Homework 2 ANLY-601

Norman Hong

February 11, 2020

Worked with Arshia Singh and Jack Hart

Question 1: Convexity

Suppose that x, x' are in R and that $K(x, x') = K(x - x') = k(d)$

Which of the following functions of $f(x)$ are convex. Explain your rationale.

1.) $f(x) = \sum_{i=1}^{\infty} ||x||_p$ **for** $p > 0$

All p-norm are convex because the functoin is always positive or equal to 0.

2.) $f(d) = k(x, x') - k'(x, x')$ **where** $k(x, x') = k(x - x') = k(d)$ **is also stationary and a positive definite kernel**

All positive definite kernels are strictly convex. Therefore, $k(x, x')$ is strictly convex. This implies that the derivative, $k'(x, x')$ is monotone increasing, which means $-k'(x, x')$ is monotone decreasing. Not all monotone functions are convex. Straight positive or negative line or affine transformations are monotone functions that are convex. The derivative of a convex function is not a straight line, which means $k'(x, x')$ is not convex. The sum of a convex and non-convex function is not convex. Therefore, $f(d)$ is not convex.

3.) $f(d) = k(x, x') \cdot k'(x, x') - b$ **for some** $b \in R$

Any postive constant rescaling of a convex function is convex. However, $k'(x, x')$ is not constant. Therefore, $f(d)$ is not convex.

4.) $f(x) = ||x||_p - max(0, x)$ **for** $p > 0$

All p-norms are convex. $max(0, x)$ is convex, but $-max(0, x)$ is not convex. Plot the function, $y = -max(0, x)$, on the 2-d plane to see that it is not convex. Therefore, $f(x)$ is not convex because it is the sum of a convex and non-convex function.

5.) $f(x) = ||x||_p + max(0, x)$ **for** $p > 0$

The sum of 2 convex functions is also convex.

Question 2

Part 1 Plot the smoothed curve from file kernel_regression_1.csv using the following set of kernels.

1. Exponential

$$k(x, x') = exp(-3||x - x'||_1)$$

2. Radial basis function

$$k(x, x') = exp(-2||x - x'||_2^2)$$

3. Uniform

$$k(x, x') = I(||x - x'||_1 < .5$$

```
data <- read.csv("kernel_regression_1.csv")
head(data)
```

```
##       x         y
## 1 -5.00 0.3346185
## 2 -4.99 0.9743960
## 3 -4.98 0.4611669
## 4 -4.97 0.1867556
## 5 -4.96 1.2308795
## 6 -4.95 0.7614954
```

```
x_data <- data$x
y_data <- data$y
```

```
# 1 dimensional kernels
exp_kernel <- function(x1, x2){
  l1 <- abs(x1-x2)
  final <- exp(-3*l1)
  return(final)
}

radial_kernel <- function(x1, x2){
  l2 <- sqrt((abs(x1-x2))^2)
  l2_squared <- l2^2
  final <- exp(-2*l2_squared)
  return(final)
}

uniform_kernel <- function(x1, x2){
  l1 <- abs(x1-x2)
  if(l1 < .5){
    return(1)
  }
  else{
    return(0)
  }
}
```

```
# Calculate mean estimate at point x. 1-d x
mean_estimate <- function(x, kernel, x_data, y_data){
  numerator <- 0
  denominator <- 0
  for (i in 1:length(x_data)){
    numerator <- kernel(x, x_data[i])*y_data[i] + numerator
    denominator <- kernel(x, x_data[i]) + denominator
  }
  return(numerator/denominator)
}
```

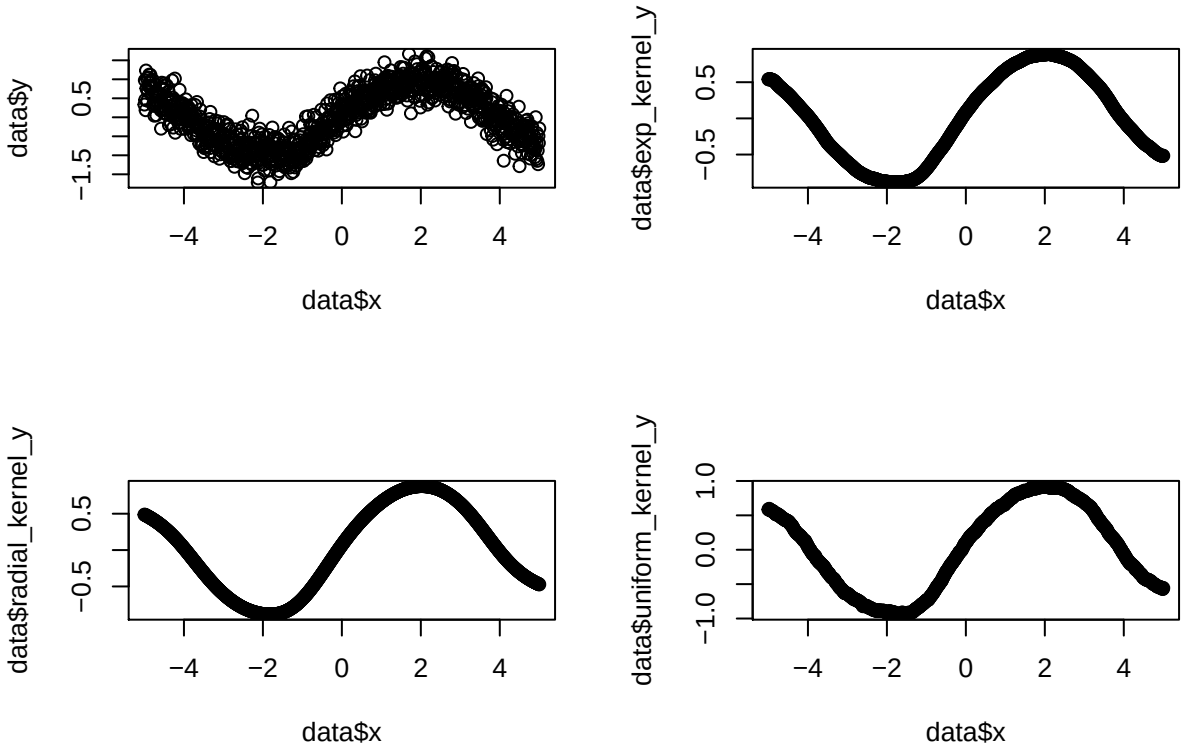
```
# exponential kernel smoothing
me <- c()
for (i in 1:length(x_data)){
```

```
  me[i] <- mean_estimate(x_data[i], exp_kernel, x_data, y_data)
}
data$exp_kernel_y <- me
```

```
# Radial basis kernel smoothing
me <- c()
for (i in 1:length(x_data)){
  me[i] <- mean_estimate(x_data[i], radial_kernel, x_data, y_data)
}
data$radial_kernel_y <- me
```

```
# Uniform kernel smoothing
me <- c()
for (i in 1:length(x_data)){
  me[i] <- mean_estimate(x_data[i], uniform_kernel, x_data, y_data)
}
data$uniform_kernel_y <- me
```

```
par(mfrow=c(2,2))
plot(data$x, data$y)
plot(data$x, data$exp_kernel_y)
plot(data$x, data$radial_kernel_y)
plot(data$x, data$uniform_kernel_y)
```



Part 2: What do you notice about the differences in the graph?

