KERNEL DENSITY ESTIMATION AND THE DIP TEST

Kernel density estimation is a large topic in statistics: Several books and quite a large number of journal articles have been written about it, since the 1940s. Comparatively little has been published about testing for unimodality (the dip test) since the original article by the Hartigans in 1985. Yet, testing for unimodality can give an interesting viewpoint on density estimation.

Let K be a probability density on the real line $\mathbb{R}$, symmetric around 0 and having a finite mean, which must be 0. For a simple choice, for $h > 0$ let $K = K_h$ be the $U[-h/2, h/2]$ density, so that $K_h(x) = 1/h$ for $|x| \leq h/2$ and 0 for $|x| > h/2$. Given observations $X_1, ..., X_n$, assumed to be i.i.d. with an unknown density f, let $f_n(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - x_j)$. Then f_n is a probability density which may serve as an estimate of the unknown f. How to choose h ("bandwidth selection") is a problem in itself. One would take $h = h_n \to 0$ as $n \to \infty$; Venables and Ripley, p. 127, suggest h_n of order $n^{-1/5}$.

In the R library "MASS", assembled by Venables and Ripley in relation to their book, is a data set "galaxies" consisting of $n = 82$ observed redshifts of galaxies from a paper of K. Roeder (1990), coming in turn from a paper by several astrophysicists. The data set and density estimation from it are treated on pp. 129-135 of Venables and Ripley, who say there are "at least four peaks" in the data; diagrams on pp. 130 and 133 do show about four peaks, but do not themselves show statistical significance of these peaks; some seem rather weak.

For Roeder's paper, she had applied the dip test to the data and found unimodality was rejected, indicating at least two peaks. Unfortunately however, at the time (1990 or a year or so earlier), there was a bug in the R software for the test. Now that it has been corrected, one can find in detail that unimodality is not rejected: PS3, problem 4. So, only one peak is statistically significant. Looking at the apparent five peaks in Fig. 5.10 p. 130 of Venables and Ripley, the fourth peak is just the right shoulder of the main, third peak; the second peak reflects very few (two?) data points; the fifth peak is too small to be significant (one sees that only a few data points contribute to it). It is not so obvious from looking at the diagram and data whether the

Date: 18.465, Feb. 22, 2015.

first, leftmost peak is significantly high and separated from the main, middle peak; for that one must do calculations in the dip test.