STAT 5460:
Nonparametric Methods in Statistics

Kris De Brabanter

kbrabant@iastate.edu

Department of Statistics & Department of Industrial and Manufacturing Systems Engineering

2419 Snedecor Hall

Ames, IA, 50011-1210, USA

When & Where: MWF: 12:05pm - 12:55pm, 3121 Snedecor Hall

(3-0) Cr. 3. Alt. F., offered even-numbered years. Prereq: STAT 5100, STAT 5420 Overview of parametric versus nonparametric methods of inference; introduction to rank-based tests and/or nonparametric smoothing methods for estimating density and regression functions; smoothing parameter selection; applications to semi-parametric models and goodness-of-fit tests of a parametric model.

Fall 2025

Contents

Chapter 1

Introduction to nonparametric statistics

1.1 Historical evolution and general background

The regression estimation problem has a long history. Already in 1632 Galileo Galilei used a procedure which can be interpreted as fitting a linear relationship to contaminated observed data. Such fitting of a line through a cloud of points is the classical linear regression problem. Roughly 125 years later, Roger Joseph Boscovich (1757) addressed the fundamental mathematical problem of determining the parameters which best fits observational equations to data. Since then, a large number of estimation methods have been developed for linear regression. Four of the most commonly used methods are the least absolute deviations, least squares, trimmed least squares and M-regression.

Probably the most well-known method is the method of least squares, although Boscovich (1757) first considered least absolute deviations. The method of least squares was first published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies around the sun. Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss-Markov theorem. The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean). For Galton, regression only had this biological meaning, but his work was later extended by Yule (1897) and Pearson (1903) to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss' formulation of 1821.

At some point in time, it became clear that it is not always easy to find a suitable parametric (linear or nonlinear) model to explain some phenomena. One was searching for a more flexible method where "the data would speak for themselves". For this reason, nonparametric smoothing methods were invented. Smoothing methods also have a long tradition. In the nineteenth century the nonparametric approach has been used as a major tool for empirical analysis: in 1857 the Saxonian economist Engel found the famous Engelsches Gesetz by constructing a curve which we would nowadays call a regressogram. The nonparametric smoothing approach has then long been neglected and the mathematical development of statistical theory in the first half of this century has mainly suggested a purely parametric approach for its simplicity in computation, its compatibility with model assumptions and also for its mathematical convenience.

The real breakthrough of these methods dates back to 1950s and early 1960s with pioneering articles of Rosenblatt (1956) and Parzen (1962) in the density estimation setting and with Nadaraya (1964) and Watson (1964) in the regression setting. Ever since, these methods are gaining more and more attention and popularity. Mainly, this is due to the fact that statisticians realized that pure parametric thinking in curve estimations often does not meet the need for flexibility in data analysis. Also the development of hardware created the demand for theory of now computable nonparametric estimates. However, nonparametric techniques have no intention of replacing parametric techniques. In fact, a combination of both can lead to the discovery of many interesting results which are difficult to accomplish by a single method e.g. semiparametric regression (Ruppert et al., 2003).

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in

which the predictor or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

The increasing importance of regression estimation is also stimulated by the growth of information technology in the past twenty years. The demand for procedures capable of automatically extracting information from massive high-dimensional data sets is rapidly growing. Usually there is no prior knowledge available, leaving the data analyst with no other choice but a nonparametric approach. Often these nonparametric techniques are pushed towards and possibly even over their limits because of their extreme flexibility. Therefore, caution is still advised when applying these techniques. Properties such as (universal) consistency (Stone, 1977) and rate of convergence may be not be neglected.

1.2 Smoothing: general concepts

To estimate a curve, such as a probability density function f_X or a regression function m, we must smooth the data in some way. The rest of this course is devoted to smoothing methods. Next, we discuss some general issues related to smoothing. There are mainly two types of problems we will study. The first is density estimation in which we have a random sample $X_1, \ldots, X_n$ from a distribution F_X with density f_X, written

$$X_1, \ldots, X_n \sim f_X,$$

and we want to estimate the probability density function f_X. The second is regression in which we have pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$Y_i = m(X_i) + e_i,$$

with $\mathbf{E}[e \mid X] = 0$, and we want to estimate the regression function m. To illustrate density and regression estimation consider the following two examples.

Example 1.1 (Density estimation) *This data set illustrates the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Figure 1.1 shows a histogram density estimator (left) and the more sophisticated kernel density estimator. Both methods will be discussed in Chapter 2.*

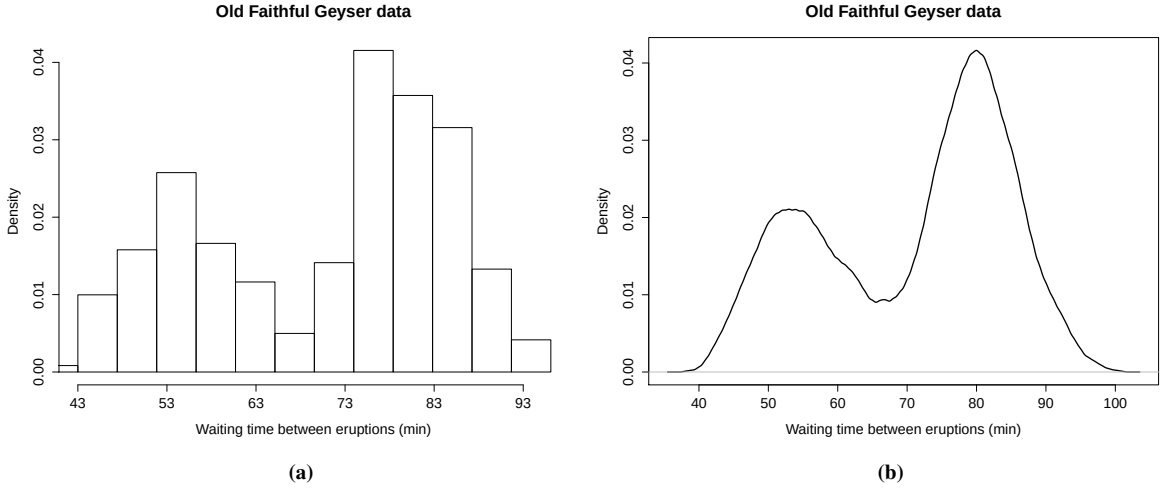


Figure 1.1: Old Faithful Geyser data. (a) Histogram. (b) Kernel density estimate

Example 1.2 (Regression) *Ruppert et al. (2003) describe data from a light detection and ranging (LIDAR) experiment. LIDAR is used to monitor pollutants. Figure 4.7 shows 221 observations. The response is the log of the ratio of*

light received from two lasers. The frequency of one laser is the resonance frequency of mercury while the second has a different frequency. The estimate shown here is the local linear (left) and local cubic (right) kernel regression (Fan and Gijbels, 1996). This method will be discussed in Chapter 4.

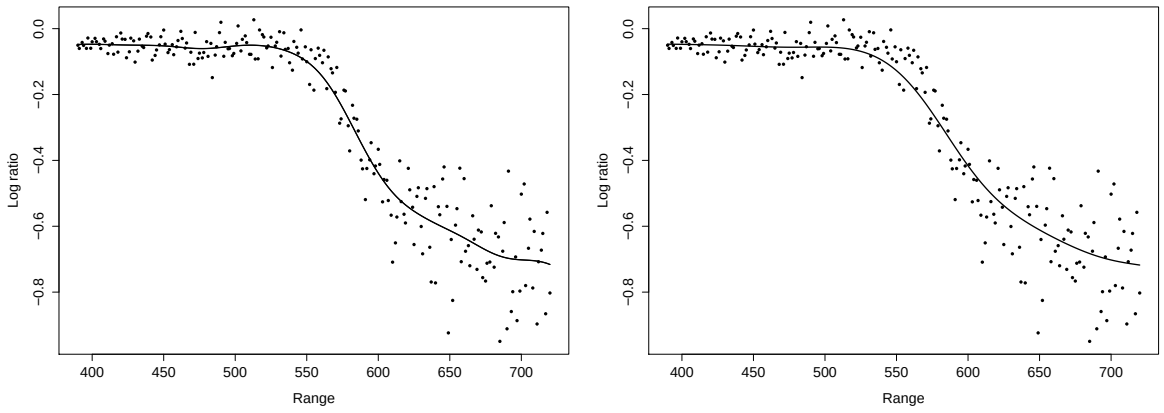


Figure 1.2: LIDAR data. (a) Local linear kernel regression. (b) Local cubic kernel regression

1.3 Some concepts on continuous random variables

1.3.1 Density and cumulative distribution

A random variable X is called **continuous** if there exists a nonnegative function f_X, called the **probability density function** of X or PDF such that

$$\mathbf{P}[X \in B] = \int_B f_X(x)\,dx,$$

for every subset B of the real line. The integral is to be interpreted in the usual Riemann sense and we implicitly assume it is well defined. In particular, the probability that the value of X falls within an interval is

$$\mathbf{P}[a \leq X \leq b] = \int_a^b f_X(x)\,dx,$$

and can be interpreted as the area under the graph of the PDF. For a single value a, we have $\mathbf{P}[X = a] = \int_a^a f_X(x)\,dx = 0$. For this reason, including or excluding the endpoints of an interval has no effect on its probability:

$$\mathbf{P}[a \leq X \leq b] = \mathbf{P}[a < X < b] = \mathbf{P}[a \leq X < b] = \mathbf{P}[a < X \leq b].$$

In order to qualify as a PDF, a function f_X must be nonnegative i.e., $f_X(x) \geq 0, \forall x$ and must also have the normalized property

$$\int f_X(x)\,dx = 1.$$

Examples of densities are the uniform, normal, Gamma, Cauchy, etc.

Since a density does not always exists, it would be desirable to describe all kinds of random variables with a single mathematical concept. This is accomplished with the **cumulative distribution function** or CDF. The CDF of a random variable X is denoted by F_X and provides the probability $\mathbf{P}[X \leq x]$. In particular for every x we have in the continuous case

$$F_X(x) = \mathbf{P}[X \leq x] = \int_{-\infty}^x f_X(x)\,dx$$

and for those x at which the PDF is continuous it follows that

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

The CDF for continuous random variables has the following properties

- F_X is monotonically nondecreasing: if $x \leq y$, then $F_X(x) \leq F_X(y)$
- $F_X(x)$ tends to 0 as $x \to -\infty$ and to 1 as $x \to \infty$
- $F_X(x)$ is a continuous function of x

1.3.2 Expectation of a continuous random variable and its properties

The **expected value** of a continuous random variable X is defined by

$$\mathbf{E}[X] = \int x f_X(x)\, dx.$$

One has to deal with the possibility that the integral $\int x f_X(x)\, dx$ is infinite or undefined. The expectation is well-defined if $\int |x| f_X(x)\, dx < \infty$. In that case, $\int x f_X(x)\, dx < \infty$. If X is a continuous random variable with given PDF, any real-valued function $Y = g(X)$ of X is also a random variable. Then,

$$\mathbf{E}[g(X)] = \int g(x) f_X(x)\, dx.$$

The rth **moment** of a continuous random variable X is defined as $\mathbf{E}[X^r]$, the expected value of the random variable X^r. The **variance**, denoted by $\mathbf{Var}[X]$, is defined as the expected value of the random variable $(X - \mathbf{E}[X])^2$. It follows immediately that the variance is also given by

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \int (x - \mathbf{E}[X])^2 f_X(x)\, dx.$$

Sometimes it can be more convenient to use the following formula for the variance

$$0 \leq \mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

If a and b are given scalars and $Y = aX + b$, then

$$\mathbf{E}[Y] = a\,\mathbf{E}[X] + b \qquad\qquad \mathbf{Var}[Y] = a^2\,\mathbf{Var}(X).$$

1.3.3 Expectation of a vector or matrix of random variables and its properties

First, consider the following definitions.

Definition 1.1 (Mean vector)

$$\mathbf{E}[\mathbf{Y}] = \begin{bmatrix} \mathbf{E}[Y_1] \\ \mathbf{E}[Y_2] \\ \vdots \\ \mathbf{E}[Y_n] \end{bmatrix}$$

Definition 1.2 (Covariance matrix) *The **covariance** matrix of $\mathbf{Y}$, denoted Σ, is an $n \times n$ matrix with the ijth element* $\mathbf{Cov}[Y_i, Y_j]$. *Σ is a symmetric matrix.*

If $\mathbf{Z} = \mathbf{c} + \mathbf{A}\mathbf{Y}$, where $\mathbf{Y}$ is a random vector and $\mathbf{A}$ is a fixed matrix and $\mathbf{c}$ is a fixed vector, then

$$\mathbf{E}[\mathbf{Z}] = \mathbf{c} + \mathbf{A}\,\mathbf{E}[\mathbf{Y}].$$

The ith component of $\mathbf{Z}$ is

$$Z_i = c_i + \sum_{j=1}^{n} a_{ij} Y_j$$

and hence

$$\mathbf{E}[Z_i] = c_i + \sum_{j=1}^{n} a_{ij}\, \mathbf{E}[Y_j].$$

Writing these equations in matrix form:

$$\begin{bmatrix} \mathbf{E}[Z_1] \\ \mathbf{E}[Z_2] \\ \vdots \\ \mathbf{E}[Z_n] \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{E}[Y_1] \\ \mathbf{E}[Y_2] \\ \vdots \\ \mathbf{E}[Y_n] \end{bmatrix}$$

Theorem 1.1 *If the covariance matrix of* $\mathbf{Y}$ *is* Σ_{YY}*, then the covariance matrix of* $\mathbf{Z} = \mathbf{c} + \mathbf{A}\mathbf{Y}$ *is*

$$\boxed{\Sigma_{ZZ} = \mathbf{A}\Sigma_{YY}\mathbf{A}^T}$$

Theorem 1.2 *Let* $\mathbf{X}$ *be a random* n *vector with mean* μ *and covariance* Σ *and let* $\mathbf{A}$ *be a fixed matrix. Then*

$$\boxed{\mathbf{E}[\mathbf{X}^T\,\mathbf{A}\,\mathbf{X}] = \mathrm{tr}(\mathbf{A}\Sigma) + \mu^T\,\mathbf{A}\,\mu}$$

Theorem 1.3 *Let* $\mathbf{X}$ *be a random vector with covariance matrix* Σ_{XX}*. If*

$$\mathbf{Y} = \underset{p\times n}{\mathbf{A}}\,\mathbf{X} \qquad \textit{and} \qquad \mathbf{Z} = \underset{m\times n}{\mathbf{B}}\,\mathbf{X}$$

where $\mathbf{A}$ *and* $\mathbf{B}$ *are fixed matrices. Then, the cross-covariance matrix of* $\mathbf{Y}$ *and* $\mathbf{Z}$ *is*

$$\Sigma_{YZ} = \mathbf{A}\Sigma_{XX}\mathbf{B}^T.$$

1.3.4 Describing the limiting behavior of a function

Big O

The big O notation describes the limiting behavior of a function when the argument tends towards a particular value or infinity, usually in terms of simpler functions. It is a member of a larger family of notations that is called Landau notation or Bachmann-Landau notation. Let f and g be two functions defined on some subset of the real numbers. One writes

$$f(x) = O(g(x)) \qquad \text{as } x \to \infty$$

if and only if there is a positive constant M such that for all sufficiently large values of x, $f(x)$ is at most M multiplied by the absolute value of $g(x)$. That is, $f(x) = O(g(x))$ if and only if there exists a positive real number M and a real number x_0 such that

$$|f(x)| \le M|g(x)| \qquad \text{for all } x \ge x_0.$$

Big O can also be used to describe the error term in an approximation to a mathematical function. The most significant terms are written explicitly, and then the least-significant terms are summarized in a single big O term. For example, in the exponential series

$$e^x = 1 + x + \frac{x^2}{2} + O(x^3) \qquad \text{as } x \to 0,$$

expresses the fact that the error, the difference $e^x - 1 - x - \frac{x^2}{2}$, is smaller in absolute value than some constant times $|x^3|$ when x is close enough to 0.

