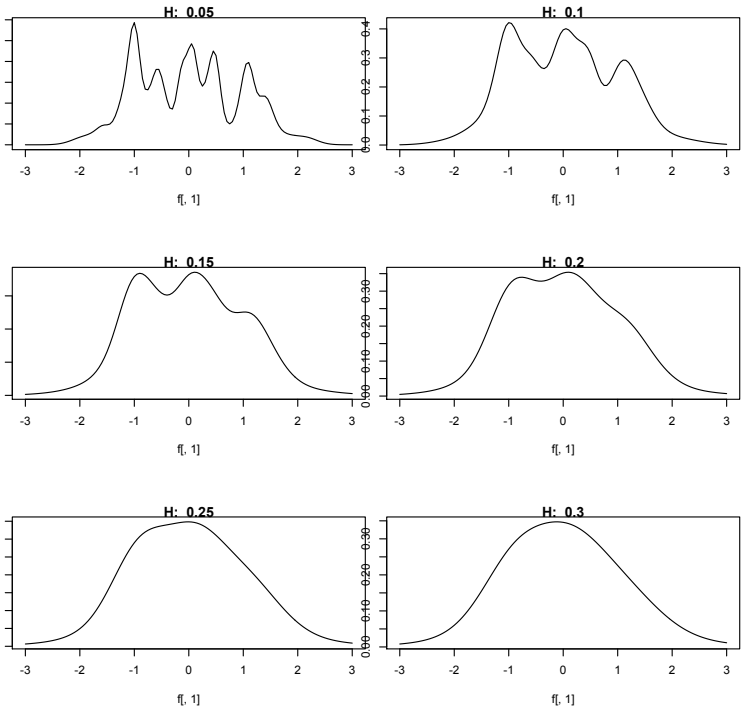
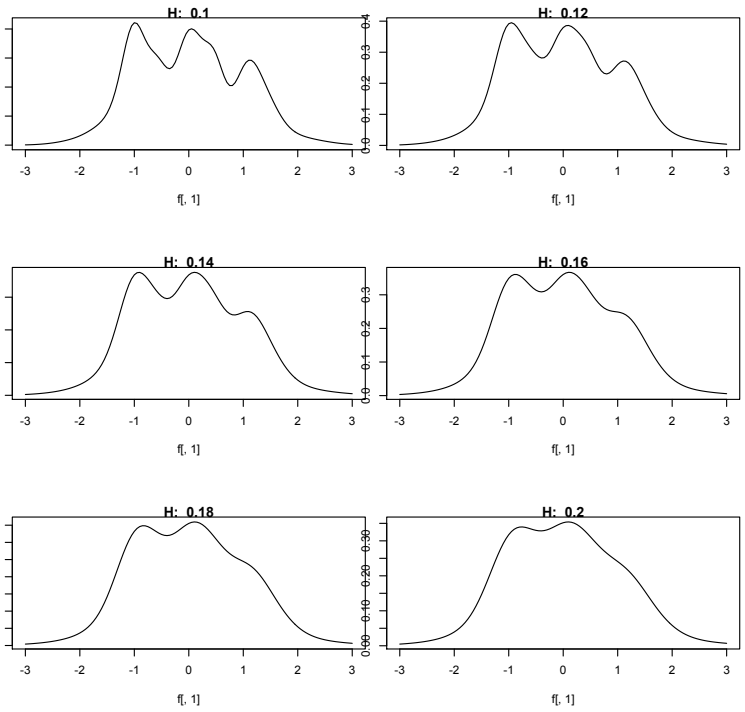
1. (a) We begin implementation procedures by assuming that K is a Gaussian kernel and our initial estimate of the density, $\widetilde{f}$, is the density estimate from the normal kernel with the rule of thumb bandwidth ($\hat{h} = 1.06\hat{\sigma}n^{-1/n}$). Since our estimate of $\widetilde{f}$ will change, the functionality to change these assumptions is included. The function *f_var_bw* has four arguments - X the data, k the name of the kernel to be used for $\widetilde{f}$, bw the bandwidth to be used for $\widetilde{f}$, and h the bandwidth for the estimation of f. For details, see the code that is attached.