The kernel transformation plots look a lot smoother than the original plot because there is less noise. The difference between exponential and radial basis kernel is not pronouced. It looks like the radial transformation is slightly smoother than the exponential and uniform kernel. All transformed plots follow the same trend.

Part 3: Now consider kernel_regression_2.csv which is a 3d data set

$$(x_1, x_2, y)$$

1. Extend the kernels above and generate a smoothed surface.

What happens if you increase the radial basis functions bandwidth from 3 to 8?

The plot becomes more linear. If the bandwidth is high, the kernel fits less locally. This could have applications to reducing overfitting.

```
data <- read.csv("kernel_regression_2.csv")
x_data <- data.matrix(data[c('x', 'y')])
y_data <- data$z
```

```
# 2 dimensional kernels --> x1 and x2 are 2 dimensional points.
exp_kernel_2d <- function(x1, x2){
  l1 <- abs(x1[1]-x2[1]) + abs(x1[2] - x2[2])
  final <- exp(-3*l1)
  return(final)
}

radial_kernel_2d <- function(x1, x2){
  l2 <- sqrt((x1[1]-x2[1])^2 + (x1[2] - x2[2])^2)
  l2_squared <- l2^2
  final <- exp(-2*l2_squared)
  return(final)
}

uniform_kernel_2d <- function(x1, x2){
  l1 <- (abs(x1[1]-x2[1]) + abs(x1[2] - x2[2]))
  if (l1 < .5){
    return(1)
  }else{
    return(0)
  }
}
```

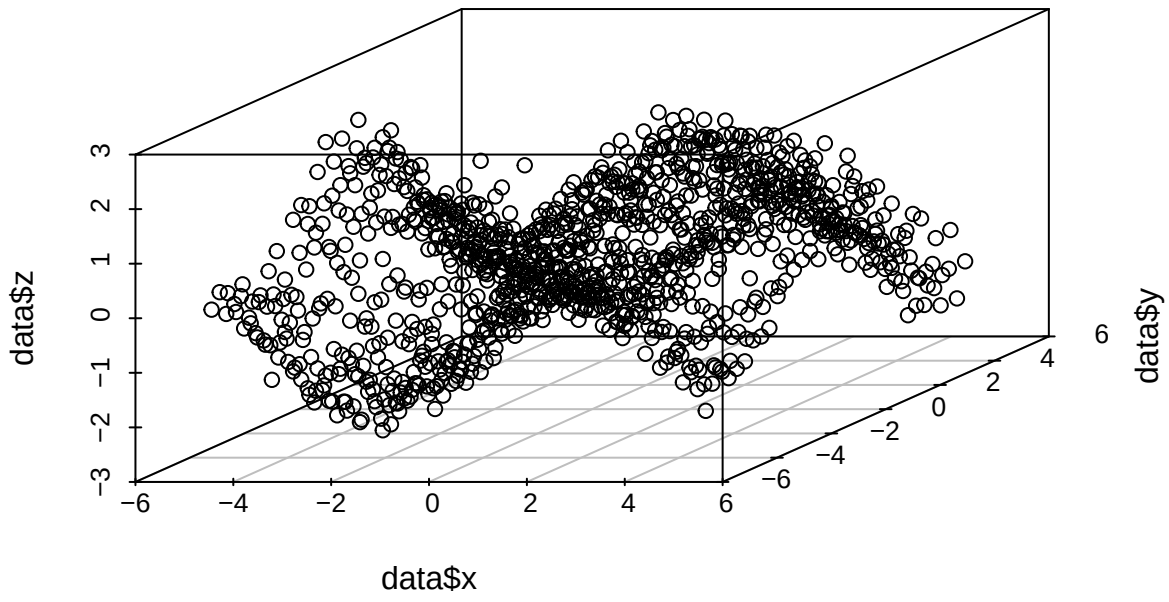
```
# Calculate mean estimate at point x. 2-d x
mean_estimate_2d <- function(x, kernel, x_data, y_data){
  numerator <- 0
  denominator <- 0
  for (i in 1:nrow(x_data)){
    numerator <- kernel(x, x_data[i,])*y_data[i] + numerator
    denominator <- kernel(x, x_data[i,]) + denominator
  }
  return(numerator/denominator)
}
```

```r
# exponential kernel smoothing
# nrow = number of rows for matrix; note that dataframe takes to long. matrix much faster.
me <- c()
for (i in 1:nrow(x_data)){
  me[i] <- mean_estimate_2d(x_data[i,], exp_kernel_2d, x_data, y_data)
}
data$exp_kernel_y <- me
```

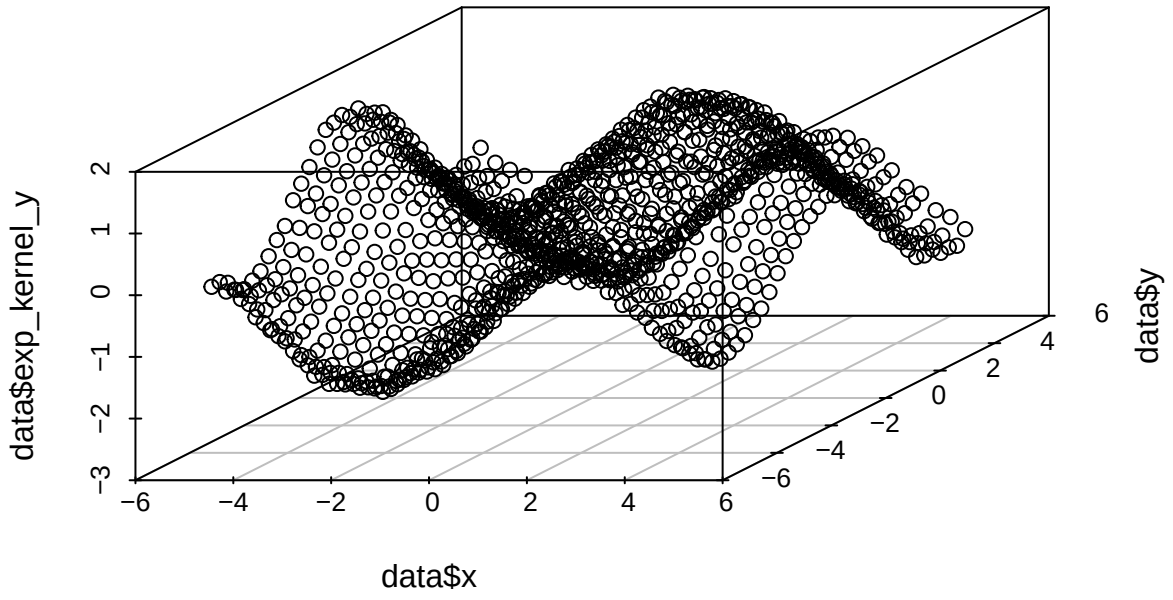
```r
# Radial basis kernel smoothing
me <- c()
for (i in 1:nrow(x_data)){
  me[i] <- mean_estimate_2d(x_data[i,], radial_kernel_2d, x_data, y_data)
}
data$radial_kernel_y <- me
```

```r
# Uniform kernel smoothing
me <- c()
for (i in 1:nrow(x_data)){
  me[i] <- mean_estimate_2d(x_data[i,], uniform_kernel_2d, x_data, y_data)
}
data$uniform_kernel_y <- me
```

