Chapter 2

Nonparametric Density Estimation

2.1 Introduction

For a moment, we will go back to simple data structures: we have observations which are realizations of univariate random variables,

$$X_1, \ldots, X_n \text{ i.i.d. } \sim F,$$

where F denotes an unknown cumulative distribution function. The goal is to estimate the distribution F. In particular, we are interested in estimating the density $f = F'$, assuming that it exists.

Instead of assuming a parametric model for the distribution (e.g. Normal distribution with unknown expectation and variance), we rather want to be "as general as possible": that is, we only assume that the density exists and is suitably smooth (e.g. differentiable). It is then possible to estimate the unknown density **function** $f(\cdot)$. Mathematically, a function is an **infinite-dimensional** object. Density estimation will become a "basic principle" how to do estimation for infinite-dimensional objects. We will make use of such a principle in many other settings such as nonparametric regression with one predictor variable (Chapter 3) and flexible regression and classification methods with many predictor variables (Chapter 7).

2.2 Estimation of a density

We consider the data which records the duration of eruptions of "Old Faithful", a famous geysir in Yellowstone National Park (Wyoming, USA). You can watch it via web-cam on `http://www.nps.gov/yell/oldfaithfulcam.htm`

2.2.1 Histogram

The histogram is the oldest and most popular density estimator. We need to specify an *"origin"* x_0 and the *class width* h for the specifications of the intervals

$$I_j = (x_0 + j \cdot h, x_0 + (j+1)h] \ \ (j = \ldots, -1, 0, 1, \ldots)$$

for which the histogram counts the number of observations falling into each I_j: we then plot the histogram such that the area of each bar is proportional to the number of observations falling into the corresponding class (interval I_j).

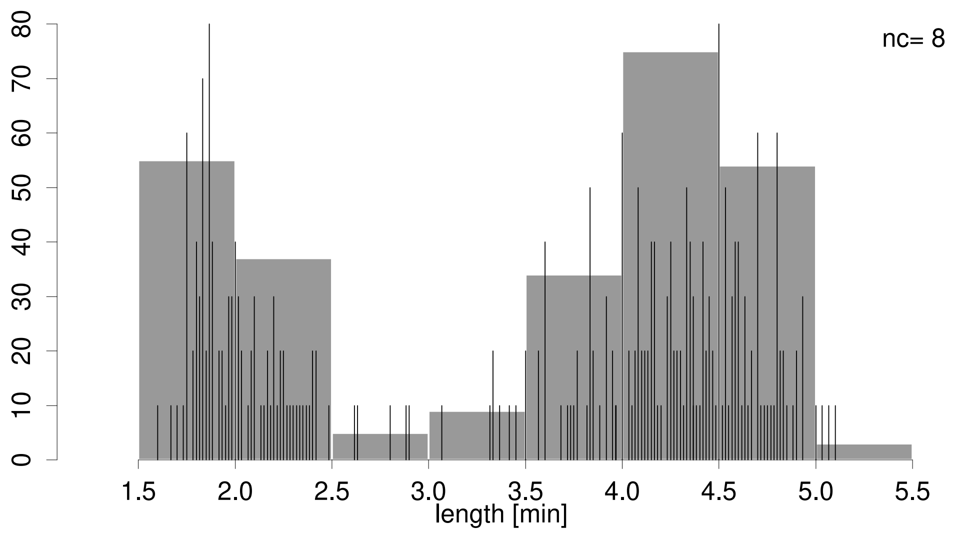
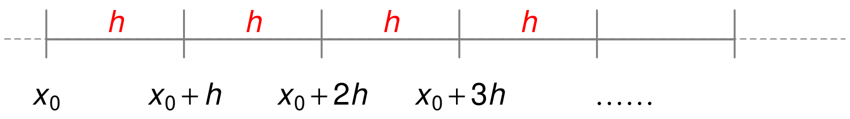


Figure 2.1: Histograms (different class widths) for durations of eruptions of “Old Faithful” geysir in Yellowstone Park ($n = 272$, `data(faithful)`).