Little o

This means that $g(x)$ grows much faster than $f(x)$, or similarly, the growth of $f(x)$ is nothing compared to that of $g(x)$. Formally, $f(x) = o(g(x))$ as $x \to \infty$ means that for *every* positive constant ϵ there exists a constant N such that

$$|f(n)| \leq \epsilon |g(n)| \qquad \text{for all } n \geq N.$$

Note the difference between the earlier formal definition for the big-O notation, and the present definition of little-o: while the former has to be true for at least one constant M the latter must hold for every positive constant ϵ, however small. In this way little-o notation makes a stronger statement than the corresponding big-O notation: every function that is little-o of g is also big-O of g, but not every function that is big-O g is also little-o of g. If $g(x)$ is nonzero, or at least becomes nonzero beyond a certain point, the relation $f(x) = o(g(x))$ is equivalent to

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0.$$

Consider the following three examples:

$$\begin{aligned} 2x &= o(x^2) \\ 2x^2 &\neq o(x^2) \\ \frac{1}{x} &= o(1). \end{aligned}$$

Big O_p

The notation

$$X_n = O_p(a_n),$$

means that the set of values X_n/a_n is stochastically bounded. That is, for any $\epsilon > 0$, there exists a finite $M > 0$ such that

$$\mathbf{P}[|X_n/a_n| \geq M] < \epsilon, \forall n.$$

Little o_p

It is also convenient to have short expressions for terms that converge in probability to zero. The notation $o_p(1)$ is short for a sequence of random vectors that converges to zero in probability. Further,

$$X_n = o_p(a_n)$$

means that the set of values X_n/a_n converges to zero in probability as n approaches an appropriate limit. Equivalently, $X_n = o_p(a_n)$ can be written as $X_n/a_n = o_p(1)$, where $X_n = o_p(1)$ is defined as

$$\lim_{n \to \infty} P(|X_n| \geq \epsilon) = 0,$$

for every positive ϵ.

There are many rules of calculus with o and O symbols. For instance, for a random sequence R_n

$$\begin{aligned} o_p(1) + o_p(1) &= o_p(1) \\ o_p(1) + O_p(1) &= O_p(1) \\ O_p(1)o_p(1) &= o_p(1) \\ (1 + o_p(1))^{-1} &= O_p(1) \\ o_p(R_n) &= R_n o_p(1) \\ O_p(R_n) &= R_n O_p(1) \\ o_p(O_p(1)) &= o_p(1). \end{aligned}$$

The previous rules are also valid for nonrandom sequences. In that case, the subscript p can be omitted. See van der Vaart (1998) for more on asymptotic statistics.

Chapter 2

Density estimation

2.1 Error criteria for density estimates

2.1.1 MSE, MISE and some other criteria

In classical statistics it is common to measure the closeness of an estimator $\hat{\theta}$ to its target parameter θ by the size of the **mean squared error** (MSE)

$$\text{MSE}(\hat{\theta}) = \mathbf{E}[\hat{\theta} - \theta]^2.$$

An appealing feature of MSE is its simple decomposition into variance and squared bias

$$\begin{aligned}
\mathbf{E}[\hat{\theta} - \theta]^2 &= \mathbf{E}\left[\hat{\theta} - \mathbf{E}[\hat{\theta}] + \mathbf{E}[\hat{\theta}] - \theta\right]^2 \\
&= \mathbf{E}\left[\hat{\theta}^2 - 2\hat{\theta}\,\mathbf{E}[\hat{\theta}] + (\mathbf{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbf{E}[\hat{\theta}])(\mathbf{E}[\hat{\theta}] - \theta) + (\mathbf{E}[\hat{\theta}] - \theta)^2\right] \\
&= \mathbf{E}[\hat{\theta}^2] - (\mathbf{E}[\hat{\theta}])^2 + (\mathbf{E}[\hat{\theta}] - \theta)^2 \\
&= \mathbf{Var}[\hat{\theta}] + \text{bias}^2[\hat{\theta}].
\end{aligned}$$

For pointwise estimation of a density function by an estimator $\hat{f}_X(x)$ this results into looking at

$$\begin{aligned}
\text{MSE}(\hat{f}_X(x)) &= \mathbf{E}[\hat{f}_X(x) - f_X(x)]^2 \\
&= \mathbf{Var}[\hat{f}_X(x)] + \text{bias}^2[\hat{f}_X(x)],
\end{aligned}$$

with $\text{bias}[\hat{f}_X(x)] = \mathbf{E}[\hat{f}_X(x)] - f_X(x)$. This equation treats the nonparametric density estimation problem as a standard point estimation problem with unknown parameter $\theta = f_X(x)$. The criterion of *local performance* is often preferred to other criteria for local performance such as the **mean absolute deviation** (MAD)

$$\text{MAD}(\hat{\theta}) = \mathbf{E}[|\hat{\theta} - \theta|] \leq \sqrt{\mathbf{Var}[\hat{\theta}]} + |\,\text{bias}[\hat{\theta}]|,$$

since it is mathematically simpler to work with. The bias-variance decomposition allows easier analysis and interpretation of the estimators as we will see later. Figure 2.1 shows a graphical interpretation regarding bias and variance. Can you determine which subfigure represents low/high bias and low/high variance?

To evaluate the *global performance* of a density estimate, the most intuitively appealing global criterion is the L_∞ norm

$$\sup_x |\hat{f}_X(x) - f_X(x)|.$$

At the other end of the spectrum is the L_1 norm

$$\int |\hat{f}_X(x) - f_X(x)|\,dx.$$

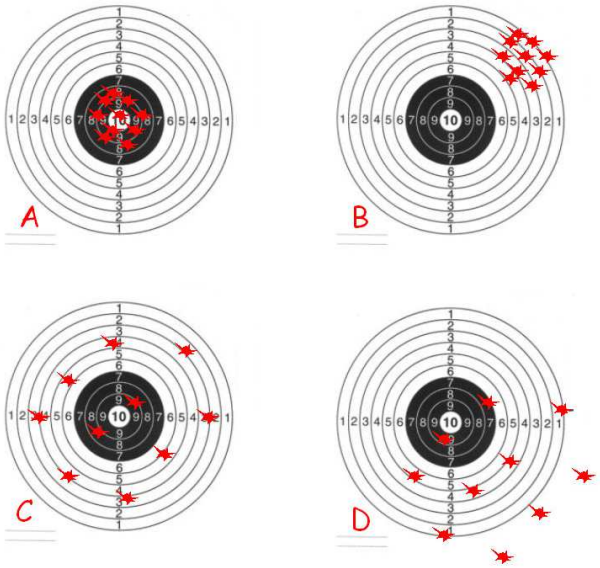


Figure 2.1: Graphical interpretation of bias and variance.

The L_1 nor the L_∞ criterion is as easily manipulated as the L_2 norm, which is referred to as **integrated squared error** (ISE)

$$\text{ISE}(\hat{f}_X) = \int [\hat{f}_X(x) - f_X(x)]^2 \, dx,$$

which is a random variable, depending on the particular realization $X_1, \ldots, X_n$. For most purposes, it will be sufficient to examine the average of the ISE over these realizations i.e, the mean of the random variable or **mean integrated squared error** (MISE)

$$\begin{aligned} \text{MISE}(\hat{f}_X) = \mathbf{E}[\text{ISE}(\hat{f}_X)] &= \mathbf{E} \int [\hat{f}_X(x) - f_X(x)]^2 \, dx \\ &\overset{\text{Fubini}}{=} \int \mathbf{E}[\hat{f}_X(x) - f_X(x)]^2 \, dx \\ &= \int \text{MSE}(\hat{f}_X(x)) \, dx \\ &= \text{IMSE}(\hat{f}_X). \end{aligned}$$

The last quantity is called the **integrated mean squared error** (IMSE). Thus, the MISE error criterion has two different though equivalent interpretations: it is a measure of both the average global error and the accumulated pointwise error. It is also possible to include a weight function, that would emphasize the tails for example or a local interval. Another measure for global performance of $\hat{f}_X$ is the **mean integrated absolute deviation** (MIAD)

$$\text{MIAD}(\hat{f}_X) = \mathbf{E} \int |\hat{f}_X(x) - f_X(x)| \, dx.$$

Other possible candidates to measure global error include

- L_p distance

$$L_p(\hat{f}_X, f_X) = \begin{cases} \left[\int |\hat{f}_X(x) - f_X(x)|^p \, dx \right]^{1/p} & 0 < p < \infty \\ \sup_x |\hat{f}_X(x) - f_X(x)| & p = \infty \end{cases}$$

- Hellinger distance

$$H_p(\hat{f}_X, f_X) = \left[\int \left(\hat{f}_X(x)^{1/p} - f_X(x)^{1/p} \right)^p dx \right]^{1/p} \qquad p > 0$$

- Kullback-Leibler information number or divergence (not a distance)

$$K(\hat{f}_X,f_X) = \int \hat{f}_X(x) \log\left(\frac{\hat{f}_X(x)}{f_X(x)}\right) dx.$$

- Total variation

$$TV(\hat{f}_X,f_X) = \sup_A \left| \int_A \hat{f}_X(x)\,dx - \int_A f_X(x)\,dx \right|$$

2.1.2 Some remarks about L_1 and L_2 criteria

- L_1 vs. L_2
  - The L_1 criterion $\int |\hat{f}_X(x) - f_X(x)|\,dx$ puts more emphasis on the tails of a density than the L_2 criterion. The latter de-emphasizes the relatively small density values by squaring.
  - Note that

$$\int |\hat{f}_X(x) - f_X(x)|\,dx \leq \int |\hat{f}_X(x)|\,dx + \int |f_X(x)|\,dx \leq 2$$

    if $\hat{f}_X$ is a density. Hence, it follows that

$$0 \leq \int |\hat{f}_X(x) - f_X(x)|\,dx \leq 2.$$

    For the L_2 criterion, we have that

$$0 \leq \int [\hat{f}_X(x) - f_X(x)]^2\,dx \leq +\infty.$$

  - In practical situations, the estimators that optimize these criteria are similar.
  - The analytical simplicity of squared error and its adequacy in practical applications makes the L_2 criterion often the criterion of choice.
- L_1 error is invariant under any smooth monotone transformation
- Scheffé's lemma (Scheffé, 1947; Devroye and Györfi, 1985). For all densities f and g on $\mathbb{R}^d$

$$\boxed{\int |f(x) - g(x)|\,dx = 2TV(f,g) = 2\int \max(f(x) - g(x),0)\,dx = 2\int \max(g(x) - f(x),0)\,dx}$$

- The result in Scheffé's lemma provides a connection with statistical classification.

2.2 Histogram estimation of a density

2.2.1 The histogram estimator

In statistics, a histogram is a graphical representation of the distribution of data (see Figure 2.18a). It is an estimate of the probability distribution of a continuous variable and was first introduced by Pearson (1895). A histogram is a representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area proportional to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1. The categories are usually specified as consecutive, non-overlapping intervals of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the same size. The rectangles of a histogram are drawn so that they touch each other to indicate that the original variable is continuous.

Choose an origin t_0 and a bin width $h > 0$, where the bin width is the width of the classes (i.e. bins). The kth bin is given by

$$B_k = [t_k, t_{k+1}[, \quad k \in \mathbb{Z}$$

with

$$t_{k+1} = t_k + h, \quad k \in \mathbb{Z}.$$

Denote by ν_k the *bin count* of the kth bin i.e., the number of sample points falling in the bin B_k. The histogram estimator is defined as

$$\boxed{\hat{f}_X(x) = \frac{\nu_k}{nh} = \frac{1}{nh} \sum_{i=1}^{n} I_{[t_k, t_{k+1}]}(X_i) \qquad \text{for } x \in B_k} \tag{2.1}$$

The histogram estimator is a very elementary estimator, but it can give the first good idea about the underlying unknown density function. But if one wants to work further with the density estimate (discriminant analysis, hazard function estimation, ...) then a more accurate estimator is needed. The histogram is a discontinuous function (a step function), and hence the density is estimated with a step function. However, there are two unknown quantities in (2.1) i.e.,

- the bin width h
- the origin t_0 (position of the edges of the bins).

Both choices can have a significant effect on the resulting histogram. These effects are illustrated by the following figures.

Example 2.1 (Buffalo snowfall data for different bin widths) *Annual snowfall in Buffalo (NY) from 1910 to 1972 in inches for different bin widths. The data is available in the R package "gss". The bin widths are (from left to right) 30, 15, 10, 7.5, 6 and 5 respectively.*

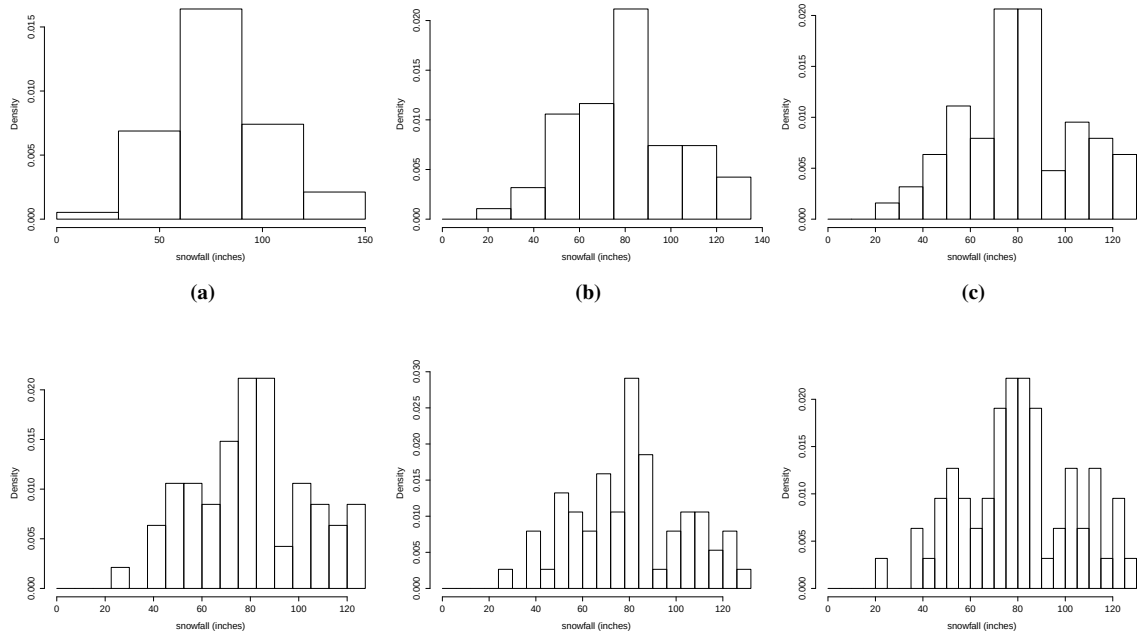


Figure 2.2: Buffalo snowfall data with bin origin $t_0 = 0$ for bin widths (a) 30, (b) 15, (c) 10, (d) 7.5, (e) 6 and (f) 5 inches respectively.

Example 2.2 (Buffalo snow data for different origins t_0) *Annual snowfall in Buffalo (NY) from 1910 to 1972 in inches for different bin widths. The data is available in the R package "gss". All bins have width 13.5 inches but different origins t_0 i.e, (from left to right): 0, 2.5, 5, 7.5, 10 and 12.5 respectively.*

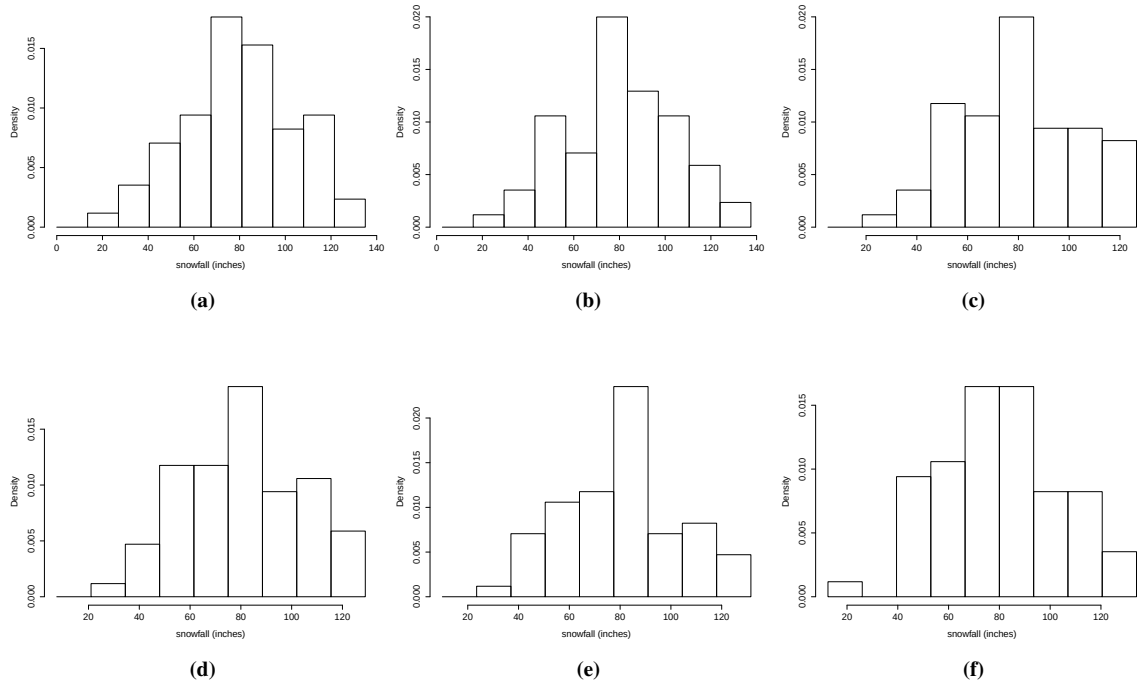


Figure 2.3: Buffalo snowfall data with different bin origins t_0. (a) 0, (b) 2.5, (c) 5, (d) 7.5, (e) 10 and (f) 12.5 respectively and bin width of 13.5 inches.

