Importance Sampling, Accept-Reject Sampling

(K Ch. 4, 12; KPT Ch. 11, Gelman Ch. 11)

Matlab scripts: `mod14_IS, mod14_ISkernel, mod14_AR, mod14_AR2`

The Gibbs Sampler rests on the assumption that we can break the full posterior kernel into conditional components with well-understood properties. This is not possible for a variety of important models. As we have learned, the Metropolis-Hastings (MH) algorithm can help in such cases by approximating unknown conditionals via a draws from a "candidate generating density (CGD)" and "acceptance probabilities". However, in some cases it is difficult to find a suitable CGD, with the result that few draws are ever accepted. This can be very costly from a computational perspective. Furthermore, even if all conditionals are known, the GS can be inefficient if there is strong correlation between draws from the different conditional components.

If we are primarily interested in the *moments* of the posterior density, or wish to evaluate it at a certain point, Importance Sampling (IS) is an alternative method that produces reliable results in many settings. If we want to take *draws* from the unknown density, Accept-Reject Sampling (AR), often also called Acceptance Sampling or Rejection sampling, is a simple alternative that works well in many applications.

Importance Sampling

The general idea behind IS is to draw from a well-known, convenient density, called the importance function, then average the draws to obtain the desired moment or ordinate of interest for the unknown density. This "unknown density" can be the entire posterior density, or any conditional component thereof. It can also simply be a complex likelihood function that one wants to evaluate at a certain point, or any desired posterior predictive construct. Simply think of IS as a method to obtain moments or density ordinates for an underlying density that has no known form.

Let's denote the importance function as $q(\theta)$, where θ can be a scalar or a vector of parameters. Further assume that the (unknown) posterior density (or other density of interest) is given by $p(\theta|y)$, as usual. A common situation where IS is useful is as an alternative to Monte Carlo integration when $p(\theta|y)$ is unknown. Specifically, assume you're interested in the *expectation* of some (potentially nonlinear) function $g(\theta)$. We learned that MCI would approximate this expectation as

$$E\left(g(\theta)|y\right) \approx \frac{1}{R}\sum g(\theta_r) \tag{1}$$

where θ_r is a draw from the underlying density of interest, such as a marginal posterior. However, now we assume that this underlying density is unknown, so we can't draw from it. Can we still get an estimate for $E\left(g(\theta)|y\right)$? Importance Sampling uses the following basic intuition to tackle this problem:

$$\begin{aligned} &E\left(g(\theta)|y\right) = E_{p(\theta|y)}\left(g(\theta)\right) = \int g(\theta)p(\theta|y)d\theta = \\ &\int \frac{g(\theta)p(\theta|y)}{q(\theta)}q(\theta)d\theta = E_{q(\theta)}\left(\frac{g(\theta)p(\theta|y)}{q(\theta)}\right) = E_{q(\theta)}\left(g(\theta)w(\theta)\right) \end{aligned} \tag{2}$$

Thus, we start out with a sought expectation of a function, where this expectation is based on the underlying (unknown) posterior, and we end with the expectation of a different function, but now the expectation is based on a known (= chosen) importance function. Naturally, the last term still includes the unknown density $p(\theta|y)$. The way out of this dilemma is to let the unknown density of interest be factored into a normalizing constant (not containing θ), and a kernel, i.e.

$$p(\theta|y)=\frac{p^*(\theta|y)}{\int p^*(\theta|y)d\theta}, \qquad \text{thus} \qquad p^*(\theta|y)\propto p(\theta|y) \tag{3}$$

We have already used this notion of "proportionality" in previous application. For example, for a generic univariate normal density, i.e. $\theta \sim n(\mu,\sigma^2)$, we have $\int p^*(\theta|y)d\theta=(2\pi\sigma^2)^{1/2}$, and $p^*(\theta|y)=\exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)$.

