Statistics for Data Science B - Exercise Set 6

1. Consider the kernel estimate of the second derivative $f''(x)$ of the density function:

$$\hat{f}''(x) = \frac{1}{nh^3}\sum_{i=1}^{n} K''\left(\frac{X_i - x}{h}\right).$$

Assume for simplicity that $K(u) = 0$ for $|u| > 1$ (this condition can be relaxed). Derive expressions for the bias, variance, MSE, and IMSE of the estimate. Derive a formula for the optimal bandwidth $h_{opt,2}$ for estimating $f''(x)$. The formula will involve $\int (f^{(4)}(x))^2 dx$. The value of this integral for the case where f is the $N(\mu, \sigma^2)$ density is $1.8512\sigma^{-9}$. Based on this, present a formula for the optimal bandwidth $\tilde{h}_{opt,2}$ for estimating $f''(x)$ in the case where f is (close to) a normal density and for the optimal IMSE that results.

Hint: Use the integration by parts formula

$$\int a\, db = ab - \int b\, da$$

Solution

We have

$$E[\hat{f}''(x)] = \frac{1}{h^3}\int_{-\infty}^{\infty} K''\left(\frac{y - x}{h}\right) f(y) dy.$$

We apply to the integral on the right side the integration by parts formula

$$\int a\, db = ab - \int b\, da$$

with $a = f(y), db = h^{-1}K''((y-x)/h)dy, b = K'((y-x)/h), da = f'(y)dy$. We get

$$E[\hat{f}''(x)] = \frac{1}{h^2}\left[K'\left(\frac{y - x}{h}\right) f(y)\right]\Big|_{-\infty}^{\infty} - \frac{1}{h^2}\int_{-\infty}^{\infty} K'\left(\frac{y - x}{h}\right) f'(y) dy.$$

The first term on the right side is 0 because of the assumption that $K(u) = 0$ for $|u| > 1$ (note that since K' is differentiable, it is continuous, so that $K'(-1) = K'(1) = 0$).

We now apply integration by parts another time to the second term. We get

$$\begin{aligned}
E[\hat{f}''(x)] &= -\frac{1}{h}\left[K\left(\frac{y-x}{h}\right)f'(y)\right]\Big|_{-\infty}^{\infty} + \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{y-x}{h}\right)f''(y)dy \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{y-x}{h}\right)f''(y)dy \\
&= \int_{-\infty}^{\infty} K(v)f''(x+hv)dv \\
&\doteq \int_{-\infty}^{\infty} K(v)\left[f''(x) + f'''(x)hv + \tfrac{1}{2}f^{(4)}(x)h^2v^2\right]dv \\
&= f''(x) + \tfrac{1}{2}f^{(4)}(x)\tau^2(K)h^2.
\end{aligned}$$

More precisely,

$$\text{Bias}(\hat{f}''(x)) = E[\hat{f}''(x)] - f''(x) = \tfrac{1}{2}f^{(4)}(x)\tau^2(K)h^2(1+o(1))$$

provided there exist $L > 0$ and $\alpha \in (0,1]$ such that $|f^{(4)}(x) - f^{(4)}((x'))| \leq L|x - x'|$.

Next, $\text{Var}(\hat{f}''(x)) = (nh^4)^{-1}\text{Var}(Z_i)$ with

$$Z_i = \frac{1}{h}K''\left(\frac{X_i - x}{h}\right).$$

We have

$$\begin{aligned}
E[Z_i^2] &= \frac{1}{h}\left[\frac{1}{h}\int_{-\infty}^{\infty} K''\left(\frac{y-x}{h}\right)^2 f(y)dy\right] \\
&= \frac{\|K''\|_2^2}{h}\left[\frac{1}{h}\int_{-\infty}^{\infty} Q\left(\frac{y-x}{h}\right)f(y)dy\right] \\
&= \left(\frac{\|K''\|_2^2}{h}\right)f(x)(1+o(1))
\end{aligned}$$

with $Q(u) = (K''(u))^2/\|K''\|_2^2$, and the term in the brackets in the second line is equal to $f(x)(1+o(1))$ by the argument used to compute $E[\hat{f}(x)]$. And we have $(E[Z_i])^2 = h^4 f''(x)^2(1+o(1))$, which is dominated by $E[Z_i^2]$. So we get

$$\text{Var}(\hat{f}''(x)) = \frac{\|K''\|_2^2 f(x)}{nh^5}(1+o(1)).$$

Putting the squared bias and the variance together, we get

$$\text{MSE}(\hat{f}''(x)) \doteq \tfrac{1}{4}f^{(4)}(x)^2\tau^4(K)h^4 + \frac{\|K''\|_2^2 f(x)}{nh^5}.$$