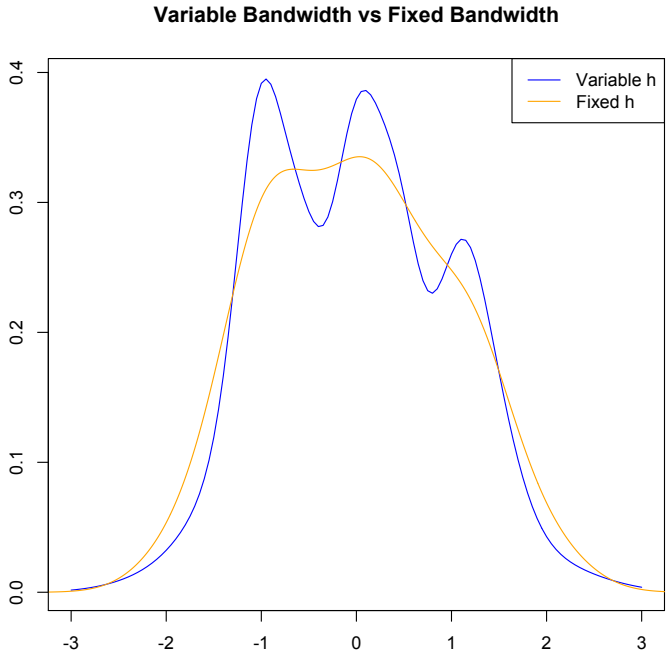
   (b) After successfully implementing the procedure, we looked to find a reasonable value for h. We began by searching over the candidate grid $\mathcal{H} = \{0.05, 0.1, 0.15, 0.25, 0.3\}$. The density estimates associated with these bandwidths are given in the figure below.



It appears that for $h > 0.2$ we over smooth and suppress underling modes in the data. For $h = 0.05$ it looks like the density is under smoothed with many sharp turns. For $0.05 < h \leq 0.2$ is is hard to determine which bandwidth offers optimal smoothing. For this reason, we reduce our search to $\mathcal{H} = \{0.1, 0.12, 0.14, 0.16, 0.18, 0.2\}$ The plots of these kernel estimate is given below. Here, $h > .14$ seems to oversmooth the data, leaving $h = 0.1, 0.12, 0.14$. It appears again that $h = 0.12$ offers the best fit. This bandwidth successfully smooths out unwanted noise while identifying modes inherent in the data. To see the variable kernel estimate with $h = 0.12$ against the fixed bandwidth model, consider the figure below.

Here we see that the fixed bandwidth kernel fails to find any of the modes in the distribution while the variable bandwidth manages to recover three modes.