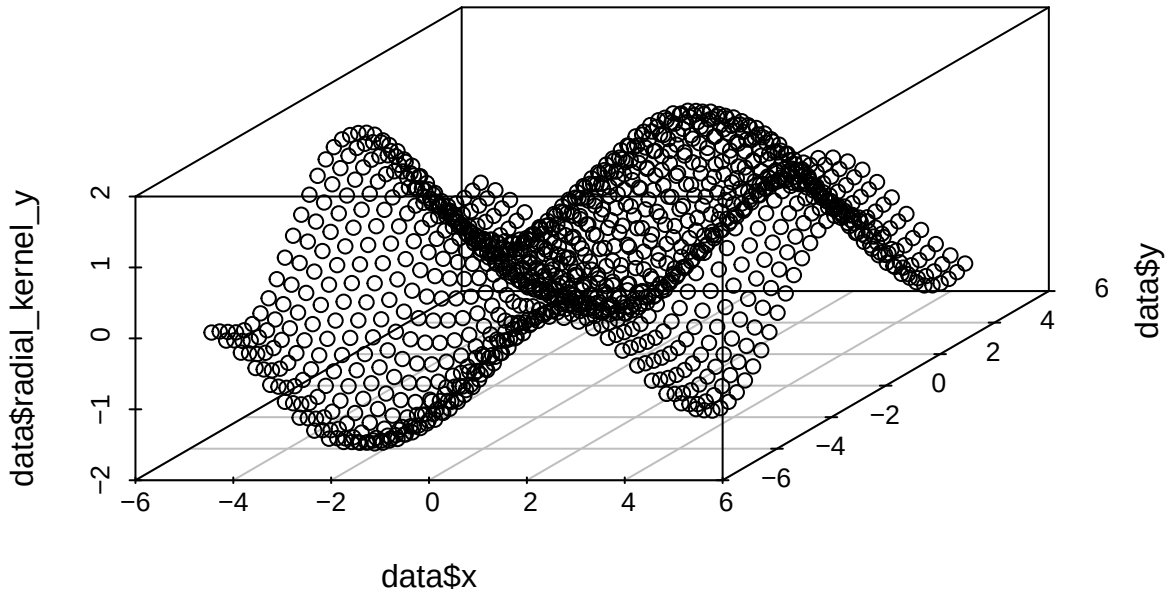
```r
scatterplot3d(data$x, data$y, data$z)
```



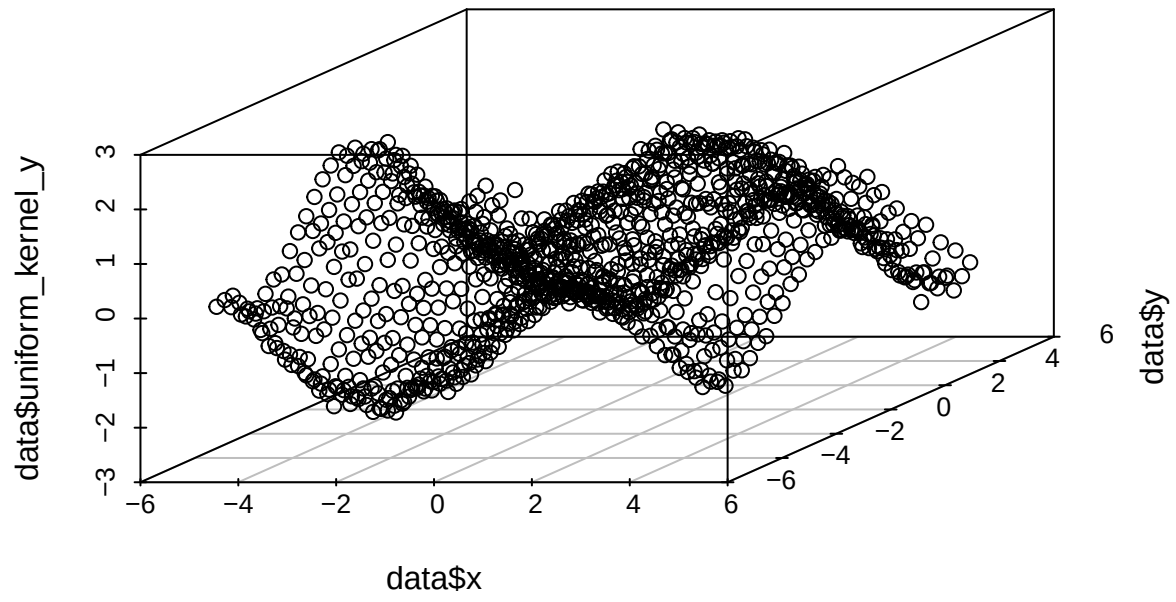

```r
scatterplot3d(data$x, data$y, data$exp_kernel_y)
```




```
scatterplot3d(data$x, data$y, data$radial_kernel_y)
```



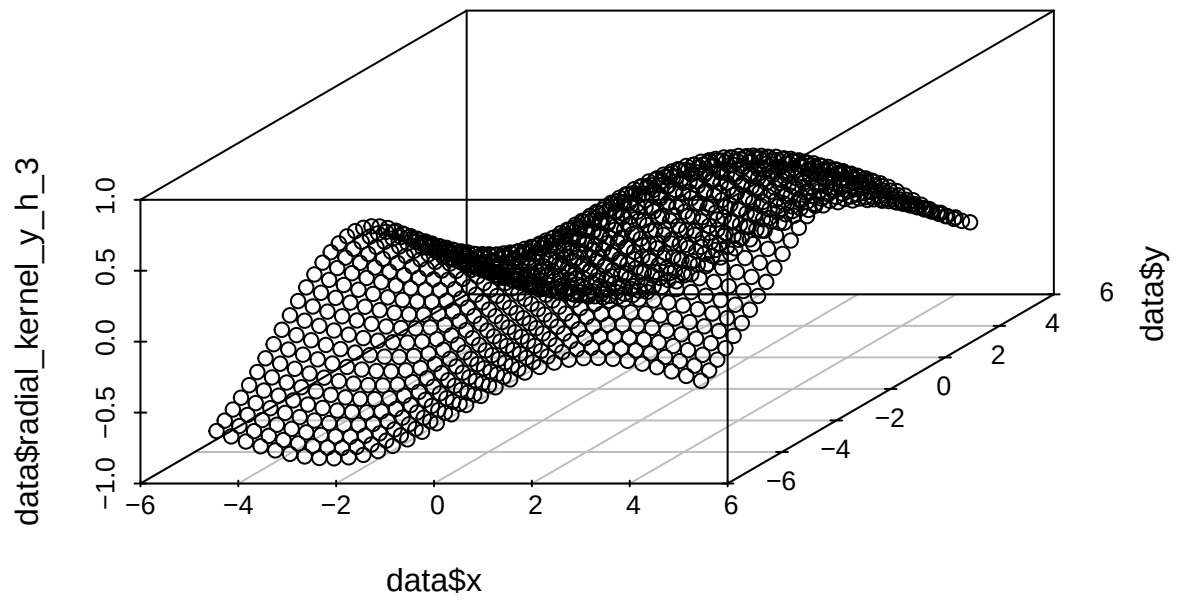

```
scatterplot3d(data$x, data$y, data$uniform_kernel_y)
```