The choice of the “origin” x_0 is highly arbitrary, whereas the role of the class width is immediately clear for the user. The form of the histogram depends very much on these two tuning parameters.

2.2.2 Kernel estimator

The naive estimator

Similar to the histogram, we can compute the relative frequency of observations falling into a small region. The density function $f(\cdot)$ at a point x can be represented as

$$f(x) = \lim_{h \to 0} \frac{1}{2h} \mathbb{P}[x - h < X \le x + h]. \tag{2.1}$$

The naive estimator is then constructed without taking the limit in (2.1) and by replacing probabilities with relative frequencies:

$$\hat{f}(x) = \frac{1}{2hn} \, \#\{i; X_i \in (x - h, x + h]\}. \tag{2.2}$$

This naive estimator is only piecewise constant since every X_i is either in or out of the interval $(x - h, x + h]$. As for histograms, we also need to specify the so-called bandwidth h; but in contrast to the histogram, we do not need to specify an origin x_0.

An alternative representation of the naive estimator (2.2) is as follows. Define the weight function

$$w(x) = \begin{cases} 1/2 & \text{if } |x| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

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

If we choose instead of the rectangle weight function $w(\cdot)$ a general, typically more smooth kernel function $K(\cdot)$, we have the definition of the kernel density estimator

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

$$K(x) \geq 0, \quad \int_{-\infty}^{\infty} K(x)dx = 1, \;\; K(x) = K(-x). \tag{2.3}$$

The estimator depends on the *bandwidth* $h > 0$ *which acts as a tuning parameter*. For large bandwidth h, the estimate $\hat{f}(x)$ tends to be very slowly varying as a function of x, while small bandwidths will produce a more wiggly function estimate. The positivity of the kernel function $K(\cdot)$ guarantees a positive density estimate $\hat{f}(\cdot)$ and the normalization $\int K(x)dx = 1$ implies that $\int \hat{f}(x)dx = 1$ which is necessary for $\hat{f}(\cdot)$ to be a density. Typically, the kernel function $K(\cdot)$ is chosen as a probability density which is symmetric around 0.

The smoothness of $\hat{f}(\cdot)$ is inherited from the smoothness of the kernel: if the rth derivative $K^{(r)}(x)$ exists for all x, then $\hat{f}^{(r)}(x)$ exists as well for all x (easy to verify using the chain rule for differentiation).

Popular kernels are the Gaussian kernel

$$K(x) = \varphi(x) = (2\pi)^{-\frac{1}{2}} e^{-x^2/2} \text{ (the density of the } \mathcal{N}(0,1) \text{ distribution)}$$

or a kernel with finite support such as $K(x) = \frac{\pi}{4}\cos(\frac{\pi}{2}x)I(|x| \leq 1)$. The Epanechnikov kernel, which is optimal with respect to mean squared error, is

$$K(x) = \frac{3}{4}\left(1 - |x|^2\right) I(|x| \leq 1).$$

But far more important than the kernel is the bandwidth h, see figure 2.2: its role and how to choose it are discussed below.