A possible approach to overcome the problem of how to choose the origin is to work with averaged shifted histograms (Scott, 1992) or kernel density estimation (see Section 2.3).

2.2.2 Asymptotic analysis of the histogram estimator

The analysis of the histogram random variable $\widehat{f}_X$ is quite simple once one recognizes that the bin counts are Binomial random variables. For the bin count of the kth bin

$$\nu_k \sim \text{Bin}(n,p_k) \qquad \text{where} \qquad p_k = \int_{B_k} f(t)\,dt.$$

Hence, we have

$$\mathbf{E}[\nu_k] = n \cdot p_k \qquad \text{and} \qquad \mathbf{Var}[\nu_k] = n \cdot p_k \cdot (1-p_k)$$

and therefore, for $x \in B_k$ (see also Figure 2.4)

$$\mathbf{E}[\widehat{f}_X(x)] = \frac{p_k}{h} \qquad \text{and} \qquad \mathbf{Var}[\widehat{f}_X(x)] = \frac{p_k \cdot (1-p_k)}{nh^2}.$$

The exact bias of the histogram estimator is

$$\mathbf{E}[\widehat{f}_X(x)] - f_X(x) = \frac{p_k}{h} - f_X(x).$$

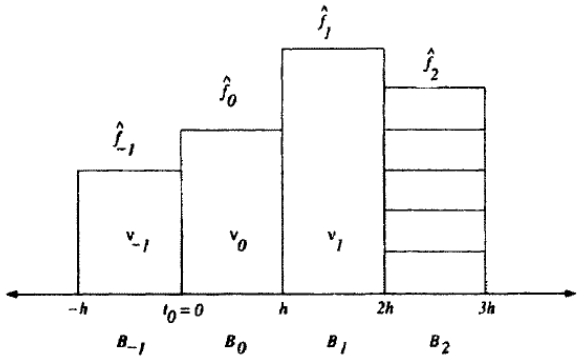


Figure 2.4: Construction of an equally-spaced histogram. Taken from Scott (1992).

In order to continue and without loss of generality consider a typical bin $B_0 = [0,h)$. Then the probability content of that bin, p_0 is given by

$$\begin{aligned} p_0 &= \int_0^h f_X(t)\,dt \\ &= \int_0^h \left[f_X(x) + (t-x) f'_X(x) + \frac{(t-x)^2}{2} f''_X(x) + \ldots \right] dt \\ &= h f_X(x) + h\left(\frac{h}{2} - x\right) f'_X(x) + O(h^3) \end{aligned}$$

so that

$$\text{bias}[\widehat{f}_X(x)] = \frac{h f_X(x) + h\left(\frac{h}{2} - x\right) f'_X(x) + O(h^3)}{h} - f_X(x) = \left(\frac{h}{2} - x\right) f'_X(x) + O(h^2)$$

implying that the bias is of order $O(h^2)$ at the center of the bin, when $x = h/2$ (see Figure 2.5).

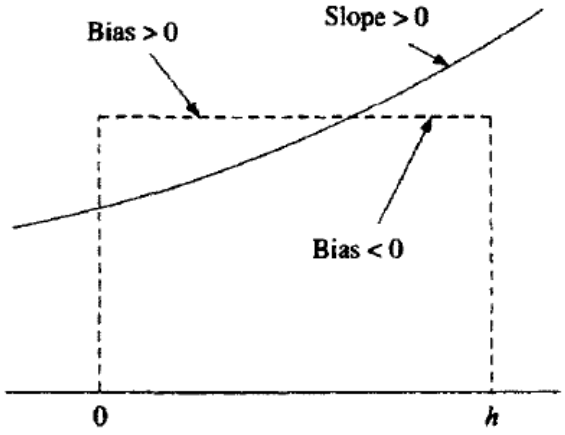


Figure 2.5: Bias of the histogram estimator in a typical bin. Taken from Scott (1992).

Using the generalized mean value theorem (or sometimes called the second mean value theorem for integrals), the leading term of the integrated squared bias for this bin is

$$\int_{B_0} \left(\left(\frac{h}{2} - x \right)^2 f'^2_X(x) + O(h^3) \right) dx = f'^2_X(\eta_0) \int_{B_0} \left(\left(\frac{h}{2} - x \right)^2 + O(h^3) \right) dx = \frac{h^3}{12} f'^2_X(\eta_0) + O(h^4), \qquad (2.2)$$

for some $\eta_0 \in B_0$. This results generalizes to other bins for some collection of points $\eta_k \in B_k$. The previous result holds for 1 bin. To obtain the total integrated squared bias (TISB) for all bins we simply sum up all contributions i.e., (2.2), over all bins, yielding

$$\begin{aligned}
\text{TISB}[\widehat{f}_X] &= \sum_{k=-\infty}^{+\infty} \frac{h^3}{12} f_X'^2(\eta_k) + O(h^4) \\
&= \frac{h^2}{12} \sum_{k=-\infty}^{+\infty} \left[h f_X'^2(\eta_k) + O(h^2) \right] \\
&= \frac{h^2}{12} \int_{-\infty}^{+\infty} f_X'^2(x)\, dx + o(h^2)
\end{aligned}$$

where the latter follows from standard Riemannian convergence of sums to integrals. However, if we assume that $f_X'^2$ has bounded total variation the remainder term becomes $O(h^3)$. As for the integrated variance for a typical bin, we have

$$\int_{B_0} \mathbf{Var}[\widehat{f}_X(x)]\, dx = \frac{p_0 \cdot (1-p_0)}{nh},$$

then summing over all bins gives the total integrated variance (TIV) and using $p_k = \int_{B_k} f_X(t)\, dt = h f_X(\xi_k)$ for some $\xi_k \in B_k$

$$\begin{aligned}
\text{TIV}[\widehat{f}_X] &= \sum_{k=-\infty}^{+\infty} \frac{p_k \cdot (1-p_k)}{nh} \\
&= \frac{1}{nh} - \frac{\int f_X^2(x)\, dx}{n} + o\left(\frac{1}{n}\right) \\
&= \frac{1}{nh} + O\left(\frac{1}{n}\right).
\end{aligned}$$

Finally, the MISE of the histogram estimator is given by

$$\text{MISE}(\hat{f}_X) = \frac{1}{nh} + \frac{1}{12} h^2 \int f_X'^2(x)\, dx + O\left(\frac{1}{n} + h^3\right).$$

2.2.3 Choice of the bin width

In the previous section we have seen that the MISE of the histogram estimator is given by (see also (Scott, 1992))

$$\text{MISE}(\hat{f}_X) = \frac{1}{nh} + \frac{1}{12} h^2 \int f_X'^2(x)\, dx + O\left(\frac{1}{n} + h^3\right), \tag{2.3}$$

for $h = h_n \to 0$ and a square integrable first order derivative density function f_X'. A way to choose the bin width $h = h_n$ is to minimize the MISE. The subscript n in the bin width is used to indicate that the bin width depends on the sample size n. Denote the optimal bin width by $h_{n,\text{MISE}}$. Then, an approximation for $h_{n,\text{MISE}}$ is obtained by minimizing the asymptotic expression for $\text{MISE}(\hat{f}_X)$ given in (2.3):

$$\begin{aligned}
h_{n,\text{AMISE}} &= \arg\min_h \text{AMISE}(\hat{f}_X) \\
&= \arg\min_h \left[\frac{1}{nh_n} + \frac{1}{12} h_n^2 \int f_X'^2(x)\, dx \right] \\
&= \left[\frac{6}{\int f_X'^2(x)\, dx} \right]^{1/3} n^{-1/3}.
\end{aligned} \tag{2.4}$$

With the above optimal choice, the minimum value of AMISE($\hat{f}_X$) becomes

$$\frac{3}{2}\frac{1}{\sqrt[3]{6}}\left[\int f'_X(x)\,dx\right]^{1/3} n^{-2/3}$$

and hence the L_2 rate of convergence of a histogram estimator is $n^{-2/3}$ or for the optimal bin width

$$\inf_{h>0} \text{MISE}(\hat{f}_X) \sim \frac{3}{2}\frac{1}{\sqrt[3]{6}}\left[\int f'^2_X(x)\,dx\right]^{1/3} n^{-2/3}.$$

However, (2.4) is not useful in practice since it depends on the true unknown density f_X. A quick and simple bin width selection rule is obtained by referring to a normal density. If $f_X = N(\mu,\sigma^2)$, then

$$\int f'^2_X(x)\,dx = \frac{1}{4\sqrt{\pi}\sigma^3}$$

and consequently

$$\hat{h}_{n,\text{AMISE}} = \left[\frac{24\sqrt{\pi}\hat{\sigma}^3}{n}\right]^{1/3} \approx 3.5\hat{\sigma}n^{-1/3},$$

where $\hat{\sigma}$ is a consistent estimator of σ e.g., $\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2}$ with $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. This bin width selector is the so-called rule-of-thumb bin width selector. More sophisticated choices can found in Scott and Terrell (1987), Sheather and Jones (1991) and Wand and Jones (1995).

2.2.4 Boundary discontinuities in the density

The approximation properties of a histogram are not affected by a simple jump in the density, if the jump occurs at the boundary of a histogram. Discontinuities can adversely affect all density estimators. This effect can be demonstrated by the following example. Consider $f_X(x) = e^{-x}, x \geq 0$ and mesh $t_k = kh$ for integer $k \geq 0$. Then the MISE (2.3) holds on the interval $(0, +\infty)$ on which $R(f'_X) = \int f'^2_X(x)\,dx = 1/2$. Hence,

$$\hat{h}_{n,\text{AMISE}} = (12/n)^{1/3} \qquad \text{and} \qquad \text{AMISE}(\widehat{f}_X) = 0.6552n^{-2/3}.$$

Suppose the discontinuity at zero was not know a priori and the mesh $t_k = (k - \frac{1}{2})h$ was chosen. Then the focus is on the bias in bin $B_0 = [-h/2,h/2)$ where the density is discontinuous (see Figure 2.6).

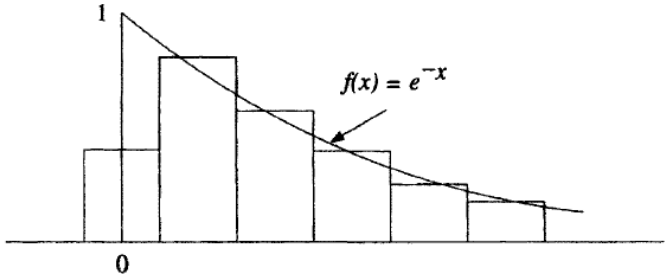


Figure 2.6: Illustration of discontinuity boundary bin problem. Taken from Scott (1992).

The probability mass in bin $B_0 = (-h/2,h/2]$ is

$$p_0 = \int_0^{h/2} e^{-x}\,dx = 1 - e^{-h/2}.$$

Now, $\mathbf{E}\,\hat{f}_X(x) = p_0/h$, $e^{-x} = 1 - x + x^2/2 + O(x^3)$, $x \to 0$ and therefore

$$\begin{aligned}\int_{-h/2}^{h/2} \text{bias}^2(\widehat{f}_X(x))\,dx &= \int_{-h/2}^{0}\left(\frac{p_0}{h} - 0\right)^2 dx + \int_0^{h/2}\left(\frac{p_0}{h} - e^{-x}\right)^2 dx \\ &= \frac{1-e^{-h}}{2} + \frac{2e^{-h/2} - e^{-h} - 1}{h} \\ &\overset{\text{Taylor series}}{=} \frac{h}{4} - \frac{h^2}{8} + O(h^3).\end{aligned}$$

Over the interval $(-h/2, +\infty)$, the total integrated squared bias is

$$\text{TISB}[\widehat{f}_X] = \frac{h}{4} - \frac{h^2}{8} + O(h^3) + \frac{h^2}{12}\int_{h/2}^{+\infty} f_X'^2(x)\,dx + o(h^2).$$

The worst outcome has been realized; the TISB is entirely dominated by the contribution from the bin containing the discontinuity at $x = 0$. Indeed, the order of the TISB is $O(h)$ instead of $O(h^2)$. Further, we have that (the variance is unchanged)

$$\text{AMISE}(\widehat{f}_X) = \frac{1}{nh} + \frac{h}{4} \Rightarrow \hat{h}_{n,\text{AMISE}} = \frac{2}{\sqrt{n}} \quad \text{and} \quad \text{AMISE}(\widehat{f}_X) = O(n^{-1/2}),$$

which is significantly worse than $O(n^{-2/3})$. This rate is a s slow as for bivariate data! The histogram tries to accommodate the discontinuity by choosing a narrower bin width. Figure 2.7 illustrates the potential impact on AMISE.

	$X > 0$ Known		$X > 0$ Unknown		
n	h^*	AMISE*	h^*	AMISE*	Error Ratio
10	1.063	0.14116	0.3162	0.31623	2.24
100	0.493	0.03041	0.1	0.1	3.29
1,000	0.229	0.00655	0.0316	0.03162	4.83
10,000	0.106	0.00141	0.01	0.01	7.09
100,000	0.049	0.00030	0.0032	0.00316	10.54

Figure 2.7: Potential impact on AMISE of lack of knowledge of boundary discontinuities. Taken from Scott (1992).

2.3 Kernel density estimation

2.3.1 The naive or simple estimator

For those x at which the PDF is continuous we know that $f_X(x) = \frac{dF_X(x)}{dx}$. From the definition of the derivative, it follows that

$$\begin{aligned}f_X(x) &= \lim_{h\to 0}\frac{F_X(x+h) - F_X(x-h)}{2h} \\ &= \lim_{h\to 0}\frac{\mathbf{P}[X \le x+h] - \mathbf{P}[X \le x-h]}{2h} \\ &= \lim_{h\to 0}\frac{1}{2h}\mathbf{P}[x-h \le X \le x+h].\end{aligned} \tag{2.5}$$

An estimator for f_X can be obtained as follows

$$\begin{aligned}\hat{f}_X(x) &= \frac{1}{2h}\frac{\#\{i : x-h \le X_i \le x+h\}}{n} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} I\{x-h \le X_i \le x+h\} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} I\left\{-1 \le \frac{x - X_i}{h} \le 1\right\},\end{aligned}$$

where $I(A)$ denotes the indicator function for the event A i.e., if A is true, $I(A) = 1$ and 0 otherwise. The latter can be written as

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

with

$$w(u) = \begin{cases} 1/2 & \text{if } u \in [-1,1] \\ 0 & \text{elsewhere.} \end{cases}$$

The estimator (2.8) has the following properties

- no problem anymore with the choice of the origin as in the case of an histogram
- $\hat{f}_X(x) = \frac{1}{2nh}\sum_{i=1}^{n} I\{x - h \leq X_i \leq x + h\} = \frac{1}{2nh}\sum_{i=1}^{n} I\{X_i - h \leq x \leq X_i + h\}$. Therefore $\hat{f}_X(\cdot)$ is an estimator which is discontinuous at $X_i \pm h$ and is constant between those values.