Thus we can modify (2) by writing

$$E_{p(\theta|y)}(g(\theta))=\int g(\theta)p(\theta|y)d\theta=\int g(\theta)\frac{p^*(\theta|y)}{\int p^*(\theta|y)d\theta}d\theta=\frac{\int g(\theta)p^*(\theta|y)d\theta}{\int p^*(\theta|y)d\theta}=$$

$$\frac{\int \frac{g(\theta)p^*(\theta|y)}{q(\theta)}q(\theta)d\theta}{\int \frac{p^*(\theta|y)}{q(\theta)}q(\theta)d\theta}=\frac{E_{q(\theta)}\left(\frac{g(\theta)p^*(\theta|y)}{q(\theta)}\right)}{E_{q(\theta)}\left(\frac{p^*(\theta|y)}{q(\theta)}\right)}=\frac{E_{q(\theta)}\left(g(\theta)w^*(\theta)\right)}{E_{q(\theta)}\left(w^*(\theta)\right)}$$

Using the law of large numbers as for MCI we get

$$E(g(\theta)|y)\approx\frac{\frac{1}{R}\sum_{r=1}^{R}g(\theta_r)w^*(\theta_r)}{\frac{1}{R}\sum_{r=1}^{R}w^*(\theta_r)}=\left(\frac{w^*(\theta_1)}{\sum_{r=1}^{R}w^*(\theta_r)}\right)g(\theta_1)+\left(\frac{w^*(\theta_2)}{\sum_{r=1}^{R}w^*(\theta_r)}\right)g(\theta_2)+\cdots+\left(\frac{w^*(\theta_R)}{\sum_{r=1}^{R}w^*(\theta_r)}\right)g(\theta_R) \tag{4}$$

where $w^*(\theta_r)=\frac{p^*(\theta_r|y)}{q(\theta_r)}$ is referred to as the "importance weight" and $\theta_r, r=1\ldots R$ are a set of random draws from the importance function $q(\theta)$.

Thus, IS can be interpreted as producing a weighted average of evaluations of $g(\theta)$, where each weight is based on the ratio between the ordinate of the full posterior at θ_r and the ordinate of the importance function at θ_r. Thus, if this ratio is large, a given draw will receive relatively larger weight in the computation of $E(g(\theta)|y)$ and vice versa. The intuition is as follows: If $q(\theta_r)$ is small compared to $p(\theta_r|y)$ we take it as evidence that θ_r lies in an area where $p(\theta_r|y)$ has considerable density mass.

Thus, we assign a higher weight to this draw. Conversely, if $q(\theta_r) >> p(\theta_r \mid y)$, we conclude that we're "out in the tails" for $p(\theta_r \mid y)$ (or in some other area of low density), and assign a low weight to that draw.

Again, the important part is that the algorithm can be implemented in practice without knowing the full $p(\theta_r \mid y)$ since the normalizing constants drop out of the formula. Thus, we can estimate each importance weight using the (known) posterior kernel rather than the full (unknown) posterior.

Two (relatively weak) conditions are needed for the estimated $E\left(g(\theta) \mid y\right)$ to converge to the true $E\left(g(\theta) \mid y\right)$ as $R \to \infty$: (i) the support of the importance function must include the support of the density it is trying to approximate (here $p(\theta \mid y)$), and (ii) $E\left(g(\theta) \mid y\right)$ must exist.

As you can imagine, the accuracy of this approach under finite sample size hinges crucially on a clever choice of the importance function. If $q(\theta_r) >> p(\theta_r \mid y)$ for a large range of θ, many weights will be essentially zero, and only a handful of draws effectively enter the computation of $E\left(g(\theta) \mid y\right)$. Thus, R would have to be very large to achieve a reasonable approximation. Particularly, if θ is high-dimensional, it can be very difficult to find a good $q(\theta_r)$. That's why IS is generally used primarily for lower-dimensional problems.

Application 1: Importance Sampling and its sensitivity to the choice of $q(\theta_r)$
(based on KPT Exercise 11.4)
Matlab script `mod14_IS, mod14_ISkernel`

Preliminaries:

It follows directly from the discussion above that the *posterior mean* for a given parameter θ can be approximated via IS as

$$E(\theta \mid y) \approx \frac{\sum_{r=1}^{R} w^*(\theta_r)\,\theta_r}{\sum_{r=1}^{R} w^*(\theta_r)} \tag{5}$$

since in that case we have $g(\theta_r) = \theta_r$. To approximate the posterior standard deviation we use again the generic relationship

$$std(\theta \mid y) = \sqrt{V(\theta \mid y)} = \sqrt{E\left(\theta^2 \mid y\right) - \left(E(\theta \mid y)\right)^2} \tag{6}$$

with

$$E\left(\theta^2 \mid y\right) \approx \frac{\sum_{r=1}^{R} w^*\left(\theta_r\right)\theta_r^2}{\sum_{r=1}^{R} w^*\left(\theta_r\right)} \tag{7}$$