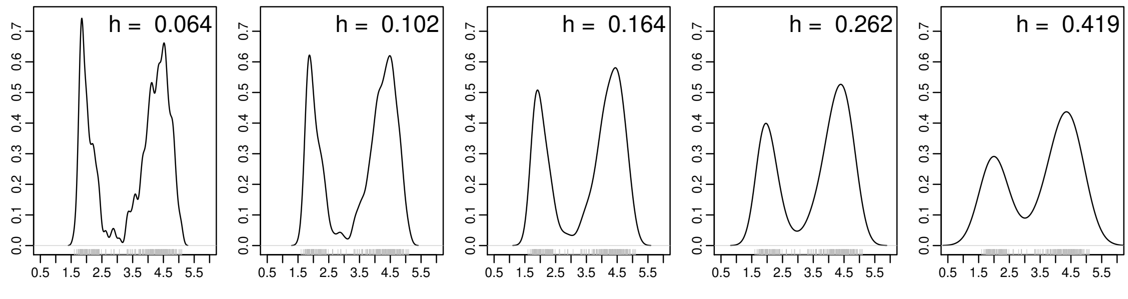


Figure 2.2: *kernel density estimates of the "Old Faithful" eruption lengths; Gaussian kernel and bandwidths* $h = 0.04 \cdot 1.6^{1,2,\ldots,5}$.

2.3 The role of the bandwidth

The bandwidth h is often also called the "smoothing parameter": a moment of thought will reveal that for $h \to 0$, we will have "δ-spikes" at every observation X_i, whereas $\hat{f}(\cdot)$ becomes smoother as h is increasing.

2.3.1 Variable bandwidths: k nearest neighbors

Instead of using a global bandwidth, we can use locally changing bandwidths $h(x)$.

The general idea is to use a large bandwidth for regions where the data is sparse. The k-nearest neighbor idea is to choose

$$h(x) = \text{ Euclidean distance of } x \text{ to the } k\text{th nearest observation},$$

where k is regulating the magnitude of the bandwidth. Note that generally, $\hat{f}(\cdot)$ will not be a density anymore since the integral is not necessarily equal to one.

2.3.2 The bias-variance trade-off

We can formalize the behavior of $\hat{f}(\cdot)$ when varying the bandwidth h in terms of bias and variance of the estimator. It is important to understand heuristically that

$$\boxed{\text{the (absolute value of the) bias of } \hat{f} \text{ increases and the variance of } \hat{f} \text{ decreases as } h \text{ increases.}}$$

Therefore, if we want to minimize the mean squared error MSE(x) at a point x,

$$\mathrm{MSE}(x) = \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] = \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2 + \mathrm{Var}(\hat{f}(x)),$$

we are confronted with a **bias-variance trade-off**. As a consequence, this allows - at least conceptually - to optimize the bandwidth parameter (namely to minimize the mean squared error) in a well-defined, coherent way. Instead of optimizing the mean squared error at a point x, one may want to optimize the integrated mean squared error (IMSE)

$$\mathrm{IMSE} = \int \mathrm{MSE}(x)dx$$

which yields an integrated decomposition of squared bias and variance (integration is over the support of X). Since the integrand is non-negative, the order of integration (over the support of X and over the probability space of X) can be reversed, denoted as MISE (mean integrated squared error) and written as

$$\mathrm{MISE} = \mathbb{E}[\int \left(\hat{f}(x) - f(x)\right)^2 dx] \tag{2.4}$$

2.3.3 Asymptotic bias and variance

It is straightforward (using definitions) to give an expression for the exact bias and variance:

$$\begin{aligned}
&\mathbb{E}[\hat{f}(x)] = \int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy \\
&\mathrm{Var}(\hat{f}(x)) \\
&= \frac{1}{nh^2}\mathrm{Var}\left(K\left(\frac{x-X_i}{h}\right)\right) = \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{x-X_i}{h}\right)^2\right] - \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{x-X_i}{h}\right)\right]^2 \\
&= n^{-1}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - n^{-1}\left(\int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy\right)^2.
\end{aligned} \tag{2.5}$$

For the bias we therefore get

$$\begin{aligned}\text{Bias}(x) &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\ \underbrace{=}_{\text{change of variable}} & \int K(z) f(x-hz)dz - f(x) = \int K(z) \left(f(x-hz) - f(x)\right) dz. \quad (2.6)\end{aligned}$$