Figure 2.8 shows the estimator (bold line) and the "boxes" (thin lines) over each observation. The R code is given below.

```
> x <- c(0, 1.5, 1.7, 1.9)
> n <- length(x)
> h <- 0.3
> xgrid <- seq(from = min(x) - 1, to = max(x) + 1, by = 0.001)
> bumps <- sapply(x, function(a) (1/(n*h))*(abs((xgrid-a)/h) <= 1) * 0.5)
> plot(xgrid, rowSums(bumps), ylab = expression(hat(f)(x)),type = "l", xlab = "x", lwd = 3)
> rug(x, lwd = 2)
> out <- apply(bumps, 2, function(b) lines(xgrid, b))
```

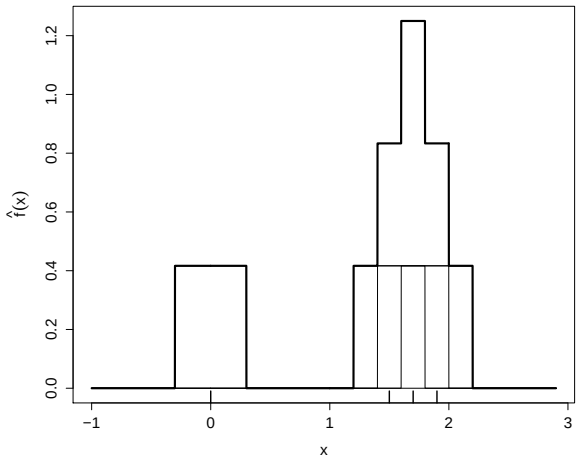


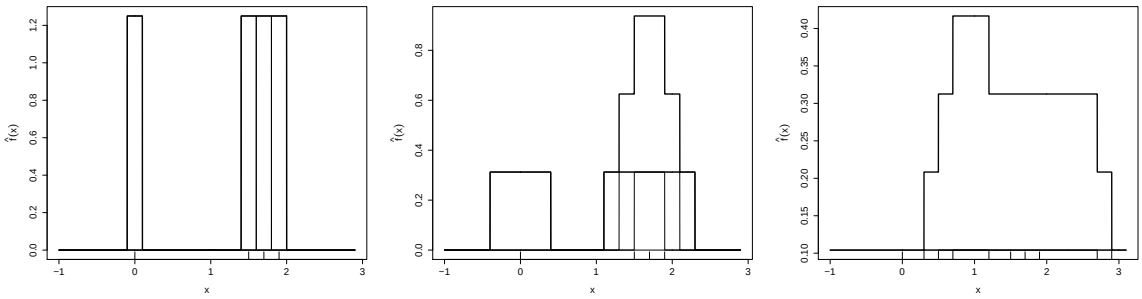
Figure 2.8: Construction of the naive density estimator with $h = 0.3$

The effect of the bandwidth h is shown in Figure 2.9. The larger the bandwidth h the smoother the fit and details are starting to disappear. The smaller the bandwidth the more spurious effects are visible and the estimate becomes more wiggly. In the limit for $h \to 0$, we have a sum of delta Dirac functions at the observations.

2.3.2 Kernel density estimation

In order the improve the estimate, one can replace the function w by a more general weight function K to obtain the kernel density estimator proposed by Rosenblatt (1956) and Parzen (1962)

$$\boxed{\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)} \tag{2.6}$$

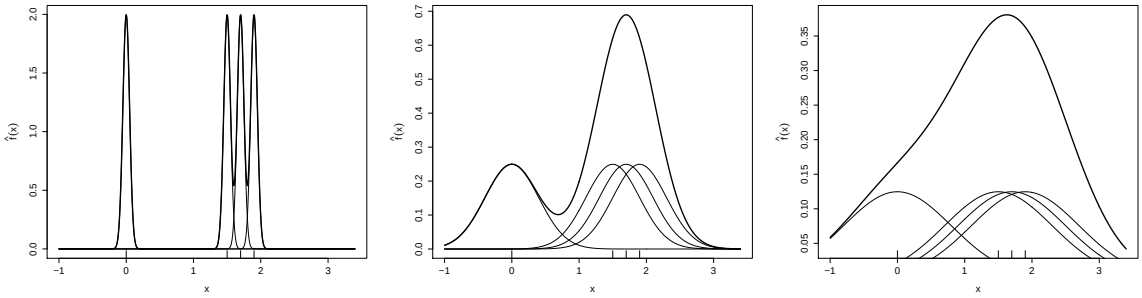


(a) $h = 0.1$ **(b)** $h = 0.4$ **(c)** $h = 1.2$

Figure 2.9: Effect of the bin width on the estimate.

where K is called the kernel or weight function. Often K is taken to be a symmetric probability density. The kernel K determines the form of the bumps and h determines their width. The mass $\frac{1}{n}$ at each data point is smoothly redistributed to its vicinity. Addition of the redistributed masses leads to the final kernel density estimate (2.6). Figure 2.10 shows the estimator (bold line) and the kernel function K (thin lines) over each observation. In this example K was taken to be $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ i.e., the Gaussian kernel. The bandwidth h has a similar effect on the kernel density estimate as the bin width on the histogram estimate. The R code is given below.

```
> x <- c(0, 1.5, 1.7, 1.9)
> n <- length(x)
> h <- 0.4
> xgrid <- seq(from = min(x) - 1, to = max(x) + 1.5, by = 0.001)
> # Gaussian kernel
> gauss <- function(x) 1/sqrt(2*pi) * exp(-(x^2)/2)
> bumps <- sapply(x, function(a) gauss((xgrid - a)/h)/(n * h))
> plot(xgrid, rowSums(bumps), ylab = expression(hat(f)(x)),type = "l", xlab = "x", lwd = 3)
> rug(x, lwd = 2)
> out<-apply(bumps, 2, function(b) lines(xgrid, b))
```



(a) $h = 0.05$ **(b)** $h = 0.4$ **(c)** $h = 0.8$

Figure 2.10: Effect of the bandwidth on the estimate.

We conclude with some remarks regarding this estimator.

- If K is a probability density function, then $\hat{f}_X$ is also a probability density (prove this)
- $\hat{f}_X$ will inherit all the continuity and differentiability properties of the kernel K
- If K may take negative values so does $\hat{f}_X$
- There is need for a bandwidth which adapts to location and the data

- What about the theoretical properties?

There are more ways to estimate a density e.g., nearest neighbor methods, orthogonal series, wavelets, polygons, maximum penalized likelihood estimation, local likelihood estimation etc. See Silverman (1986) and Scott (1992) for these other density estimation methods. In our example we have used the Gaussian kernel. Of course, many other kernels exists. Figure 2.11 illustrates a variety of different kernels. All these kernels are densities. For more information regarding kernel types: `http://en.wikipedia.org/wiki/Kernel_(statistics)`. It can be shown that the Epanechnikov (or Bartlett) kernel is the optimal kernel in L_2 sense (blue line in Figure 2.11) and possibly also in L_1 sense (Devroye and Györfi, 1985).

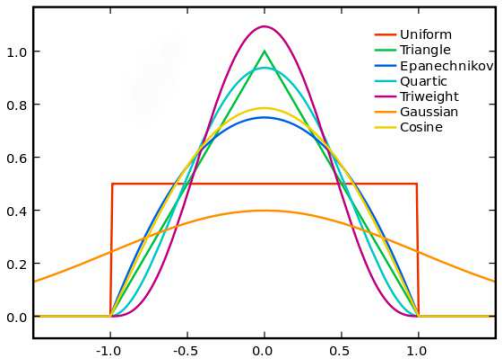


Figure 2.11: Different kernel types.

2.3.3 Theoretical analysis for kernel density estimation

In this paragraph we will study the asymptotic approximations for MSE and MISE in case of the kernel density estimate (2.6). Although possible to obtain exact expression for MSE and MISE, they are not attractable since they depend on the bandwidth in a fairly complicated way. This is the reason why we will only focus on the asymptotic expressions which depend on the bandwidth in a much simpler way.

In order to have a simpler notation, consider the following alternative way of writing (2.6)

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

where

$$K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right).$$

Asymptotic approximations for the bias and variance can now be constructed (under necessary conditions on f_X and K). Let $X_1,\ldots,X_n$ be an independent and identically distributed (i.i.d.) sample from f_X. Then by using a Taylor expansion it follows that

$$\begin{aligned}
\mathbf{E}[\hat{f}_X(x)] &= \mathbf{E}[K_h(x-X)] \\
&= \int K_h(x-y)f_X(y)\,dy \\
&= \int K(u)f_X(x-uh)\,du \\
&= \int K(u)\left[f_X(x) - f_X'(x)uh + \frac{1}{2}f_X''(x)u^2h^2 + o\big(h^2\big)\right]du \\
&= f_X(x)\int K(u)\,du - f_X'(x)h\int uK(u)\,du + \frac{1}{2}f_X''(x)h^2\int u^2K(u)\,du + o\big(h^2\big).
\end{aligned}$$

If

$$K \geq 0 \qquad \int K(u)\,du = 1 \qquad \int uK(u)\,du = 0 \qquad 0 < \int u^2 K(u)\,du < \infty, \tag{2.7}$$

and the density f_X is twice continuously differentiable, then

$$\text{bias}[\hat{f}_X(x)] = \mathbf{E}[\hat{f}_X(x)] - f_X(x) = \frac{1}{2} f_X''(x) h^2 \int u^2 K(u)\,du + o\big(h^2\big).$$

Next,

$$\begin{aligned}
\mathbf{Var}[\hat{f}_X(x)] &= \frac{1}{n}\left\{\mathbf{E}[K_h^2(x-X)] - (\mathbf{E}[K_h(x-X)])^2\right\} \\
&= \frac{1}{n}\left\{\frac{1}{h^2}\int K^2\left(\frac{x-y}{h}\right) f_X(y)\,dy - (\mathbf{E}[\hat{f}_X(x)])^2\right\} \\
&= \frac{1}{nh}\int K^2(u)\left(f_X(x) + o\big(1\big)\right)\,du - \frac{1}{n}(f_X(x) + o\big(1\big))^2 \\
&= \frac{1}{nh} f_X(x)\int K^2(u)\,du + o\left(\frac{1}{nh}\right).
\end{aligned}$$

To summarize, we have that

$$\text{bias}[\hat{f}_X(x)] = \frac{1}{2} f_X''(x) h^2 \mu_2 + o\big(h^2\big), \qquad \mu_2 = \int u^2 K(u)\,du$$

and

$$\mathbf{Var}[\hat{f}_X(x)] = \frac{1}{nh} f_X(x) R(K) + o\left(\frac{1}{nh}\right), \qquad R(K) = \int K^2(u)\,du.$$

Finally, one can state the following important result:

If $h = h_n \to 0$ as $n \to \infty$, then bias$[\hat{f}_X(x)] \to 0$ as $n \to \infty$
If $nh_n \to \infty$ as $n \to \infty$, then $\mathbf{Var}[\hat{f}_X(x)] \to 0$ as $n \to \infty$

The role of the bandwidth is clear: it balances bias and variance

- if $h \searrow$ then $(\text{bias})^2 \searrow$ and variance $\nearrow$
- if $h \nearrow$ then $(\text{bias})^2 \nearrow$ and variance $\searrow$

Recall that these expressions where derived under the assumptions (2.7) and that the underlying unknown density f_X is twice continuously differentiable. From Section 2.1, we can now establish the MSE and MISE of the kernel density estimator. It follows that

$$\begin{aligned}
\text{MSE}(\hat{f}_X(x)) &= \frac{1}{4} h^4 \mu_2^2 (f_X''(x))^2 + \frac{1}{nh} f(x) R(K) + o\left(h^4 + \frac{1}{nh}\right) \\
\text{MISE}(\hat{f}_X) &= \frac{1}{4} h^4 \mu_2^2 \int (f_X''(x))^2\,dx + \frac{1}{nh} R(K) + o\left(h^4 + \frac{1}{nh}\right),
\end{aligned}$$

under the assumption that the second derivative of the unknown density is square integrable. Further, denote the asymptotic MSE and MISE by

$$\text{AMSE}(\hat{f}_X(x)) = \frac{1}{4} h^4 \mu_2^2 (f_X''(x))^2 + \frac{1}{nh} f(x) R(K) \tag{2.8}$$

$$\text{AMISE}(\hat{f}_X) = \frac{1}{4} h^4 \mu_2^2 \int (f_X''(x))^2\,dx + \frac{1}{nh} R(K). \tag{2.9}$$

h	constant or global bandwidth
$h(x)$	local variable bandwidth
$h(X_i)$	global variable bandwith

2.3.4 Theoretical optimal bandwidth choices

For the bandwidth h we can distinguish between leading to the following estimators

$$\begin{aligned}
\widehat{f}_X(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \\
\widehat{f}_{X,LV}(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(x)}K\left(\frac{x-X_i}{h(x)}\right) \\
\widehat{f}_{X,GV}(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(X_i)}K\left(\frac{x-X_i}{h(X_i)}\right).
\end{aligned}$$

With the choice $h(x)$, a different basic kernel estimator is used at each point (see Figure 2.12a). The idea behind $\widehat{f}_{X,GV}$ is that the kernel centered on the observation X_i has its own scale parameter $h(X_i)$. The interpretation and construction are shown in Figure 2.12b.

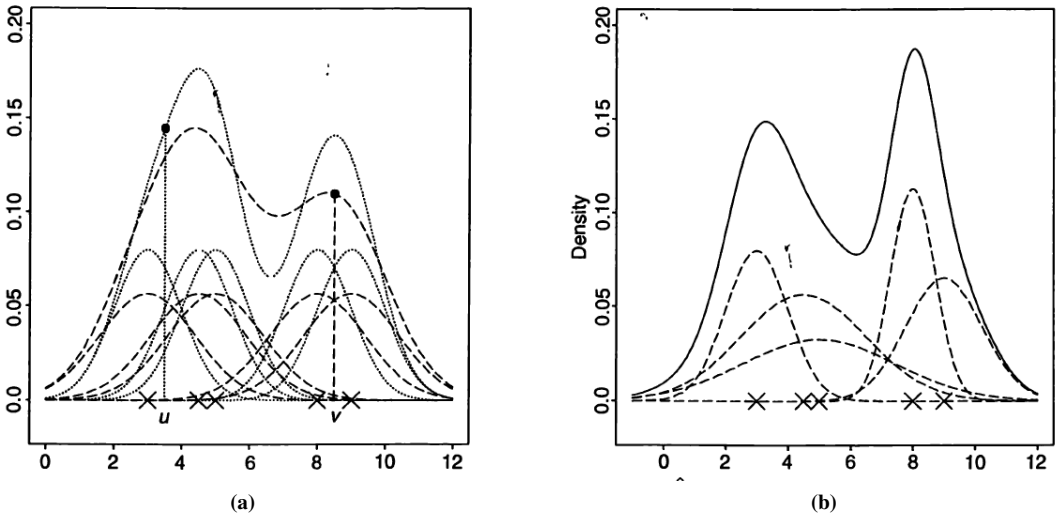


Figure 2.12: Interpretation and construction of KDE with (a) local variable bandwidth and (b) global variable bandwidth . Taken from Wand and Jones (1995).

I refer the interested reader to Wand and Jones (1995) for a general description about these three bandwidth types. See Loftsgaarden and Quesenberry (1965) and Abramson (1982) for the local variable and global variable bandwidth respectively.

A possible criterion for selecting a constant bandwidth parameter h is the MISE. The theoretical optimal bandwidth is then the one minimizing MISE. We denote this bandwidth by $h_{n,\text{MISE}}$. An asymptotic approximation is denoted by $h_{n,\text{AMISE}}$. From (2.9) it easy to show that

$$\boxed{h_{n,\text{AMISE}} = \left[\frac{R(K)}{\mu_2^2 R(f_X'')}\right]^{1/5} n^{-1/5}}$$

with $R(f_X'') = \int (f_X''(x))^2\,dx$. It now immediately follows that

$$h_{n,\text{MISE}} \sim \left[\frac{R(K)}{\mu_2^2 R(f_X'')}\right]^{1/5} n^{-1/5},$$

where $\alpha \sim \beta$ means that α and β are asymptotically equivalent i.e.,

$$\lim_{n\to\infty} \frac{\alpha}{\beta} = 1$$

or in our case

$$\lim_{n\to\infty} \frac{h_{n,\text{MISE}}}{h_{n,\text{AMISE}}} = 1.$$

It is interesting to see that if f_X fluctuates rapidly, $R(f_X'')$ is big (why?) and consequently $h_{n,\text{AMISE}}$ is small.

