Homework #5: Density Estimation

Hyunsuk Ko

Due: Wed Oct 12 | 11:45am

DS 6030 | Fall 2022 | University of Virginia

Required R packages and Directories

```
data.dir = 'https://mdporter.github.io/DS6030/data/' # data directory
library(R6030)     # functions for DS 6030
library(ks)        # functions for KDE
library(tidyverse) # functions for data manipulation
```

Problem 1 Geographic Profiling

```
set.seed(2019)
n = 283
sd = 2.1
x = sqrt(rnorm(n, sd=sd)^2 + rnorm(n, sd=sd)^2)

geo = read.csv("../data/geo_profile.csv", header = FALSE)
colnames(geo) = c("dist")
```

Geographic profiling, a method developed in criminology, can be used to estimate the home location (roost) of animals (https://www.sciencedirect.com/science/article/pii/S0022519305004157) based on a collection of sightings. The approach requires an estimate of the distribution the animal will travel from their roost to forage for food.

A sample of 283 distances that pipistrelle bats traveled (in meters) from their roost can be found at:

- **Bat Data**: https://mdporter.github.io/DS6030/data//geo_profile.csv (https://mdporter.github.io/DS6030/data//geo_profile.csv)

One probability model for the distance these bats will travel is:

$$f(x;\theta) = \frac{x}{\theta}\exp\left(-\frac{x^2}{2\theta}\right)$$

where the parameter $\theta > 0$ controls how far they are willing to travel.

a. Derive the MLE for θ (i.e., show the math).

$L = \frac{x}{\theta}\exp\left(-\frac{x^2}{2\theta}\right)$

$logL$ = $\sum_{i=1}^{n} log(\frac{x}{\theta} * \exp\left(-\frac{x^2}{2\theta}\right)) = \sum_{i=1}^{n} logx - log\theta + log(\exp\left(-\frac{x^2}{2\theta}\right))$

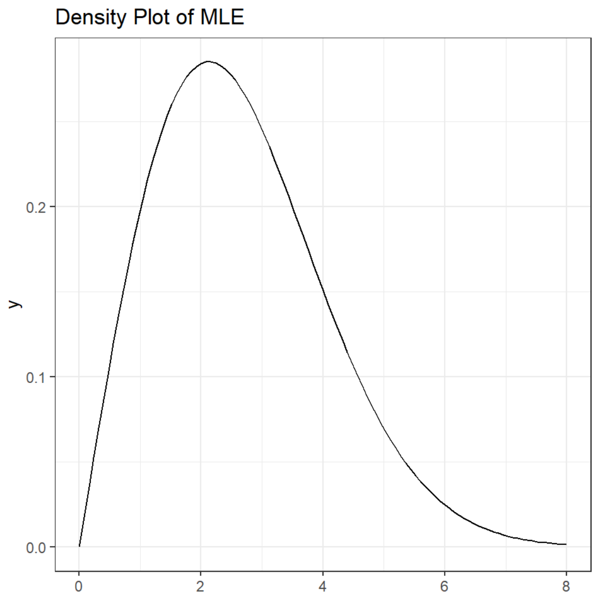
b. What is the MLE of θ for the bat data? (Use results from a, or use computational methods.)

$\frac{\partial logL}{\partial\theta} = \sum_{i=1}^{n} -\frac{1}{\theta} + \frac{1}{\exp\left(-\frac{x^2}{2\theta}\right)} * \exp\left(-\frac{x^2}{2\theta}\right) * \frac{x^2}{2\theta^2} = \sum_{i=1}^{n} -\frac{1}{\theta} + \frac{x^2}{2\theta^2} = -\frac{n}{\theta} + \frac{n}{2\theta^2} * \sum_{i=1}^{n} x_i^2 = \frac{-2n\theta + n * \sum_{i=1}^{n} x_i^2}{2\theta^2} = 0$