Integrating (and using the fact that $f(x)$ integrates to 1), we get

$$\text{IMSE}(\hat{f}'') \doteq \tfrac{1}{4}\tau^4(K)\left[\int_{-\infty}^{\infty} f^{(4)}(x)^2 dx\right]h^4 + \frac{\|K''\|_2^2}{nh^5} = Ah^4 + B(nh^5)^{-1},$$

where

$$A = \tfrac{1}{4}\tau^4(K)\left[\int_{-\infty}^{\infty} f^{(4)}(x)^2 dx\right]$$

and $B = \|K''\|_2^2$. Minimizing the IMSE involves minimizing the function $g(h) = Ah^4 + B(nh^5)^{-1}$. We differentiate $g(h)$ and set the derivative to zero. We have $g'(h) = 4Ah^3 - 5B(nh^6)^{-1}$. Setting this to zero leads to

$$h_{opt,2} = \left(\frac{5B}{4A}\right)^{1/9} n^{-1/9} = \left(\frac{5\|K''\|_2^2}{\tau^4(K)E_4(f)}\right)^{1/9} n^{-1/9} \tag{1}$$

with $E_4(f) = \int f^{(4)}(x)^2 dx$.

The corresponding optimal IMSE is

$$\begin{aligned}\mathrm{MSE}_{opt}(\hat{f}''(x)) = g(h_{opt,2}) &= \left(\frac{5\|K''\|_2^2}{\tau^4(K)E_4(f)}\right)^{4/9}\left(A + B\left(\frac{5\|K''\|_2^2}{\tau^4(K)E_4(f)}\right)^{1/9}\right) n^{-4/9} \\ &= \left(\frac{5\|K''\|_2^2}{\tau^4(K)E_4(f)}\right)^{4/9}\left(\tfrac{1}{4}\tau^4(K)E_4(f) + \|K''\|_2^2\left(\frac{5\|K''\|_2^2}{\tau^4(K)E_4(f)}\right)^{1/9}\right) n^{-4/9}\end{aligned}$$

It is given in the question that for the $N(\mu, \sigma^2)$ density, we have $E_4(f) = 1.8512\sigma^{-1/9}$. Plugging this into the formula (1) for $h_{opt,2}$, we get

$$\tilde{h}_{opt,2} = \left(\frac{5\|K''\|_2^2}{\tau^4(K)E_4(f)}\right)^{1/9} n^{-1/9} = 1.1167\left(\frac{\|K''\|_2^2}{\tau^4(K)}\right)^{1/9} \sigma\, n^{-1/9}.$$

2. Consider the following procedure:

a. Estimate $f''(x)$ using the Gaussian kernel and the bandwidth $\tilde{h}_{opt,2}$ derived in Question 1. Denote the estimate by $\tilde{f}''(x)$. You may use the fact that for the $N(0,1)$ density φ we have

$$\int_{-\infty}^{\infty} (\varphi''(u))^2 du = \frac{3}{8\sqrt{\pi}}$$

b. Numerically compute the integral

$$\int_{-\infty}^{\infty} (\tilde{f}''(x))^2 dx.$$

This can be done in a number of ways. One simple method runs as follows. Make the change of variable

$$y = \exp((x - \bar{X})/s)/(1 + \exp((x - \bar{X})/s)),$$

$$x = \bar{X} + s(\log y - \log(1 - y)),$$

$$dx = s\, dy/[y(1 - y)],$$

where $\bar{X}$ and s are, respectively, the sample mean and SD of the data. Write the integral as

$$\int_0^1 (\tilde{f}''(\bar{X} + s(\log y - \log(1-y)))^2 [y(1-y)]^{-1} s \, dy.$$