Another possible criterion for selecting a local variable bandwidth $h(x)$ is to consider the local performance measure $\text{MSE}(\hat{f}_X(x))$. Based on $h(x)$, the kernel density estimator takes the following form

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

Denote

$$\begin{aligned} h_{n,\text{MSE}}(x) &= \arg\min_h \text{MSE}(\hat{f}_{X,L}(x)) \\ h_{n,\text{AMSE}}(x) &= \arg\min_h \text{AMSE}(\hat{f}_{X,L}(x)). \end{aligned}$$

From (2.8), we have

$$\boxed{h_{n,\text{AMSE}}(x) = \left[\frac{f_X(x)R(K)}{\mu_2^2 (f_X''(x))^2}\right]^{1/5} n^{-1/5}}$$

provided $f_X''(x) \neq 0$. With this choice of $h_{n,\text{AMSE}}(x)$, then

$$\begin{aligned} \text{AMSE}(\hat{f}_{X,L}(x)) &= \frac{1}{4} h_{n,\text{AMSE}}(x)^4 \mu_2^2 (f_X''(x))^2 + \frac{1}{n h_{n,\text{AMSE}}(x)} f(x) R(K) \\ &= \frac{1}{4}\{\mu_2^2 R^4(K)\}^{1/5}\{(f_X^2 f_X'')^2(x)\}^{1/5} n^{-4/5} + \{\mu_2^2 R^4(K)\}^{1/5}\{(f_X^2 f_X'')^2(x)\}^{1/5} n^{-4/5} \\ &= \frac{5}{4}\{\mu_2^2 R^4(K)\}^{1/5}\{(f_X^2 f_X'')^2(x)\}^{1/5} n^{-4/5}, \end{aligned}$$

and therefore

$$\text{AMISE}(\hat{f}_{X,L}) = \frac{5}{4}\{\mu_2^2 R^4(K)\}^{1/5} R\left\{(f_X^2 f_X'')^{1/5}\right\} n^{-4/5}.$$

In case we did not choose a bandwidth h depending on x i.e. a constant or global bandwidth, we have

$$\text{AMISE}(\hat{f}_X) = \frac{5}{4}\{\mu_2^2 R^4(K)\}^{1/5} \{R(f_X'')\}^{1/5} n^{-4/5}.$$

Show that using $h(x)$ will always lead to an improvement i.e.

$$R\left\{(f_X^2 f_X'')^{1/5}\right\} \leq \{R(f_X'')\}^{1/5}.$$

The above inequality shows that using $h(x)$ will always lead to an improvement (asymptotically), at least if $h(x)$ is chosen optimal.

The choices $h_{n,\text{AMISE}}$ and $h_{n,\text{AMSE}}(x)$ are theoretical choices and of not much practical use since they depend on the unknown quantities f_X and f_X''.

2.3.5 Choice of kernel function

Substituting $h_{n,\text{AMISE}}$ into AMISE($\hat{f}_X$) yields

$$\frac{5}{4}[\mu_2^2 R^4(K)]^{1/5}[R(f_X'')]^{1/5} n^{-4/5}$$

which depends on the kernel K via

$$C(K) \equiv [\mu_2^2 R^4(K)]^{1/5}.$$

By minimizing the above quantity we can determine the optimal choice for the kernel (in L_2 sense). First, notice that $C(K)$ is invariant under rescaling

$$K_\delta(\cdot) = \frac{1}{\delta} K\big(\frac{\cdot}{\delta}\big), \delta > 0$$

of the kernel K. Indeed,

$$\int K_\delta(u) u^2 \, du = \frac{1}{\delta} \int K\big(\frac{u}{\delta}\big) u^2 \, du = \delta^2 \int K(v) v^2 \, dv = \delta^2 \mu_2$$

and

$$R(K_\delta) = \int K_\delta^2(u) \, du = \frac{1}{\delta^2} \int K^2\big(\frac{u}{\delta}\big) = \frac{1}{\delta} R(K).$$

It immediately follows that

$$C(K_\delta) = [\delta^4 \mu_2^2 \frac{1}{\delta^4} R^4(K)]^{1/5} = C(K).$$

The optimal kernel K is the one that satisfies the following optimization problem

$$\begin{cases} \min \int K^2(u) \, du \\ \text{s.t.} \qquad K \geq 0, \int K(u) \, du = 1, \int u K(u) \, du = 0, \int u^2 K(u) \, du = a^2, \end{cases}$$

where $a \neq 0$ is fixed. Using Lagrange multipliers, this leads to the famous Epanechnikov kernel (Epanechnikov, 1969)

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{if } |u| \leq 1 \\ 0 & \text{if } |u| > 1. \end{cases}$$

One can prove that this kernel is a solution to the above minimization problem and is unique.

What is the effect of using the Epanechnikov kernel over other kernels in practice? Define the efficiency of a kernel K relative to the Epanechnikov kernel (K_0) as

$$\text{eff}(K) = \left\{ \frac{C(K_0)}{C(K)} \right\}^{5/4}.$$

This ratio represents the sample sizes necessary to obtain the same minimum AMISE (for a given f_X) when using K_0 as when using K. If $\text{eff}(K) = 0.95$, this indicates that the Epanechnikov kernel (K_0) can achieve the same minimum AMISE using 95% of the data used with K. Table 2.1 represents some relative efficiencies for other popular choices for K. It is clear that by using so-called "suboptimal" kernels one loses very little in terms of performance.

2.3.6 Bias reduction and higher order kernels

In the previous paragraphs we derived the asymptotic approximations for the bias and the variance of the kernel estimator and showed that the convergence rate of the kernel density estimator is of order $n^{-4/5}$. Remember that these were obtained under the assumptions that

$$K \geq 0 \qquad \int K(u) \, du = 1 \qquad \int u K(u) \, du = 0 \qquad 0 < \int u^2 K(u) \, du < \infty, \tag{2.10}$$

and necessary assumptions on f_X''. If K is constrained to be a probability density function then it is necessarily true that

$$\mu_2(K) = \int u^2 K(u) \, du > 0.$$

Kernel	Form	eff(K)
Epanechnikov	$\frac{3}{4}(1-u^2)I\{u \in [-1,1]\}$	1
Biweight	$\frac{15}{16}(1-u^2)^2 I\{u \in [-1,1]\}$	0.994
Triweight	$\frac{35}{32}(1-u^2)^3 I\{u \in [-1,1]\}$	0.987
Normal	$\frac{1}{\sqrt{2\pi}} e^{-(1/2)u^2}$	0.951
Triangular	$(1-\lvert u\rvert)I\{u \in [-1,1]\}$	0.986
Uniform	$\frac{1}{2}I\{u \in [-1,1]\}$	0.930

Table 2.1: Efficiencies of several popular kernels compared to the optimal kernel.

However, without this restriction, it is possible to construct K so that $\mu_2(K) = 0$ which will have the effect of reducing the bias to be of order h^4, provided that the assumption of a continuous square integrable second derivative is strengthened to f_X having a continuous square integrable fourth order derivative.

We will discuss what happens with the asymptotic approximations if we work with kernels which are not constrained by (2.10). A kernel satisfying

$$\begin{cases} \int K(u)\,du = 1 \\ \int u^j K(u)\,du = 0 \quad j = 1,\ldots,k-1 \\ \int u^k K(u)\,du \neq 0 \end{cases}$$

is called a kth order kernel.

If we require that K is symmetric, then this implies that k is even. The kernel satisfying the assumptions (2.10) is a second order kernel. Every symmetric probability density is a second order kernel.

Denote by $K_{[k]}$ a kth order kernel. Assuming the necessary conditions on $f_X^{(k)}$, the kth order derivative of the density function f_X, and using arguments similar to those of the previous paragraph, the expectation of the kernel density estimator can be approximated as follows

$$\begin{aligned} \mathbf{E}[\hat{f}_X(x)] &= (K_h * f_X)(x) \qquad \text{with } K = K_{[k]} \\ &= \int K_h(x-y) f_X(y)\,dy \qquad \text{with } K = K_{[k]} \\ &= \int K_{[k]}(u) f_X(x-uh)\,du \\ &= \int K_{[k]}(u) \sum_{l=0}^{k} (-uh)^l \frac{1}{l!} f_X^{(l)}(x)\,du + o\big(h^k\big) \\ &= f_X(x) + (-1)^k \frac{1}{k!} \mu_k(K_{[k]}) f_X^{(k)}(x) h^k + o\big(h^k\big) \end{aligned}$$

where

$$\mu_k(K_{[k]}) = \int u^k K_{[k]}(u)\,du \neq 0.$$

We have the following asymptotic approximations for bias and variance based on a kth order kernel

$$\begin{aligned} \text{bias}[\hat{f}_X(x)] &= (-1)^k \frac{1}{k!} \mu_k(K_{[k]}) f_X^{(k)}(x) h^k + o\big(h^k\big) \\ \mathbf{Var}[\hat{f}_X(x)] &= \frac{1}{nh} f_X(x) R(K_{[k]}) + o\left(\frac{1}{nh}\right). \end{aligned}$$

This automatically leads to the mean integrated squared error expression

$$\text{AMISE}[\hat{f}_X] = \frac{1}{(k!)^2} \mu_k^2(K_{[k]}) R(f_X^{(k)}) h^{2k} + \frac{1}{nh} R(K_{[k]}).$$

Then, the AMISE optimal bandwidth is given by

$$\boxed{h_{n,\text{AMISE}} = \left[\frac{(k!)^2 R(K_{[k]})}{2k\mu_k^2(K_{[k]})R(f_X^{(k)})}\right]^{1/(2k+1)} n^{-1/(2k+1)}}$$

With this optimal bandwidth the AMISE error is of order $n^{-2k/(2k+1)}$. For a fourth order kernel, the bias is of order h^4 (and not h^2 as for a second order kernel), the AMISE optimal bandwidth would be of order $n^{-1/9}$ which results in an AMISE of order $n^{-8/9}$. This should be compared with the convergence rate $n^{-4/5}$ of a kernel estimator based on a second-order kernel. Note that the convergence rate $n^{-2k/(2k+1)}$ approaches n^{-1} as k becomes larger. Hence, for sufficiently smooth densities, the convergence rate can be made arbitrarily close to n^{-1}, the parametric convergence rate. These are of course asymptotic considerations, which do not imply that for sample sizes usually encountered in practice, a higher-order kernel will necessarily improve the error.

There are several rules for constructing higher-order kernels. See for example Jones and Foster (1993). We will only discuss one rule here. Let $K_{[k]}$ be a kth order symmetric kernel (k even) which is assumed to be differentiable. Then

$$K_{[k+2]}(u) = \frac{k+1}{k}K_{[k]}(u) + \frac{1}{k}uK'_{[k]}(u),$$

is a $(k+2)$th order kernel. This formula can thus be used to generate higher-order kernels. Consider for example the standard normal density function which is a second-order kernel. Then a fourth-order kernel can be obtained via

$$\begin{aligned} K_{[4]}(u) &= \frac{3}{2}\varphi(u) + \frac{1}{2}u\varphi'(u) \\ &= \frac{1}{2}(3-u^2)\varphi(u), \qquad \varphi'(u) = -u\varphi(u), \end{aligned}$$

where φ is the standard normal density. Figure 2.13 shows the standard normal density φ together with the fourth order kernel $K_{[4]}$ derived from it.

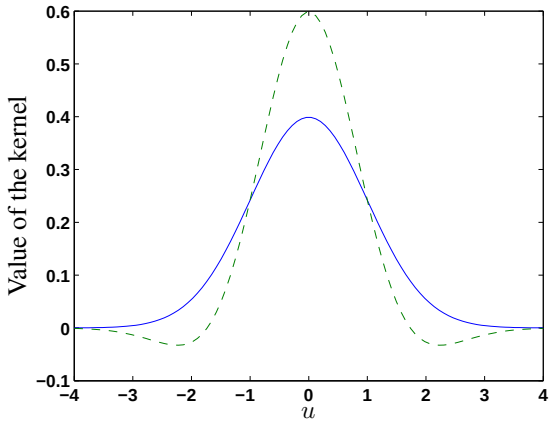


Figure 2.13: Second order kernel (full) and its fourth order kernel (dashed).

2.4 Asymptotic properties of a kernel density estimator

Let $X_1, \ldots, X_n$ be an i.i.d. sample from f_X and consider the kernel density estimator

$$\widehat{f}_X(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

with bandwidth h_n depending on n. To ease the notation, the dependence on the sample size will be omitted.

One of the main tools to establish asymptotic results of the kernel density estimator can be found in Bochner (1955).

Lemma 2.1 (Bochner (1955)) *Suppose that the kernel K satisfies the following properties*

(A1) $\int |K(u)|\,du < \infty$

(A2) $\lim_{|u|\to\infty} |uK(u)| = 0$.

Let a function g satisfy $\int |g(u)|\,du < \infty$ and let $\{h\}$ be a sequence of positive constants such that $\lim_{n\to\infty} h = 0$. Define

$$g_n(x) = \frac{1}{h}\int K\left(\frac{u}{h_n}\right) g(x-u)\,du,$$

then at every point of continuity of g we have that

$$\lim_{n\to\infty} g_n(x) = g(x)\int K(u)\,du.$$

We can now use Lemma 2.1 to obtain the asymptotic unbiasedness (Theorem 2.1) and limit for the variance (Theorem 2.2) for the kernel density estimator.

Theorem 2.1 (Parzen (1962)) *Let the kernel K satisfy assumptions A1, A2 and $\int K(u)\,du = 1$. If $h \to 0$ as $n \to \infty$ then*

$$\lim_{n\to\infty} \mathbf{E}[\widehat{f}_X(x)] = f_X(x)$$

in all points x where f_X is continuous.

PROOF. By applying Lemma 2.1 we have that

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E}[\widehat{f}_X(x)] &= \lim_{n\to\infty} \frac{1}{h}\int K\left(\frac{x-y}{h}\right) f_X(y)\,dy \\
&= \lim_{n\to\infty} \frac{1}{h}\int K\left(\frac{u}{h}\right) f_X(x-u)\,du \\
&= f_X(x)\int K(u)\,du \\
&= f_X(x).
\end{aligned}$$

■

Theorem 2.2 (Parzen (1962)) *Let K satisfy assumptions A1 and A2. Further, assume that $\int K(u)\,du = 1$ and $\sup_u |K(u)| < \infty$. If $h \to 0$ as $n \to \infty$ then*

$$\lim_{n\to\infty} nh\,\mathbf{Var}[\widehat{f}_X(x)] = f_X(x)\int K^2(u)\,du,$$

provided f_X is continuous in x.

PROOF. Note that

$$\begin{aligned}
\mathbf{Var}[\widehat{f}_X(x)] &= \frac{1}{n^2h^2}\,\mathbf{Var}\left[\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] \\
&\overset{\text{i.i.d}}{=} \frac{n}{n^2h^2}\,\mathbf{Var}\left[K\left(\frac{x-X}{h}\right)\right] \\
&= \frac{1}{nh^2}\left(\mathbf{E}\left[K\left(\frac{x-X}{h}\right)\right]^2 - \left\{\mathbf{E}\left[K\left(\frac{x-X}{h}\right)\right]\right\}^2\right) \\
&= \frac{1}{n}\left(\mathbf{E}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]^2 - \left\{\mathbf{E}[\widehat{f}_X(x)]\right\}^2\right).
\end{aligned}$$

By Theorem 2.1 and $h \to 0$ as $n \to \infty$, it follows that

$$\begin{aligned}\lim_{n\to\infty} nh\,\mathbf{Var}[\widehat{f}_X(x)] &= \lim_{n\to\infty} \frac{1}{h}\int K^2\left(\frac{x-y}{h}\right) f_X(y)\,dy - \lim_{n\to\infty} h\{\mathbf{E}[\widehat{f}_X(x)]\}^2 \\ &= \lim_{n\to\infty} \frac{1}{h}\int K^2\left(\frac{x-y}{h}\right) f_X(y)\,dy.\end{aligned}$$

In order to use Lemma 2.1, we need to verify the following conditions on K^2 (see A1 and A2 in Lemma 2.1)

(C1)

$$\begin{aligned}\int K^2(y)\,dy &= \int |K(y)||K(y)|\,dy \\ &\leq \sup_y |K(y)| \int |K(y)|\,dy < \infty\end{aligned}$$

(C2)

$$0 \leq |yK^2(y)| \leq \underbrace{\sup_y |K(y)|}_{<\infty} \cdot \underbrace{|yK(y)|}_{\to 0 \text{ if } |y|\to\infty}.$$

Applying Lemma 2.1 yields

$$\lim_{n\to\infty} nh\,\mathbf{Var}[\widehat{f}_X(x)] = \lim_{n\to\infty} \frac{1}{h}\int K^2\left(\frac{x-y}{h}\right) f_X(y)\,dy = \lim_{n\to\infty} \frac{1}{h}\int K^2\left(\frac{u}{h}\right) f_X(x-u)\,du = f_X(x)\int K^2(u)\,du.$$

■

The weak consistency of the kernel density estimator now follows easily.

