Homework 11 Solutions
Statistics 200B
Due May 2, 2019

1. Prove Theorem 20.7 from Wasserman: that the following identity holds for the cross-validation estimator of rish for a histogram:

$$\int \left(\hat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\hat{p}_j^2$$

Proof:

Since

$$\hat{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} I(x \in B_j),$$

and

$$\hat{f}_{(-i)}(X_i) = \sum_{j=1}^{m} \frac{\hat{p}_{j(-i)}}{h} I(X_i \in B_j),$$

where $\hat{p}_j = \nu_j/n$ and $\hat{p}_{j(-i)} = \nu_{j(-i)}/(n-1)$. ν_j is the number of observations in B_j, and $\nu_{j(-i)}$ is the number of observations after omitting X_i in B_j. If $X_i \in B_j$, $\nu_{j(-i)} = \nu_j - 1$. Otherwise, $\nu_{j(-i)} = \nu_j$.

Hence,

$$\hat{f}_{(-i)}(X_i) = \sum_{j=1}^{m} \frac{\nu_j - 1}{(n-1)h} I(X_i \in B_j).$$

From the above, and using the fact that $I(x \in B_j)I(x \in B_{j'}) = 0$ for $j \neq j'$,"

$$\int \left(\hat{f}_n(x)\right)^2 dx = \int \left(\sum_{j=1}^{m} \frac{\hat{p}_j}{h} I(x \in B_j)\right)^2 dx = \int \sum_{j=1}^{m} \frac{\hat{p}_j^2}{h^2} I(x \in B_j) dx$$
$$= \sum_{j=1}^{m} \int_{B_j} \frac{\hat{p}_j^2}{h^2} dx = \sum_{j=1}^{m} \int_{\frac{j-1}{m}}^{\frac{j}{m}} \frac{\hat{p}_j^2}{h^2} dx = \sum_{j=1}^{m} \frac{\hat{p}_j^2}{h^2} h = \frac{1}{h}\sum_{j=1}^{m} \hat{p}_j^2.$$

$$\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\nu_j - 1}{(n-1)h} I(X_i \in B_j) = \sum_{j=1}^{m} \frac{\nu_j - 1}{(n-1)h} \left(\sum_{i=1}^{n} I(X_i \in B_j) \right)$$
$$= \sum_{j=1}^{m} \frac{\nu_j - 1}{(n-1)h} \nu_j = \sum_{j=1}^{m} \frac{\nu_j^2}{(n-1)h} - \sum_{j=1}^{m} \frac{\nu_j}{(n-1)h} = \frac{n^2}{(n-1)h} \sum_{j=1}^{m} \hat{p}_j^2 - \frac{n}{(n-1)h}.$$

Therefore,

$$\int \left(\hat{f}_n(x)\right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i) = \frac{1}{h} \sum_{j=1}^{m} \hat{p}_j^2 - \frac{2}{n} \left(\frac{n^2}{(n-1)h} \sum_{j=1}^{m} \hat{p}_j^2 - \frac{n}{(n-1)h} \right)$$
$$= \left(\frac{1}{h} - \frac{2n}{(n-1)h} \right) \sum_{j=1}^{m} \hat{p}_j^2 + \frac{2}{(n-1)h}$$
$$= \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h} \sum_{j=1}^{m} \hat{p}_j^2.$$

2. Calculate $E[\hat{f}_n(x)]$ and $V[\hat{f}_n(x)]$ when $X_1, \ldots, X_n$ are *iid* random variables with PDF f and
$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h} \right)$$
for some kernel function K. Express the variance using the form $V[Z] = E[Z^2] - (E[Z])^2$, rather than $V[Z] = E(Z - E[Z])^2$.

