a very small number of weights are very large and dominate the rest, due to poor choice of \( g(\theta) \).

Some simple examples:
• Supposes \( p(\theta) \propto p_0(\theta)I(\theta \in A) \) where \( p_0 \) is the p.d.f. of a known, standard distribution, such as a multivariate T distribution, and where \( A \) is some specified set of values. For example, a univariate T distribution for \( \theta \) conditional on \( \theta > m \) for some specified constant \( m \). In this case an obvious importance sampling distribution is \( g(\theta) = p_0(\theta) \). Then \( g(\theta)/p(\theta) = 1 \) if \( \theta \in A \) and zero otherwise. This may be very efficient, and will be when the target distribution concentrates heavily on the set \( A \).

• The above example and that below are instances of the following setup. Suppose that we know the form of \( p(\theta) = a(\theta)g(\theta) \) where \( g(\theta) \) is a p.d.f. with the desirable characteristics for an importance sampler, and \( a(\theta) \) is some easily evaluated function, perhaps a contribution to the likelihood for \( \theta \) from a statistical model under with \( p(\theta) \) is the posterior density - but a factor that complicates the posterior analysis. Then, we may use “part” of the posterior for simulation, and correct based on the weight \( w(\theta) \propto a(\theta) \).

• In the simple AR(1) model we have a posterior \( p(\phi, v|x_{1:n}) \) that has a slightly complicated mathematical form and cannot be directly sampled. We have, however, the conditional normal/inverse gamma posterior for \( (\phi, v) \) arising from the AR model conditioning on \( x_1 \) and ignoring the contribution to the posterior from that initial observation; this conditional posterior is easily sampled and evaluated.

See the worked Matlab example on this interesting application of importance sampling.

1.10.1 Resampling for Inference and Prediction

• Treating the importance weighted samples as a discrete approximation to the posterior \( p(\theta) \) allows for MC estimation of integrals that include, for example, aspects of predictive distributions. For predicting a “future” random quantity \( y \) based on a model \( p(y|\theta) \), we use the predictive density

\[
p(y) = \int p(y|\theta)p(\theta)d\theta.
\]

Forecasting the future of a time series is one example. In such contexts, predictive expectations follow from the importance sampler via integration; e.g.; \( E(y) = \int h(\theta)p(\theta)d\theta \) where \( h(\theta) = E(y|\theta) \).

• Often we want to simulate predictive distributions. This is easily done via composition, treating the importance sampling weights as a discrete probability distribution on the sampled \( \theta_i \) values:

- Sample from the set \( \{\theta_i\} \) according to the multinomial distribution defined by weights \( \{w_i\} \); then
- Sample \( y \sim p(y|\theta_i) \).

• Resampling is a device sometimes used to convert the discrete distribution into a uniform discrete distribution, so generating an approximate random sample from \( p(\theta) \), although with duplicate values. This is useful when we can easily generate very large MC importance samples, and then resample to “flatten out” the weights and produce an equally weighted version, then use simple Monte Carlo approximations without worrying about weights. This has the feature that points \( \theta_i \) with very low weights will be disregarded, and is useful generally when the importance sampler is effective and generates well-balanced weights in the first place.

See the worked Matlab example on resampling in the AR(1) model analysis.

1.11 Simple Accept/Reject

Simple accept/reject sampling also uses an importance sampling function and the implied importance ratio to correct samples from the wrong distribution. This method is very widely used in simple random variate
generation, especially univariate distributions. It utilizes the same importance ratio ideas, but leads to exact
corrections and so exact samples from \( p(\theta) \). The use nowadays is restricted to univariate or low dimensional
problems where it can be the most efficient approach, but in even modestly complicated distributions it can he
difficult to implement, since it involves knowledge of an upper bound on the importance ratio. It is
particularly hard to implement when the target distribution of interest is known only up to a constant of nor-
malization, the common situation in posterior analysis, and in higher dimensions. More recently introduced
and advanced methods, including so-called slice-sampling and related methods, offer some real practical
advantages in, again, lower dimensional and mathematically relatively tractable examples. However, the
principles underlying accept/reject are, as with importance sampling, simply critical principles in moving to
more comprehensive approaches involving accept/reject methods in MCMC algorithms.

