1. 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 $b \geq 0$. The density in question is

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

with $f(x) = 0$ for $x < b$ where $\Phi(x)$ is the standard Normal distribution function; note that the Half-normal distribution corresponds to $b = 0$. Consider rejection sampling using the shifted Exponential proposal density

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

(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

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

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

$T = \frac{f(Y)}{Mg(Y)}$. Then if $U \geqslant T$, set $X = Y$

so, if $U \geqslant \frac{f(x+\frac{Y}{b})}{Mg(x+\frac{Y}{b})}$ then set $X = b + \frac{Y}{b}$

Since $\frac{f(x)}{g(x)} \leq M$, M is the largest value of $\frac{f(x)}{g(x)}$

$$\frac{f(x)}{g(x)} = \frac{\dfrac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{b\,e^{-b(x-b)}} = \frac{e^{-\frac{x^2}{2}+bx-b^2}}{b\sqrt{2\pi}(1-\Phi(b))} \quad \text{for } x \geqslant b$$

need to find max of $e^{-\frac{x^2}{2}+bx-b^2}$ since denominator is constant.

Find where the $\frac{d}{dx} = 0$

$$\left(e^{-\frac{x^2}{2}+bx-b^2}\right)\frac{d}{dx} = (-x+b)\left(e^{-\frac{x^2}{2}+bx-b^2}\right)$$

The derivative is 0 when $x = b$, hence $M = \frac{f(b)}{g(b)}$

Then, $U \leq \frac{f(b+\frac{Y}{b})}{\frac{f(b)}{g(b)}\,g(b+\frac{Y}{b})}$

$$U \leq \frac{\dfrac{e^{-\frac{(b+\frac{Y}{b})^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{\left(\dfrac{\frac{e^{-\frac{b^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{b\,e^{-b(b-b)}}\right)\left(b\,e^{-b(b+\frac{Y}{b}-b)}\right)}$$

$$U \leq \frac{e^{-\frac{(b+\frac{Y}{b})^2}{2}}}{e^{-\frac{b^2}{2}}\cdot e^{-Y}}$$

$$U \leq \frac{e^{-\frac{b^2}{2}-Y-\frac{Y^2}{2b^2}}}{e^{-\frac{b^2}{2}-Y}}$$

$$U \leq e^{-\frac{Y^2}{2b^2}}$$

$\ln(U) \leq -\frac{Y^2}{2b^2}$ ln both sides

Then, $-2\ln(U) \geqslant \frac{Y^2}{b^2}$

so if $X = b + \frac{Y}{b}$ then $-2\ln(U) \geqslant \frac{Y^2}{b^2}$

■

(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 $M = \max f(x)/g(x)$.)

This is shown in a)

$T = \frac{f(Y)}{Mg(Y)}$. Then if $U \geqslant T$, set $X = Y$

so, if $U \geqslant \frac{f(x+\frac{Y}{b})}{Mg(x+\frac{Y}{b})}$ then set $x = b + \frac{Y}{b}$

Since $\frac{f(x)}{g(x)} \leq M$, M is the largest value of $\frac{f(x)}{g(x)}$

$$\frac{f(x)}{g(x)} = \frac{\dfrac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{b\,e^{-b(x-b)}} = \frac{e^{-\frac{x^2}{2}+bx-b^2}}{b\sqrt{2\pi}(1-\Phi(b))} \quad \text{for } x \geqslant b$$

need to find max of $e^{-\frac{x^2}{2}+bx-b^2}$ since denominator is constant.