Theorem 2.3 (Parzen (1962)) *Suppose K satisfies the conditions of Theorem 2.2. Assume that $h \to 0$ as $n \to \infty$ such that $nh \to \infty$. Then, if f_X is continuous in x, we have*

$$\widehat{f}_X(x) \xrightarrow{p} f_X(x) \quad \textit{or} \quad \lim_{n\to\infty} \mathbf{P}(|\widehat{f}_X(x) - f_X(x)| \geq \epsilon) = 0 \textit{ for any } \epsilon > 0.$$

PROOF. This is a direct consequence of Chebyshev's inequality, Theorem 2.1 and Theorem 2.2. ■

The following properties (asymptotic normality, uniform weak consistency and uniform strong consistency) will be given without proof.

Theorem 2.4 (Asymptotic normality, Parzen (1962)) *Suppose K satisfies the conditions of Theorem 2.2. Assume that $h \to 0$ as $n \to \infty$ such that $nh \to \infty$. Then, if f_X is continuous in x, we have*

$$\frac{\widehat{f}_X(x) - \mathbf{E}[\widehat{f}_X(x)]}{\sqrt{\mathbf{Var}[\widehat{f}_X(x)]}} \xrightarrow{d} N(0,1).$$

In what follows, we use the notation

$$k(u) = \int e^{-iuy} K(y)\,dy$$

for the Fourier transform of the kernel function K.

Theorem 2.5 (uniform weak consistency, Parzen (1962)) *Suppose K satisfies the conditions of Theorem 2.2, K is symmetric and $\int |k(u)|\,du < \infty$. Further assume f_X is uniformly continuous and that $h \to 0$ as $n \to \infty$ such that $nh^2 \to \infty$. Then,*

$$\sup_x |\widehat{f}_X(x) - f_X(x)| \xrightarrow{p} 0.$$

Theorem 2.6 (uniform strong consistency, Nadaraya (1965); Schuster (1969); Van Ryzin (1969)) *Let K be a probability density function satisfying conditions A1 and A2 (see Lemma 2.1) and assume K is of bounded variation. Further assume f_X is uniformly continuous and that*

$$\sum_{n=1}^{\infty} \exp(-\gamma n h^2) < \infty$$

for all $\gamma > 0$. Then,

$$\sup_x |\widehat{f}_X(x) - f_X(x)| \stackrel{a.s.}{\to} 0.$$

Silverman (1978) weakened the condition $\sum_{n=1}^{\infty} \exp(-\gamma n h^2) < \infty$ for all $\gamma > 0$ and showed that for suitable kernels and uniformly continuous densities the conditions $h \to 0$ and $(nh)^{-1} \log n \to 0$ are sufficient for uniform strong consistency of the kernel density estimate.

2.5 Density estimation at the boundaries

Kernel density estimation can fail dramatically when the region of definition of the data at hand is not unbounded e.g. for support on $[0,\infty]$. We will discuss 2 methods, boundary kernels and transformation of the KDE, to cope with this problem. The latter is not particularly designed for the boundary region. It also works in the interior. Other methods do exist e.g. reflection of data (Cline and Hart, 1991) and pseudo-data methods (Cowling and Hall, 1996). Many boundary corrections now exist, but almost all those with good theoretical performance allow the corrected estimator to become negative. An exception is provided by a sophisticated transformation methodology (Jones and Foster, 1996). It is clear that a variety of boundary correction methods for kernel density estimation now exists, and most are referred to in Jones (1993).

2.5.1 Boundary kernels

Consider the following mine accident example.

Example 2.3 *The data are time intervals (in days) between accidents resulting in fatalities in mines in Division 5 of the Great-Britain National Coal Board over a 245 day period in 1950. The kernel density estimate (KDE) is based on a biweight kernel with bandwidth $h = 10.5$, peaks around 3 days between accidents, even though roughly 35% of the data points fall in the range $[0,3)$. That is, the KDE is biased downward near the boundary, see Figure 2.14.*

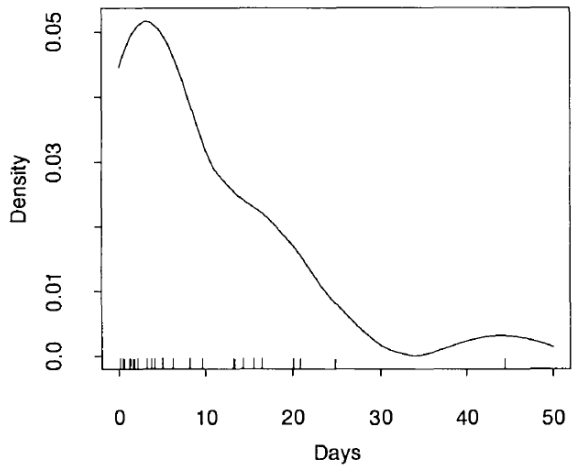


Figure 2.14: Kernel density estimate of mine accident data. Taken from Simonoff (1996).

The reason for this is from the original definition of the estimator, given in (2.5), which motivates the kernel estimate with uniform kernel:

$$g(x) \equiv \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}.$$

Typically we have that $g(x) = f_X(x)$, the density function. Say that the density has a lower boundary at 0, and f_X is being estimated at some x within the boundary region, i.e. $x = ph$ with $0 \le p < 1$. Then (2.5) becomes

$$g(x) \equiv \lim_{h\to 0} \frac{F(x+h) - F(0)}{2h}$$

since $x < h$. This immediately implies

$$\begin{aligned} g(x) &= \lim_{h\to 0} \frac{x+h}{2h} \frac{F(x+h) - F(0)}{x+h} \\ &= \lim_{h\to 0} \frac{p+1}{2} \frac{F(x+h) - F(0)}{x+h} \quad \text{since } x = ph \\ &= \frac{p+1}{2} \lim_{h\to 0} \frac{F\left(\frac{x+h}{2} + \frac{p+1}{2}\cdot h\right) - F\left(\frac{x+h}{2} - \frac{p+1}{2}\cdot h\right)}{(p+1)h} \\ &\approx \frac{p+1}{2} f\left(\frac{x}{2}\right). \end{aligned}$$

The usual kernel formulation is estimating a value that is biased downward near the boundary, unless $f_X(x) = 0$ there.

Taylor series approximations similar to the ones we did earlier (for the interior region) formalize this effect more in depth. Consider a kernel K with support on $]-1,1[$ and f_X has support on $[0, +\infty[$. Note that this follows immediately due to the support of K

$$-1 \le \frac{x-y}{h} \le 1.$$

This implies that $x - h \le y \le x + h$. Also, since $x = ph$ we have that $x - h = (p-1)h < 0$. Since the density only has support on $[0, +\infty[$, the limits of integration become

$$\begin{aligned} \mathbf{E}[\widehat{f}_X(x)] &= \frac{1}{h}\int_0^{x+h} K\left(\frac{x-y}{h}\right) f_X(y)\, dy \\ &= \int_{-1}^{x/h} K(u) f_X(x-uh)\, du \\ &= \int_{-1}^{p} K(u)\big(f_X(x) - f'_X(x)uh + \frac{1}{2}f''_X(x)u^2h^2 + o(h^2)\big)\, du \\ &= f_X(x)\int_{-1}^{p} K(u)\, du - f'_X(x)h\int_{-1}^{p} uK(u)\, du + \frac{1}{2}f''_X(x)h^2\int_{-1}^{p} u^2K(u)\, du + o(h^2) \\ &= a_0(p)f_X(x) - hf'_X(x)a_1(p) + \frac{h^2}{2}a_2(p)f''_X(x) + o(h^2), \end{aligned}$$

and

$$\mathbf{Var}[\widehat{f}_X(x)] = \frac{f_X(x)b(p)}{nh} + o\left(\frac{1}{nh}\right)$$

with

$$\begin{aligned} a_l(p) &= \int_{-1}^{p} u^l K(u)\, du \\ b(p) &= \int_{-1}^{p} K^2(u)\, du. \end{aligned}$$

Away from the boundary i.e., $p \geq 1$, these expressions reduce to the usual ones, but near the boundary the KDE is not even consistent, unless $f_X(x) = 0$ (since $a_0(p) < 1$). Even if the kernel is locally normalized to integrate to 1 (by dividing $\widehat{f}_X(ph)$ by $a_0(p)$), the bias in the boundary region is still $O(h)$ rather than $O(h^2)$ in the interior region.

One way of correcting the boundary bias of the KDE is by using special kernels called *boundary kernels*. These kernels are only used within the boundary region (and the usual kernel K is used in the interior). Define

$$c_l(p) = \int_{-1}^{p} u^l L(u)\, du,$$

with L some kernel function closely related to K but of a different form. Let $\widehat{f}_{X,K}$ and $\widehat{f}_{X,L}$ be the KDE based on the kernels K and L respectively. The idea is this. Take a linear combination of K and L in such a way that the resulting kernel has the desired properties $a_0(p) = 1$ and $a_1(p) = 0$. Such a procedure is called generalized jackknifing (Jones, 1993). This procedure seeks a linear combination

$$\widehat{f}_{X,B} = \alpha \widehat{f}_{X,K} + \beta \widehat{f}_{X,L} \tag{2.11}$$

with good asymptotic bias properties for the final KDE $\widehat{f}_{X,B}$ with (for now) unknown kernel B. We immediately have that

$$\begin{aligned} \mathbf{E}[\widehat{f}_{X,B}] &= \alpha\, \mathbf{E}[\widehat{f}_{X,K}] + \beta\, \mathbf{E}[\widehat{f}_{X,L}] \\ &= \alpha\big(a_0(p) f_X(x) - h f'_X(x) a_1(p) + O(h^2)\big) + \beta\big(c_0(p) f_X(x) - h f'_X(x) c_1(p) + O(h^2)\big). \end{aligned}$$

In order to make the boundary bias $O(h^2)$, we need to solve the following system of equations w.r.t. α and β

$$\begin{cases} \alpha a_0(p) + \beta c_0(p) = 1 \\ \alpha a_1(p) + \beta c_1(p) = 0. \end{cases}$$

The solution is given by

$$\alpha = \frac{c_1(p)}{a_0(p)c_1(p) - a_1(p)c_0(p)} \qquad \text{and} \qquad \beta = \frac{-a_1(p)}{a_0(p)c_1(p) - a_1(p)c_0(p)}.$$

Plugging in the above results in (2.11) yields

$$\begin{aligned} \widehat{f}_{X,B} &= \frac{c_1(p)}{a_0(p)c_1(p) - a_1(p)c_0(p)} \widehat{f}_{X,K} + \frac{a_1(p)}{a_0(p)c_1(p) - a_1(p)c_0(p)} \widehat{f}_{X,L} \\ &= \frac{1}{nh} \sum_{i=1}^{n} \left(\frac{c_1(p)}{a_0(p)c_1(p) - a_1(p)c_0(p)} K\left(\frac{x - X_i}{h}\right) - \frac{a_1(p)}{a_0(p)c_1(p) - a_1(p)c_0(p)} L\left(\frac{x - X_i}{h}\right) \right). \end{aligned}$$

Consequently, the kernel B to be used in the boundary region is given by

$$B(x) = \frac{c_1(p)K(x) - a_1(p)L(x)}{a_0(p)c_1(p) - a_1(p)c_0(p)}$$

and leads to

$$\mathbf{E}[\widehat{f}_{X,B}] = f(x) + O(h^2).$$

This means that the bias in the boundary is restored to the $O(h^2)$ level in the interior. Many different forms of L are possible, leading to different boundary kernels (or family of boundary kernels). One useful form is to take

$$L(x) = xK(x)$$

resulting in the boundary kernel

$$B(x) = \frac{[c_1(p) - a_1(p)x]K(x)}{a_0(p)c_1(p) - a_1(p)c_0(p)}.$$

Figure 2.15 shows the above kernel B for various values of p for the biweight kernel.

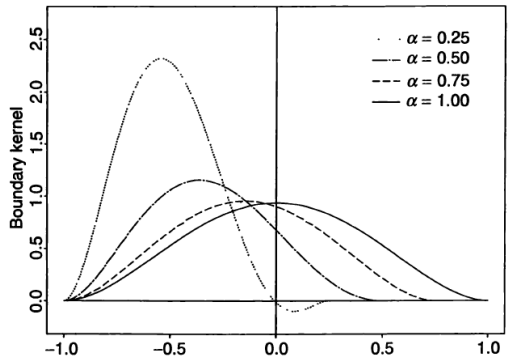


Figure 2.15: B(x) based on the biweight kernel for $p = 1/4$ (solid curve), $p = 1/2$ (dashed curve), $p = 3/4$ (dotted curve) and $p = 1$ (do-dashed curve). In this figure, α in the legend is equal to p from the notes. Taken from Wand and Jones (1995)

Example 2.4 *Recall the mine accident data earlier in this section. Figure 2.16 shows the boundary kernel estimate of this data set.*

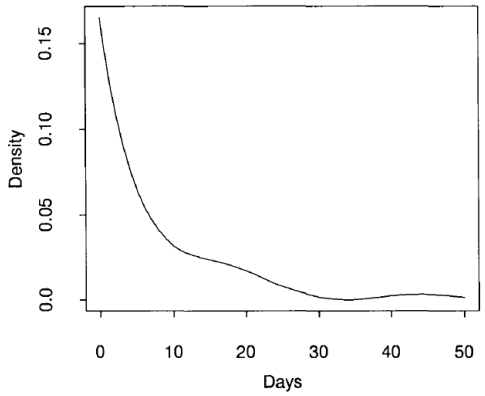


Figure 2.16: Boundary kernel estimate of the mine accident data. Taken from Simonoff (1996)

However, this bias correction comes at a cost: the reduction in bias comes with an increase in variance. Figure 4.3 gives the 95% variability plot of the boundary kernel estimate for the mine data. The most striking property is the great widening of the envelope as the boundary is approached.

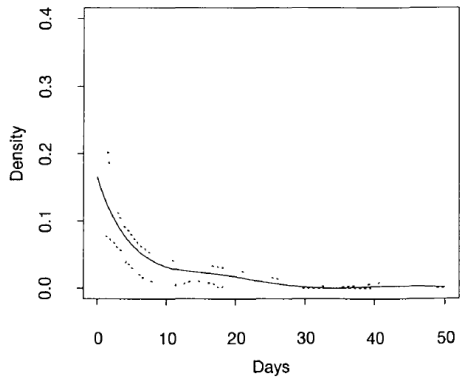


Figure 2.17: Variability plot for boundary kernel estimate of the mine accident data. Taken from Simonoff (1996)

Example 2.5 *Consider the kernel density estimate of the exponential density based on a sample of* $n = 1000$. *The solid curve is the density estimate and the dashed curve is the true density.*

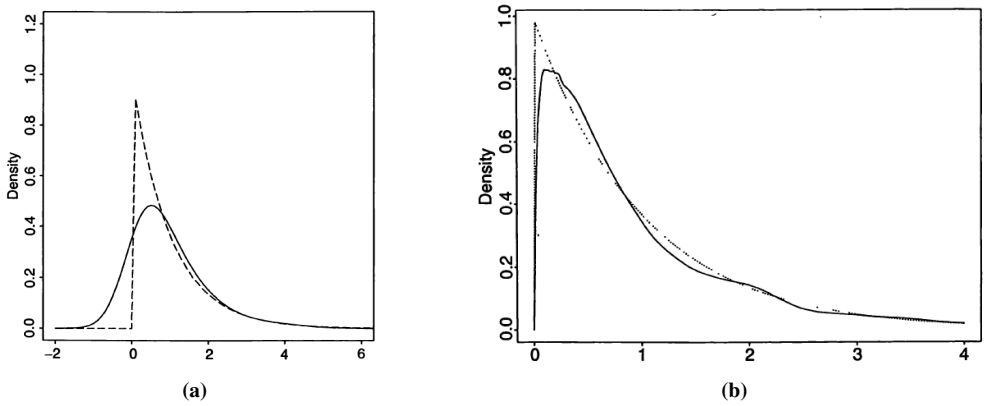


Figure 2.18: kernel density estimate of the exponential density based on a sample of $n = 1000$ and with boundary kernel based on the biweight kernel. The solid curve is the density estimate and the dashed curve is the true density. Taken from Wand and Jones (1995)

2.5.2 Transformation of kernel density estimators

If the random sample $X_1, \ldots, X_n$ has a density f_X that is difficult to estimate then another possibility is to apply a transformation to the data to obtain a new sample $Y_1, \ldots, Y_n$ having a density g_Y that can be more easily estimated using KDE. One would then *backtransform* the estimate g_Y to obtain the estimate of f_X. Suppose the transformation is given by $Y_i = t(X_i)$ where t is an increasing differentiable function on the support of f_X. Then from statistical distribution theory it follows that

$$f_X(x) = g_Y(t(x))t'(x).$$

The KDE of f_X is then given by

$$\widehat{f}_{X,T}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{t(x) - t(X_i)}{h}\right)(t'(x)/h). \tag{2.12}$$

The transformation KDE is neither a local nor a variable KDE. An application of the mean value theorem to (2.12) gives

$$\widehat{f}_{X,T}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{t'(\xi_i)(x - X_i))}{h}\right)(t'(x)/h)$$

where ξ_i lies between x and X_i.