Define

$$y_j = (2j-1)/(2J)$$

where J is a (large) integer. Then approximate the integral as

$$\frac{1}{2J} \sum_{j=1}^{J} g(y_j),$$

where $g(y) = s(\tilde{f}''(\bar{X} + s(\log y - \log(1-y))^2 [y(1-y)]^{-1}$. Keep increasing J until the result does not change by more than 1%.

c. Plug the resulting value of $\int (\tilde{f}''(x))^2 dx$ into the formula for the optimal bandwidth h_{opt} for estimating the density $f(x)$ using the Gaussian kernel. Then compute the density estimate with this bandwidth.

Write code to implement the above procedure. Use the code to compute a density estimate for the data in the file `ex6dat.txt`. In parallel, use the R function `density` with the Sheather-Jones bandwidth to compute a density estimate. Compare the density estimate you obtained by implementing the above procedure with the density estimate produced by `density`.

Solution

To compute $\tilde{h}_{opt,2}$ we need $\tau^2(K)$ and $\|K''\|_2^2$ for the normal kernel. Now, $\tau^2(K)$ is just the variance associated with the kernel, which for the normal kernel is equal to 1. It is stated in the question that for the normal kernel, $\|K''\|_2^2 = 3/(8\sqrt{\pi})$. Substituting these results into the formula for $\tilde{h}_{opt,2}$, we get $\tilde{h}_{opt,2} = 0.9397\sigma n^{-1/9}$. Following a recommendation of Scott, we'll replace σ by the minimum of the SD of the data and the IQR of the data divided by the IQR of the $N(0,1)$ distribution, which is 1.349.

I separated the R code into two parts. Part 1 computes the estimate of $f''(x)$ and, with this, carries out the computation of the integral $\int (\hat{f}''(x))^2 dx$ for various values of J. Part 2 computes the estimate of $f(x)$. In computing the integral, I took J values ranging from 10 to 300 in increments of 10. The results were as follows:

