Addendum to the appendix in Tarboton (1994) on Non-Parametrix density estimation

This addendum provides derivations, corrections and clarifications to the method presented in the appendix to the paper.

Determination of data adaptive bandwidths

Following equation (A2) in the paper the variable bandwidth kernel density estimate is written

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

Following suggestions in Silverman (1986) data adaptive bandwidths are used based upon k nearest neighbors. We take

$$h_i = h_{ref} \frac{d_{ki}}{\overline{d_{ki}}}$$

where d_{ki} is the distance to the k^{th} nearest neighbor from x_i, $k=n^{0.8}$, $\overline{d_{ki}}$ is the geometric mean of all k^{th} nearest neighbor distances and h_{ref} is a reference bandwidth given by equation (A3)

$$h_{ref} = 0.9n^{-0.2} \min(\text{standard deviation}, (\text{interquartile range}/1.34))$$

Integration to determine equations (A5) and (A6) from equation (A4), including corrections to equations (A4), (A6) and (A8).

In general a kernel density estimate is

$$f(x) = \sum_{i=1}^{n} \frac{1}{nh_i} K\left(\frac{x - x_i}{h_i}\right) \tag{1}$$

The x_i are discrete due to rounding so are assumed to be uniformly distributed in the interval $(x_i-\delta, x_i+\delta)$ i.e.

$$U(t) = \begin{cases} \frac{1}{2\delta} & t \in (x_i - \delta, x_i + \delta) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

f(x) is evaluated from (1) by taking the expectation over this uniform distribution for each x_i

$$f(x) = \sum_{i=1}^{n} \frac{1}{nh_i} \int_{x_i-\delta}^{x_i+\delta} U(t) K\left(\frac{x - t}{h_i}\right) dt = \sum_{i=1}^{n} \frac{1}{nh_i} \int_{x_i-\delta}^{x_i+\delta} \frac{1}{2\delta} K\left(\frac{x - t}{h_i}\right) dt \tag{3}$$

This is a correction to equation (A4) in the paper that omitted the 1/2δ. Equation (3) can be written

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} I_i(x) \tag{4}$$

where for each data point $I_i(x)$ is defined as

$$I_i(x) = \frac{1}{h_i}\int_{x_i-\delta}^{x_i+\delta} \frac{1}{2\delta} K\left(\frac{x-t}{h_i}\right) dt \tag{5}$$

Now with **Gaussian Kernel**

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \tag{6}$$

Equation (5) becomes

$$I_i(x) = \frac{1}{h_i}\int_{x_i-\delta}^{x_i+\delta} \frac{1}{2\delta}\frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x-t}{h_i}\right)^2/2} dt \tag{7}$$

Use the change of variables

$$v = \frac{x-t}{\sqrt{2}h_i} \quad dv = \frac{-dt}{\sqrt{2}h_i} \quad dt = -\sqrt{2}h_i dv \tag{8}$$

Then

$$I_i(x) = \frac{1}{h_i}\int_{\frac{x-x_i+\delta}{\sqrt{2}h_i}}^{\frac{x-x_i-\delta}{\sqrt{2}h_i}} \frac{1}{2\delta}\frac{1}{\sqrt{2\pi}} e^{-v^2}(-\sqrt{2}h_i)dv = -\frac{1}{2\delta\sqrt{\pi}}\int_{\frac{x-x_i+\delta}{\sqrt{2}h_i}}^{\frac{x-x_i-\delta}{\sqrt{2}h_i}} e^{-v^2} dv = \frac{1}{2\delta\sqrt{\pi}}\int_{\frac{x-x_i-\delta}{\sqrt{2}h_i}}^{\frac{x-x_i+\delta}{\sqrt{2}h_i}} e^{-v^2} dv \tag{9}$$

The integration limits are now in terms of the variable of integration v and were obtained by substituting the limits in equation (7) which were terms of the variable t into (8).