Solution:

$$E[\hat{f}_n(x)] = E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h} \right) \right] = \frac{1}{nh} \sum_{i=1}^{n} E\left[K\left(\frac{x - X_i}{h} \right) \right]$$
$$(X_1, \ldots, X_n \text{ are } iid) = \frac{1}{h} E\left[K\left(\frac{x - X_1}{h} \right) \right]$$
$$= \frac{1}{h} \int K\left(\frac{x - y}{h} \right) f(y) dy.$$

$$
\begin{aligned}
V[\hat{f}_n(x)] &= V\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] \\
(X_1,\ldots,X_n \text{ are } iid) &= \frac{1}{n^2h^2}\sum_{i=1}^{n}V\left[K\left(\frac{x-X_i}{h}\right)\right] \\
&= \frac{1}{nh^2}V\left[K\left(\frac{x-X_1}{h}\right)\right] \\
&= \frac{1}{nh^2}\left\{E\left[K^2\left(\frac{x-X_1}{h}\right)\right]-\left(E\left[K\left(\frac{x-X_1}{h}\right)\right]\right)^2\right\} \\
&= \frac{1}{nh^2}\left\{\int K^2\left(\frac{x-y}{h}\right)f(y)dy-\left(\int K\left(\frac{x-y}{h}\right)f(y)dy\right)^2\right\}.
\end{aligned}
$$

3. Let $b(x) = E[\hat{f}_n(x)] - f(x)$, using the same setup as in Problem 1. Show that

$$
b(x) \approx \frac{1}{2}h^2 f''(x)\sigma_K^2
$$

where $\sigma_K^2 = \int x^2 K(x)dx$. *Hint: Your expression for* $E[\hat{f}_n(x)]$ *will contain a term like* $\frac{x-y}{h}$. *Make the change of variable* $y = x - ht$, *then use the Taylor expansion*

$$
f(x-ht) \approx f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x).
$$

Solution:

$$
\begin{aligned}
E[\hat{f}_n(x)] &= \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{x-y}{h}\right)f(y)dy \\
(\text{Let } y = x-ht) &= \frac{1}{h}\int_{+\infty}^{-\infty} K(t)f(x-ht)(-h)dt \\
&= \int_{-\infty}^{+\infty} K(t)f(x-ht)dt \\
&\approx \int_{-\infty}^{+\infty} K(t)\left[f(x)-htf'(x)+\frac{1}{2}h^2t^2f''(x)\right]dt \\
&= f(x)\int_{-\infty}^{+\infty} K(t)dt - hf'(x)\int_{-\infty}^{+\infty} tK(t)dt + \frac{1}{2}h^2f''(x)\int_{-\infty}^{+\infty} t^2K(t)dt
\end{aligned}
$$

$$\left(\text{Since } \int_{-\infty}^{+\infty} K(t)dt = 1; \int_{-\infty}^{+\infty} tK(t)dt = 0; \int_{-\infty}^{+\infty} t^2K(t)dt = \sigma_K^2.\right)$$
$$= f(x) + \frac{1}{2}h^2 f''(x)\sigma_K^2.$$

Therefore,

$$b(x) = E[\hat{f}_n(x)] - f(x) \approx \frac{1}{2}h^2 f''(x)\sigma_K^2.$$

4. Read the file `glass.dat` from bCourse into `R`. Estimate the density of the first variable (refractive index) using a histogram and using a kernel density estimator with Normal kernel. Use cross-validation to choose the amount of smoothing in each case. For both the histogram and kernel density estimator, turn in plots showing (a) the estimated risk for different choices of smoothing, and (b) the estimator using the optimal choice of smoothing. Also turn in your code.

   Solution:

