Density estimation: kernel methods

Juneteenth, 2020

Given a sample of data $X_1, ...X_n$, we would like to estimate its underlying probability density function f.

$$P(a < X < b) = \int_a^b f(x)dx \tag{1}$$

Histograms

Considering that for a window width (or a bin width) h

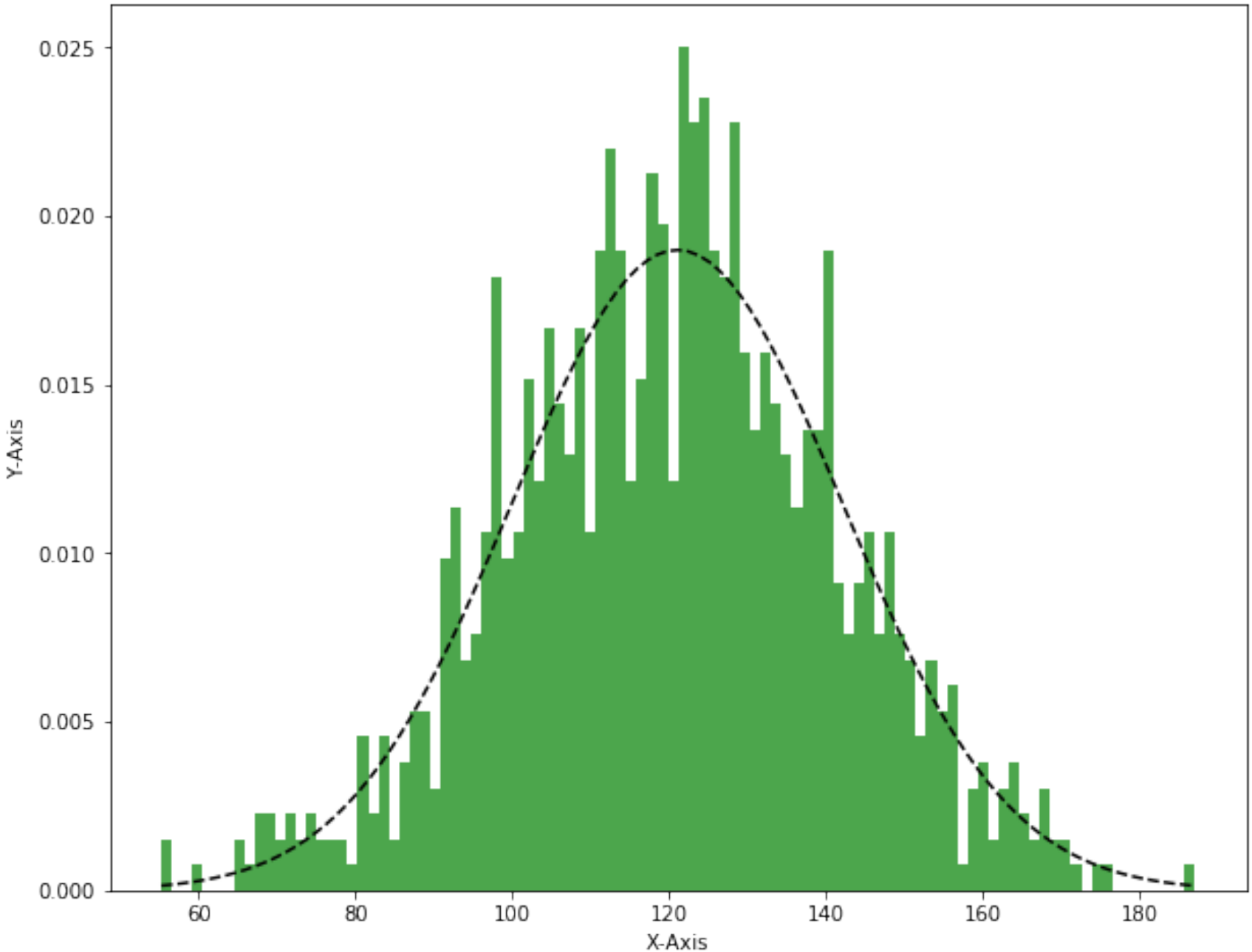
$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h) \tag{2}$$