To approximate this expression in general, we invoke an asymptotic argument. We assume that $h \to 0$ as sample size $n \to \infty$, that is:

$$\boxed{h = h_n \to 0 \text{ with } nh_n \to \infty.}$$

This will imply that the bias goes to zero since $h_n \to 0$; the second condition requires that h_n is going to zero more slowly than $1/n$ which turns out to imply that also the variance of the estimator will go to zero as $n \to \infty$. To see this, we use a Taylor expansion of f, assuming that f is sufficiently smooth:

$$f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}h^2 z^2 f''(x) + \ldots$$

Plugging this into (2.6) yields

$$\begin{aligned}\text{Bias}(x) &= -hf'(x) \underbrace{\int zK(z)dz}_{=0} + \frac{1}{2}h^2 f''(x) \int z^2 K(z)dz + \ldots \\ &= \frac{1}{2}h^2 f''(x) \int z^2 K(z)dz + \text{higher order terms in } h.\end{aligned}$$

For the variance, we get from (2.5)

$$\begin{aligned}\text{Var}(\hat{f}(x)) &= n^{-1} \int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)dy - n^{-1}(f(x) + \text{Bias}(x))^2 \\ &= n^{-1}h^{-1} \int f(x-hz)K(z)^2 dz - \underbrace{n^{-1}(f(x) + \text{Bias}(x))^2}_{=O(n^{-1})} \\ &= n^{-1}h^{-1} \int f(x-hz)K(z)^2 dz + O(n^{-1}) = n^{-1}h^{-1} f(x) \int K(z)^2 dz + o(n^{-1}h^{-1}),\end{aligned}$$

assuming that f is smooth and hence $f(x-hz) \to f(x)$ as $h_n \to 0$.
In summary: for $h = h_n \to 0,\ h_n n \to \infty$ as $n \to \infty$

$$\boxed{\begin{array}{lcll}\text{Bias}(x) & = & h^2 f''(x) \int z^2 K(z)\, dz/2 + o(h^2) & (n \to \infty) \\ \text{Var}(\hat{f}(x)) & = & (nh)^{-1} f(x) \int K(z)^2 dz + o((nh)^{-1}) & (n \to \infty)\end{array}}$$

The optimal bandwidth $h = h_n$ which minimizes the leading term in the asymptotic MSE(x) can be calculated straightforwardly by solving $\frac{\partial}{\partial h}\text{MSE}(x) = 0$,

$$h_{\text{opt}}(x) = n^{-1/5} \left(\frac{f(x) \int K^2(z)dz}{(f''(x))^2 (\int z^2 K(z)dz)^2}\right)^{1/5}. \quad (2.7)$$

Since it's not straightforward to estimate and use a *local* bandwidth $h(x)$, one rather considers minimizing the MISE, i.e., $\int MSE(x)\,dx$ which is *asymptotically*

$$\text{asympt. MISE} = \int \text{Bias}(x)^2 \; + \; \text{Var}(\hat{f}(x))\; dx = \frac{1}{4}h^4 R(f'')\,\sigma_K^2 + R(K)/(nh), \qquad (2.8)$$

where $R(g) = \int g^2(x)\,dx$, $\sigma_K^2 = \int x^2 K(x)\,dx$, and the "global" asymptotically optimal bandwidth becomes

$$h_{\text{opt}} = n^{-1/5}\left(R(K)/\sigma_K^2 \times 1/R(f'')\right)^{1/5}. \qquad (2.9)$$

The optimal rate for the MISE and MSE(x) is therefore $O(n^{-4/5})$.

2.3.4 Estimating the bandwidth

As seen from (2.9), the asymptotically best bandwidth depends on $R(f'') = \int f''^2(x)\,dx$ which is unknown (whereas as $R(K)$ and σ_K^2 are known). It is possible to estimate the f'' again by a kernel estimator with an "initial" bandwidth h_{init} (sometimes called a pilot bandwidth) yielding $\hat{f}''_{h_{\text{init}}}$. Plugging this estimate into (2.9) yields an estimated bandwidth $\hat{h}$ for the density estimator $\hat{f}(\cdot)$ (the original problem): of course, $\hat{h}$ depends on the initial bandwidth h_{init}, but choosing h_{init} in an ad-hoc way is less critical for the density estimator than choosing the bandwidth h itself. Furthermore, methods have been deviced to determine h_{init} and h simultaneously (e.g., "Sheather-Jones", in R using `density(*, bw="SJ")`).