To assess the suitability of the importance function it is also useful to capture the mean and standard deviation of the *R* weight terms via

$$E\left(w^*\left(\theta_r\right)\right) \approx \frac{1}{R}\sum_{r=1}^{R} w^*\left(\theta_r\right) \text{ and}$$

$$std\left(w^*\left(\theta_r\right)\right) \approx \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(w^*\left(\theta_r\right)\right)^2 - \left(\frac{1}{R}\sum_{r=1}^{R} w^*\left(\theta_r\right)\right)^2} \tag{8}$$

In the Matlab script we assume (for illustrative purposes) that we know the underlying density of interest, i.e. $\theta \mid y \sim n(0,1)$. We use four different importance functions:

$t(0,1,2.5), \quad t(0,1,100), \quad t(3,1,2.5), \quad t(3,1,100)$

A t-density with small degrees of freedom (last number) has fat tails. As the degrees of freedom increase, the t converges to the normal. With DoF=100, the t (with mean 0, scale 1) is virtually identical to the standard normal. Thus, we would expect the second importance function to perform best. As we move the mean far away from the truth (last 2 cases), fat tails actually become a virtue, as they reach more of the higher density region of the underlying density than a tight importance function with the wrong mean.

You should first run the script with the default setting of R=1000 draws. With that many draws, the choice of importance function is not very critical – they all perform reasonably well. Then reduce R to, say, 100. Now the choice of importance function matters a lot more.

<u>Common rule:</u> The importance function should have relatively fat tails (just in case we're way off with the mean).

Note: In script `mod14_IS` we use the full correct pdf of the target function in the sampling process. In contrast, in `mod14_ISkernel` we use only the kernel of the target function. As you can see from the results, the estimated moments for the target are essentially identical in either case.

However, the average weight is close to 1 ("perfect importance function") when we use the correct target pdf, but much further from 1 using only the kernel. So, on average, using only the kernel of the target the importance function is further off the mark, but still covers the entire distribution of the correct target sufficiently to yield acceptable estimates of the underlying moments.

Practical application of importance sampling:

IS can be handy within the *Chib* routine to approximate the marginal likelihood in the last step, where we need to *evaluate the likelihood function at the posterior mean*. Often times the full likelihood will be complex, especially in the presence of hierarchical distributions for model coefficients, which usually implies multi-dimensional integrals for the likelihood. Example: Panel count data models (Chib & Winkelman, 1998).

Accept-Reject Sampling

Assume again you have some target density $f(\theta)$ with support Θ for which you only know its kernel $\tilde{f}(\theta)$ (but this kernel doesn't point towards a unique density, as would be the case in a well-behaved Gibbs Sampler). You would like to obtain draws from the target density. AR Sampling starts by defining a source density $s(\theta)$, with kernel $\tilde{s}(\theta)$, and support Θ^*. Thus, we have

$$
\begin{aligned}
& f(\theta) = c_f \tilde{f}(\theta) \quad with \quad c_f = \left(\int_{\Theta} \tilde{f}(\theta) d\theta \right)^{-1} \\
& s(\theta) = c_s \tilde{s}(\theta) \quad with \quad c_s = \left(\int_{\Theta^*} \tilde{s}(\theta) d\theta \right)^{-1}
\end{aligned}
\tag{9}
$$

For AR to work, the importance ratio $f(\theta)/s(\theta)$ must have a known bound, i.e. there must exist some constant M for which $f(\theta)/s(\theta) \leq M, \forall\theta$, or, alternatively, $Ms(\theta) \geq f(\theta), \forall\theta$. Also, the support of the source density must encompass the support of the target density, i.e $\Theta \subseteq \Theta^*$.

The AR Sampler then proceeds as follows:

1. Draw a random uniform variate U from $U(0,1)$.
2. Draw a candidate θ_c from the source density.
3. If $U \leq \frac{\tilde{f}(\theta_c)}{\tilde{M}\tilde{s}(\theta_c)}$ accept the draw as a draw from the target density. Else reject the draw.
4. Repeat 1-3 as many times as needed to receive the desired number of draws.