we can define the estimate $\hat{f}$ as

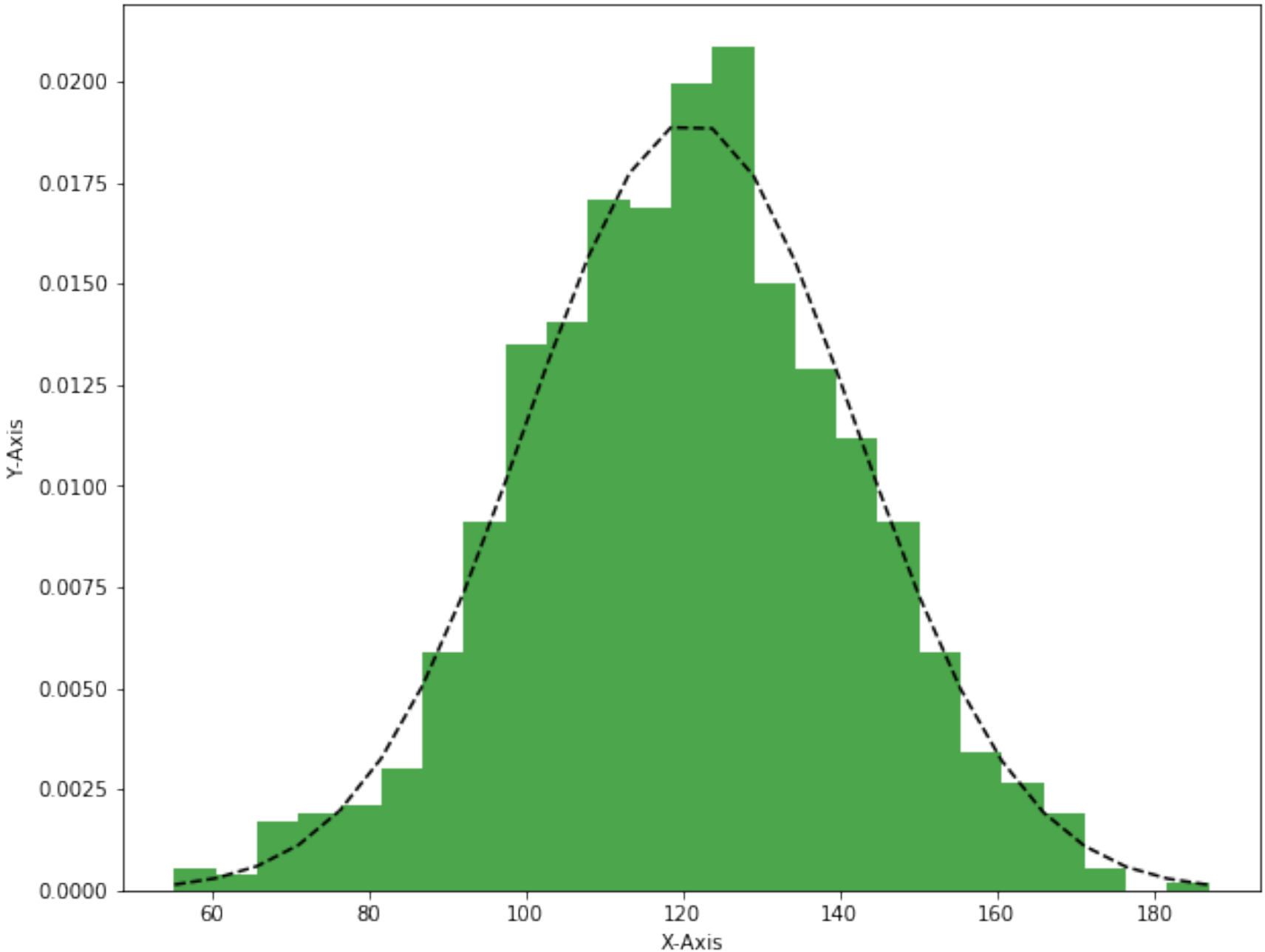
$$\hat{f}(x) = \frac{1}{2hn} m \tag{3}$$

with m being the amount of samples X_i falling in $[x - h, x + h]$

Histograms with 100 bins



Histograms with 25 bins



Nearest neighbours

Define

$$d_1(t) \leq d_2(t) \leq ... \leq d_n(t)$$

as the distance from a point t to the points of the sample.

For each $r > 0$, you expect $2rnf(t)$ samples to fall in $[t - r, t + r]$. So in $[t - d_k(t), t + d_k(t)]$ you have k samples by definition.