Find where the $\frac{d}{dx} = 0$

$$\left(e^{-\frac{x^2}{2}+bx-b^2}\right)\frac{d}{dx} = (-x+b)\left(e^{-\frac{x^2}{2}+bx-b^2}\right)$$

The derivative is 0 when $x = b$, hence $M = \frac{f(b)}{g(b)}$

Note: We know this is the max value since derivative at $x > b$ is negative

Probability of acceptance is given by: $P\left(U \leq \frac{f(Y)}{Mg(Y)}\right) = \frac{1}{M}$

Hence, $\frac{1}{M} = \frac{g(b)}{f(b)} = \frac{\sqrt{2\pi}b(1-\Phi(b))}{e^{(-b^2/2)}}$

What happens to probability of large value of b?

$$\lim_{b \to \infty} \frac{\sqrt{2\pi}\, b(1-\Phi(b))}{e^{(-b^2/2)}}$$

using L'hopital's rule This equals $\frac{\sqrt{2\pi}(1-\Phi(b)) + \sqrt{2\pi}b(-\Phi'(b))}{-b\,e^{-b^2/2}}$

Using L'hopital's rule again this is $= \frac{-\sqrt{2\pi}\Phi'(b) + \sqrt{2\pi}(-\Phi'(b)) + \sqrt{2\pi}b(-\Phi''(b))}{-e^{-b^2/2} + b^2 e^{-b^2/2}}$

Notice that $\Phi(b) = P(Z \leq b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-\frac{x^2}{2}} \, dx$

Then, $\Phi'(b) = \frac{1}{\sqrt{2\pi}} e^{-\frac{b^2}{2}}$

Then, $\Phi''(b) = \frac{-b}{\sqrt{2\pi}} e^{-\frac{b^2}{2}}$

So,

$$\frac{-\sqrt{2\pi}\,\Phi'(b) + \sqrt{2\pi}\left(-\Phi'(b)\right) + \sqrt{2\pi}\, b\left(-\Phi''(b)\right)}{-e^{-b^2/2} + b^2 e^{-b^2/2}} = \frac{-2\sqrt{2\pi}\,\Phi'(b) - \sqrt{2\pi}\, b\, \Phi''(b)}{-\exp(-b^2/2) + b^2 \exp(-b^2/2)}$$

Then this equals to $\frac{-2e^{-b^2/2} + b^2 e^{-\frac{b^2}{2}}}{-\exp(-b^2/2) + b^2\exp(-b^2/2)}$

Then, $\frac{e^{-b^2/2}(b^2-2)}{e^{-b^2/2}(b^2-1)} = \frac{b^2-2}{b^2-1}$

So,

$$\lim_{b \to \infty} \frac{\sqrt{2\pi}\, b\left(1-\Phi(b)\right)}{e^{(-b^2/2)}} = \lim_{b \to \infty} \frac{b^2-2}{b^2-1} = 1$$

$\therefore$ For large values of b, the probability of acceptance is 1. ∎

(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 $\lambda = \lambda(b)$ 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.)

$$P(U \leq f(Y)/Mg(Y)) = \frac{1}{M} = \frac{f(x)}{g_\lambda(x)}$$

$$= \frac{\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{\lambda e^{-\lambda(x-b)}}$$

Deriving with respect to λ to find min value

$$\ln\left(\frac{\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{\lambda e^{-\lambda(x-b)}}\right)\frac{d}{d\lambda} = \ln\left(\frac{e^{-\frac{x^2}{2}+\lambda(x-b)}}{\lambda\sqrt{2\pi}(1-\Phi(b))}\right)\frac{d}{d\lambda}$$

$$= \left[\left(-\frac{x^2}{2} + \lambda(x-b)\right) - \left(\ln(\lambda) + \ln(\sqrt{2\pi}(1-\Phi(b)))\right)\right]\frac{d}{d\lambda}$$

$$= (x-b) - \frac{1}{\lambda}$$

Need to find λ s.t the derivative is 0

Then $(x-b) - \frac{1}{\lambda} = 0$

$$(x-b) = \frac{1}{\lambda}$$

$$\lambda = \frac{1}{x-b}$$

Then $\frac{1}{M} = \frac{f(x)}{g_{\frac{1}{x-b}}(x)}$

$$= \frac{\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{\lambda e^{-\lambda(x-b)}}$$

$$= \frac{\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}(1-\Phi(b))}}{\frac{1}{x-b} e^{-1}}$$