Gelman p. 285 shows a graph that provides good intuition for this sampler. The approximation function ($Ms(\theta)$) lies above the target density for all values of θ, as required. If, for a given θ, $Ms(\theta)$ is very close to $f(\theta)$, the ratio $\frac{f(\theta_c)}{Ms(\theta_c)}$ will be relatively large (i.e. close to 1 from below), and there is a good chance the draw will get accepted. If the approximation is poor at θ, i.e. $Ms(\theta)$ lies far above $f(\theta)$, the ratio will be small, and the draw will likely get rejected.

Naturally, this implies that the choice of a well-fitting source density is key. If it doesn't approximate the target well, a large *M* will have to be chosen to assure that $Ms(\theta) \geq f(\theta), \forall\theta$, which will make the ratio $\frac{f(\theta_c)}{Ms(\theta_c)}$ very small for many draws of θ, and many draws will get rejected. In other words, an *efficient* AR sampler is one that accepts a large number of draws.

As indicated in step 3 above, we need to work with the target kernel $\tilde{f}(\theta)$ since the normalizing constant c_f and thus the full target density are unknown, by assumption. In the following, we will show that by using the decision rule $U \leq \frac{\tilde{f}(\theta_c)}{\tilde{M}\tilde{s}(\theta_c)}$ we obtain indeed draws from $f(\theta)$. First, we need to find a suitable quantity $\tilde{M}$ (let's call it the "importance multiplier") such that $\frac{\tilde{f}(\theta_c)}{\tilde{M}\tilde{s}(\theta_c)} \leq 1 \,\forall \theta_c$. A safe default choice for $\tilde{M}$ is the maximum over all $\theta's$ of the ratio of the target kernel to the source kernel, i.e.

$$\tilde{M} = \max_{\theta}\left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)}\right) \tag{10}$$

since in that case

$$\frac{\tilde{f}(\theta_c)}{\tilde{M}\tilde{s}(\theta_c)} = \frac{\tilde{f}(\theta_c)}{\left\{\max_{\theta}\left(\frac{\tilde{f}(\theta_c)}{\tilde{s}(\theta_c)}\right)\right\}\tilde{s}(\theta_c)} = \left\{\max_{\theta}\left(\frac{\tilde{f}(\theta_c)}{\tilde{s}(\theta_c)}\right)\right\}^{-1}\left(\frac{\tilde{f}(\theta_c)}{\tilde{s}(\theta_c)}\right) \leq 1 \,\forall \theta_c \text{ guaranteed.}$$

The overall acceptance rate can be formally derived as follows:

$$\begin{aligned}
&\Pr\left(U \leq \tilde{f}(\theta)/\tilde{M}\tilde{s}(\theta)\right) = \int_{\Theta}\left(\int_{0}^{\tilde{f}(\theta)/\tilde{M}\tilde{s}(\theta)} 1\,dU\right)s(\theta)\,d\theta = \\
&\int_{\Theta}\left(\tilde{f}(\theta)/\tilde{M}\tilde{s}(\theta)\right)s(\theta)\,d\theta = \frac{1}{\tilde{M}}\int_{\Theta}\frac{s(\theta)}{\tilde{s}(\theta)}\tilde{f}(\theta)\,d\theta = \frac{c_s}{\tilde{M}}\int_{\Theta}\tilde{f}(\theta)\,d\theta = \frac{c_s}{\tilde{M}c_f}
\end{aligned} \tag{11}$$

where we use the property of the standard uniform density of $\int_0^a f(U)\,dU = \int_0^a 1\,dU = a$. Note that the expected acceptance rate can be interpreted as the expectation over $s(\theta)$ of $\frac{\tilde{f}(\theta)}{\tilde{M}\tilde{s}(\theta)}$. This shows again that if this ratio is small for most θs (i.e the source density approximates the target poorly), the expected acceptance rate will be small as well. Naturally, this acceptance probability is usually unknown ex ante, since it includes the unknown normalization constant of the target density, c_f. However, this result is useful in the following proof that shows that the AR indeed produces draws from $f(\theta)$:

First note the following joint probability for any subset A of Θ:

$$\Pr(\theta \text{ accepted, } \theta \in A) = \int_A\left(\int_{\Theta}\frac{\tilde{f}(\theta)}{\tilde{M}\tilde{s}(\theta)}s(\theta)\,d\theta\right)s(\theta)\,d\theta = \int_A\frac{\tilde{f}(\theta)}{\tilde{M}\tilde{s}(\theta)}s(\theta)\,d\theta = \frac{c_s}{\tilde{M}}\int_A\tilde{f}(\theta)\,d\theta \tag{12}$$