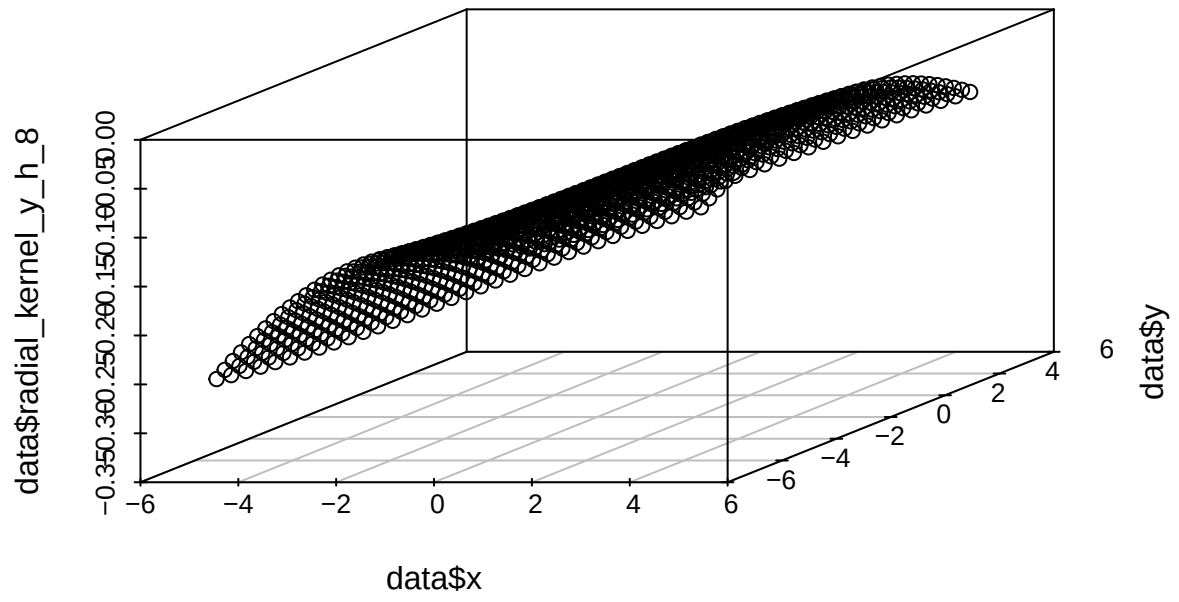

```
radial_kernel_2d <- function(x1, x2, h){
  l2 <- sqrt((x1[1]-x2[1])^2 + (x1[2] - x2[2])^2)
  l2_squared <- l2^2
  final <- exp(-l2_squared/(h^2))

  return(final)
}

# Calculate mean estimate at point x. 2-d x
temp <- function(x, kernel, x_data, y_data){
  numerator <- 0
  denominator <- 0
  for (i in 1:nrow(x_data)){
    numerator <- kernel(x, x_data[i,], 3)*y_data[i] + numerator
    denominator <- kernel(x, x_data[i,], 3) + denominator
  }
  return(numerator/denominator)
}

# Radial basis kernel smoothing
me <- c()
for (i in 1:nrow(x_data)){
  me[i] <- temp(x_data[i,], radial_kernel_2d, x_data, y_data)
}
data$radial_kernel_y_h_3 <- me
```

```r
# Calculate mean estimate at point x. 2-d x
temp <- function(x, kernel, x_data, y_data){
  numerator <- 0
  denominator <- 0
  for (i in 1:nrow(x_data)){
    numerator <- kernel(x, x_data[i,], 8)*y_data[i] + numerator
    denominator <- kernel(x, x_data[i,], 8) + denominator
  }
  return(numerator/denominator)
}

# Radial basis kernel smoothing
me <- c()
for (i in 1:nrow(x_data)){
  me[i] <- temp(x_data[i,], radial_kernel_2d, x_data, y_data)
}
data$radial_kernel_y_h_8 <- me
```

```r
scatterplot3d(data$x, data$y, data$radial_kernel_y_h_3)
```




```r
scatterplot3d(data$x, data$y, data$radial_kernel_y_h_8)
```



Problem 4.

For the following, suppose we have n observations. Calculate the posterior distribution for:

1.) $\mu\ N(\tau, \nu), \sigma^2\ InverseGamma(\alpha, \beta)$**, where** $X\ N(\mu, \sigma^2)$ **and** τ, ν, α, β **are all constant**

The following 3 equations are the pdfs of σ^2, μ, and the data generating mechanism

$$f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}exp[-\beta/\sigma^2] \tag{1}$$

$$f(\mu) = \frac{1}{(2\pi\nu)^{1/2}}exp\left[\frac{-(\mu-\tau)^2}{2\nu}\right] \tag{2}$$

$$\prod_{i=1}^{n} f(x|\mu, \sigma^2) = \left(\frac{1}{\sigma(2\pi)^{1/2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n} x_i^2 - 2\mu x_i + \mu^2}{2\sigma^2}\right] \tag{3}$$

The posterior equation is as follows:

$$f(\mu, \sigma^2|x) = \prod_{i=1}^{n} \frac{f(x_i|\mu, \sigma^2)f(\mu)f(\sigma^2)}{f(x_i)} \tag{4}$$

where $f(x_i)$ is a constant because the data is not a random variable under a bayesian framework. Therefore,

$$f(\mu, \sigma^2|x) \propto \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) f(\mu) f(\sigma^2) \tag{5}$$

Plugging in the likelihood and pdf into equation 5,

$$f(\mu, \sigma^2|x) \propto \left(\frac{1}{\sigma(2\pi)^{1/2}}\right)^n exp\left[-\frac{\sum_{i=1}^{n} x_i^2 - 2\mu x_i + \mu^2}{2\sigma^2}\right] \frac{1}{(2\pi\nu)^{1/2}} exp\left[\frac{-(\mu-\tau)^2}{2\nu}\right] \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} exp[-\beta/\sigma^2] \tag{6}$$

Looking at equation 6, notice that

$$\frac{1}{(2\pi\nu)^{1/2}} \left(\frac{1}{(2pi)^{1/2}}\right) (\beta^\alpha/\Gamma(\alpha))$$

is constant. Therefore, can ignore these. The posterior now becomes,

$$f(\mu, \sigma^2|x) \propto \left(\frac{1}{\sigma}\right)^n exp\left[-\frac{\sum_{i=1}^{n} x_i^2 - 2\mu x_i + \mu^2}{2\sigma^2}\right] exp\left[\frac{-(\mu-\tau)^2}{2\nu}\right] (\sigma^2)^{-\alpha-1} exp[-\beta/\sigma^2] \tag{7}$$

Grouping like terms.

