Visualization and Learning Structure: Problem Session 7/2/16

There are two kinds of problems to explore below: visualizing distributions of OkCupid profiles and looking at some of the theoretical properties of the kernel density estimates (motivated by all the good questions after today's class).
You should adapt today's class code posted on the website.

1. In class today, we explored the age and smoke status of the OkCupid users; now we'll look at their height and drug use status.

   (a) First, explore the effect of different histogram bin widths on the distribution of height. Note that the height variable has some questionable values in it. You can choose to include them or not. See `help(hist)` for how to change the number/location of the bins.

   Specifically, you may want to compare using the Sturges' default rule to using smaller or larger bins. Does your height data look normal?

   (b) Now let's look at height by drug use status (" " the blank group, "never", "sometimes", "often"). Try at least three different methods of comparing the conditional distributions of height given the drug status.

   - If you use kernel density estimates in some form, what kernel and bandwidth do you use?
   - Describe any differences you see in the height distributions across subgroups.

   (c) Compare the conditional distributions of drug use status as a function of height (conditional density plot; `help(cdplot)`). Pick a reasonable bandwidth.

   At what heights do we see big changes in the distribution of drug use status?

2. We can estimate the true density $f(x)$ using a kernel density estimate $\hat{f}(x)$:

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

   where n is the number of observations, h is the bandwidth, and K is the choice of kernel function. Kernels are often chosen for ease of computation and to reduce bias as much as possible (i.e. $E[\hat{f}(x)] \approx f(x)$).

   Three useful properties are:

$$\int K(t)dt = 1 \qquad \int tK(t)dt = 0 \qquad \int t^2 K(t)dt = \sigma_K^2$$

These properties are more standard/well-known for the Gaussian "bell" kernel of $K_G(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$, but you can use any appropriate function for which they hold.

Verify that these properties hold for the Uniform [0,1] kernel, $K_U(t) = \frac{1}{2}$, and the Epanechnikov "bubble" kernel, $K_E(t) = \frac{3}{4}(1 - t^2)$. Both compact on $t \in [-1, 1]$. (Note that $\sigma^2_{K_U} = \frac{1}{3}$ and $\sigma^2_{K_E} = \frac{1}{5}$.)

3. We can measure how well our density estimate $\hat{f}(x)$ approximates the truth $f(x)$ by the asymptotic mean integrated square error (ASIME) for $\hat{f}(x)$ as an approximation of $f(x)$ is:

$$AMISE = MISE(h)_{n\to\infty} = E_{n\to\infty} \int (\hat{f}_h - f)^2 = \frac{R(K)}{nh} + \frac{1}{4}\sigma^4_K h^4 R(f^{''})$$

The first term is the dominating term of the integrated variance, the second the integrated square bias. Also note that $R()$ refers to the "roughness" of the function; we can measure it by the integral of its square, i.e. $R(\psi(t)) = \int \psi(t)^2 dt$

(a) Show that the optimal bandwidth, h^*, that minimizes the $AMISE$ is

$$h^* = \left[\frac{R(K)}{\sigma^4_K R(f'')}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

(b) Given this h^*, find an expression for the optimal AMISE (i.e. $AMISE^*$) as a function of the kernel choice, the sample size, and the true unknown density.

Note that this calculation will also verify that the kernel's contribution to the AMISE is a function of $\sigma_K \cdot R(K)$, i.e. the variance and roughness of the kernel, and allows us to compute relative efficiencies across different kernels.

4. Unfortunately, we don't usually know the true $f(x)$ which means we would have to approximate $R(f^{''})$ to actually find the best bandwidth. Scott and Silverman derived some rules of thumb to help us approximate h:

$h_{Sc} = 1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}$ $h_{Si1} = 0.79 \cdot IQR \cdot n^{-\frac{1}{5}}$ $h_{Si2} = 0.9 \cdot \min(\hat{\sigma}, \frac{IQR}{1.34}) n^{-\frac{1}{5}}$

Calculate the three above bandwidths for our OkCupid heights.
Find and graph three kernel density estimates using these bandwidths (using a Gaussian kernel). How different are they? What effects did the different bandwidths have on your density estimates (if any)?

Useful R functions: `stdev(heights); summary(heights)` (for the IQR)