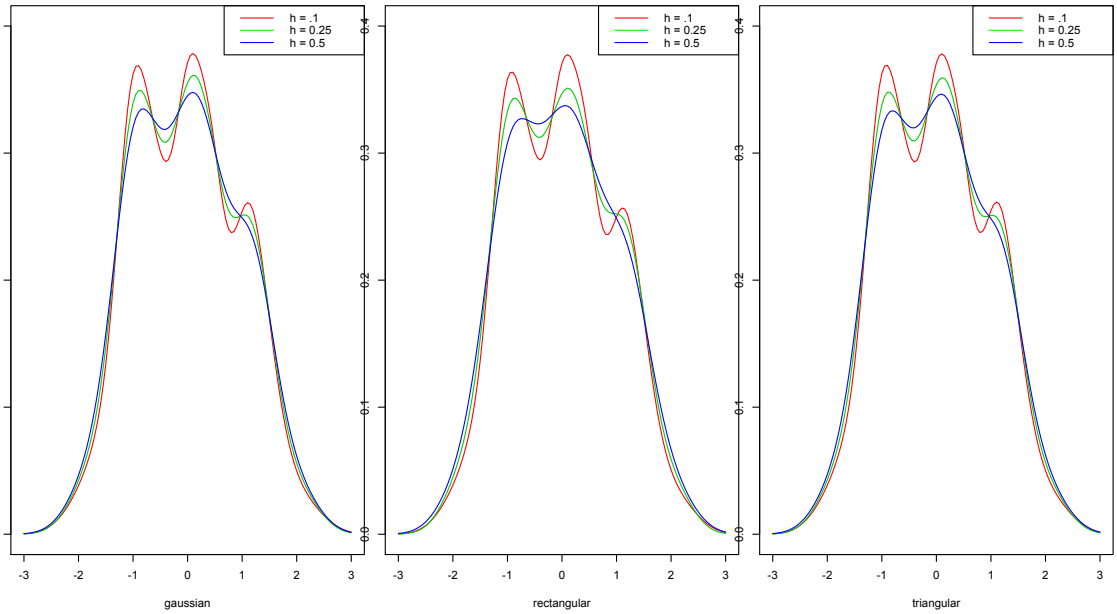




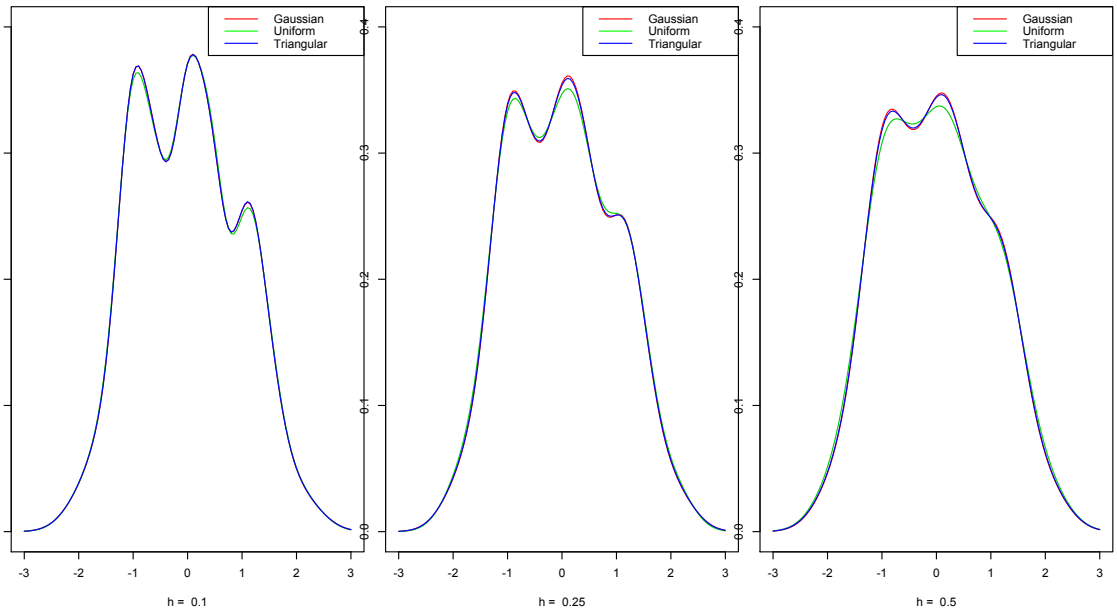
This is a drastic improvement on the global bandwidth model.

Having attained an optimal bandwidth h, we now turn our attention to what affect the bandwidth and kernel choices of $\widetilde{f}$ have on our estimate of $\widehat{f}$. We con-

sider the uniform, triangular, and Gaussian kernel as well as bandwidths of sizes of $g = 0.1, 0.25, 0.5$. To see the affect of bandwidth of $\widetilde{f}$ on $\widehat{f}$ consider the figure below.



Here we see that the bandwidth of $\widetilde{f}$ greatly impacts the performance of our procedure for any kernel. For higher bandwidths of g, we begin to remove modes in the density while lower values of g we see a more modes. For the affect of the kernel on the estimation consider the figure below.



For any bandwidth, it is almost impossible to distinguish between the density estimates. Perhaps the uniform kernel smooths more so than the other two methods, but the difference is negligible. For this reason, we choose to use a Gaussian kernel for $\widetilde{f}$ and focus on choosing an appropriate value of g.