$$f(\mu, \sigma^2|x) \propto \left(\frac{1}{\sigma^2}\right)^{n/2} (\sigma^2)^{-\alpha-1} exp\left[-\frac{\sum_{i=1}^{n} x_i^2 - 2\mu x_i + \mu^2}{2\sigma^2} - \frac{(\mu-\tau)^2}{2\nu} - \beta/\sigma^2\right] \tag{8}$$

$$f(\mu, \sigma^2|x) \propto \left(\sigma^2\right)^{(-n/2)-\alpha-1} exp\left[-\left[\frac{\sum_{i=1}^{n} x_i^2 - 2\mu x_i + \mu^2}{2\sigma^2} + \frac{(\mu-\tau)^2}{2\nu} + \beta/\sigma^2\right]\right] \tag{9}$$

The priors μ and σ^2 are independent from each other. The only way to introduce a dependence on each other is through the parameters. In this case, the parameters of each prior is constant and is not a random variable. For example, the mean parameter for μ is a constnat and not some function that depends on σ^2. This means that $f(\mu|\sigma^2) = f(\mu)$ and vise versa. This implies that $\mu|\sigma^2, x = \mu|x$ and $\sigma^2|\mu, x = \sigma^2|x$

Notice that

$$\sum_{i=1}^{n} x_i = n\bar{x} \tag{10}$$

and

$$\frac{\mu^2 - 2\tau\mu}{2\nu} + \frac{-2\mu n\bar{x} + n\mu^2}{2\sigma^2} = \frac{\mu^2}{2\nu} + \frac{n\mu^2}{2\sigma^2} - \frac{2\tau\mu}{2\nu} - \frac{2\mu n\bar{x}}{2\sigma^2} = \frac{\mu^2\left(\frac{1}{\nu} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{\tau}{\nu} + \frac{n\bar{x}}{\sigma^2}\right)}{2} \tag{11}$$

Using equation 11 and completing the square, equation 12 follows from equation 9.

$$f(\mu|\sigma^2, x) \propto exp\left[-\left[\frac{(\mu - \frac{\tau/\nu + n\bar{x}/\sigma^2}{1/\nu + n/\sigma^2})^2}{2/(1/\nu + n/\sigma^2)}\right]\right] + c \tag{12}$$

where c are a combination of constant terms and terms that contain σ^2

From equation 12, it can be seen that $\mu|\sigma^2, x \sim N\left(\frac{\tau/\nu + n\bar{x}/\sigma^2}{1/\nu + n/\sigma^2}, (1/\nu + n/\sigma^2)\right)$

The posterior $\sigma^2|\mu, x$ can be obtained by rerranging equation 9 by isolating the σ^2 terms.

$$f(\sigma^2|\mu, x) \propto \left(\sigma^2\right)^{(-n/2)-\alpha-1} exp\left[-\left[\frac{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} + \beta}{\sigma^2}\right]\right] exp\left[-\frac{(\mu-\tau)^2}{2\nu}\right] \tag{13}$$

Looking at equation 13, can see that $\sigma^2|\mu, x \sim InverseGamma\Big((n/2)+\alpha, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}+\beta\Big)$

2.) $p \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_k)$ **where** $X \sim Multinomial(p)$ **where** $p = (p_1, p_2, .., p_k)$**.**

The prior p is

$$f(p) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

where p is a vector, $\sum_{i=1}^{k} x_i = 1$, $x_i \geq 0$ for all i, $k \geq 2$ is the number of categroies, and $\alpha_i > 0$.

The data generating distribution $x|p$ is

$$f(x|p) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

where $\sum_{i=1}^{k} x_i = n$, $n > 0$ is the number of trials, $p_1, p_2, ..., p_k$ are the ith event probabilities and $\sum_{i=1}^{k} p_i = 1$.

The support of $f(p)$ is $p_i \in (0, 1)$ and $\sum_i p_i = 1$. Support of $f(x|p)$ is $x_i \in (0, 1, 2, ..., n)$ for $i = 0, 1, ..., k$ and $\sum_i x_i = n$.

Since there are n observations of $X|p$,

$$\prod_{i=1}^{n} f(x|p) = \frac{(n!)^n}{x_1!x_2!...x_k!} \cdot p_1^{\sum_{i=1}^{n} x_{1i}} p_2^{\sum_{i=1}^{n} x_{2i}} ... p_k^{\sum_{i=1}^{n} x_{ki}}$$

Then, the posterior is

$$f(p|x) \propto \frac{(n!)^n}{x_1!x_2!...x_k!} \cdot p_1^{\sum_{i=1}^{n} x_{1i}} p_2^{\sum_{i=1}^{n} x_{2i}} ... p_k^{\sum_{i=1}^{n} x_{ki}} \cdot \left(\frac{1}{\beta(\alpha)}\right) \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

$$f(p|x) \propto \underbrace{\left(\frac{1}{\beta(\alpha)}\right) \frac{(n!)^n}{\prod_{i=1}^{k} x_i!}}_{constant} \cdot \underbrace{\prod_{a=1}^{k} p_a^{\sum_{i=1}^{n} x_{ai} + \alpha_a - 1}}_{dirilectkernel}$$

Therefore,

$$f(p|x) \sim Dir\Big(\alpha_a = \sum_{i=1}^{n} x_{ai} + a_a - 1\Big)$$

for $a = 1, 2, 3, ..., k$.

3.) $\lambda \sim Gamma(\alpha, \beta)$ **where** $X \sim Poisson(\lambda)$**.**

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} exp\Big[-\beta\lambda\Big] \tag{1}$$

$$\prod_{i=1}^{n} f(x|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \tag{2}$$

$$f(\lambda|x) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} exp\Big[-\beta\lambda\Big] \tag{3}$$

$$f(\lambda|x) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)\prod_{i=1}^n x_i!}\right)\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \cdot \lambda^{\alpha-1}e^{-\beta\lambda} \tag{4}$$

$$f(\lambda|x) = \underbrace{\left(\frac{\beta^\alpha}{\Gamma(\alpha)\prod_{i=1}^n x_i!}\right)}_{constant} \cdot \underbrace{\lambda^{\sum_{i=1}^n x_i+\alpha-1}e^{-(n+\beta)\lambda}}_{gammakernel} \tag{6}$$

$$\lambda|x \sim Gamma(n\bar{x}+\alpha, \beta+n)$$

Problem 5

show that in logistic regression:

1.) The L2 penalty (ridge) is equivalent to a normal prior

Logistic regression assumes the data generating process is bernoulli distributed. let $f(y|p)$ denote the likelihood or pdf of a bernoulli distribution. Then,

$$f(y|p) = p^y(1-p)^{1-y} \tag{1}$$

where p is the probability of observing 1.
Assume there are n observations. Then the likelihood $f(y|p)$ of n observations is

$$\prod_{i=1}^n f(y_i|p) = \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i} \tag{2}$$

In a logistic regression model, p is assumed to be a function that is dependent on inputs x. Formally,

$$p = P(Y = y_i = 1) = \frac{1}{1+exp[-\beta x_i]} \tag{3}$$

where β is a matrix with dimensions $n \times k$ and x is a column vector with dimension that equals the number of predictor variables k.
Rewriting equation 2 in terms of β yields equation 4

$$\prod_{i=1}^n f(y_i|\beta) = \prod_{i=1}^n \Big(\frac{1}{1+exp[-\beta x_i]}\Big)^{y_i}\Big(1-\frac{1}{1+exp[-\beta x_i]}\Big)^{1-y_i} \tag{4}$$