$$= \frac{(x-b)e^{-\frac{x^2}{2}+1}}{\sqrt{2\pi}(1-\Phi(b))}$$

Now derive with respect to x

ignored denominator as it's a constant

$$\frac{(x-b)e^{-\frac{x^2}{2}+1}}{\sqrt{2\pi}(1-\Phi(b))}\frac{d}{dx} = e^{-\frac{x^2}{2}+1} + (-x)(x-b)e^{-\frac{x^2}{2}+1}$$

x value that makes derivative $=0$

$$e^{-\frac{x^2}{2}+1}\left(1 + (-x)(x-b)\right) = 0$$

$$-x^2 + xb + 1 = 0$$

$$x = \frac{-b \pm \sqrt{b^2+4}}{-2}$$

Then we have $\lambda = \frac{1}{x-b}$

and $x = \frac{-b \pm \sqrt{b^2+4}}{2}$

So $\lambda_1 = \frac{1}{\frac{-b+\sqrt{b^2+4}}{-2} - b} = \frac{1}{\frac{-b+\sqrt{b^2+4}\ +2b}{-2}} = \frac{-2}{b+\sqrt{b^2+4}}$

$\lambda_2 = \frac{-2}{b-\sqrt{b^2+4}}$

So $\lambda = \begin{cases} \frac{-2}{b+\sqrt{b^2+4}} \\ \frac{-2}{b-\sqrt{b^2+4}} \end{cases}$

but since $\lambda > 0$ then $b+\sqrt{b^2+4} < 0$ or $b-\sqrt{b^2+4} < 0$ Since numerator is negative

but $b \geqslant 0 \implies b+\sqrt{b^2+4} > 0$

Hence λ is $\frac{-2}{b-\sqrt{b^2+4}}$

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 and some a and b then $\widehat{\theta}_i = y_i$ for all i.

proof: Assume $\{Y_i\}$ are linear. i.e) $Y_i = a \times i + b$, $Y_i = \theta_i + \varepsilon_i$ $(i = 1, \cdots, n)$

Then $\hat{\theta}_i = Y_i \ \forall i$

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

So, we will evaluate the following,

$$\sum_{i=1}^{n}(Y_i - Y_i)^2 + \lambda \sum_{i=2}^{n-1}(Y_{i+1} - 2Y_i + Y_{i-1})^2$$ where $Y_i = ai + b$ (Assumption)

$$= \lambda \sum_{i=2}^{n-1}\left(a(i+1) + b - 2(ai + b) + a(i-1) + b\right)^2$$

$$= \lambda \sum_{i=2}^{n-1}\left(ai + a + b - 2ai - 2b + ai - a + b\right)^2$$ They cancel each other out

$$= \lambda \sum_{i=2}^{n-1}(0)^2$$

Then, we can see that the objective function is 0 and is minimized

Thus, for $\hat{\theta}_i = Y_i$ minimizes the objective function assuming Y_i is linear ■

(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?

We wish to write $\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n-1}(\theta_{i+1} - 2\theta_i + \theta_{i-1})^2$ in matrix form.

Consider this $y^* = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ first n elements are $(y_1, \cdots, y_n)$ and rest 0

vector length is $2n-2$

Then let X be the matrix of size $(2n-2) \times n$ s.t

first n rows and n columns is the I_n (identity of $n \times n$ size)

The rows proceeding will contain $\sqrt{\lambda}$ $-2\sqrt{\lambda}$ $\sqrt{\lambda}$ at some point depending on the row

This will allow for computation of $\lambda(\theta_{i+1} - 2\theta_i + \theta_{i-1})^2$

(for the last row, 1 -2 1 will be the last 3 elements)

$$X = \begin{pmatrix} 1 & 0 & - & - & - & - & - \\ 0 & 1 & 0 & - & - & - & - \\ 0 & 0 & 1 & - & - & - & - \\ \vdots & & & \ddots & & & \\ 0 & - & - & - & - & - & 1 \\ \sqrt{\lambda} & -2\sqrt{\lambda} & \sqrt{\lambda} & 0 & - & - & - \\ 0 & \sqrt{\lambda} & -2\sqrt{\lambda} & \sqrt{\lambda} & 0 & - & - \\ \vdots & & & \ddots & & & \\ 0 & 0 & - & - & \sqrt{\lambda} & -2\sqrt{\lambda} & \sqrt{\lambda} \end{pmatrix}$$