(c) Suppose that $c_i \equiv \widetilde{f}(x_i)^{1/2}$ is constant. Then we look to use $MISE(h)$ as our decision function. We begin by finding the bias of the estimate.

$$\begin{aligned}
\mathbb{E}(\widehat{f}(x)) &= \mathbb{E}\Big[\frac{1}{nh}\sum_{i=1}^{n} c_i K\Big(\frac{(x-x_i)c_i}{h}\Big)\Big] \\
&= \frac{1}{nh}\sum_{i=1}^{n} c_i \mathbb{E}\Big[K\Big(\frac{(x-x_i)c_i}{h}\Big)\Big] \\
&= \frac{1}{nh}\sum_{i=1}^{n} c_i \int K\Big(\frac{(x-w)c_i}{h}\Big) f(w)dw
\end{aligned}$$

Now completing a change of variable for $z = \frac{(x-w)c_i}{h}$, followed by a Taylor series expansion we have

$$\begin{aligned}
&= \frac{1}{nh}\sum_{i=1}^{n} c_i \int K(z) f(x - zh/c_i)\frac{h}{c_i}dz \\
&= \frac{1}{n}\sum_{i=1}^{n} \int K(z)\Big\{f(x) - f'(x)\frac{zh}{c_i} + f''(x)\frac{z^2h^2}{2c_i^2} + O(h^3)\Big\}dz \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big\{f(x)\int K(z)dz - \frac{f'(x)h}{c_i}\int zK(z)dz + f''(x)\frac{h^2}{2c_i^2}\int z^2K(z)dz + o(h^2)\Big\} \\
&= f(x) + f''(x)h^2\mu_2(K)\frac{1}{2n}\sum_{i=1}^{n} c_i^{-2} + o(h^2)
\end{aligned}$$

Using this see the bias is given by

$$Bias(\widehat{f}) = f''(x)\frac{h^2}{2}\mu_2(K)\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}$$

Now, we look to find the form of the variance of our estimator.

$$\begin{aligned}
Var\Big(\widehat{f}(x)\Big) &= \frac{1}{n^2h^2}\sum_{i=1}^{n} c_i^2 Var\Big[K\Big(\frac{(x-x_i)c_i}{h}\Big)\Big] \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} c_i^2\Big[\mathbb{E}\Big(K\Big[\frac{(x-x_i)c_i}{h}\Big]^2\Big) - \mathbb{E}\Big(K\Big[\frac{(x-x_i)c_i}{h}\Big]\Big)^2\Big] \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} c_i^2\mathbb{E}\Big(K\Big[\frac{(x-x_i)c_i}{h}\Big]^2\Big) - \sum_{i=1}^{n}\Big[\frac{c_i}{nh}\mathbb{E}\Big(K\Big[\frac{(x-x_i)c_i}{h}\Big]\Big)\Big]^2
\end{aligned}$$

Focusing on the first term above we have

$$\begin{aligned}
\frac{1}{n^2h^2}\sum_{i=1}^{n} c_i^2\mathbb{E}\Big(K\Big[\frac{(x-x_i)c_i}{h}\Big]^2\Big) &= \frac{1}{n^2h}\sum_{i=1}^{n} c_i^2\int K\Big(\frac{(x-w)c_i}{h}\Big)^2 f(w)dw \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} c_i^2\int K(z)^2\Big\{f(x)+O(1)\Big\}dz \\
&= \frac{1}{n^2h}\sum_{i=1}^{n} c_i\Big\{f(x)||K||_2^2+O(1)\Big\} \\
&= \frac{f(x)||K||_2^2}{nh}\frac{1}{n}\sum_{i=1}^{n} c_i + O(1/n^2h)
\end{aligned}$$

Now focusing in the second term, we see that

$$\begin{aligned}
\sum_{i=1}^{n}\Big[\frac{c_i}{nh}\mathbb{E}\Big(K\Big(\frac{(x-x_i)c_i}{h}\Big)\Big)\Big]^2 &= \sum_{i=1}^{n}\Big[\frac{c_i}{nh}\int K\Big(\frac{(x-w)c_i}{h}\Big)f(w)dw\Big]^2 \\
&= \sum_{i=1}^{n}\Big[\frac{1}{n}\int K(z)f(x-hz/c_i)dz\Big]^2 \\
&= \sum_{i=1}^{n}\Big[\frac{1}{n}\int K(z)\Big\{f(x)+O(1)\Big\}dz\Big]^2 \\
&= \sum_{i=1}^{n}\Big[\frac{1}{n}f(x)+O(1/n)\Big]^2 \\
&= \frac{1}{n}f^2(x)+O(1/n^2)
\end{aligned}$$