Define [1]

$$\hat{f}(t) = \frac{k}{2nd_k(t)} \tag{4}$$

[1] *"A non-parametric estimate of a multivariate density function"*, by Loftsgaarden and Quesenberry, 1965

Nearest neighbours

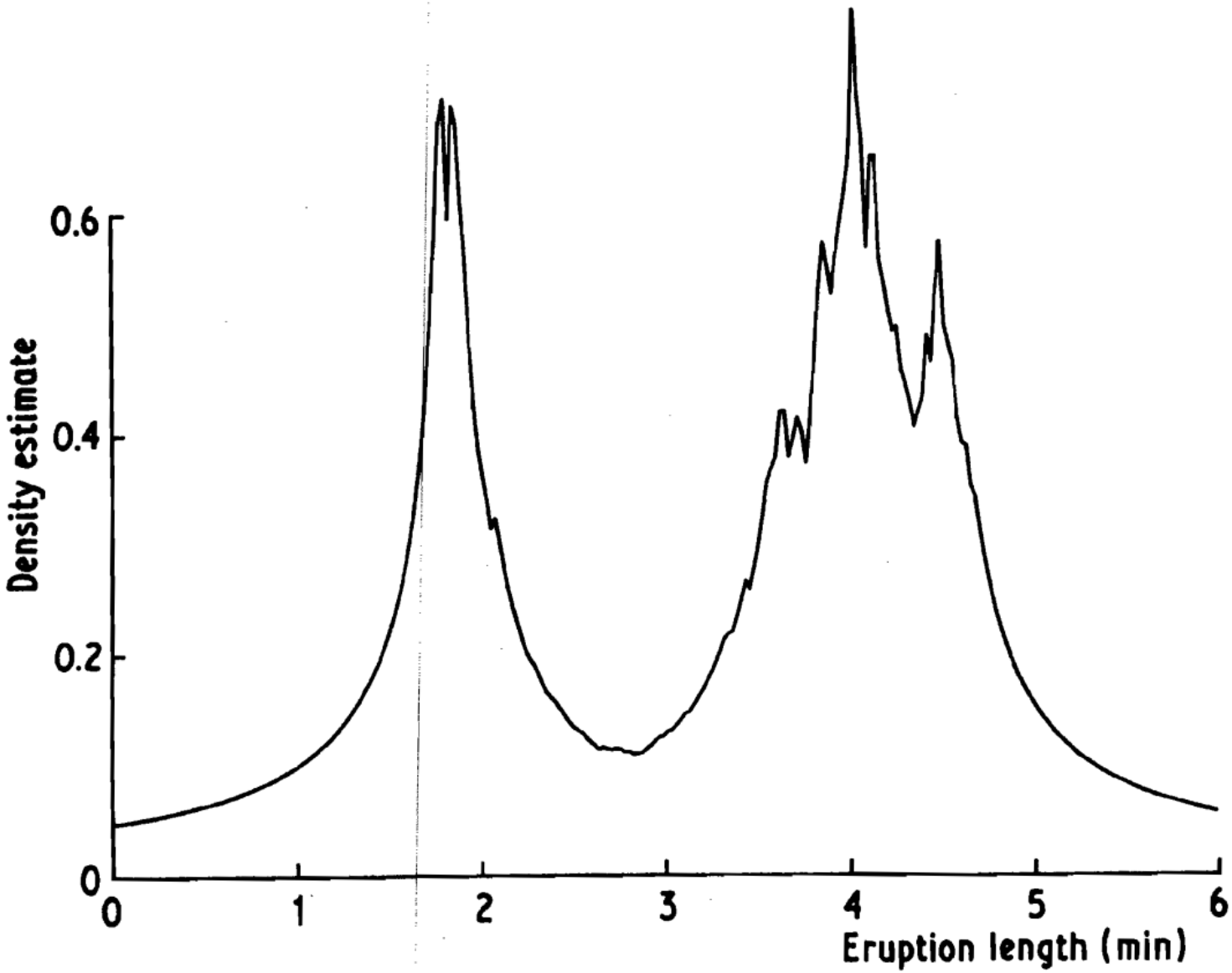


Fig. 2.10 *Nearest neighbour estimate for Old Faithful geyser data, $k = 20$.*

The Kernel Estimator

Assume K a function satisfying

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{5}$$

The kernel estimator is defined [2] as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{6}$$