Then:

$$pr(\ \theta \in A \,|\, \theta \text{ accepted}) = \frac{pr(\theta \text{ accepted}, \theta \in A)}{pr(\theta \text{ accepted})} = \frac{\frac{c_s}{\tilde{M}} \int_A \tilde{f}(\theta) d\theta}{\frac{c_s}{\tilde{M} c_f}} = \tag{13}$$

$$c_f \int_A \tilde{f}(\theta) d\theta = \int_A c_f \tilde{f}(\theta) d\theta = \int_A f(\theta) d\theta = pr(\theta \in A)$$

This says that the probability that θ is from a subset A of it's original support Θ , conditional on being accepted by the AR algorithm, is equal to the unconditional probability, *taken over the target density*, that θ lies in A. This is a fancy way of saying that "if θ is accepted, it is indeed a draw from $f(\theta)$. Note when $A = \Theta$, we immediately get:

$$pr(\ \theta \in \Theta \,|\, \theta \text{ accepted}) = \int_\Theta f(\theta) d\theta = 1 \tag{14}$$

Application: Approximating the Triangular

Matlab script `mod14_AR`
(based on KPT Ex. 11.22-11.24)

The triangular density has a lower and upper bound of a and b, respectively. It is given by

$$f(\theta) = 1 - |\theta| I(\theta \in [a,b]) \tag{15}$$

where *I(.)* is an indicator function that takes a value of 1 if the condition it carries is satisfied, and 0 otherwise. We need to keep this restriction on the support of θ in mind, but we can ignore it for mathematical purposes in the following computations. Since this simple density cannot be factored into a normalizing constant other than "1", the full density is de facto also its kernel, or

$$c_f = 1 \quad \tilde{f}(\theta) = 1 - |\theta| = f(\theta) \tag{16}$$

For the following cases, assume $a = -1, b = 1$.

<u>Case 1:</u>
Assume you choose as a source density a uniform with the same bounds, i.e.

$$s(\theta) = U[-1,1] = (1-(-1))^{-1} I(\theta \in [-1,1]) = \tfrac{1}{2} I(\theta \in [-1,1]) \tag{17}$$

Here, the entire density is a constant. In that case we interpret it as the kernel, s.t.:

$$c_s = 1 \quad \tilde{s}(\theta) = \tfrac{1}{2} \tag{18}$$

We first need to compute the importance multiplier $\tilde{M}$ that ensures that the importance ratio $\frac{\tilde{f}(\theta)}{\tilde{M}\tilde{s}(\theta)}$ remains at or below one. Building on the expression for $\tilde{M}$ from above, we get

$$\tilde{M} = \max_{\theta}\left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)}\right) = \max\left(\frac{(1-|\theta|)}{\frac{1}{2}}\right) = 2\max(1-|\theta|) = 2 .$$

This allows us to compute an analytical expected acceptance rate of

$$\Pr(\theta_c \text{ accepted}) = \frac{c_s}{\tilde{M}c_f} = \frac{1}{2} \tag{19}$$

The decision rule is based on "accept if $U \le \frac{\tilde{f}(\theta_c)}{\tilde{M}\tilde{s}(\theta_c)} = 1-|\theta|$"

Case 2:
Now change your source function to a standard normal. This implies:

$$s(\theta) = (2\pi)^{-1/2}\exp\left(-\tfrac{1}{2}\theta^2\right), \qquad c_s = (2\pi)^{-1/2}, \quad \tilde{s}(\theta) = \exp\left(-\tfrac{1}{2}\theta^2\right)$$

and

$$\tilde{M} = \max_{-1\le\theta\le1}\left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)}\right) = \max_{-1\le\theta\le1}\left(\frac{(1-|\theta|)}{\exp\left(-\frac{1}{2}\theta^2\right)}\right) = \max_{-1\le\theta\le1}\left((1-|\theta|)\exp\left(\tfrac{1}{2}\theta^2\right)\right).$$

$$\max\left((1-\theta)\exp\left(\tfrac{1}{2}\theta^2\right)\right)$$

$$\begin{aligned} FOC: \quad & (1-\theta)\exp\left(\tfrac{1}{2}\theta^2\right)\theta + \exp\left(\tfrac{1}{2}\theta^2\right)(-1) = 0 \\ & \exp\left(\tfrac{1}{2}\theta^2\right)\left((1-\theta)\theta - 1\right) = 0 \\ & \theta^2 - \theta + 1 = 0 \\ & \theta = \frac{1 \pm \sqrt{1-4}}{2} \end{aligned}$$

