Assignment #2 STA410H1F/2102H1F

due Friday October 18, 2019

Instructions: Solutions to problems 1 and 2 are to be submitted on Quercus (PDF files only). You are strongly encouraged to do problems 3–6 but these are not to be submitted for grading.

1. An interesting variation of rejection sampling is the ratio of uniforms method. We start by taking a bounded function h with $h(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} h(x)\,dx < \infty$. We then define the region

$$\mathcal{C}_h = \left\{(u,v) : 0 \leq u \leq \sqrt{h(v/u)}\right\}$$

and generate (U,V) uniformly distributed on $\mathcal{C}_h$. We then define the random variable $X = V/U$.

(a) The joint density of (U,V) is

$$f(u,v) = \frac{1}{|\mathcal{C}_h|} \text{ for } (u,v) \in \mathcal{C}_h$$

where $|\mathcal{C}_h|$ is the area of $\mathcal{C}_h$. Show that the joint density of (U,X) is

$$g(u,x) = \frac{u}{|\mathcal{C}_h|} \text{ for } 0 \leq u \leq \sqrt{h(x)}$$

and that the density of X is $\gamma\, h(x)$ for some $\gamma > 0$.

(b) The implementation of this method is somewhat complicated by the fact that it is typically difficult to sample (U,V) from a uniform distribution on $\mathcal{C}_h$. However, it is usually possible to find a rectangle of the form $\mathcal{D}_h = \{(u,v) : 0 \leq u \leq u_+, v_- \leq v \leq v_+\}$ such that $\mathcal{C}_h$ is contained within $\mathcal{D}_h$. Thus to draw (U,V) from a uniform distribution on $\mathcal{C}_h$, we can use rejection sampling where we draw proposals (U^*,V^*) from a uniform distribution on the rectangle $\mathcal{D}_h$; note that the proposals U^* and V^* are independent random variables with Unif$(0,u_+)$ and Unif(v_-,v_+) distributions, respectively. Show that we can define u_+, v_- and v_+ as follows:

$$u_+ = \max_x \sqrt{h(x)} \quad v_- = \min_x x\sqrt{h(x)} \quad v_+ = \max_x x\sqrt{h(x)}.$$

(Hint: It suffices to show that if $(u,v) \in \mathcal{C}_h$ then $(u,v) \in \mathcal{D}_h$ where $\mathcal{D}_h$ is defined using u_+, v_-, and v_+ above.)

(c) Implement (in R) the method above for the standard normal distribution taking $h(x) = \exp(-x^2/2)$. In this case, $u_+ = 1$, $v_- = -\sqrt{2/e} = -0.8577639$, and $v_+ = \sqrt{2/e} = 0.8577639$. What is the probability that proposals are accepted?

2. Suppose we observe $y_1, \cdots, y_n$ where

$$y_i = \theta_i + \varepsilon_i \quad (i = 1, \cdots, n)$$

where $\{\varepsilon_i\}$ is a sequence of random variables with mean 0 and finite variance representing noise. We will assume that $\theta_1, \cdots, \theta_n$ are smooth in the sense that $\theta_i = g(i)$ for some continuous and differentiable function g. The least squares estimates of $\theta_1, \cdots, \theta_n$ are trivial — $\widehat{\theta}_i = y_i$ for all i — but we can modify least squares in a number of ways to accommodate the "smoothness" assumption on $\{\theta_i\}$. In this problem, we will consider estimating $\{\theta_i\}$ by minimizing

$$\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n-1}(\theta_{i+1} - 2\theta_i + \theta_{i-1})^2$$

where $\lambda > 0$ is a tuning parameter that controls the smoothness of the estimates $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$. (This method is known as Whittaker graduation in actuarial science and the Hodrick-Prescott filter in economics.)

(a) Show that if $\{y_i\}$ are exactly linear, i.e. $y_i = a \times i + b$ for all i then $\widehat{\theta}_i = y_i$ for all i.