[2] *"Remarks on some non-parametric estimates of a density function"*, by Rosenblatt, 1956

And also

The variable kernel estimator

$$\hat{f}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h d_{j,k}} K\left(\frac{t - X_j}{h d_{j,k}} \right) \tag{7}$$

or the generalized k-th nearest neighbour estimate

$$\hat{f}(t) = \frac{1}{n d_k(t)} \sum_{i=1}^{n} K\left(\frac{t - X_i}{d_k(t)} \right) \tag{8}$$

Or using orthogonal functions

For ϕ_ν a set of orthogonal functions,

$$\hat{f}(t) = \sum_{\nu=0}^{M} \hat{f}_\nu \phi_\nu(t) \tag{9}$$

with

$$\hat{f}_\nu = E\phi_\nu(X) = \frac{1}{n}\sum_{i=1}^{n} \phi_\nu(X_i) \tag{10}$$

MSE and MISE, bias and variance

$$MSE(\hat{f}) = E[\hat{f}(x) - f(x)]^2 = [E\hat{f}(x) - f(x)]^2 + var\hat{f}(x) \quad (11)$$

$$MISE(\hat{f}) = \int MSE dx^3 \quad (12)$$

$$E\hat{f}(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y) dy \quad (13)$$

$$nvar\hat{f}(x) = \int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y) dy - \left(\int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y) dy\right)^2 \quad (14)$$

[3] Also Rosenblatt, 1956

Approximating MSE and MISE

Suppose that the kernel satisfies

$$\int K(t)dt = 1 \tag{15}$$

$$\int tK(t)dt = 0 \tag{16}$$

$$\int t^2 K(t)dt = k_2 \neq 0 \tag{17}$$

The bias

Doing a change of variables $y = x - ht$

$$bias = E\hat{f}(x) - f(x) = \int K(t)[f(x - ht) - f(x)]dt \tag{18}$$

Then use a Taylor's expansion on f to find

$$bias = \frac{1}{2}h^2 f''(x)k_2 + \text{higher order terms in h} \tag{19}$$

Integrating

$$\int bias(x)^2 dx \approx \frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx \tag{20}$$

The variance

Using $y = x - ht$ and the bias approximation from the last slide

$$var\hat{f}(x) = \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}[f(x) + bias(x)]^2$$
$$\approx \frac{1}{hn}\int f(x-ht)K(t)^2 dt - \frac{1}{n}[f(x) + O(h^2)] \qquad (21)$$

Assuming h is small, n is large and using Taylor's expansion,

$$var\hat{f}(x) \approx \frac{1}{nh}f(x)\int K(t)^2 dt \qquad (22)$$

And integrating

$$\int var\hat{f}(x)dx \approx \frac{1}{nh}\int K(t)^2 dt \qquad (23)$$

Ideal Kernel

So far, we have

$$MISE \approx \frac{1}{4} h^4 k_2^2 \int f''(x)^2 dx + \frac{1}{nh} \int K(t)^2 dt \tag{24}$$

Somehow there is an optimal value of h that minimizes $MISE$, [4]

$$h_{opt} = k_2^{-2/5} n^{-1/5} \left(\int K(t)^2 dt \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5} \tag{25}$$

Substituting,

$$MISE(h_{opt}) = \frac{4}{5} C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5} \tag{26}$$

$$\text{with } C(K) = k_2^{2/5} \left(\int K(t)^2 dt \right)^{4/5} \tag{27}$$

[4] *"On estimation of a probability density function and mode"*, by Parzen, 1962 (Lemma 4A)

Ideal Kernel

So choose a Kernel that minimizes $C(K)$. Assume $k_2 = 1$ (or rescale) to see this as minimization problem with constraints. It has a solution.