Now the error function is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-v^2} dv \tag{10}$$

Equation (9) then becomes

$$I_i(x) = \frac{1}{2\delta\sqrt{\pi}} \int_0^{\frac{x-x_i+\delta}{\sqrt{2}h_i}} e^{-v^2} dv - \frac{1}{2\delta\sqrt{\pi}} \int_0^{\frac{x-x_i-\delta}{\sqrt{2}h_i}} e^{-v^2} dv = \frac{1}{4\delta}\left(\mathrm{erf}\left(\frac{x-x_i+\delta}{\sqrt{2}h_i}\right) - \mathrm{erf}\left(\frac{x-x_i-\delta}{\sqrt{2}h_i}\right)\right) \tag{11}$$

Substituting this in to (4) gives the result that is equation (A5) in the paper

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{4\delta}\left(\mathrm{erf}\left(\frac{x-x_i+\delta}{\sqrt{2}h_i}\right) - \mathrm{erf}\left(\frac{x-x_i-\delta}{\sqrt{2}h_i}\right)\right) \tag{12}$$

Now with **Epanechnikov Kernel**

$$K(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1-u^2/5) & -\sqrt{5} < u < \sqrt{5} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

It is important to account for the finite support of this Kernel when evaluating (5).

$$I_i(x) = \frac{1}{h_i} \int_{x_i-\delta}^{x_i+\delta} \frac{1}{2\delta} K\left(\frac{t-x}{h_i}\right) dt \tag{14}$$

Note that here the symmetry of the Kernel has been used to switch x and t. This makes keeping track of the integration limits a bit more direct. Recognizing that between (13) and (14) the

$$u = \frac{t-x}{h_i} \quad t = x + h_i u \tag{15}$$

K(u) is nonzero over the range $-\sqrt{5} < u < \sqrt{5}$ for u. This corresponds to $x - \sqrt{5}\,h_i < t < x + \sqrt{5}\,h_i$. The lower limit on the integral in (14) should therefore be taken as the larger of $x_i - \delta$ and $x - \sqrt{5}h_i$. Denote this a:

$$a = \max(x - \sqrt{5}h_i, x_i - \delta) \tag{16}$$

Similarly the upper limit should be taken as the smaller of $x_i + \delta$ and $x + \sqrt{5}h_i$. The upper limit should also not be less than the lower limit. The upper limit denoted b can be written:

$$b = \max(\min(x + \sqrt{5}h_i, x_i + \delta), a) \tag{17}$$

This corrects equation (A8) in the paper. With these (14) becomes

$$I_i(x) = \frac{1}{h_i}\int_a^b \frac{1}{2\delta} K\left(\frac{t-x}{h_i}\right) dt \tag{18}$$

Now changing variables

$$u = \frac{t-x}{h_i} \quad t = x + h_i u \quad dt = h_i du \tag{19}$$

we get

$$I_i(x) = \frac{1}{h_i}\int_{\frac{x-a}{h_i}}^{\frac{x-b}{h_i}} \frac{1}{2\delta} K(u) h_i du = \frac{1}{2\delta}\int_{\frac{x-a}{h_i}}^{\frac{x-b}{h_i}} K(u) du \tag{20}$$

Now with K(u) given in the positive part of (13)

$$\int_L^U K(u)du = \frac{3}{4\sqrt{5}}\int_L^U (1-u^2/5)du = \frac{3}{4\sqrt{5}}(U-L) - \frac{1}{20\sqrt{5}}(U^3 - L^3) \tag{21}$$

Therefore

$$\begin{aligned} I_i(x) &= \frac{1}{2\delta}\left(\frac{3}{4\sqrt{5}}\left(\frac{x-b}{h_i} - \frac{x-a}{h_i}\right) - \frac{1}{20\sqrt{5}}\left(\left(\frac{x-b}{h_i}\right)^3 - \left(\frac{x-a}{h_i}\right)^3\right)\right) \\ &= \frac{3}{8\delta\sqrt{5}}\left(\frac{x-b}{h_i} - \frac{x-a}{h_i}\right) - \frac{1}{40\delta\sqrt{5}}\left(\left(\frac{x-b}{h_i}\right)^3 - \left(\frac{x-a}{h_i}\right)^3\right) \end{aligned} \tag{22}$$

with this

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{3}{8\delta\sqrt{5}}\left(\frac{x-b}{h_i} - \frac{x-a}{h_i}\right) - \frac{1}{40\delta\sqrt{5}}\left(\left(\frac{x-b}{h_i}\right)^3 - \left(\frac{x-a}{h_i}\right)^3\right) \tag{23}$$

This corrects equation (A6) in the paper to what I now believe is the correct result, but if it is not I am sure someone will tell me.

Citations

Silverman, B. W., (1986), Density Estimation for Statistics and Data Analysis, Chapman and Hall, 175 p.

Tarboton, D. G., (1994), "The Source Hydrology of Severe Sustained Drought in the Southwestern United States," Journal of Hydrology, 161: 31-69.