$\therefore \theta = \frac{\sum_{i=1}^{n} x_i^2}{2n}$

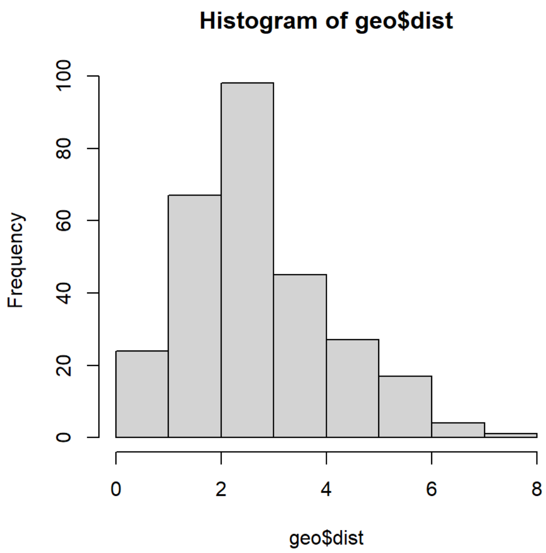
c. Using the MLE value of θ from part b, compute the estimated density at a set of evaluation points between 0 and 8 meters. Plot the estimated density.

```
values = geo$dist
mle_theta = sum(values ^ 2)/ (2*n)

function_x = function(x) {
  result = (x / mle_theta) * exp(-(x^2) / (2 * mle_theta))
  return (result)
}

ggplot() +
  xlim(0,8) +
  geom_function(fun = function_x) +
  labs(title = "Density Plot of MLE")
```

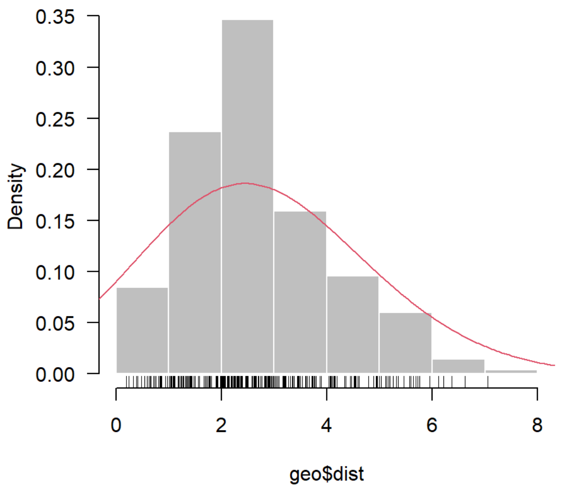


d. Estimate the density using KDE. Report the bandwidth you chose and produce a plot of the estimated density.

```
bw = 5
bks = seq(0,8)
hh = hist(geo$dist, breaks = bks)
```




```
f = kde(geo$dist, h = bw/3)

plot(hh,freq=FALSE,ylim=c(0,max(c(hh$density,f$estimate))), las=1,main='',border='white',col='grey75')
rug(jitter(geo$dist))
lines(f$eval.points,f$estimate,col=2,lwd=1.25)
```



e. Which model do you prefer, the parametric or KDE?

As KDE gives similar distribution as the true distribution, I prefer KDE.

Problem 2: Interstate Crash Density

Interstate 64 (I-64) is a major east-west road that passes just south of Charlottesville. Where and when are the most dangerous places/times to be on I-64? The crash data (link below) gives the mile marker and fractional time-of-week for crashes that occurred on I-64 between mile marker 87 and 136 in 2016. The time-of-week data takes a numeric value of *<dow>.<hour/24>*, where the dow starts at 0 for Sunday (6 for Sat) and the decimal gives the time of day information. Thus `time=0.0417` corresponds to Sun at 1am and `time=6.5` corresponds to Sat at noon.