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right) & \text{if } -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{otherwise.} \end{cases}$$

K_e is called the *Epanechnikov* [5] [6] kernel and it is used to measure efficiency of different kernels by computing

$$eff(K) = \left(\frac{C(K_e)}{C(K)}\right)^{5/4} \tag{28}$$

[5] "*Nonparametric estimation of a multidimensional probability density*", by Epanechnikov, 1969

[6] "*The efficiency of some nonparametric competitors of the t-test*" by Hodges and Lehmann, 1956

Ideal Kernel

Table 3.1 *Some kernels and their efficiencies*

Kernel	$K(t)$	*Efficiency (exact and to 4 d.p.)*
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert<\sqrt{5}$, 0 otherwise	1
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert<1$, 0 otherwise	$\left(\frac{3087}{3125}\right)^{1/2}\approx 0.9939$
Triangular	$1-\lvert t\rvert$ for $\lvert t\rvert<1$, 0 otherwise	$\left(\frac{243}{250}\right)^{1/2}\approx 0.9859$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-(1/2)t^2}$	$\left(\frac{36\pi}{125}\right)^{1/2}\approx 0.9512$
Rectangular	$\frac{1}{2}$ for $\lvert t\rvert<1$, 0 otherwise	$\left(\frac{108}{125}\right)^{1/2}\approx 0.9295$

Ideal h and Least Squares Cross Validation

Look at ISE [7][8]

$$\int (\hat{f} - f)^2 = \int \hat{f}^2 - 2\int \hat{f} f + \int f^2. \tag{29}$$

Define

$$R(\hat{f}) = \int \hat{f}^2 - 2\int \hat{f} f \tag{30}$$

$$\hat{f}_{-i}(x) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x - Xj}{h}\right) \tag{31}$$

$$M_0(h) = \int \hat{f}^2 - \frac{2}{n} \sum_i \hat{f}_{-i}(X_i) \tag{32}$$

---

[7] *"Empirical choice of histograms and kernel density estimators"*, by Rudemo, 1982

[8] *"An alternative method of cross-validation for the smoothing of density estimates"*,by Bowman, 1984

Ideal h and Least Squares Cross Validation

Noting that

$$E\frac{1}{n}\sum_i \hat{f}_{-i}(X_i) = E\hat{f}_{-n}(X_n) = E\int \hat{f}(x)f(x)dx \tag{33}$$

implies

$$EM_0(h) = ER(\hat{f}). \tag{34}$$

So $M_0(h) + \int f^2$ is an unbiased estimator for $MISE$. Assuming the minimizers of M_0 and EM_0 are close, a good choice of h is the one that minimizes M_0.

M_0 can be rewritten into a more computer-friendly manner.

Some results

Assuming that [9]

- K (a bounded Borel function) satisfies $\int |K(t)|dt < \infty$ and $\int K(t)dt = 1$
- $|tK(t)| \to 0$ as $|t| \to \infty$
- h (that depends on n) satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$

Then $\hat{f}(x) \to f(x)$ in probability as $n \to \infty$

[9] *"On estimation of a probability density function and mode"*, by Parzen, 1962

Some results

Assuming that [10]

- K is bounded, has bounded variation, the set of discontinuities has measure zero
- f is uniformly continuous on $(-\infty, \infty)$
- h (that depends on n) satisfies $h \to 0$ and $nh(\log n)^{-1} \to \infty$ as $n \to \infty$

Then $\sup_x |\hat{f}(x) - f(x)| \to 0$ with probability 1 $n \to \infty$. Conditions are necessary as well as suficient.

[10] *"Convergence uniforme d'un estimateur de la densité par la méthod de noyau"*, by Bertrand-Retali, 1978

Some results

Assuming that [11]

- K is a non-negative Borel function that integrates to 1
- h (that depends on n) satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$

Then $\int |\hat{f}(x) - f(x)|dx \to 0$ with probability 1 $n \to \infty$. Conditions are necessary as well as suficient and no assumption is made on f.

[11] "*Non-parametric density estimation: The L_1 view*", by Devroye and Györfi, 1985

Remarks

To choose the ideal window width h, you can make an assumption over f'' and draw an estimate from the h_{opt} expression, but that leads to massive oversmoothing. [12]

If the number of samples n is fixed there is no guarantee of convergence of any kind. Using Gaussian kernels, for any f in $L^2(\mathbb{R})$ and any $\epsilon > 0$ there exists $t > 0$, $N \in \mathbb{N}$ and $a_n \in \mathbb{R}$ such that [13]

$$\left\| f - \sum_{n=0}^{N} a_n e^{-(x-nt)^2} \right\|_2 < \epsilon \tag{35}$$

[12] *"Density Estimation for Statistics and Data Analysis"*, by Silverman, 1986

[13] *"Approximating with Gaussians"*, by Calcaterra and Boldt, 2008