Lets assume that the prior distribution of β is multivariatenormal. Then, the pdf of β is

$$f(\beta) = \Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big)exp\Big[-\frac{(\beta-\mu)^T(\beta-\mu)}{2\Sigma}\Big] \tag{5}$$

where μ and Σ are constant. Then the posterior $f(\beta|x)$ is as follows:

$$f(\beta|x) \propto \prod_{i=1}^n \Big(\frac{1}{1+exp[-\beta x_i]}\Big)^{y_i}\Big(1-\frac{1}{1+exp[-\beta x_i]}\Big)^{1-y_i} \cdot \Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big)exp\Big[-\frac{(\beta-\mu)^T(\beta-\mu)}{2\Sigma}\Big] \tag{6}$$

The MAP for $f(\beta|x) \propto f(x|\beta)f(\beta)$ is also the MAP for $log(f(\beta|x)) \propto log(\prod_{i=1}^n f(y_i|\beta)f(\beta))$ because log transformation is a 1 to 1 montone increasing function. It can be shown that the MAP is equal to the MLE when n approaches infinity, so the finding the MAP is equal to finding the MLE in the logistic regression.

$$log(f(x|\beta)) \propto log(\prod_{i=1}^{n} f(y_i|\beta)) + log(f(\beta)) \tag{7}$$

Notice that

$$log(\prod_{i=1}^{n} f(y_i|\beta)) = \sum_{i=1}^{n} y_i log(p(x_i)) + (1-y_i) log(1-p(x_i)) \tag{8}$$

$$log(\prod_{i=1}^{n} f(y_i|\beta)) = \sum_{i=1}^{n} log[1-p(x_i)] + \sum_{i=1}^{n} y_i log[p(x_i)/(1-p(x_i))] \tag{9}$$

Pluging in $p(x_i)$ and $log[p(x_i)/(1-p(x_i))]$ into equation 9 yields

$$log(\prod_{i=1}^{n} f(y_i|\beta)) = \sum_{i=1}^{n} log(1) + e^{\beta x_i} + \sum_{i=1}^{n} y_i(\beta x_i)$$

Also,

$$log(f(\beta)) = log\Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big) - \Big[\frac{(\beta-\mu)^T(\beta-\mu)}{2\Sigma}\Big]$$

$$log(f(\beta)) = log\Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big) - \frac{(\beta-\mu)^T\beta - (\beta-\mu)^T\mu}{2\Sigma} \tag{10}$$

$$log(f(\beta)) = log\Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big) - \frac{(\beta^T-\mu^T)\beta - (\beta^T-\mu^T)\mu}{2\Sigma}$$

$$log(f(\beta)) = log\Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big) - \frac{\beta^T\beta - \mu^T\beta - \beta^T\mu - \mu^T\mu}{2\Sigma} \tag{11}$$

Then the posterior becomes

$$log(f(\beta|x)) \propto \sum_{i=1}^{n} log(1) + e^{\beta x_i} + \sum_{i=1}^{n} y_i(\beta x_i) + log\Big(\frac{1}{det(\Sigma)^{1/2}(2\pi)^{k/2}}\Big) - \frac{\beta^T\beta - \mu^T\beta - \beta^T\mu - \mu^T\mu}{2\Sigma} \tag{12}$$

The goal is to maximize equation 12, which will be achived with the MAP. Therefore, we can see that the $\beta^T\beta$ term resembles the l2-norm by definition and it causes a decrease to the likelihood if coefficients are large.

The l1 penalty is a LaPlace prior

Assume that β is from a laplace distribution. Then the prior distribution $f(\beta)$ is

$$f(\beta) = \frac{1}{2b} exp\Big[-\frac{|\beta-\mu|}{b}\Big]$$

and

$$log(f(\beta)) = log\Big(\frac{1}{2b}\Big) - \frac{|\beta-\mu|}{b}$$

Therefore, the posterior distribution becomes

$$log(f(\beta|x)) \propto \sum_{i=1}^{n} log(1) + e^{\beta x_i} + \sum_{i=1}^{n} y_i(\beta x_i) + log\Big(\frac{1}{2b}\Big) - \frac{|\beta-\mu|}{b}$$

This is the 1-d case, and it can be seen that the $-|\beta|/b$ term acts as a force that minimizes the likelihood if the coefficients are large. In addition, by definiton it resembles the l1-norm. For the k-d case where k is the number of predictors, the same result follows if assuming each β_i prior is independent of each other. In this case, the product of the priors simply result in a sum of $|\beta|/b$ terms, which is by definition the l1-norm.

Problem 6

1.) Explain in your own words the difference between the posterior distribution and posterior predictive distribution.

The posterior predictive distribution is the posterior distribution averaged over all possible parameters. Essentially, the posterior predictive distribution accounts for uncertainty about the parameter.

2.) Which one would you use to predict future values of X? Explain your rationale.

The posterior predictive distribution should be used for predicting future values. It should provide a better estimate of future values than plugging in the single best estimate for the parameter, like the map or mean, because it accounts for the randomness of generating the parameter.

3.) Show that as $n \to \infty$ for a $X \sim N(\mu, \sigma^2)$ where $\mu \sim N(\alpha, \beta)$, $\sigma^2 \sim InverseGamma(\tau, \nu)$, that $\mu_{MAP} \to \mu_{MLE}$ and $\sigma^2_{MAP} \to \sigma^2_{MLE}$.

The posteriors were found in problem 4 to be

$$\mu|\sigma^2, x \sim N\left(\frac{\tau/\nu + n\bar{x}/\sigma^2}{1/\nu + n/\sigma^2}, (1/\nu + n/\sigma^2)\right) \tag{2}$$

$$\sigma^2|\mu, x \sim InverseGamma\left((n/2) + \alpha, \frac{\sum_{i=1}^n (x_i - \mu)^2}{2} + \beta\right) \tag{4}$$

The MAP is the mode of the posterior distribution, which means

$$\mu_{MAP} = \frac{\tau/\nu + n\bar{x}/\sigma^2}{1/\nu + n/\sigma^2} = \frac{\tau/v + (\sum_{i=1}^n x_i)/\sigma^2}{1/\nu + n/\sigma^2} = \frac{\tau\sigma^2/\nu}{\sigma^2/\nu + n} + \frac{\sum_{i=1}^n x_i}{\sigma^2/\nu + n} \tag{5}$$

and

$$\sigma^2_{MAP} = \frac{\frac{\sum_{i=1}^n (x_i - \mu)^2}{2} + \beta}{(n/2) + \alpha + 1} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n + 2\alpha + 2} + \frac{\beta}{n/2 + \alpha + 1} \tag{6}$$

$$\lim_{n\to\infty} \mu_{MAP} = \frac{\sigma^2\tau}{\sigma^2 + n\nu} + \frac{\sum_{i=1}^n x_i}{\sigma^2/\nu + n} \tag{7}$$

Looking at equation 7, it can be seen that the term of the left tends to 0 as n approaches infinity, which means the term on the right dominates. Because σ^2, ν, τ are constant, $\frac{\sum_{i=1}^n x_i}{n}$ dominates. This means the limit approaches μ_{MLE}.