As we argued in class, we see that this term goes to zero rapidly. Therefore, our estimate of the variance of $\hat{f}$ is given by

$$Var(\widehat{f}(x)) = \frac{f(x)||K||_2^2}{nh}\frac{1}{n}\sum_{i=1}^{n} c_i + O(1/n^2h)$$

(Notice that if $c_i = 1$, for $i = 1, 2, \ldots n$, we simply get the mean and variance of the unweighted kernel). Using this we can define the MSE as

$$MSE(\widehat{f}) = \frac{f(x)||K||_2^2}{nh}\frac{1}{n}\sum_{i=1}^{n} c_i + f''(x)^2\frac{h^4}{4}\mu_2^2(K)\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2$$

Lastly we can find the MISE as follows

$$\begin{aligned} MISE(\widehat{f}) &= \int MSE(\widehat{f})dx \\ &= \int \frac{f(x)||K||_2^2}{nh}\frac{1}{n}\sum_{i=1}^{n} c_i dx + \int f''(x)^2 \frac{h^4}{4}\mu_2^2(K)\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2 dx \\ &= \frac{||K||_2^2}{nh}\frac{1}{n}\sum_{i=1}^{n} c_i \int f(x)dx + \frac{h^4}{4}\mu_2^2(K)\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2 \int f''(x)^2 dx \\ &= \frac{||K||_2^2}{nh}\frac{1}{n}\sum_{i=1}^{n} c_i + \frac{h^4}{4}\mu_2^2(K)||f''||_2^2\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2 \end{aligned}$$

We now look to minimize MISE as a function of h.

$$\frac{\partial}{\partial h}MISE(\widehat{f}) = -\frac{||K||_2^2}{nh^2}\frac{1}{n}\sum_{i=1}^{n} c_i + h^3\mu_2^2(K)||f''||_2^2\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2 \stackrel{set}{=} 0$$

This implies

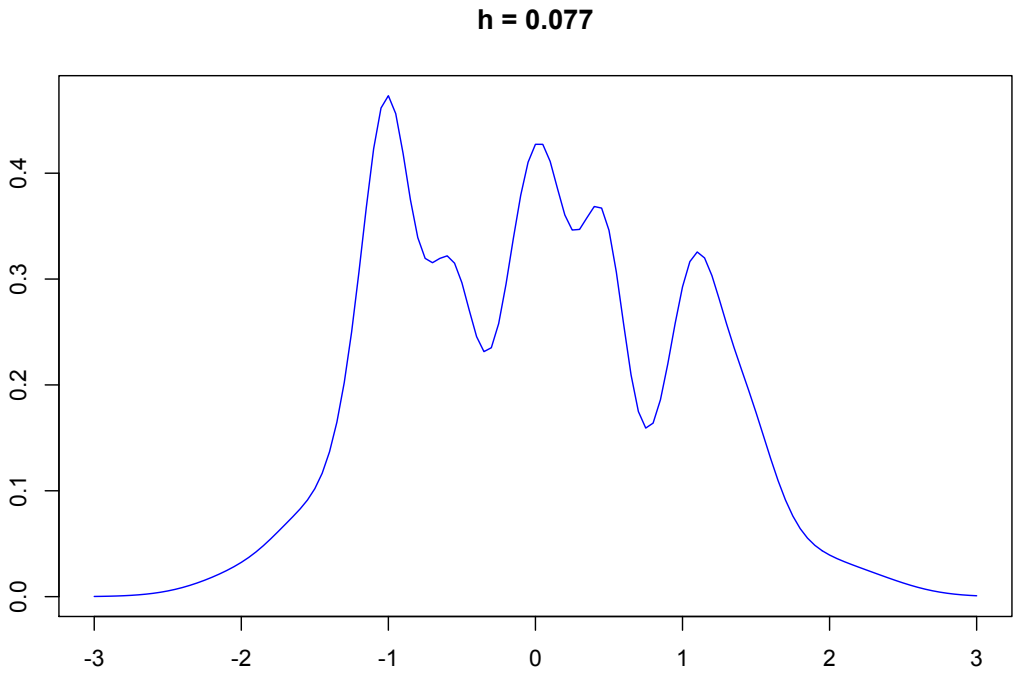
$$h_{MISE} = \left[\frac{||K||_2^2 \frac{1}{n}\sum_{i=1}^{n} c_i}{\mu_2^2(K)||f''||_2^2\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2}\right]^{1/5} n^{-1/5}$$