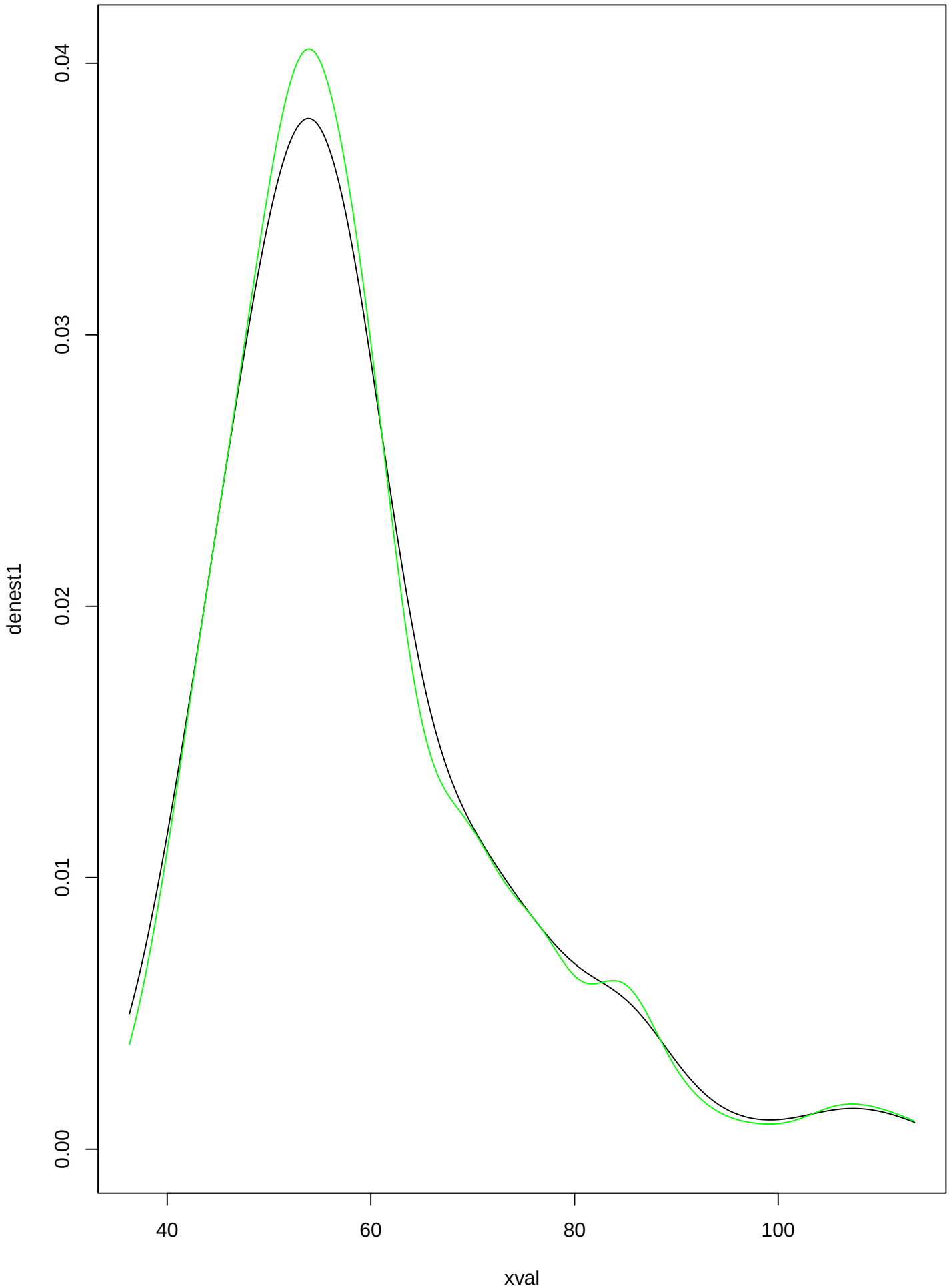
```
source("C:/Users/owner/Dropbox/WORK/Zucker/Stat for DS/Targilim/5786/targ6q2a.r")
[1]  1.009331e-06 9.844186e-07 9.744038e-07 9.769656e-07 9.780772e-07 9.767793e-07 9.756426e-07
```

```
[8]  9.754566e-07 9.757983e-07 9.762111e-07 9.764687e-07 9.765376e-07 9.764847e-07 9.763929e-07
[15] 9.763164e-07 9.762741e-07 9.762624e-07 9.762688e-07 9.762819e-07 9.762944e-07 9.763030e-07
[22] 9.763076e-07 9.763090e-07 9.763086e-07 9.763075e-07 9.763062e-07 9.763051e-07 9.763044e-07
[29] 9.763040e-07 9.763038e-07
```

So, we see that the integral is equal, to 3 significant digits, to 9.76×10^{-7}.

The R codes are in the attached files `targ8q2a.r` and `targ8q2b.r`. The next page shows a plot of the density estimate I got along with the density estimate produced by the R function `density` with the Sheather-Jones bandwidth selector. The estimates are very close, with the main difference being that the estimate produced by `density` was slightly higher at the peak. The bandwidth I obtained with my code was $h = 4.34$, while the bandwidth that `density` obtained was $h = 3.27$, so that my h is about 33% higher than that produced by `density`, which I regard as being not too far off (recall that the Sheather-Jones approach uses a more sophisticated method of choosing the bandwidth in the preliminary estimation of $f''(x)$).

Remark: Minimizing the IMSE treats the bias and the standard deviation on an equal footing. In many applications, it will be desirable to choose h so as to make the bias converge to 0 faster than the standard deviation. This is needed, for example, in computing a confidence interval for $f(x)$ at a given x. One simple way of achieving this is to take the $h = n^{-\delta} h_{SJ}$, where h_{SJ} is the bandwidth produced by the Sheather-Jones method and δ is a small positive number. I don't know of any general recommendation for the value of δ.


denest1
0.00
0.01
0.02
0.03
0.04
40
60
80
100
xval

3. Let $X_1, \ldots, X_n$ be iid with unspecified distribution. Consider the nonparametric kernel density estimate we presented in class:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