(b) In principal, we could compute $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$ using ordinary least squares estimation. Show that $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_n)^T$ minimizes

$$\|\boldsymbol{y}^* - X\boldsymbol{\theta}\|^2$$

where $\boldsymbol{y}^*$ is a vector of length $2n-2$ and X is an $(2n-2) \times n$ matrix. What are $\boldsymbol{y}^*$ and X?

(c) When n is large, computing $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$ directly, for example, using the OLS formulation in part (b) is computationally expensive when n is large. Alternatively, we could use the Gauss-Seidel algorithm but it converges slowly, particularly for larger values of λ. One possible alternatively is a randomized modification of the Gauss-Seidel algorithm, which at each stage solves a $p(\ll n)$ variable least squares problem.
The basic algorithm is as follows:

0. Initialize $\widehat{\boldsymbol{\theta}}$.

1. Randomly sample a subset w of size p from the integers $1, \cdots, n$. Define X_w to be the submatrix of X with column indices w and $X_{\bar{w}}$ to be the submatrix of X with column indices in the complement of w; define $\boldsymbol{\theta}_w$ and $\boldsymbol{\theta}_{\bar{w}}$ analogously so that $X\boldsymbol{\theta} = X_w\boldsymbol{\theta}_w + X_{\bar{w}}\boldsymbol{\theta}_{\bar{w}}$.

2. Define $\widehat{\boldsymbol{\theta}}_w$ to minimize

$$\left\|\boldsymbol{y}^* - X_{\bar{w}}\widehat{\boldsymbol{\theta}}_{\bar{w}} - X_w\boldsymbol{\theta}_w\right\|$$

with respect to $\boldsymbol{\theta}_w$.

3. Repeat steps 1 and 2 until convergence.

Show that the objective function is non-increasing from one iteration to the next.

(d) On Quercus, there is a function `HP` in a file `HP.txt`, which implements the randomized block Gauss-Seidel algorithm outlined in part (c). This function can be used as follows

```
> r <- HP(x,lambda=2000,p=20,niter=100)
```

where `lambda` is the value of λ, `p` is the value of p, and `niter` specifies the number of iterations of the algorithm. The output of the function (contained here in `r`) consists of two components: the estimates of $\{\theta\}$ in `r$xhat` and the values of the objective function for each iteration in `r$ss`.
Use this function (with $\lambda = 2000$) on data on monthly yields on short-term British securities over a 21 year period (252 months). Try out various values of p between 5 and 50 (using 1000 iterations). Describe how the objective function decreases with each iteration as p varies between 5 and 50.

(e) (Optional) Methods such as the randomized block Gauss-Seidel algorithm are essential in problems where the number of unknown parameters is extremely large. The goal is not to minimize the objective function but merely to find a solution that's close to the minimum.
In the context of the randomized block Gauss-Seidel algorithm, what factors should you consider in choosing p, the numbers of parameters being optimized at each step? Keep in mind that for least squares, the number of floating point operations needed increases with p like p^2.

Supplemental problems (not to hand in):

3. To generate random variables from some distributions, we need to generate two independent two independent random variables Y and V where Y has a uniform distribution over some finite set and V has a uniform distribution on $[0, 1]$. It turns out that Y and V can be generated from a single Unif$(0, 1)$ random variable U.

(a) Suppose for simplicity that the finite set is $\{0, 1, \cdots, n-1\}$ for some integer $n \geq 2$. For $U \sim \text{Unif}(0, 1)$, define

$$Y = \lfloor nU \rfloor \text{ and } V = nU - Y$$

where $\lfloor x \rfloor$ is the integer part of x. Show that Y has a uniform distribution on the set $\{0, 1, \cdots, n-1\}$, V has a uniform distribution on $[0, 1]$, and Y and V are independent.

(b) What happens to the precision of V defined in part (a) as n increases? (For example, if U has 16 decimal digits and $n = 10^6$, how many decimal digits will V have?) Is the method in part (a) particularly feasible if n is very large?