As in the MISE for the fixed bandwidth, we require that we have some estimate of f''. One method was simply to refer to the normal density and calculate $||\phi||_2^2$. We can improve this estimate however. We needed an initial kernel estimate $\widetilde{f}$ is aid in our variable bandwidth procedure. In theory, this kernel should approximate the global features of f and will serve as a more informative estimate of $||f''||_2^2$. Of course, we will need to estimate this value. But as we discussed in class, there are several powerful methods for this estimation procedure. During implementation, we use a helper package, *kedd*, to estimate, $\widetilde{f}''$ from $\widetilde{f}$ and then carry out our own approximation of $||\widetilde{f}''||_2^2$. (Again for implementation details, see the code attached below). Using this as our plug-in estimator, our estimate of h can be written as

$$\hat{h}_{MISE} = \left[\frac{||K||_2^2 \frac{1}{n}\sum_{i=1}^{n} c_i}{\mu_2^2(K)||\widetilde{f}''||_2^2\left(\frac{1}{n}\sum_{i=1}^{n} c_i^{-2}\right)^2}\right]^{1/5} n^{-1/5}$$

For our sample data, we calculate that $\hat{h}_{MISE} = 0.077$ which is smaller than our original estimate of h but was still considered a reasonable value after our initial grid search. The kernel density estimate given from this procedure can be seen below.

This estimate appears to under smooth the data. But recall that the data we were given has four modes, not three that our initial estimate identified. Here, this kernel estimate appears to have found three, possible four modes. In any case, $\hat{h}_{MISE}$ and the variable bandwidth procedure clearly out preform the global bandwidth procedure.



2. (a) Let $\widehat{f'}$ be given by

$$\widehat{f'}(x) = \frac{1}{h^2}\sum_{i=1}^{n} K'\Big(\frac{x - X_i}{h}\Big)$$

where f is defined on $[0, \infty)$ and K on $[-1, 1]$. Suppose further that x is a boundary point. That is for some $0 \leq p < 1$, $x = ph$. We can then calculate the expected value of this estimate near the boundary as follows.

$$\begin{aligned}
\mathbb{E}(\widehat{f'}) &= \mathbb{E}\left[\frac{1}{nh^2}\sum_{i=1}^{n} K'\Big(\frac{x - X_i}{h}\Big)\right] \\
&= \frac{1}{h^2}\mathbb{E}\left[K'\Big(\frac{x - X}{h}\Big)\right] \\
&= \frac{1}{h^2}\int_0^{\infty} K'\Big(\frac{x - w}{h}\Big) f(w)dw \\
&= -\frac{1}{h^2}\int_p^{-\infty} K'(z) f(x - zh) h dz
\end{aligned}$$

The the last equality was due to the substitution $z = \frac{x-w}{h}$ with $-hdz = dw$ and $w = x - zh$. Now recall that K was only defined on $[-1, 1]$ so K' is only defined on $[-1, 1]$. This gives

$$
\begin{aligned}
&= \frac{1}{h}\int_{-\infty}^{p} K'(z)f(x-zh)dz \\
&= \frac{1}{h}\int_{-1}^{p} K'(z)f(x-zh)dz \\
&\overset{Taylor}{=} \frac{1}{h}\int_{-\infty}^{p} K'(z)\Big\{f(x) - zhf'(x) + \frac{(zh)^2}{2}f''(x) + O(h^3)\Big\}dz \\
&= \frac{f(x)}{h}\int_{-1}^{p} K'(z)dz - f'(x)\int_{-1}^{p} zK'(z)dz + \frac{hf''(x)}{2}\int_{-1}^{p} z^2K''(z)dz + O(h^2) \\
&= \frac{1}{h}a_0^1(p)f(x) - a_1^1(p)f'(x) + \frac{h}{2}a_2^1(p)f''(x) + O(h^2)
\end{aligned}
$$