- Assume that \( w(\theta) = p(\theta)/g(\theta) < M \) for some constant \( M \). With normalized densities we must have
  \( M > 1 \), so that \( 1/M < 1 \). If \( g(\theta) \) represents a “good” potential importance sampler, then \( w(\theta) \) will
decay in the tails rapidly, and if it is a “good” approximation to \( p(\theta) \) then \( M \) should not be too far from 1.
- Generate a candidate value \( \theta \sim g(\theta) \). If \( w(\theta) \) is large, then the candidate seems to represent \( p(\theta) \);
  otherwise, \( \theta \) is an unlikely value under \( p(\theta) \).

The accept/reject theory formalizes this:

1. Accept \( \theta \) with probability \( w(\theta)/M \) : if accepted, it is a draw from \( p(\theta) \); otherwise reject and try again.
2. Equivalently, and operationally, generate \( u \sim U(0, 1) \) independently of \( \theta \). Then accept \( \theta \) as a draw
   from \( p(\theta) \) if, and only if, \( u < w(\theta)/M \).

The theory underlying this is simple. We look at the distribution of a random quantity \( \theta \) that is generated
by this algorithm. We need to show simply that the distribution of \( \theta \) conditional on it having been accepted,
i.e., conditional on \( u < w(\theta)/M \), is in fact \( P(\theta) \). By Bayes’ theorem, the p.d.f. of \( (\theta | u < w(\theta)/M) \) is just
\[
f(\theta | u < w(\theta)/M) = \frac{g(\theta)Pr(u < w(\theta)/M | \theta)}{Pr(u < w(\theta)/M)}.
\]

Now

- \( Pr(u < w(\theta)/M | \theta) = w(\theta)/M \) since \( u \sim U(0, 1) \), and
- \( Pr(u < w(\theta)/M) = \int Pr(u < w(\theta)/M | \theta)g(\theta)d\theta = \int w(\theta)g(\theta)d\theta/M = \int p(\theta)d\theta/M = 1/M. \)

Thus
\[
f(\theta | u < w(\theta)/M) = g(\theta)w(\theta) = p(\theta)
\]
as required: accepted values have p.d.f. \( p(\theta) \).

Rejected values are ... rejects. One key question is the efficiency, based on the proportion of rejects. If we
want a sample of size 10,000 and have to simulated millions due to a high rejection rate, we will probably try
another approach. Implicit in the proof of the accept/reject theory above is the fact that \( Pr(u < w(\theta)/M) = 1/M. \) That is, a pair \( (\theta, u) \) generated this way has a probability \( 1/M \) of acceptance. An importance sampler
\( g(\theta) \) such that \( M \) is close to 1 is, naturally, and efficient accept/reject sampler and leads to a high probability
of acceptance.

- One interesting example of generating \( Ga(a, 1) \) distributions for any positive value of \( a \) based on
  samples from the \( Ga(\lfloor a \rfloor, 1) \) where \( \lfloor a \rfloor \) is the integer part of \( a \) (Robert & Casella, example 2.3.4).
  More general rejection methods for gamma distributions are also available, as discussed in Gamerman &
  Lopes (§1.5).
• **Normal from Cauchy:** One simple and illustrative example is simulating the normal from the Cauchy. Take \( p(\theta) \) to be standard normal and \( g(\theta) = \{\sqrt{\pi}(1 + x^2/s)\}^{-1} \). The Cauchy is fatter tailed than the normal so we know the importance ratio will decay fast, and that it should deliver an effective importance sampling approach, so it should be of use for rejection too. In this example, you can easily show that \( w(\theta) < M_s \) where \( M_s = \sqrt{(2\pi/s) \exp(s/2 - 1)} \). It is then trivial to also show that \( M \) is minimized at \( s = 1 \), with \( M_1 = \sqrt{(2\pi/e)} \), and so \( 1/M_1 \approx 0.66 \), so about a 2 in 3 acceptance probability.

*See the worked Matlab example.*