4. One issue with rejection sampling is a lack of efficiency due to the rejection of random variables generated from the proposal density. An alternative is the acceptance-complement (A-C) method described below.
Suppose we want to generate a continuous random variable from a density $f(x)$ and that $f(x) = f_1(x) + f_2(x)$ (where both f_1 and f_2 are non-negative) where $f_1(x) \leq g(x)$ for some density function g. Then the A-C method works as follows:

1. Generate two independent random variables $U \sim \text{Unif}(0, 1)$ and X with density g.
2. If $U \leq f_1(X)/g(X)$ then return X.
3. Otherwise (that is, if $U > f_1(X)/g(X)$) generate X from the density

$$f_2^*(x) = \frac{f_2(x)}{\displaystyle\int_{-\infty}^{\infty} f_2(t)\, dt}.$$

Note that we must be able to easily sample from g and f_2^* in order for the A-C method to be efficient; in some cases, they can both be taken to be uniform distributions.

(a) Show that the A-C method generates a random variable with a density f. What is the probability that the X generated in step 1 (from g) is "rejected" in step 2?

(b) Suppose we want to sample from the truncated Cauchy density

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

using the A-C method with $f_2(x) = k$, a constant, for $-1 \le x \le 1$ (so that $f_2^*(x) = 1/2$ is a uniform density on $[-1, 1]$) with

$$f_1(x) = f(x) - f_2(x) = f(x) - k \quad (-1 \le x \le 1).$$

If $g(x)$ is also a uniform density on $[-1, 1]$ for what range of values of k can the A-C method be applied?

(c) Defining f_1, f_2, and g as in part (b), what value of k minimizes the probability that X generated in step 1 of the A-C algorithm is rejected?

5. Suppose we want to generate a random variable X from the tail of a standard normal distribution, that is, a normal distribution conditioned to be greater than some constant b. The density in question is

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}(1 - \Phi(b))} \quad \text{for } x \ge b$$

with $f(x) = 0$ for $x < b$ where $\Phi(x)$ is the standard normal distribution function. Consider rejection sampling using the shifted exponential proposal density

$$g(x) = b \exp(-b(x - b)) \quad \text{for } x \ge b.$$

(This proposal density is used by the Monty Python algorithm to sample from the tail of the normal distribution.)

(a) Define Y be an exponential random variable with mean 1 and U be a uniform random variable on $[0, 1]$ independent of Y. Show that the rejection sampling scheme defines $X = b + Y/b$ if

$$-2 \ln(U) \ge \frac{Y^2}{b^2}.$$

(Hint: Note that $b + Y/b$ has density g.)

(b) Show the probability of acceptance is given by

$$\frac{\sqrt{2\pi} b (1 - \Phi(b))}{\exp(-b^2/2)}.$$

What happens to this probability for large values of b? (Hint: You need to evaluate $M = \max f(x)/g(x)$.)

(c) Suppose we replace the proposal density g defined above by

$$g_\lambda(x) = \lambda \exp(-\lambda(x-b)) \quad \text{for } x \geq b.$$

(Note that g_λ is also a shifted exponential density.) What value of λ maximizes the probability of acceptance? (Hint: Note that you are trying to solve the problem

$$\min_{\lambda>0} \max_{x \geq b} \frac{f(x)}{g_\lambda(x)}$$

for λ. Because the density $g_\lambda(x)$ has heavier tails, the minimax problem above will have the same solution as the maximin problem

$$\max_{x \geq b} \min_{\lambda>0} \frac{f(x)}{g_\lambda(x)}$$

which may be easier to solve.)

6. (a) Suppose that $E_1, E_2, \cdots$ are independent Exponential random variables with density $f(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$. Then the distribution of $T_k = E_1 + \cdots + E_k$ is a Gamma distribution whose density is

$$g_k(x) = \frac{\lambda^k x^{k-1} \exp(-\lambda x)}{(k-1)!} \quad \text{for } x \geq 0.$$

Show that

$$P(T_k \geq 1) = \int_1^\infty g_k(x)\, dx = \sum_{j=0}^{k-1} \frac{\lambda^j \exp(-\lambda)}{j!}$$

(Hint: Use integration by parts.)
(b) How might you use the result of part (a) to generate random variables from a Poisson distribution with mean λ?