(b) Let $a_j^1(p) = \int u^j K(u)du$, $c_j^1(p) = \int u^j L(u)du$ for two kernels $K \neq L$. Then we have

$$
\begin{aligned}
\mathbb{E}(c_2^1(p)\hat{f}_K) &= \frac{1}{h}c_2^1(p)a_0^1(p)f(x) - c_2^1(p)a_1^1(p)f'(x) + \frac{h}{2}c_2^1(p)a_2^1(p)f''(x) + O(h^2) \\
\mathbb{E}(a_2^1(p)\hat{f}_L) &= \frac{1}{h}a_2^1(p)c_0^1(p)f(x) - a_2^1(p)c_1^1(p)f'(x) + \frac{h}{2}a_2^1(p)c_2^1(p)f''(x) + O(h^2)
\end{aligned}
$$

Define $B(u) = c_2^1(p)K(u) - a_2^1(p)L(u)$. Then we have that

$$
\mathbb{E}(\widehat{f}_B(x)) = \frac{1}{h}\{a_2^1(p)c_0^1(p) - c_2^1(p)a_0^1(p)\}f(x) - \{c_2^1(p)a_1^1(p) - a_2^1(p)c_1^1(p)\}f'(x) + O(h^2)
$$

Now, by letting $b_a^c(p) = a_2^1(p)c_0^1(p) - c_2^1(p)a_0^1(p)$ and $b_c^a(p) = c_2^1(p)a_1^1(p) - a_2^1(p)c_1^1(p)$ we can write

$$
\mathbb{E}(\widehat{f}_B(x)) = \frac{1}{h}b_a^c(p)f(x) - b_c^a(p)f'(x) + O(h^2)
$$

Notice that we can repeat this exact process for two different starting kernels $K' \neq L'$ to attain that

$$
\mathbb{E}(\widehat{f}_D(x)) = \frac{1}{h}d_e^f(p)f(x) - d_f^e(p)f'(x) + O(h^2)
$$

Using the kernels B, D we can then remove the bias in the f' term by the following weights.

$$
\begin{aligned}
\mathbb{E}(d_f^e\widehat{f}_B(x)) &= \frac{1}{h}d_f^e b_a^c(p)f(x) - d_f^e b_c^a(p)f'(x) + O(h^2) \\
\mathbb{E}(b_c^a\widehat{f}_D(x)) &= \frac{1}{h}b_c^a d_e^f(p)f(x) - b_c^a d_e^f(p)f'(x) + O(h^2)
\end{aligned}
$$

So defining the kernel $F(u) = \frac{d_f^e(p)B(u) - b_c^a(p)D(u)}{\frac{1}{h}(d_f^e b_a^c - b_c^a d_e^f)}$ implies that

$$
\mathbb{E}(\widehat{f}_F(x)) = f(x) + O(h^2)
$$

Therefore, we see that by weighting two kernels, we can successfully remove the f'' bias term and then combining a *combination* of kernels, we can remove the bias of the f' term. Therefore, we can remove the bias in the boundary case of f' by properly weighting, then combining four separate kernels.

In the paper, *Bayesian Approach to the Choice of Smoothing Parameter in Kernel Density Estimation*, authors Gangopadhyay and Cheung investigate the bandwidth selection problem for variable bandwidth kernels. The paper first discusses work done in global bandwidth selection procedures such as Least Squares Cross Validation (LSCV) and improvements with Biased Cross Validation (BCV) techniques using the leave one kernel density estimate. They note, however, that choosing a bandwidth globally fundamentally contradicts the essence of nonparametric smoothing. If the data is collected at a mode of the underlying distribution, the amount of smoothing needed will be less than at points far from the modes. This motivates variable bandwidth procedures which was introduced by Fan et al. 1996. The authors note that this procedure does not preform well in moderate sample sizes. In their paper, Gangopadhyay and Cheung attempt to address this issue by introducing a prior distribution of the bandwidth h that will replace any missing information that the sample fails to capture.