Estimating local bandwidths

Note that the $h_{\text{opt}}(x)$ bandwidth selection in (2.7) is more problematical mainly because $\hat{f}_{\hat{h}_{\text{opt}}(x)}(x)$ will not integrate to one without further normalization. On the other hand, it can be important to use *locally varying* bandwidths instead of a single global one in a kernel estimator at the expense of being more difficult. The plug-in procedure outlined above can be applied *locally*, i.e., conceptually for each x and hence describes how to estimate local bandwidths from data and how to implement a kernel estimator with locally varying bandwidths. In the related area of nonparametric regression, in section 3.2.1, we will show an example about locally changing bandwidths which are estimated based on an iterative version of the (local) plug-in idea above.

Other density estimators

There are quite a few other approaches to density estimation than the kernel estimators above (whereas in practice, the *fixed* bandwidth kernel estimators are used predominantly because of their simplicity). An important approach in particular aims to estimate the *log density* $\log f(x)$ (setting $\hat{f} = \exp(\widehat{\log f})$) which has no positivity constraints and whose "normal limit" is a simple quadratic. One good implementation is in Kooperberg's R package `logspline`, where spline knots are placed in a stepwise algorithm minimizing approximate BIC (or AIC). This is can be seen as another version of locally varying bandwidths.

2.4 Higher dimensions

Quite many applications involve multivariate data. For simplicity, consider data which are i.i.d. realizations of d-dimensional random variables

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \text{ i.i.d. } \sim f(x_1, \ldots, x_d) dx_1 \cdots dx_d$$

where $f(\cdot)$ denotes the multivariate density.

The multivariate kernel density estimator is, in its simplest form, defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right),$$

where the kernel $K(\cdot)$ is now a function, defined for d-dimensional $\mathbf{x}$, satisfying

$$K(\mathbf{x}) \geq 0, \int_{\mathbb{R}^d} K(\mathbf{x}) d\mathbf{x} = 1, \; K(\cdot) \text{ radially symmetric.}$$

Usually, the kernel $K(\cdot)$ is chosen as a product of a kernel K_{univ} for univariate density estimation

$$K(\mathbf{u}) = \prod_{j=1}^{d} K_{univ}(u_j).$$

2.4.1 The curse of dimensionality

In practice, multivariate kernel density estimation is often restricted to dimension $d = 2$. The reason is, that a higher dimensional space (with d of medium size or large) will be only very sparsely populated by data points. Or in other words, there will be only very few neighboring data points to any value $\mathbf{x}$ in a higher dimensional space, unless the sample size is extremely large. This phenomenon is also called the *curse of dimensionality*.

An implication of the curse of dimensionality is the following lower bound for the best mean squared error of nonparametric density estimators (assuming that the underlying density is twice differentiable): it has been shown that the best possible MSE rate is

$$O(n^{-4/(4+d)}).$$

The following table evaluates $n^{-4/(4+d)}$ for various n and d:

$n^{-4/(4+d)}$	$d=1$	$d=2$	$d=3$	$d=5$	$d=10$
$n=100$	0.025	0.046	0.072	0.129	0.268
$n=1000$	0.004	0.010	0.019	0.046	0.139
$n=100'000$	$1.0 \; 10^{-4}$	$4.6 \; 10^{-4}$	$13.9 \; 10^{-4}$	0.006	0.037

Thus, for $d = 10$, the rate with $n = 100'000$ is still 1.5 times worse than for $d = 1$ and $n = 100$.