There is no interior solution for this case. Thus, we need to focus on the boundary outcomes (-1) and 1, and the breakpoint 0. Of these three choices the objective function will be maximized at $\theta = 0$, thus we have again $\tilde{M} = 1$. The analytical expected acceptance probability is

$$\Pr(\theta_c \text{ accepted}) = \frac{c_s}{\tilde{M}c_f} = c_s = (2\pi)^{-1/2} \approx 0.4 .$$

The decision rule is based on "accept if $U \le \frac{\tilde{f}(\theta_c)}{\tilde{M}\tilde{s}(\theta_c)} = \frac{(1-|\theta_c|)}{\exp\left(-\frac{1}{2}\theta_c^2\right)}$.

Case 3:
The source density or kernel is often tailored to mimic the mean and variance of the target density or kernel. For the triangular (-1,1) we have

$$E(\theta)=\int_{-1}^{1}\theta(1-|\theta|)d\theta=\int_{-1}^{0}\theta(1+\theta)d\theta+\int_{0}^{1}\theta(1-\theta)d\theta=$$

$$\left[\tfrac{1}{2}\theta^2+\tfrac{1}{3}\theta^3\right]_{-1}^{0}+\left[\tfrac{1}{2}\theta^2-\tfrac{1}{3}\theta^3\right]_{0}^{1}=-\tfrac{1}{2}+\tfrac{1}{3}+\tfrac{1}{2}-\tfrac{1}{3}=0$$

$$E\left(\theta^2\right)=\int_{-1}^{1}\theta^2(1-|\theta|)d\theta=\int_{-1}^{0}\theta^2(1+\theta)d\theta+\int_{0}^{1}\theta^2(1-\theta)d\theta=$$

$$\left[\tfrac{1}{3}\theta^3+\tfrac{1}{4}\theta^4\right]_{-1}^{0}+\left[\tfrac{1}{3}\theta^3-\tfrac{1}{4}\theta^4\right]_{0}^{1}=\tfrac{1}{3}-\tfrac{1}{4}+\tfrac{1}{3}-\tfrac{1}{4}=\tfrac{2}{3}-\tfrac{1}{2}=\tfrac{1}{6}$$

Thus, $V(\theta)=E\left(\theta^2\right)-\left(E(\theta)^2\right)=\tfrac{1}{6}$

Thus, let's change our source function to a normal with mean 0 and variance 1/6. This implies:

$$s(\theta)=\left(\tfrac{1}{3}\pi\right)^{-1/2}\exp\left(-3\theta^2\right),\qquad c_s=\left(\tfrac{1}{3}\pi\right)^{-1/2},\quad \tilde{s}(\theta)=\exp\left(-3\theta^2\right)$$

and

$$\tilde{M}=\max_{\theta}\left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)}\right)=\max\left(\frac{(1-|\theta|)I(\theta\in[a,b])}{\exp\left(-3\theta^2\right)}\right)=\max\left((1-|\theta|)I(\theta\in[a,b])\exp\left(3\theta^2\right)\right).$$

$$\max\left((1-\theta)\exp\left(3\theta^2\right)\right)$$

$$FOC:\quad (1-\theta)\exp\left(3\theta^2\right)6\theta+\exp\left(3\theta^2\right)(-1)=0$$

$$\exp\left(3\theta^2\right)\left((1-\theta)6\theta-1\right)=0$$

$$\theta^2-\theta+\tfrac{1}{6}=0$$

$$\theta=\frac{1\pm\sqrt{1-\frac{4}{6}}}{2}\approx .789$$

Since both target and source functions are symmetric around 0, there will be a second solution of $\theta\approx-0.789$. Either one leads to $\tilde{M}\approx 1.37$, and an analytical acceptance rate of $\frac{1}{1.37\sqrt{\left(\frac{1}{3}\pi\right)}}\approx .72$

Thus, this source density leads to the most efficient AR sampler.

Example 2:

Suppose you want to generate draws from a normal density with mean=0, variance=4, **truncated** to the interval [-2,2] using AR Sampling. You want to try the following source densities:

(i) A uniform over [-2,2]

(ii) An ***un***truncated normal with mean and variance equal to the *truncated* mean and variance of the target density.