Following the motivation above, the authors try to derive the posterior distribution of h, but seeing that h is not a model parameter, they first devise a way to introduce h as a parameter to the model in question. They achieve this by introducing *truncated version of* $f(x)$ defined as

$$f_h(x) = \int f(x)K_h(x-u)du$$

h therefore acts a scale parameter to f. That is as h decreases, $f_h(x) \to f(x)$. Having introduced h as a model parameter, it is now possible to discuss the Bayesian approach to the bandwidth selection procedure. Letting $h \sim \pi(h)$ be the prior distribution of h then the posterior is given by

$$\pi(h|x) = \frac{f_h(x)\pi(h)}{\int f_h(x)\pi(h)dh}$$

We see that $\pi(h|x)$ relies on $f_h(x)$. Therefore, a plug in estimate of f is needed. The authors

cleverly notice that $f_h(x) = \mathbb{E}(K_h(x - X))$ so it is reasonable to estimate this function by

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x_i - x}{h}\Big)$$

which is just the kernel estimator of f. So the approximated posterior distribution is given by

$$\hat{\pi}(h|X_1, \ldots, X_n, x) = \frac{\widehat{f}_h(x)\pi(h)}{\int \widehat{f}_h(x)\pi(h)dh}$$

This then gives rise to the posterior mean, and hence the estimate of h,

$$h^* = \int h\hat{\pi}(h|X_1, \ldots, X_n, x)dh$$

The authors note that this procedure is quite computationally intensive in that none of the integrals need have a closed form solution. Instead of addressing these computational issues, the authors focus on a case where a closed form solution *does* exist. In the case where K is a normal kernel, then h serves as a standard deviation of $K_h(X_i - x)$. Under this assumption, h has conjugate prior given by the Inverse Gamma distribution.

$$\tau(h) = \frac{2}{\Gamma(\alpha)\beta^{\alpha}}\frac{1}{h^{\alpha+1}}\exp\big\{-\frac{1}{\beta h^2}\big\} \qquad h > 0$$

Using this nice case, they authors successfully show that

$$h^*(x) = \frac{\Gamma(\alpha)}{\sqrt{2\beta}\Gamma(\alpha + 1/2)}\frac{\sum_{i=1}^{n}[1/(\beta(X_i - x)^2 + 2)]^{\alpha}}{\sum_{i=1}^{n}[1/(\beta(X_i - x)^2 + 2)]^{\alpha+1/2}}$$

which gives a close form solution to the estimated mean of the posterior distribution of h. They go on to point out that the posterior does in fact depend on parameters of the prior, so a sensitivity analysis of the posterior's behavior for different (α, β) combinations is necessary. Moreover, they note that by deriving $\hat{\pi}(h|X_1, \ldots, X_n, x)$ a more robust sense of appropriate values of h can be achieved.

To assess the effectiveness of their method, Gangopadhyay and Cheung devise a simulation experiment in which they measure the goodness of fit of global bandwidth kernels selected by *BCV* and *LSCV* as well as the smoothing method based on *K-NN* against their newly proposed local bandwidth selection method. They measure the effectiveness of each method by the average squared error

$$ASE = \frac{1}{m}\sum_{i=1}^{m}(\widehat{f}(v_i) - f(v_i))^2$$

For the synthetic data generated from $0.6N(4,1) + 0.4N(0,1)$ they show that their method out preforms all others considered for a large set of (α, β). The authors claim that this simulation shows their method is relatively stable with respect to prior parameter choices. In their first application, for a dataset with $n = 63$, they show their method successfully identifies all three modes in comparison to the noisy estimate of *K-NN* and the over smoothed *BCV* and *LSCV*. Lastly, they show that in the Claw data example, their method is the only one to come close to identifying the underlying structure of the distribution. It should be noted, however, that the prior model parameters are not near the parameters found in the authors' simulations. Therefore while it does fit the data quite well, the authors cannot claim model stability for such large values of (α, β). Moreover, in applications we do not see a comparison to other local bandwidth estimators such as those introduced by Fan et al. 1996. I would be interested to see a comparison of the performance of variable bandwidth estimators in moderate sample sizes.

This paper improves on the variable kernel density estimation in moderate sample sizes by introducing h as a scale parameter to the kernel estimate. By developing a Bayesian framework to estimate h, the authors show that by supplying additional information in the form of the Bayesian prior $\pi(\cdot)$ may stabilize the estimation of h. Through simulation study and applications, they show that their variable bandwidth estimate outperforms global bandwidth selection procedures and is stable with respect to the prior model parameters.