Then $\|y^* - X\theta\|^2$ will give $\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n-2}(\theta_{i+1} - 2\theta_i + \theta_{i-1})^2$

We already know each $\hat{\theta}_i$ is the minimizer

so, $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_n)^T$ minimizes $\|y^* - X\theta\|^2$ ∎

(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
$$\left\| \boldsymbol{y}^* - \overset{\text{fixed}}{X_{\bar{w}}\widehat{\boldsymbol{\theta}}_{\bar{w}}} - \overset{\text{minimize}}{X_w\boldsymbol{\theta}_w} \right\|$$
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.

We want to show that $\|y^* - X_{\bar{w}}\hat{\theta}_{\bar{w}} - X_w\hat{\theta}_w\|$ is smaller at $(k+1)^{th}$ iteration than at k^{th} iteration

At k^{th} iteration, we have w and the complement $\bar{w}$

Then, we have $\hat{\theta}_w$ s.t $\|y^* - X_{\bar{w}}\hat{\theta}_{\bar{w}} - X_w\theta_w\|$ is minimized

and call $\hat{\theta}_k$ s.t $X\hat{\theta}_k = X_{\bar{w}}\hat{\theta}_{\bar{w}} + X_w\hat{\theta}_w$

Then we have $\|y^* - X_{\bar{w}}\hat{\theta}_{\bar{w}} - X_w\hat{\theta}_w\| = \|y^* - X\hat{\theta}_k\|$

Since $\hat{\theta}_k$ minimizes.

At $(k+1)^{th}$ iteration, suppose we have w' and complement $\overline{w'}$

Then, we have $\hat{\theta}_{w'}$ s.t $\|y^* - X_{\overline{w'}}\hat{\theta}_{\overline{w'}} - X_{w'}\theta_{w'}\|$

Since $\|y^* - X\theta_k\| = \|y^* - X_{\overline{w'}}\hat{\theta}_{\overline{w'}} - X_{w'}\theta_{w'}\|$ — w' and w are different subset of the column vectors. Thus we can rearrange them

And we minimize $\|y^* - X_{\overline{w'}}\hat{\theta}_{\overline{w'}} - X_{w'}\theta_{w'}\|$ with respect to $\theta_{w'}$

Call it $\hat{\theta}_{k+1}$ that minimizes it, s.t $X\hat{\theta}_{k+1} = X_{\overline{w'}}\hat{\theta}_{\overline{w'}} + X_{w'}\hat{\theta}_{w'}$

Then, $\|y^* - X_{\overline{w'}}\hat{\theta}_{\overline{w'}} - X_{w'}\hat{\theta}_{w'}\| = \|y^* - X\hat{\theta}_{k+1}\| \leq \|y^* - X_{\overline{w'}}\hat{\theta}_{\overline{w'}} - X_{w'}\theta_{w'}\| = \|y^* - X\hat{\theta}_k\|$