- **Crash Data**: https://mdporter.github.io/DS6030/data//crashes16.csv (https://mdporter.github.io/DS6030/data//crashes16.csv)

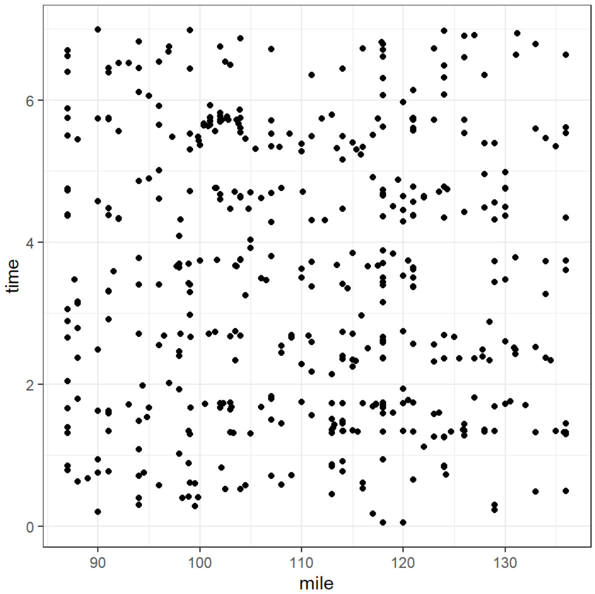
```
#readr::write_csv(tibble(x), "../data/crashes16.csv", col_names=FALSE)

crash = read.csv("../data/crashes16.csv")
crash
```

mile <dbl>	time <dbl>
87.0	6.61875
118.0	6.70347
120.0	0.05486
90.0	0.20625
124.2	0.72569
118.0	3.88125
114.0	4.46528
122.0	4.62639
122.0	4.64931
95.0	4.89861

1-10 of 456 rows Previous 1 2 3 4 5 6 ... 46 Next

a. Extract the crashes and make a scatter plot with mile marker on x-axis and time on y-axis.

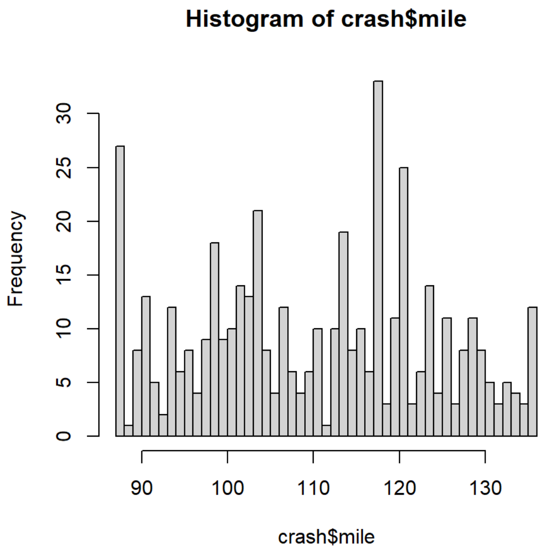
```
ggplot(crash, aes(x = mile, y = time)) +
  geom_point()
```



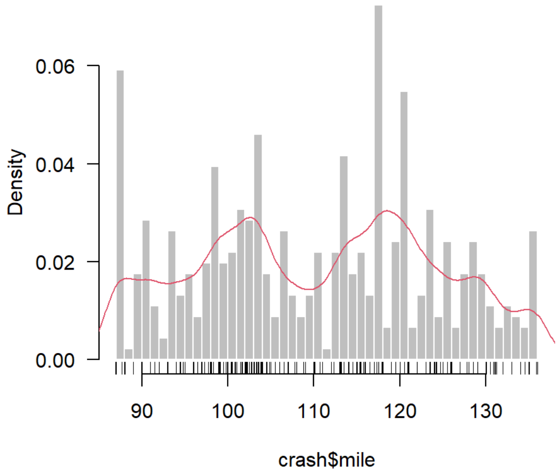
b. Use KDE to estimate the *mile marker* density.