$$\lim_{n\to\infty} \sigma^2_{MAP} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n + 2\alpha + 2} + \frac{\beta}{n/2 + \alpha + 1} \tag{8}$$

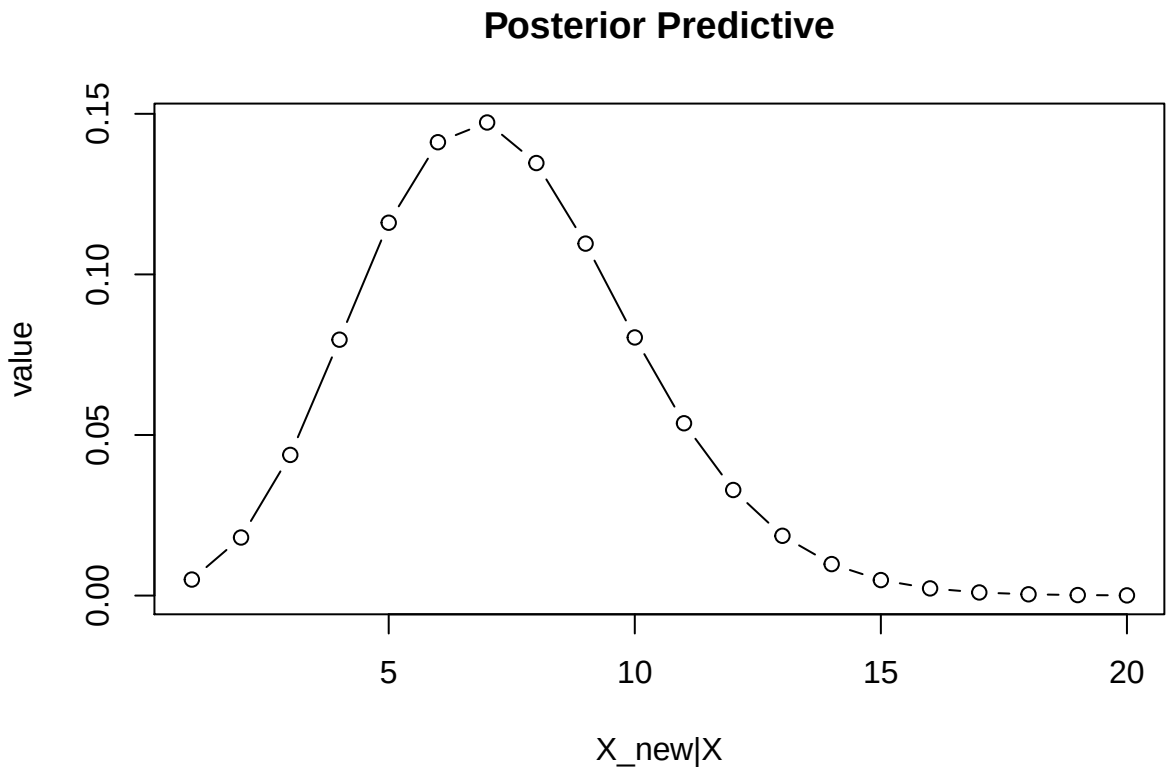
Looking at equation 8, it can be seen that the term on the right tends to 0 as n approaches infinity. Since α, β, μ are constant, $\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$ dominates, which is σ^2_{MLE}. For both MAP estimates, the bias terms approach 0, which is why the MLE results.

Problem 7

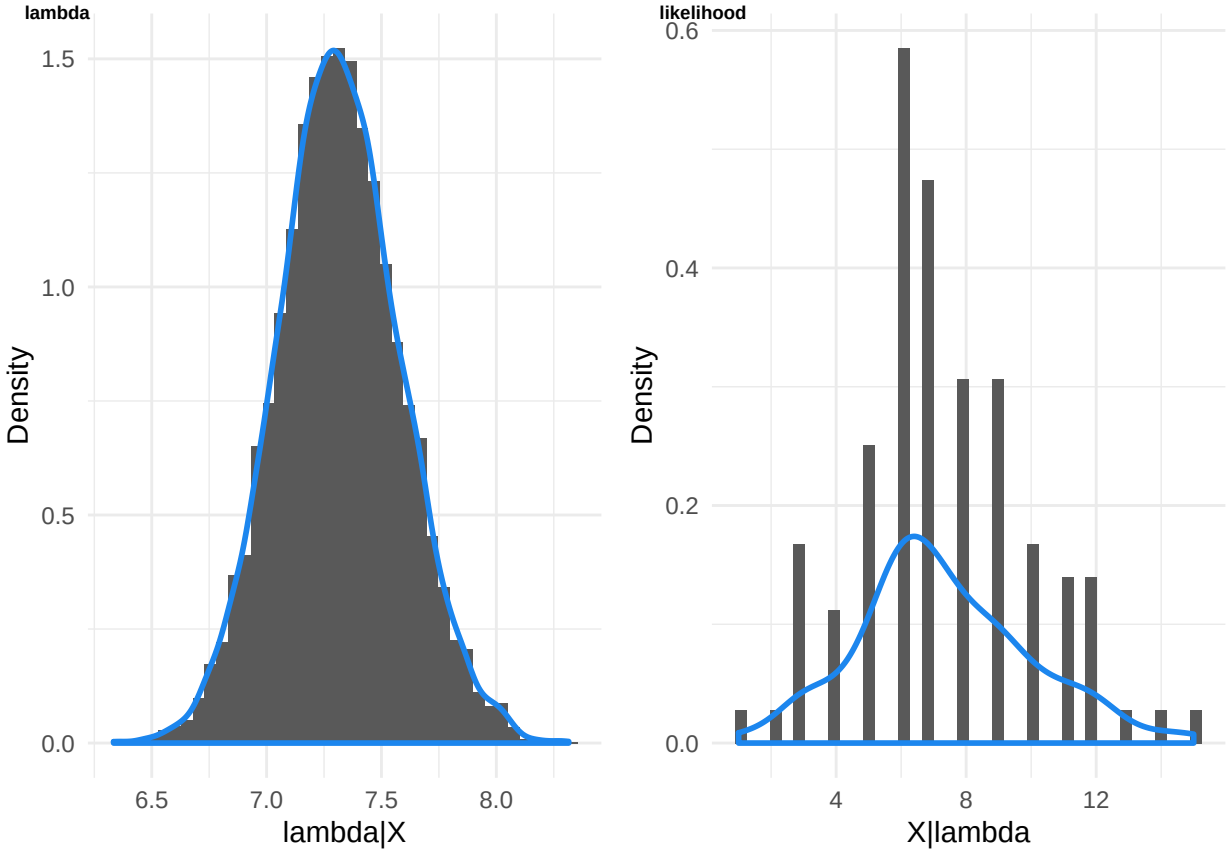
```
data_gibbs <- read.csv("gibbs.csv")
```

```
# Summary statistics for sample X
n <- nrow(data_gibbs)
sum_x <- sum(data_gibbs)

# Burn-in
burn <- 2000

# Gibbs sampler lambda | data
# shape = alpha, rate = beta
lambda <- rgamma(10000, shape=(30+sum_x), rate=(4+n))

lambda <- lambda[-(1:burn)]
```