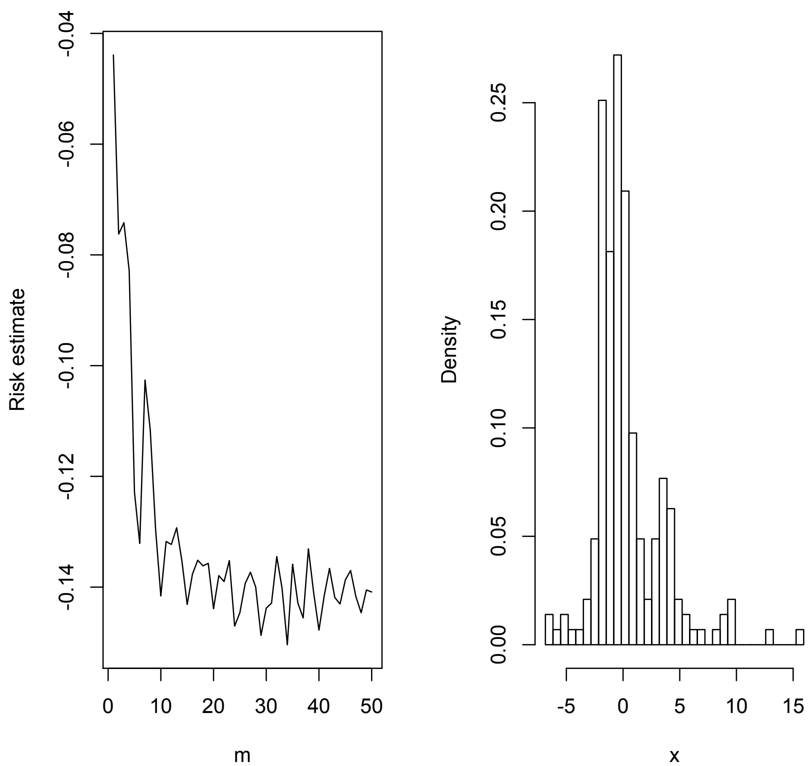
   First, read in the data.

```
glass <- read.table("glass.dat", header=TRUE)
head(glass)
x <- glass$RI # This is the data of which we want to estimate the density.
```

   Second, use a histogram. The optimal number of bins is found by minimizing the estimated cross-validation risk as in the function `jhat_hist`.

```
jhat_hist <- function(m, x){
  breaks <- seq(min(x), max(x), length = m+1) # m bins
  h <- (breaks[2] - breaks[1])
  n <- length(x)
  phat <- hist(x, breaks = breaks, plot = FALSE)$counts / n
  return(2/((n-1)*h) - (n+1)/((n-1)*h) * sum(phat^2))
}
mvals <- 1:50
risk <- sapply(mvals, jhat_hist, x = x)
par(mfrow = c(1, 2))
plot(mvals, risk, type = "l", xlab = "m", ylab = "Risk estimate")
mopt <- mvals[risk==min(risk)]
hist(x, breaks = seq(min(x), max(x), length = mopt+1), prob = TRUE, main = "")
```

The optimal number of bins is 34. The plot of the estimated risk vs. the number of bins and the histogram with the optimal number of bins is below.