Example 2.6 *A simple illustrative example of this idea comes from estimating the lognormal density. In this case the transformation* $Y_i = \ln X_i$ *was applied to the data and hence the* Y_i*'s are a sample from the* $N(0,1)$ *distribution. Figure 2.19a shows the ordinary KDE of the data after they have undergone the natural log transformation. Figure 2.19b shows the estimate of the lognormal density obtained via* (2.12)*.*

One question remains here. What is the best choice of the transformation t? That choice depends quite heavily on the shape of f_X. If f_X is a skewed unimodal density, it can be argued that t should be a convex function on the support of f_X, since it will, in a certain sense, reduce the skewness of f_X. One possible class of transformations t for heavily skewed data is the shifted power family given by

$$t(x,\lambda_1,\lambda_2) = \begin{cases} (x+\lambda_1)^{\lambda_2}\,\text{sign}(\lambda_2), & \lambda_2 \neq 0 \\ \ln(x+\lambda_1), & \lambda_2 = 0, \end{cases}$$

where $\lambda_1 > -\min(X)$ and $\min(X)$ denotes the lower endpoint of the support of f_X. This is an extension of the Box-Cox family of transformations.

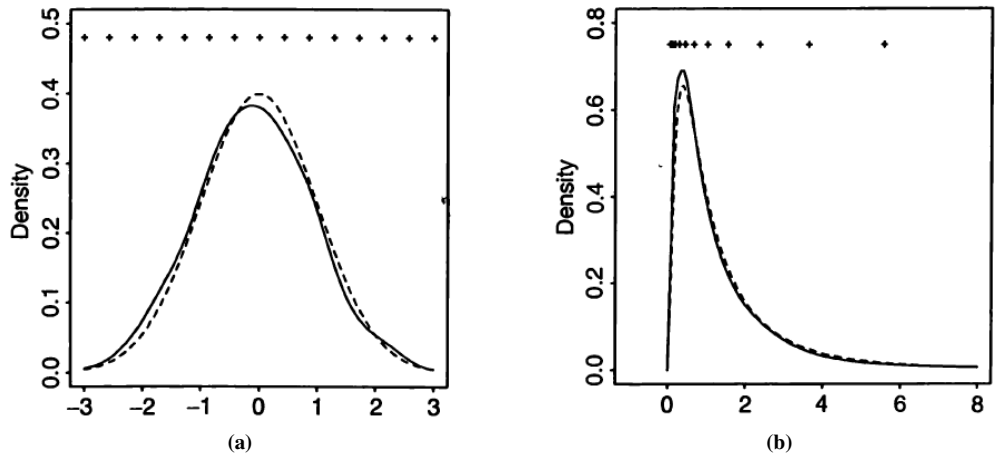


Figure 2.19: Transformation of KDE with $n = 1000$. (a) Solid curve is the ordinary KDE on the transformed scale while the solid curve in (b) is its backtransformation using $t^{-1}(x) = e^x$. The dashed curves are the true functions. Taken from Wand and Jones (1995)

2.6 Bandwidth selection for kernel density estimation

This section gives an overview of two type of bandwidth selection criteria. Of course there are many more and even better ones (e.g. biased cross-validation, direct plug-in rules, solve-the-equation rules, smoothed cross-validation and the double kernel method). See the books of Silverman (1986) and Scott (1992) for a comprehensive overview.

2.6.1 The normal reference rule

The formula for an optimal constant bandwidth was given by

$$h_{n,\text{AMISE}} = \left[\frac{R(K)}{\mu_2^2 R(f_X'')}\right]^{1/5} n^{-1/5}.$$

Now suppose f_X belongs to some standard family of distributions. For example, consider f_X to be $N(\mu,\sigma^2)$ with μ and σ^2 unspecified. Then, it easy to show that

$$R(f_X'') = \frac{1}{\sigma^5}\frac{3}{8\sqrt{\pi}}.$$

Therefore, by making a reference to a normal density, the formula for the asymptotically optimal bandwidth becomes

$$\widehat{h}_{n,\text{AMISE}} = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2^2}\right]^{1/5} \widehat{\sigma} n^{-1/5}, \tag{2.13}$$

where $\widehat{\sigma}$ is some estimator of σ. This is bandwidth selector is also called a rule-of-thumb bandwidth selector. Possible choices for $\widehat{\sigma}$ are

- Sample standard deviation
- Standardized sample interquartile range: $R/1.349$ (where R is the interquartile range)
- To accommodate long-tailed distributions and possible outliers

$$\widehat{\sigma} = \frac{\text{median}\{|X_i - \text{median}(X_1,\ldots,X_n)|\}}{0.6745}$$

It is recommended the smaller of $\widehat{\sigma}$ and $R/1.349$ i.e.,

$$\boxed{\widehat{h}_{n,\text{NR}} = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2^2}\right]^{1/5} \min\left(\widehat{\sigma}, \frac{R}{1.349}\right) n^{-1/5}}. \tag{2.14}$$

This gives for

- Gaussian kernel: $\widehat{h}_{n,\text{NR}} = 1.06 \min\left(\widehat{\sigma}, \frac{R}{1.349}\right) n^{-1/5}$
- Epanechnikov kernel: $\widehat{h}_{n,\text{NR}} = 2.34 \min\left(\widehat{\sigma}, \frac{R}{1.349}\right) n^{-1/5}$
- biweight kernel: $\widehat{h}_{n,\text{NR}} = 2.78 \min\left(\widehat{\sigma}, \frac{R}{1.349}\right) n^{-1/5}$.

This bandwidth selector is sensitive to deviations from the assumptions on the normal density e.g., multimodality, skewness and kurtosis. An improved bandwidth selection selection rule (but still sensitive to the assumptions above) can be obtained by writing an Edgeworth expansion for f_X around the Gaussian density. Such a rule is provided in Hjort and Jones (1996) and is given by

$$\widehat{h}_{n,\text{INR}} = \widehat{h}_{n,\text{NR}} \left(1 + \frac{35}{48}\widehat{\gamma}_4 + \frac{35}{32}\widehat{\gamma}_3^2 + \frac{385}{1024}\widehat{\gamma}_4^2\right)^{-1/5}$$

where $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ are the sample skewness and kurtosis respectively.

2.6.2 Oversmoothed bandwidth selection

The oversmoothing or maximal smoothing principle relies on the fact that there is an upperbound for the AMISE optimal bandwidth for estimation of densities with a fixed value of particular scale measure. Terrell (1990) showed that

$$h_{n,\text{ AMISE}} \leq \left[\frac{243R(K)}{35\mu_2^2}\right]^{1/5} n^{-1/5}\sigma.$$

for all densities having a standard deviation σ and that this bound is attained by the beta(4,4) or triweight density (see Figure 2.20). Similar results can be shown to hold for other scale measures. The above bound on h_{AMISE} motivates the oversmoothed bandwidth selector

$$\boxed{\widehat{h}_{n,\text{OS}} = \left[\frac{243R(K)}{35\mu_2^2}\right]^{1/5} n^{-1/5}S} \tag{2.15}$$

where S the sample standard deviation.

2.6.3 Least squares cross-validation

Least squares cross-validation was introduced by Rudemo (1982) and Bowman (1984). The motivation for the least squares cross-validation comes from expanding the MISE of the kernel density estimator

$$\begin{aligned}\text{MISE}\{\widehat{f}_X\} &= \mathbf{E}[\text{ISE}\{\widehat{f}_X\}] = \mathbf{E}\int [\widehat{f}_X(x) - f_X(x)]^2\,dx \\ &= \mathbf{E}\int \widehat{f}_X^2(x)\,dx - 2\,\mathbf{E}\int \widehat{f}_X(x) f_X(x)\,dx + \int f_X^2(x)\,dx.\end{aligned}$$

The term $\int f_X^2(x)\,dx$ does not depend on h and hence minimization of $\text{MISE}\{\widehat{f}_X\}$ w.r.t. to h is equivalent to minimizing

$$\text{MISE}\{\widehat{f}_X\} - \int f_X^2(x)\,dx = \mathbf{E}\left[\int \widehat{f}_X^2(x)\,dx - 2\int \widehat{f}_X(x) f_X(x)\,dx\right]$$

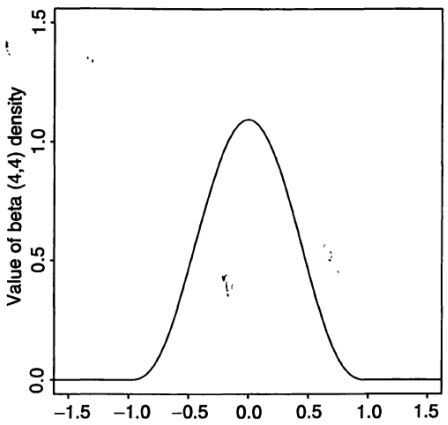


Figure 2.20: the beta(4,4) density.

It can be shown that an unbiased estimator for $\mathbf{E}\int \widehat{f}_X(x) f_X(x)\,dx$ is given by

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_X^{(-i)}(X_i),$$

where

$$\widehat{f}_X^{(-i)}(x) = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} K_h(x - X_j),$$

is the kernel density estimator based on the sample $X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n$ i.e., the sample with X_i deleted. The above estimator is often referred to as the *leave-one-out estimator*. The term cross-validation refers to the fact that part of the sample is used to obtain information about another part i.e., $X_1,\ldots,X_{i-1},X_i,X_{i+1},\ldots,X_n$ is used to get information about $f_X(X_i)$. Therefore, the least squares cross-validation quantity is defined as

$$\mathrm{LSCV}(h) = \int \widehat{f}_X^2(x)\,dx - \frac{2}{n(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n} K_h(X_i - X_j).$$

The least squares cross-validation bandwidth selector is the one that minimizes the least squares cross-validation quantity

$$\boxed{\widehat{h}_{\mathrm{LSCV}} = \arg\min_h \mathrm{LSCV}(h)} \tag{2.16}$$

One has to be careful when using LSCV for discretized data. It is recommended to minimize (2.16) over a range of values of h. Such a range can be suggested by e.g. the normal reference rules in the previous paragraph. Further, it is sometimes the case that (2.16) has more than one local minimum. Finally, the LSCV bandwidth selector is highly variable. For example, the variance of LSCV is roughly 15.7 times larger than biased cross-validation (see Jones and Kappenman (1991)). Figure 2.21 shows $\mathrm{LSCV}(h)$ vs. $\log_{10}(h)$ for two particular samples of size $n = 100$ from the standard normal density using the standard normal kernel. Figure 2.21b is an example of a LSCV function having two minima. The $\log_{10}$ of the MISE optimal bandwidth $h_{\mathrm{MISE}} \approx 0.445$ is shown by the dashed vertical line. Note that the actual minimum is much smaller than h_{MISE}. This phenomenon has led to the suggestion that $\widehat{h}_{\mathrm{LSCV}}$ be taken to correspond to the largest local minimizer of $\mathrm{LSCV(h)}$.

The rate of convergence of $\widehat{h}_{\mathrm{LSCV}}$ to h_{MISE} is rather slow. In fact, Hall and Marron (1987) showed that

$$\frac{\widehat{h}_{\mathrm{LSCV}}}{h_{\mathrm{MISE}}} = 1 + O_p(n^{-1/10}) \quad \text{as} \quad n \to \infty.$$

This slow rate is not the best possible. This best rate can be shown to be of order $n^{-1/2}$.

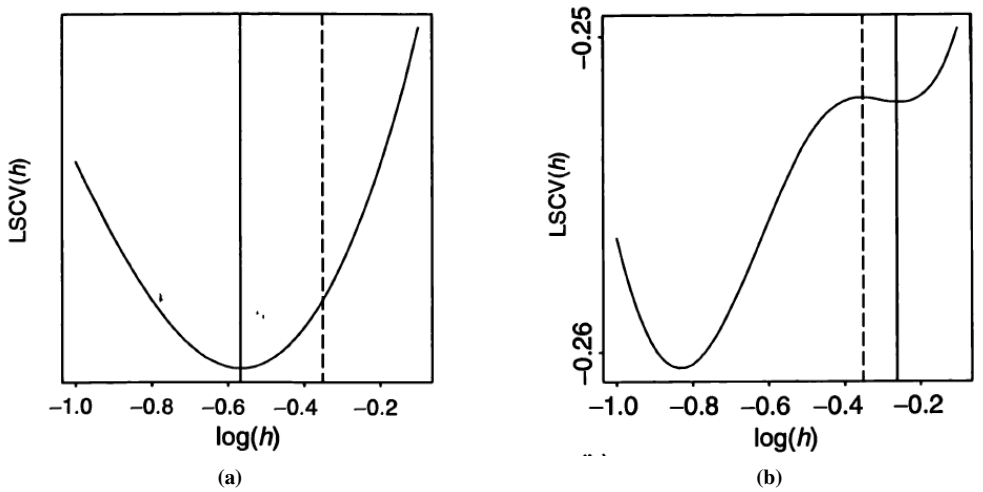


Figure 2.21: Examples of LSCV(h) for two samples of 100 N(0,1) observations. A $\log_{10}$ scale is used on the horizontal axis. The dashed vertical line shows the position of $\log_{10}(h_{\text{MISE}})$. The solid vertical lines show the position of $\log_{10}(\widehat{h}_{\text{LSCV}})$ if $\widehat{h}_{\text{LSCV}}$ is taken to correspond to the largest local minimum. The kernel is the standard normal density. Taken from Wand and Jones (1995).