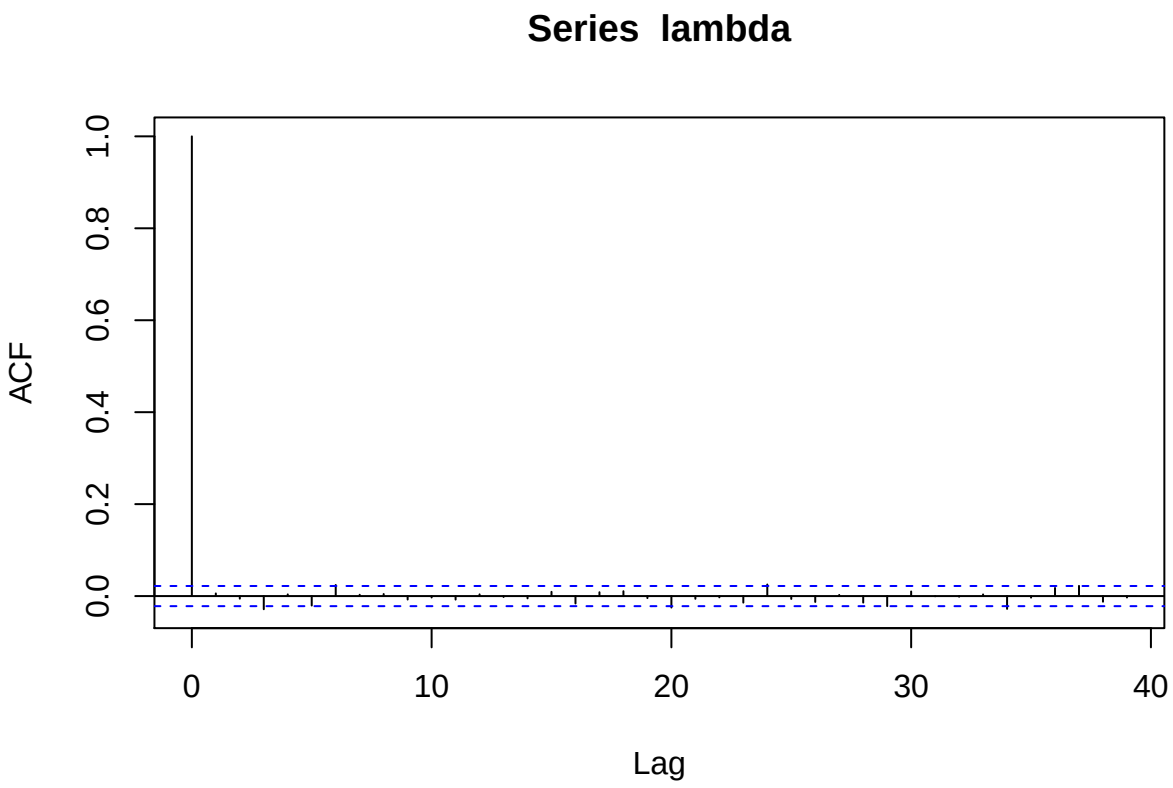
```
# Posterior Predictive
y <- rep(NA,20)
x <- seq(1:20)
for(i in 1:20){
  temp <- cumprod(761:(760+x[i]))
  num <- temp[length(temp)]
  y[i] <- ((104/105)^761)*((1/105)^x[i])*((num)/factorial(x[i]))
}
plot(y, type='b', xlab='X_new|X', ylab='value', main='Posterior Predictive')
```



```r
# Plot graph
plot_parameter <- function(data_in, parameter){
  ggplot(aes(x=parameter, stat(density)), data=tibble(parameter=data_in)) +
  geom_histogram(bins=40) + geom_density(color='dodgerblue2', size=1, alpha= 0.4) +
  theme_minimal() +
  ylab('Density') +
  xlab(parameter)
}

plot_grid(plot_parameter(lambda, parameter='lambda|X'), plot_parameter(data_gibbs$value, 'X|lambda'), la
```




```r
acf(lambda)
```



Problem 8

Referenced hw8hw9.ipynb file for the solution. Looking at the figures on the right, which are plots of data points, it can be seen that the distribution for $\lambda_1, \lambda_2, , \tau_1$ are stationary. This is further confirmed because the distribution plots converged. Lastly, tau_1 was found to be at time step 79. The same results follow for the optional component of the question.

Problem 9

Lake county is found to have the lowest amount of radon and Blue Earth County has the highest. The plot of data points from markov chain indicate that the priors, slope, intercept, and model error have converged.

Interview Question

2.) Show that, for the class of distributions in the regular exponential family, the mean update function(expected value of posterior distribution) is a weighted average of the prior distribution and observations.

Let $\theta|X_1, X_2, ..., X_n$ be the posterior distribution. Then the pdf is

$$\pi_{x_0,n0}(\theta|X_1, X_2, ..., X_n) = exp(n_0 x_0^T \theta - n_0 A(\theta))$$

where $A(\theta)$ is the moment generating function. The first derivative is

$$\Delta\pi_{x_0,n_0}(\theta) = (n_0x_0 - n_0A'(\theta))\pi(\theta)$$

where $A'(\theta)$ is the first derivative of the moment generating function.
Lets integrate both sides. Using the leibniz integral rule, we can pull the derivative operator out, and Since the integral of a probabiltiy density function is 1, the resulting equation becomes

$$\int \Delta\pi(\theta)d\theta = \Delta \int \pi(\theta)d\theta = \Delta(1) = 0$$

$$\int (n_0x_0 - n_0A'(\theta))\pi(\theta)d\theta = \int n_0x_0\pi(\theta) - \int n_0A'(\theta)\pi(\theta)d\theta = 0$$

$$\int n_0x_0\pi(\theta) = \int n_0A'(\theta)\pi(\theta)d\theta$$

$$\int n_0x_0\pi(\theta) = \int n_0A'(\theta)\pi(\theta)d\theta$$

Since $\int A'(\theta)\pi(\theta)d\theta = E[A'(\theta)]$ and n_0 is a constant,

$$\int x_0\pi(\theta) = \int A'(\theta)\pi(\theta)d\theta = E[A'(\theta)]$$

Lastly,

$$E[A'(\theta)] = x_0 \int \pi(\theta) = x_0 \cdot 1$$

The first derivative of a moment generating function of x is equal to the expected value of the x. Therefore,

$$E[A'(\theta)] = E[E(\theta|x_1, x_2, ..., x_n)] = E(\theta|x_1, x_2, ..., x_)n = x_0$$

where $x_0 = \frac{n\bar{x}}{n+n_0} + \frac{n_0x_0}{n+n0}$.

3.) Why do hierarchical models provide better model fit and regularization when data is sparse?

When there is a lot of data, no pooling works because there is a lot of data points so extremes of the data is captured in the sample, therefore can generalize better. When there is sparse data, partial pooling will be better because we can fit several distributions to subsets of the sample, and then use one of these common distributions to fit to new observations. This will provide better fit than a pooled model because these different distributions will better fit the data, which will lower the variance. Hierarchical models provide more regularization than a non-pooled model by shrinking coefficients closer to the mean, but not as much as a pooled model.