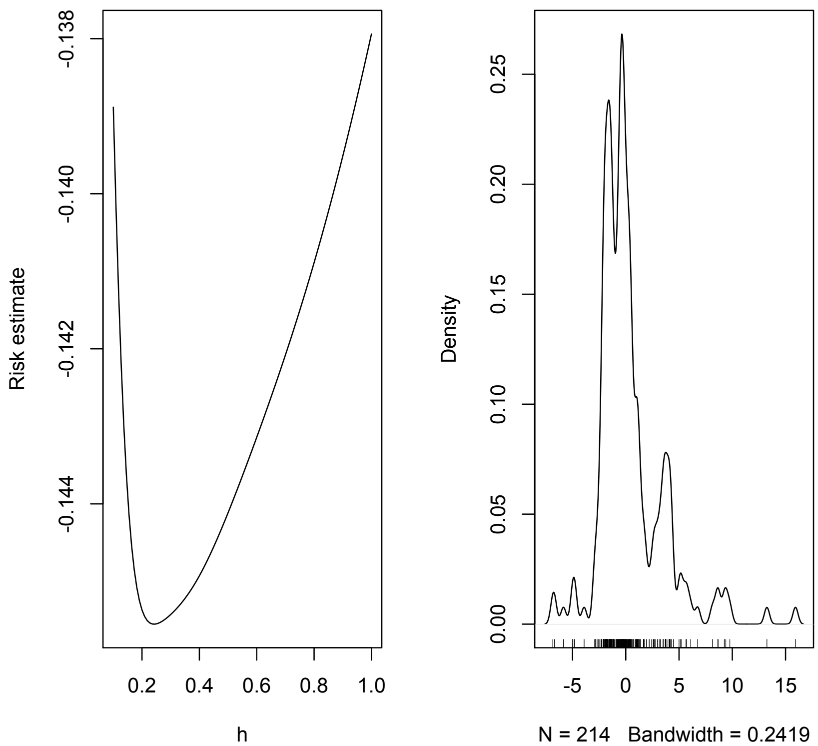


Third, use a kernel density estimate. The optimal bandwidth is found by minimizing the estimated cross-validation risk as in the function `jhat_kde`.

```
jhat_kde <- function(h, x){ # defined for normal kernel only
  n <- length(x)
  crossmat <- outer(x, x, "-")/h
  kstar <- dnorm(crossmat, sd = sqrt(2)) - 2*dnorm(crossmat)
  return(sum(kstar)/(h*n^2) + 2*dnorm(0)/(n*h))
}
hseq <- seq(0.1, 1, length = 100)
risk <- sapply(hseq, jhat_kde, x = x)
par(mfrow = c(1, 2))
plot(hseq, risk, type = "l", xlab = "h", ylab = "Risk estimate")
```

```
h.opt <- optimize(jhat_kde, lower = 0.1, upper = 1, x = x)$minimum
plot(density(x, bw = h.opt), main = "")
rug(x)
```

The optimal bandwidth is about 0.24. The plot of the estimated risk vs. the bandwidth h and the kernel density plot with the optimal bandwidth is below.



We can see that the histogram estimate and the kernel density estimate are very similar in this case.

5. Suppose we observe $X_1, \ldots, X_n$ *iid* bivariate random variables, with $X_i = (X_{i1}, X_{i2})$. Consider the two-dimensional kernel density estimator of the form

$$\hat{f}_n(x) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}\right) K\left(\frac{x_2 - X_{i2}}{h_2}\right)$$

where $x = (x_1, x_2)$ and K is the (univariate) Normal PDF with mean zero and variance one.

(a) Show this is equivalent to using a two-dimensional Normal kernel, with different standard deviations for each element and zero correlation between them.

Proof:

Define the two-dimensional Normal kernel as

$$K^*(x, y; \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right]\right),$$

which has mean 0, the standard deviation of X as σ_x, the standard deviation of Y as σ_y, and zero correlation between X and Y.

Since the univariate standard Normal kernel is

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}\right) K\left(\frac{x_2 - X_{i2}}{h_2}\right) \\
&= \frac{1}{nh_1h_2} \sum_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1 - X_{i1})^2}{2h_1^2}\right)\right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2 - X_{i2})^2}{2h_2^2}\right)\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1h_2} \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left[\frac{(x_1 - X_{i1})^2}{h_1^2} + \frac{(x_2 - X_{i2})^2}{h_2^2}\right]\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1h_2} K^*\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}; 1, 1\right).
\end{aligned}$$

Therefore, $\hat{f}_n(x)$ defined in this problem is equivalent to using a two dimensional Normal kernel, with standard deviations for each element as 1 and 0 correlation between them.

Equivalently, $K^*\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}\right)$ can be seen as the two-dimensional normal density with mean zero and standard deviations h_1 and h_2, evaluated at $x_1 - X_{i1}$ and $x_2 - X_{i2}$.