Then we have $\|y^* - X\hat{\theta}_{k+1}\| \leq \|y^* - X\hat{\theta}_k\|$

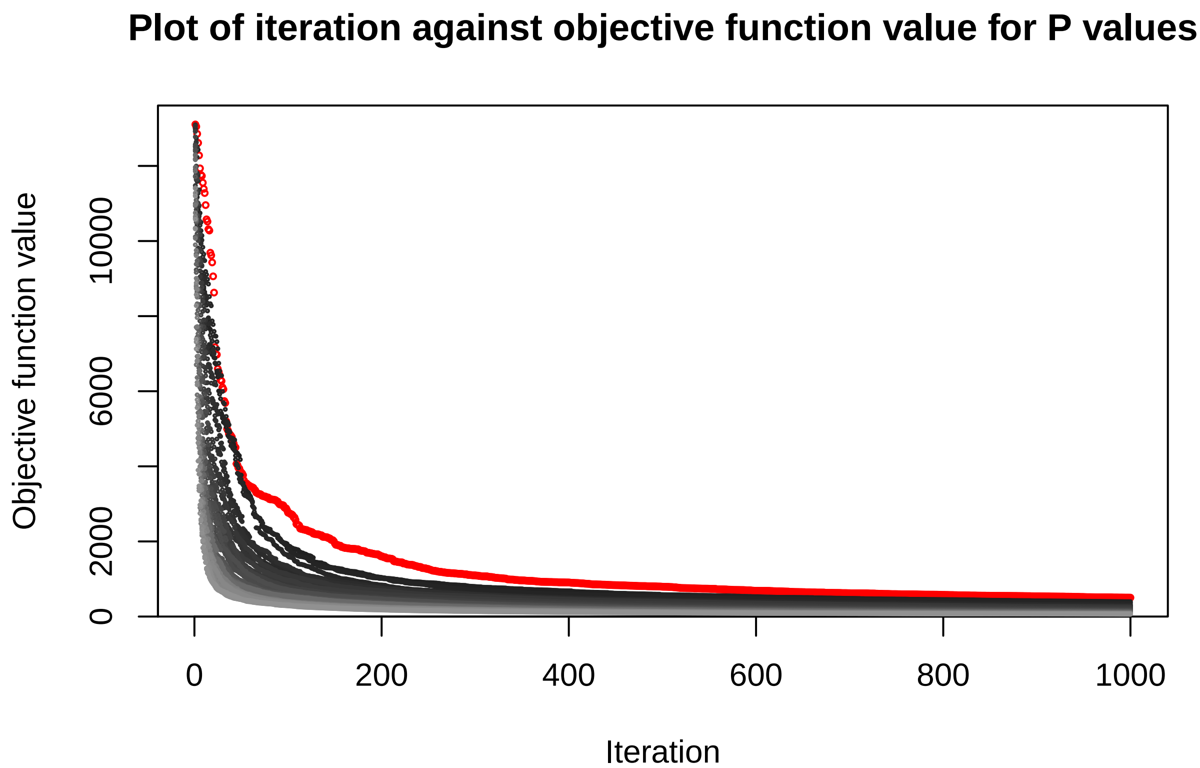
So this is non increasing from one iteration to next $\blacksquare$

STA410 A2

Harold Hyun Woo Lee

10/21/2020

```r
HP <- function(x,lambda,p=20,niter=200) {
        n <- length(x)
        a <- c(1,-2,1)
        aa <- c(a,rep(0,n-2))
        aaa <- c(rep(aa,n-3),a)
        mat <- matrix(aaa,ncol=n,byrow=T)
        mat <- rbind(diag(rep(1,n)),sqrt(lambda)*mat)
        xhat <- x
        x <- c(x,rep(0,n-2))
        sumofsquares <- NULL
        for (i in 1:niter) {
           w <- sort(sample(c(1:n),size=p))
           xx <- mat[,w]
           y <- x - mat[,-w]%*%xhat[-w]
           r <- lsfit(xx,y,intercept=F)
           xhat[w] <- r$coef
           sumofsquares <- c(sumofsquares,sum(r$residuals^2))
           }
        r <- list(xhat=xhat,ss=sumofsquares)
        r
        }

data <- scan("yield.txt")

#Calculating objective function for different values of p

#First create an empty vector and name it R
R <- vector(mode="list", length=46)

names(R) <- c((5:50))
#The for loop will store values of the objective function at its index corresponding to p value.
for (i in (5:50)){
  R[[i]] <- HP(data, lambda = 1000, p = i, niter = 1000)$ss

}

#Plot the first values of the objective function
#Then add on the other values from p = {6,...,50}
plot(R[[5]], cex = 0.4, col = "red", xlab= "Iteration", ylab="Objective function value", main="Plot of 
for (i in 6:50){
  points(R[[i]], col = colors()[160+i], cex=0.2)
}
```



In the graph, p = 5 is graphed using the color Red. As you increase the p value, you can see the change in color.
As P value increases to 50, the color becomes from black to grey. This suggests that the objective function always decreases with each iteration as p goes from 5 to 50.
Thus, suggests that the objective function is never increasing.