A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation

S. J. Sheather,M. C. Jones, 1991, J.R. Statistic Society B, Volume 53, Issue 3, 1991, Pages 683-690

for academic use:

The usual kernel density estimate $\hat{f}_h$ of a univariate density f based on a random sample $X_1, \ldots, X_n$ of size n is

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} h^{-1} K\{h^{-1}(x - X_i)\}. \quad (1)$$

The bandwidth h has already been introduced in Section 1; the function K is the kernel function which we take to be a symmetric probability density. All the data-based h selection procedures discussed in this paper are based on choosing h to (at least approximately) minimize a kernel-based estimate of mean integrated squared error (MISE) via the first two terms of its usual asymptotic expansion (AMISE) valid as $n \to \infty$ and $h = h(n) \to 0$:

$$\text{AMISE}(h) = (nh)^{-1} R(K) + \tfrac{1}{4} h^4 \sigma_K^4 R(f'') \quad (2)$$

(e.g. Silverman (1986), section 3.3). Here, the notation follows the convention $R(g) = \int g^2(x)\, dx$ and $\sigma_g^2 = \int x^2 g(x)\, dx$ for appropriate functions g, and the quantities

The bandwidth h_{2S}, for use in equation (1), is the solution to the equation

$$[R(K)/\{\sigma_K^4 \hat{S}_D(\hat{\alpha}_2(h))\}]^{1/5} n^{-1/5} - h = 0 \quad (12)$$

where

$$\hat{S}_D(\alpha) = \{n(n-1)\}^{-1} \alpha^{-5} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{iv}\{\alpha^{-1}(X_i - X_j)\}.$$

From manipulation of equation (9), we have

$$\hat{\alpha}_2(h) = 1.357\{\hat{S}_D(a)/\hat{T}_D(b)\}^{1/7} h^{5/7}.$$

Here, the constant is $D_1(\phi)/R^{1/7}(\phi)$ and the term in brackets is the second stage estimate of $R(f'')/R(f''')$;

$$\hat{T}_D(b) = -\{n(n-1)\}^{-1} b^{-7} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{vi}\{b^{-1}(X_i - X_j)\}.$$

For this estimate the bandwidths a and b are given by a normal scale model estimate of equation (9) and of the corresponding formula for estimating $R(f''')$ in Jones and Sheather (1991) respectively to be

$$a = 0.920 \hat{\lambda} n^{-1/7} \text{ and } b = 0.912 \hat{\lambda} n^{-1/9},$$

where $\hat{\lambda}$ is the sample interquartile range.

We successfully use the Newton–Raphson method to solve equation (12). A Fortran subroutine is available on request from the first author.

see Chapter 20 of
https://kdepy.readthedocs.io/en/latest/introduction.html#Selecting-a-suitable-bandwidth

A fast version implemented using FFT
https://kdepy.readthedocs.io/en/latest/introduction.html#Selecting-a-suitable-bandwidth