The moments of a $n\left(\mu,\sigma^2\right)$, truncated to $\left[a,b\right]$ are given as:

$$E\left(\theta \mid a \le \theta \le b\right) = \mu + \sigma\left(\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right)$$

$$V\left(\theta \mid a \le \theta \le b\right) = \sigma^2\left(1 + \frac{\frac{a-\mu}{\sigma}\phi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma}\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} - \left(\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right)^2\right)$$

where φ and Φ denote the pdf and cdf of the standard normal, respectively.

The kernel and normalizing constant for target and source densities are given as follows:

$$f\left(\theta\right) = \left(\Phi\left(\tfrac{2-0}{2}\right) - \Phi\left(\tfrac{-2-0}{2}\right)\right)^{-1}\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{1}{8}\theta^2\right)I\left(-2 \le \theta \le 2\right) =$$

$$\left(\Phi\left(1\right) - \Phi\left(-1\right)\right)^{-1}\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{1}{8}\theta^2\right)I\left(-2 \le \theta \le 2\right)$$

$$c_f = \left(\Phi\left(1\right) - \Phi\left(-1\right)\right)^{-1}\left(8\pi\right)^{-1/2} = 0.2922$$

$$\tilde{f}\left(\theta\right) = \exp\left(-\frac{1}{8}\theta^2\right)I\left(-2 \le \theta \le 2\right)$$

<u>uniform over [-2,2]:</u>

$$c_s = 1 \qquad \tilde{s}\left(\theta\right) = \left(2 - \left(-2\right)\right)^{-1} = \frac{1}{4}$$

<u>***Un***truncated normal with mean and variance equal to the *truncated* mean and variance of the target density:</u>

First get those moments:

$$E\left(\theta \mid -2 \le \theta \le 2\right) = \tilde{\mu} = 0 + 2\frac{\phi\left(-1\right) - \phi\left(1\right)}{\Phi\left(1\right) - \Phi\left(-1\right)} = 0$$

$$V\left(\theta \mid -2 \le \theta \le 2\right) = \tilde{V} = 4\left(1 + \frac{-\phi\left(-1\right) - \phi\left(1\right)}{\Phi\left(1\right) - \Phi\left(-1\right)} - \left(\frac{\phi\left(-1\right) - \phi\left(1\right)}{\Phi\left(1\right) - \Phi\left(-1\right)}\right)^2\right) \approx 1.1645$$

$$c_s = \frac{1}{\sqrt{2\pi\tilde{V}}} \approx 0.3697 \qquad \tilde{s}\left(\theta\right) = \exp\left(-\frac{1}{2\tilde{V}}\theta^2\right)$$

The “multiplier” factor $\tilde{M}$ and the analytical acceptance rates are:

<u>uniform over [-2,2]:</u>

$$\tilde{M} = \max_{-2\le\theta\le 2}\left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)}\right) = \max_{-2\le\theta\le 2}\left(4\exp\left(-\tfrac{1}{8}\theta^2\right)\right)$$

$$FOC: \quad \exp\left(-\tfrac{1}{8}\theta^2\right)\left(-\tfrac{1}{4}\theta\right) = 0 \rightarrow \theta^* = 0 \rightarrow \tilde{M} = 4$$

$$AR = \frac{c_s}{\tilde{M}c_f} = \frac{1}{4*0.2922} \approx 0.86$$

<u>***Un**truncated normal with mean and variance equal to the *truncated* mean and variance of the target density:*</u>

$$\tilde{M} = \max_{-2\le\theta\le 2}\left(\exp\left(-\frac{1}{8}\theta^2 + \frac{1}{2\tilde{V}}\theta^2\right)\right) = \exp\left(0.3044\theta^2\right)$$

This is maximized at the highest permissible value of $\theta = 2$, which yields $\tilde{M} \approx 3.38$ and $AR = \frac{0.3697}{3.38*0.2922} \approx 0.374$.

This example is implemented in Matlab script `mod14_AR2`. As can be seen from the resulting plot, neither approach manages to characterize the two tails of the true underlying density, though the approximation is satisfactory for most of the interior range.

<u>References:</u>

Chib, S., E. Greenberg and R. Winkelmann. 1998. Posterior simulation and Bayes Factors in panel count data models. *Journal of Econometrics* **86**: 33-54.