In all parts of this question you should use the R function `density` to compute density estimates with bandwidth given by either $h = n^{-0.025} h_{SJ}$ or $h = n^{-0.05} h_{SJ}$, where h_{SJ} is the bandwidth produced by the Sheather-Jones method (option `bw='SJ'`). Use the Epanechnikov kernel.

a. In class we presented the following formula for an asymptotic $100(1-\alpha)\%$ pointwise confidence interval for $f(x)$, based on the asymptotic distribution result that appears in the preceding question.

$$\hat{f}(x) \pm z_{1-\alpha/2} \left(\frac{\hat{f}(x)\|K\|_2^2}{nh}\right)^{1/2}$$

Generate a sample of size 300 from the $Gamma(3,2)$ distribution and compute the nonparametric density estimate and 95% confidence limits for x in the range 0 to 15 in steps of 0.02. Take $h = n^{-0.05} h_{SJ}$. Plot the estimated density and the confidence limits.

b. We can consider also a simultaneous confidence band, with the property that

$$P(f(x) \in [f_L(x), f_U(x)] \textbf{ for all } x \in [0, x^*]) \doteq 1 - \alpha$$

One way to do this is to find d that satisfies

$$P(\sup_{x \in [0,x^*]} |\hat{f}(x) - f(x)| \leq d) \doteq 1 - \alpha$$

and then take $f_L(x) = \hat{f}(x) - d$ and $f_L(x) = \hat{f}(x) + d$. The value d can be found using the nonparametric bootstrap. Carry out this procedure on the data you generated in Part (a), with $\alpha = 0.05$ and $x^* = 15$. Again take $h = n^{-0.05} h_{SJ}$. Use 1,000 bootstrap replications. Plot the estimated density and the simultaneous confidence limits. Compare these confidence limits to those you generated in Part (a) and comment.

c. Write code to use simulation to evaluate the performance of the density estimate and the confidence band by computing the mean estimated value over 1,000 replications and the percentage of these replications in which the confidence band covers the true density value over the entire range from $[0, x^*]$. Run your code on the data you generated in Part (a), with $\alpha = 0.05$ and $x^* = 15$. Again take $h = n^{-0.05} h_{SJ}$. Use 1,000 bootstrap replications. Do this for $h = n^{-0.025} h_{SJ}$ and $h = n^{-0.05} h_{SJ}$, once on the entire set of 300 observations and once on the first 150 observations only. Comment on the results.

Solution

Part (a)

The formula for the confidence interval is

$$\hat{f}(x) \pm z_{1-\alpha/2} \left(\frac{\hat{f}(x) \|K\|_2^2}{nh} \right)^{1/2}$$

For the Epanechnikov kernel, we have

$$\|K\|_2^2 = \int_{-1}^{1} \left(\frac{3}{4}\right)^2 (1-u^2)^2 \, du = \left(\frac{9}{16}\right) \int_{-1}^{1} (1 - 2u^2 + u^4) \, du = \left(\frac{9}{16}\right) \left[u - \frac{2u^3}{3} + \frac{u^5}{5} \right|_{-1}^{1} = \frac{3}{5}$$

R code to carry out the calculations is given in the file `targ6q3.r`.

Remark: In this question, we used a distribution where the range of the data is $(0, \infty)$ rather than $(-\infty, \infty)$. With kernel density estimation, care is needed in computing the estimates near the boundary. We didn't discuss this issue in class. There are various approaches to dealing with it.

Part (b)

R code to carry out the calculations is given in the file `targ6q3.r`. The simultaneous confidence band is somewhat wider than the pointwise confidence intervals. This is the price we pay for simultaneous coverage.

Part (c)

R code to carry out the calculations is given in the file `targ6q3.r`. The results I obtained are as follows:

```
 n   adjdel nrep nboot covpr mean_cihw
 150  0.025 1000  1000 0.997    0.0529
 150  0.050 1000  1000 0.999    0.0582
 300  0.025 1000  1000 0.991    0.0372
 300  0.050 1000  1000 0.998    0.0410
```

The coverage rates are considerably higher than nominal. Perhaps if we increased the number of bootstrap replications, we would obtain more accurate results. The mean confidence interval half-width is slightly smaller with $h = n^{-0.025} h_{SJ}$ than with $h = n^{-0.05} h_{SJ}$