- Report the bandwidth.
- Plot the density estimate.

```
bw = 5
bks = seq(87, 136)
hh = hist(crash$mile, breaks = bks)
```



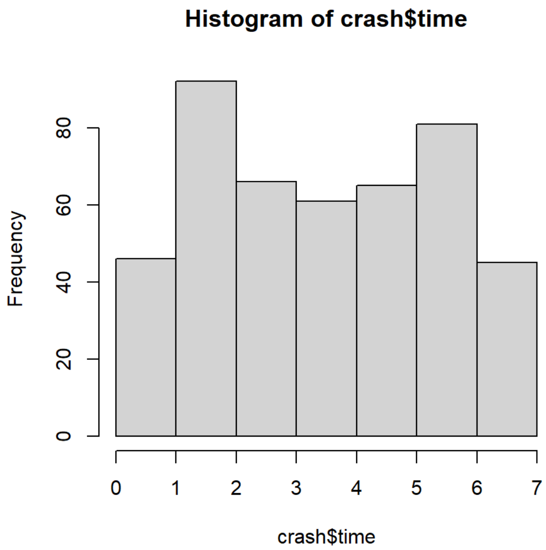

```
f = kde(crash$mile, h = bw/3)

plot(hh,freq=FALSE,ylim=c(0,max(c(hh$density,f$estimate))), las=1,main='',border='white',col='grey75')
rug(jitter(crash$mile))
lines(f$eval.points,f$estimate,col=2,lwd=1.25)
```



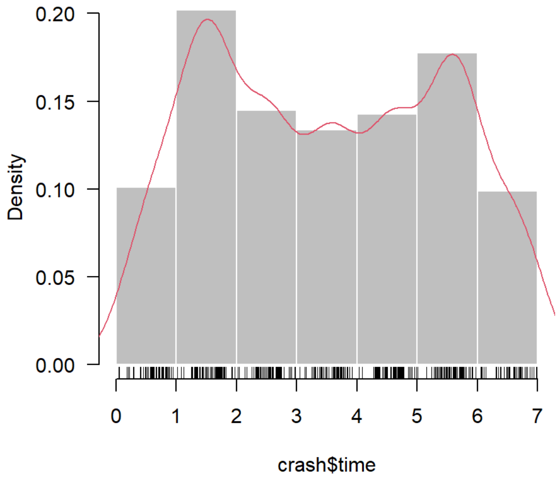
c. Use KDE to estimate the temporal *time-of-week* density.

- Report the bandwidth.
- Plot the density estimate.

```
bw = 10
bks = seq(0, 7)
hh = hist(crash$time, breaks = bks)
```



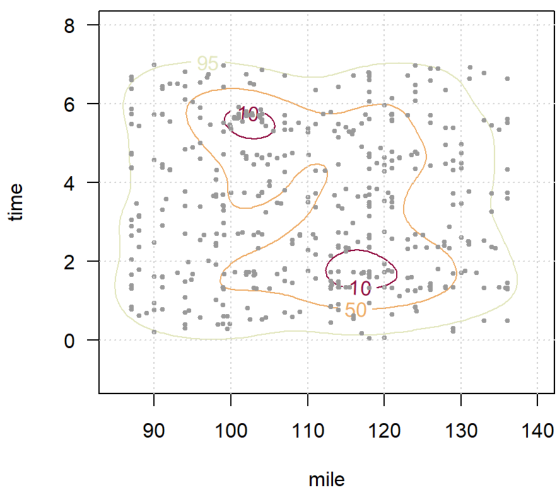

```
#f = kde(crash$time, h = bw/3)
f = kde(crash$time)

plot(hh,freq=FALSE,ylim=c(0,max(c(hh$density,f$estimate))), las=1,main='',border='white',col='grey75')
rug(jitter(crash$time))
lines(f$eval.points,f$estimate,col=2,lwd=1.25)
```



d. Use KDE to estimate the bivariate mile-time density.

- Report the bandwidth parameters.
- Plot the bivariate density estimate.

```
H1 = Hscv(crash)

#smoothed cross-validation bw estimator
f1 = kde(crash,H=H1)

#useHformultivariatedata
plot(f1, cont=c(10,50,95), las=1,xlim=c(85,140),ylim=c(-1,8))
points(crash,pch=19,cex=.5,col='grey60')
grid()
```



e. Based on the estimated density, approximate the most dangerous place and time to drive on this strech of road. Identify the mile marker and time-of-week pair.

Mile between 100 ~ 115 at Friday night, and mile between 115 ~ 120 at Monday night seems to be the most dangerous place and time to drive.