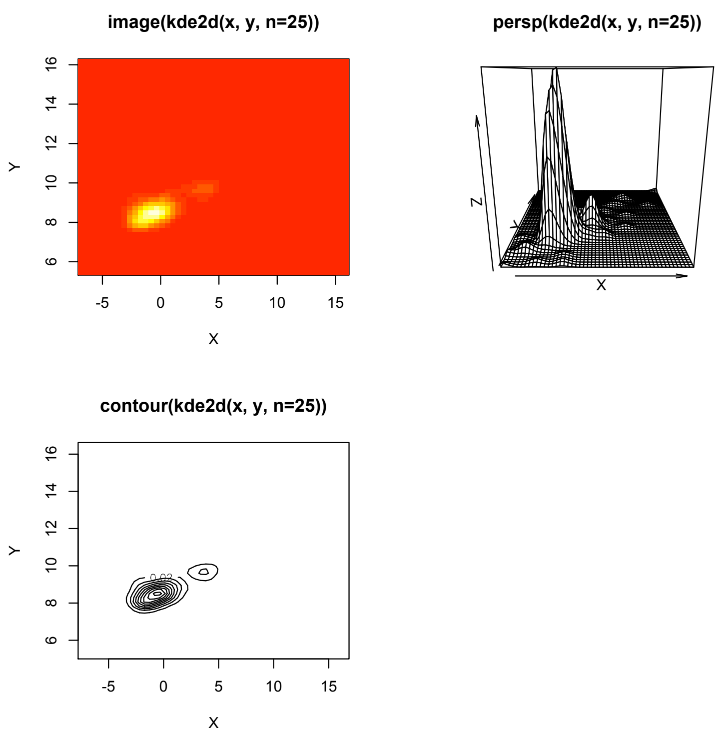
(b) This estimator is implemented in `R` by the function `kde2d`, which is part of the `MASS` package. Use this function to estimate the joint density of the first and seventh variables (`RI` and `Ca`) in the glass dataset. (You may use the function's default bandwidth for this problem.) Experiment with plotting the results, using the functions `image`, `persp`, and `contour`, and turn in the one you like best. *Hint: the argument* `n` *to* `kde2d` *changes the resolution of the resulting image.*

Solution:

The code are as follows.

```
library(MASS)
x <- glass$RI
y <- glass$Ca
result <- kde2d(x, y, n=50)
par(mfrow=c(2,2))
image(result, xlab="X", ylab="Y", main="image(kde2d(x, y, n=50))")
persp(result, xlab="X", main="persp(kde2d(x, y, n=50))")
contour(result, xlab="X", ylab="Y", main="contour(kde2d(x, y, n=50))")
```

The best plotting results are obtained with the default resolution $n = 50$. The plots are as follows.



The peak of the 2-dimensional kernel-density estimator is around (0,8).

(c) Comment on how the estimator is able to capture the correlation between the two variables, even though there is no correlation in the kernel itself.

Answer:

The correlation between X and Y corresponds to diagonally oriented contours in their joint density. Even though each kernel is a density with contours aligned with the two coordinate axes (corresponding to no correlation), their sum forms diagonal contours simply due to the diagonal orientation of the sampled (X_i, Y_i) points induced by their correlation.

Also, from the form of the kernel density estimator, we can have the following.

$$\hat{f}_n(x, y) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right),$$

X and Y are independent if and only if $K\left(\frac{x-X}{h_1}\right)$ and $K\left(\frac{y-Y}{h_2}\right)$ are independent. This is equivalent to

$$E\left[K\left(\frac{x - X}{h_1}\right) K\left(\frac{y - Y}{h_2}\right)\right] = E\left[K\left(\frac{x - X}{h_1}\right)\right] E\left[K\left(\frac{y - Y}{h_2}\right)\right].$$

Approximately, this is

$$\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right) = \left[\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_2}\right)\right]$$

$$\Leftrightarrow \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right) = \left[\frac{1}{nh_1}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right)\right]\left[\frac{1}{nh_2}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_2}\right)\right]$$

$$\Leftrightarrow \hat{f}_n(x, y) = \hat{f}_n(x)\hat{f}_n(y).$$

Otherwise, if X and Y are correlated, $\hat{f}_n(x, y) \neq \hat{f}_n(x)\hat{f}_n(y)$.
Since we know that $f(x, y) = f(x)f(y)$ iff X and Y are independent, from the above, we can see that the estimator is able to capture the correlation between the two variables.