2.6.4 Biased cross-validation

Least squares cross-validation (LSCV) relies on the exact formula for the mean integrated squared error. Biased cross-validation (BCV) is based on the expression for the asymptotic mean integrated squared error:

$$\text{AMISE}(\hat{f}_X) = \frac{1}{4}h^4\mu_2^2 \int (f_X''(x))^2\,dx + \frac{1}{nh}R(K) = \frac{1}{4}h^4\mu_2^2 R(f_X'') + \frac{1}{nh}R(K) \tag{2.17}$$

derived in Section 2.3. The BCV is obtained by replacing the unknown quantity $R(f_X'')$ in (2.17) by its estimator $\widehat{R(f_X'')}$. How to obtain such an estimator? Let's replace f_X'' by its kernel density estimate $\widehat{f}_X''$ and consider $\mathbf{E}\{R(\widehat{f}_X'')\}$. Perhaps we can find an asymptotic unbiased estimator for $R(f_X'')$. First, note that

$$\begin{aligned}
R(\widehat{f}_X'') &= \int \widehat{f}_X''^2(x)\,dx \\
&= \frac{1}{n^2h^6}\sum_{i=1}^n\sum_{j=1}^n \int K''\left(\frac{x-X_i}{h}\right)K''\left(\frac{x-X_j}{h}\right)dx \\
&= \frac{1}{n^2h^6}\sum_{i=1}^n \int \left\{K''\left(\frac{x-X_i}{h}\right)\right\}^2 dx + \frac{1}{n^2h^6}\sum\sum_{i\neq j} \int K''\left(\frac{x-X_i}{h}\right)K''\left(\frac{x-X_j}{h}\right)dx \\
&= \frac{1}{n^2h^5}\sum_{i=1}^n \int \left\{K''(u)\right\}^2 du + \frac{1}{n^2h^6}\sum\sum_{i\neq j} \int K''\left(\frac{x-X_i}{h}\right)K''\left(\frac{x-X_j}{h}\right)dx.
\end{aligned}$$

Taking expectations, assuming $K^{(r)}(\pm\infty) = 0$ for $r = 0,1$, $\int K(u)\,du = 1$, $\int uK(u)\,du = 0$ and $f^{(4)}(x) < \infty$ for all x of the support yields

$$\begin{aligned}
\mathbf{E}\{R(\widehat{f}_X'')\} &= \frac{1}{nh^5}R(K'') + \frac{1}{n^2h^6}n(n-1)\iiint K''\left(\frac{x-y}{h}\right)K''\left(\frac{x-z}{h}\right)f_X(y)f_X(z)\,dydzdx \\
&= \frac{1}{nh^5}R(K'') + \frac{n-1}{nh^6}\int\left[\left\{\int K''\left(\frac{x-y}{h}\right)f_X(y)\,dy\right\}\left\{\int K''\left(\frac{x-z}{h}\right)f_X(z)\,dz\right\}\right]dx \\
&= \frac{1}{nh^5}R(K'') + \frac{n-1}{nh^6}\int\left[\int K''\left(\frac{x-y}{h}\right)f_X(y)\,dy\right]^2 dx \\
&= \frac{1}{nh^5}R(K'') + \frac{n-1}{nh^4}\int\left[\int K''(u)f_X(x-uh)\,du\right]^2 dx \\
&= \frac{1}{nh^5}R(K'') + \frac{n-1}{n}\int\left[\int K(u)f_X''(x-uh)\,du\right]^2 dx \\
&= \frac{1}{nh^5}R(K'') + \frac{n-1}{n}\int\left[\int K(u)(f_X''(x) - uhf'''(x) + O(h^2))\,du\right]^2 dx \\
&= \frac{1}{nh^5}R(K'') + \frac{n-1}{n}R(f_X'') + O(h^2) \\
&\approx \frac{1}{nh^5}R(K'') + R(f_X'') + O(h^2).
\end{aligned}$$

Setting

$$\widehat{R(f_X'')} = R(\widehat{f}_X'') - \frac{1}{nh^5}R(K'') \tag{2.18}$$

gives an asymptotic unbiased estimator (i.e., for $h \to 0$ as $n \to \infty$) for $R(f_X'')$. This leads to the formulation of the BCV criterion

$$\text{BCV}(h) = \frac{1}{4}h^4\mu_2^2\widehat{R(f_X'')} + \frac{1}{nh}R(K).$$

For $h \sim n^{-1/5}$ the kernel estimate $R(\widehat{f}_X'')$ has a positive bias. The estimator (2.18) adjusts for this bias. The BCV bandwidth selector is the one minimizing the $\text{BCV}(h)$ quantity i.e.,

$$\boxed{\widehat{h}_{\text{BCV}} = \arg\min_h \text{BCV}(h).} \tag{2.19}$$

The main advantage of $\widehat{h}_{\text{BCV}}$ is that it is more stable than $\widehat{h}_{\text{LSCV}}$, in the sense that its asymptotic variance is considerably lower. Indeed, Jones and Kappenman (1992) showed that the ratio of the variance of LSCV and BCV for a Gaussian kernel is

$$\frac{\sigma^2_{\text{LSCV}}}{\sigma^2_{\text{BCV}}} \approx 15.7.$$

This indicates that $\widehat{h}_{\text{LSCV}}$ is considerably more variable than $\widehat{h}_{\text{BCV}}$. This is shown in Figure 2.22. The relative stability of BCV is manifest through the tightness of the distribution of the density of the $\widehat{h}_{\text{BCV}}$'s compared to the of the $\widehat{h}_{\text{BCV}}$'s. However, the fact that the BCV distribution is situated to the right of 0 indicates the positive bias present in the $\widehat{h}_{\text{BCV}}$'s, while there is no noticeable bias in the $\widehat{h}_{\text{LSCV}}$'s. This variance-bias trade-off for bandwidth selectors is also present for other types of rules. Like LSCV, the BCV criterion function occasionally has more than one local minimum, as well as being globally minimised at $h = 0$. Finally, Scott and Terrell (1987) showed that

$$n^{1/10}\left(\frac{\widehat{h}_{\text{BCV}}}{h_{\text{MISE}}} - 1\right) \xrightarrow{d} N(0,\sigma^2_{\text{BCV}}).$$

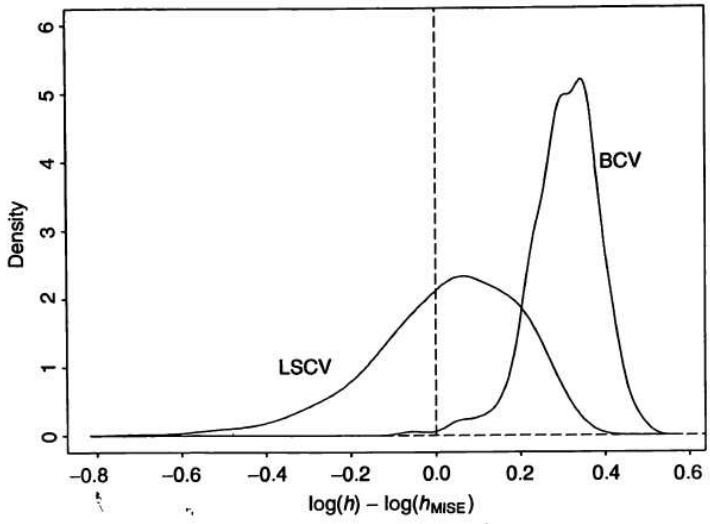


Figure 2.22: Density estimates of $\log_{10}(\widehat{h}_{\text{LSCV}}) - \log_{10}(h_{\text{MISE}})$ and $\log_{10}(\widehat{h}_{\text{BCV}}) - \log_{10}(h_{\text{MISE}})$. Selected bandwidths are based on 500 simulated samples of size n = 100 from a normal mixture density. Taken from Wand and Jones (1995).

2.6.5 Plug-in bandwidth selectors

Direct plug-in rules

Recall the formula for an optimal constant bandwidth

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2^2 R(f'')}\right]^{1/5} n^{-1/5}, \tag{2.20}$$

which contains the unknown quantity $R(f'')$. This is the so-called density functional. Under "strong" smoothness assumptions on the density ($\lim_{|x|\to\infty} f''(x) = 0$ and $\lim_{|x|\to\infty} f'''(x) = 0$), we have by using integration by parts

$$\begin{aligned} R(f'') = \int f''^2(x)\,dx &= \int f''(x)df'(x) \\ &= -\int f'''(x)f'(x)\,dx \\ &= -\int f'''(x)df(x) \\ &= \int f^{(4)}(x)f(x)\,dx \equiv \psi_4. \end{aligned}$$

It is now immediately clear that

$$R(f'') = \int f^{(4)}(x)f(x)\,dx = \mathbf{E}[f^{(4)}(X)]$$

motivating the estimator (law of large numbers)

$$\widehat{\psi}_4(g) = \frac{1}{n}\sum_{i=1}^{n} f^{(4)}(X_i) = \frac{1}{n^2 g^5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(4)}\left(\frac{X_i - X_j}{g}\right). \tag{2.21}$$

Replacing $R(f'')$ in (2.20) by the kernel estimator $\widehat{\psi}_4$ in (2.21) leads to the Direct Plug-In bandwidth selector

$$\boxed{\widehat{h}_{\text{DPI}} = \left[\frac{R(K)}{\mu_2^2 \widehat{\psi}_4(g)}\right]^{1/5} n^{-1/5}.}$$

Unfortunately, this rule is not fully automatic since $\widehat{h}_{\mathrm{DPI}}$ depends on the choice of the pilot bandwidth g. One way of choosing g is to appeal to the formula for the AMSE-optimal bandwidth for estimation of $\widehat{\psi}_4(g)$. If the same second-order kernel K is used in $\widehat{\psi}_4(g)$ then from (2.21) the AMSE-optimal bandwidth is

$$g_{\mathrm{AMSE}} = \left[\frac{2K^{(4)}(0)}{-\mu_2\psi_6}\right]^{1/7} n^{-1/7}.$$

The derivation of this result is given in Wand and Jones (1995, p. 67-70). However, this rule for choosing g has the same defect as $\widehat{h}_{\mathrm{DPI}}$ above: it depends on an unknown density functional, namely ψ_6. We could estimate ψ_6 using another kernel estimate, but its optimal bandwidth depends on ψ_8. This problem will not go away since the optimal bandwidth for estimating ψ_r depends on ψ_{r+2}.

The usual strategy for overcoming this problem is to estimate a ψ_r functional with a quick and simple estimate, such as a version of the normal scale rule (2.13). This means that we really have a family of direct plug-in bandwidth selectors that depend on the number of stages of functional estimation before a quick and simple estimate is used. Suppose that a direct plug-in rule involves l successive kernel functional estimations, with the initial bandwidth chosen via a quick and simple procedure. We will call such a rule an l-stage direct plug-in bandwidth selector and denote it by $\widehat{h}_{\mathrm{DPI,l}}$. The normal reference rule can be thought of as being a zero-stage direct plug-in bandwidth selector. Usually a 2-stage procedure is preferred. These are the four steps needed to obtain $\widehat{h}_{\mathrm{DPI,2}}$ (Sheather and Jones, 1991)

Step 1 Estimate ψ_8 using the normal reference rule: $\widehat{\psi}_8 = 105/(32\sqrt{\pi}\widehat{\sigma}^9)$.

Step 2 Estimate ψ_6 using the kernel estimator $\widehat{\psi}_6(g_1)$ with $g_1 = [-2K^{(6)}(0)/(\mu_2\widehat{\psi}_8)]^{1/9}n^{-1/9}$.

Step 3 Estimate ψ_4 using the kernel estimator $\widehat{\psi}_4(g_2)$ with $g_2 = [-2K^{(4)}(0)/(\mu_2\widehat{\psi}_6(g_1)]^{1/7}n^{-1/7}$.

Step 4 The selected bandwidth is

$$\widehat{h}_{\mathrm{DPI,2}} = \left[\frac{R(K)}{\mu_2^2\widehat{\psi}_4(g_2)}\right]^{1/5} n^{-1/5}.$$

Figure 2.23 shows the effect of the number of stages l on the bandwidth selection process. Finally, it can be shown that

$$n^{5/14}(\widehat{h}_{\mathrm{DPI,2}}/h_{\mathrm{MISE}} - 1) \xrightarrow{d} N(0,\sigma^2_{\mathrm{DPI}}).$$

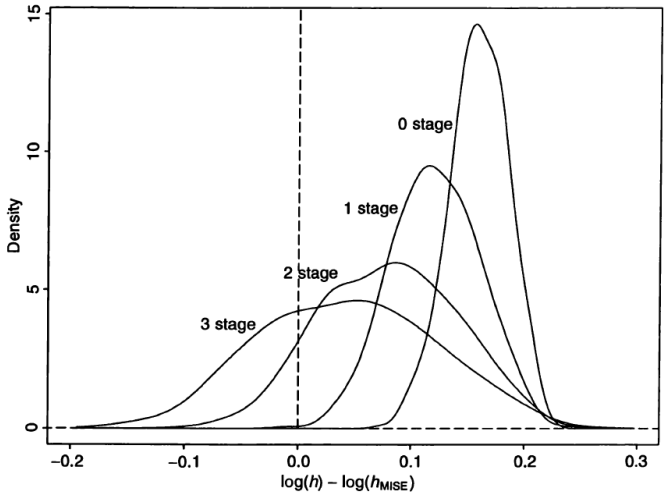


Figure 2.23: Density estimates based on values of $\log_{10}(\widehat{h}_{\mathrm{DPI,l}}) - \log_{10}(h_{\mathrm{MISE}})$ for $l = 0,1,2,3$. Selected bandwidths are based on 500 simulated samples of size $n = 100$ from the normal mixture density. Taken from Wand and Jones (1995).

Solve-the-equation rules

Motivated by the formula for the AMISE optimal bandwidth "solve-the-equation" (STE) rules require that the bandwidth h be chosen to satisfy the relationship

$$\widehat{h}_{\mathrm{STE}} = \left[\frac{R(K)}{\mu_2^2 \widehat{\psi}_4(\gamma(h))}\right]^{1/5} n^{-1/5}$$

where the pilot bandwidth for the estimation of ψ_4 is a function γ of h. The choice of γ can be motivated by noting the relationship

$$g_{\mathrm{amse}} = \left[\frac{2K^{(4)}(0)\mu_2^2}{R(K)\mu_2}\right]^{1/7} (-\psi_4/\psi_6)^{1/7} h^{5/7}.$$

This suggests taking

$$\gamma(h) = \left[\frac{2K^{(4)}(0)\mu_2^2}{R(K)\mu_2}\right]^{1/7} (-\widehat{\psi}_4(g_1)/\widehat{\psi}_6(g_2))^{1/7} h^{5/7}$$

with $\widehat{\psi}_4(g_1)$ and $\widehat{\psi}_6(g_2)$ the kernel estimates of ψ_4 and ψ_6 respectively. The choice of g_1 and g_2 is similar as in the DPI rule. These are the four steps needed to obtain $\widehat{h}_{\mathrm{STE},2}$ (Sheather and Jones, 1991)

Step 1 Estimate ψ_6 and ψ_8 using the normal reference rule: $\widehat{\psi}_6 = -15/(16\sqrt{\pi}\widehat{\sigma}^7)$ and $\widehat{\psi}_8 = 105/(32\sqrt{\pi}\widehat{\sigma}^9)$.

Step 2 Estimate ψ_4 and ψ_6 using the kernel estimator $\widehat{\psi}_4(g_1)$ and $\widehat{\psi}_6(g_2)$ with $g_1 = [-2K^{(4)}(0)/(\mu_2\widehat{\psi}_6)]^{1/7} n^{-1/7}$ and $g_2 = [-2K^{(6)}(0)/(\mu_2\widehat{\psi}_8)]^{1/9} n^{-1/9}$.

Step 3 Estimate ψ_4 using the kernel estimator $\widehat{\psi}_4(\gamma(h))$ with

$$\gamma(h) = \left[\frac{2K^{(4)}(0)\mu_2^2}{R(K)\mu_2}\right]^{1/7} (-\widehat{\psi}_4(g_1)/\widehat{\psi}_6(g_2))^{1/7} h^{5/7}.$$

Step 4 The selected bandwidth is the solution to the equation

$$h = \left[\frac{R(K)}{\mu_2^2 \widehat{\psi}_4(\gamma(h))}\right]^{1/5} n^{-1/5}.$$

2.6.6 Smoothed cross-validation bandwidth selection

Smoothed cross-validation (SCV) (Müller, 1985; Staniswalis, 1985; Hall et al., 1992) is similar to plug-in bandwidth selection in that it uses a kernel estimator with pilot bandwidth g to estimate the integrated squared bias component of $\mathrm{MISE}(\widehat{f_X})$. Because of this, the methods have similar theoretical properties. The difference is that SCV is based on the exact integrated squared bias rather than its asymptotic approximation. This has the intuitively appealing feature of having less dependence on asymptotic approximations. On the other hand, SCV is not as easy to implement as DPI and is somewhat more difficult to analyze. It can be shown that an exact MISE expression for the kernel density estimator is given by

$$\mathrm{MISE}(\widehat{f}_X) = \frac{R(K)}{nh} + \left(1 - \frac{1}{n}\right) \int (K_h * f_X)^2(x)\, dx - 2\int (K_h * f_X)(x) f(x)\, dx + \int f_X^2(x)\, dx$$

where $(g * h)(x) = \int g(x-y)h(y)\, dy$. By ignoring the asymptotically negligible $\frac{1}{n}$ in the second term yields

$$\mathrm{MISE}(\widehat{f}_X) \approx \frac{R(K)}{nh} + \int \left\{(K_h * f_X)(x) - f_X(x)\right\}^2 dx.$$

The second term is now exactly equal to the integrated squared bias (ISB) of $\widehat{f}_X(x)$, while the first term is a good approximation of the integrated variance. SCV is now obtained by replacing f_X by a pilot estimator

$$\widehat{f}_{X,g} = \frac{1}{n}\sum_{i=1}^{n} L_g(x - X_i)$$

with $L_g(x) = L(x/g)/g$ for a possible different kernel L and bandwidth g. This gives

$$\mathrm{SCV}(h) = \frac{R(K)}{nh} + \widehat{\mathrm{ISB}}(h)$$

where

$$\widehat{\mathrm{ISB}}(h) = \int \left\{ (K_h * \widehat{f}_{X,g})(x) - \widehat{f}_{X,g}(x) \right\}^2 dx$$

is an estimate of integrated squared bias (ISB). The smoothed bandwidth cross-validation selector is

$$\boxed{\widehat{h}_{\mathrm{SCV}} = \arg\min_h \mathrm{SCV}(h).}$$

The choice of the bandwidth g is of major importance. We will not elaborate further on this but refer to reader to Wand and Jones (1995) and reference therein. For a specific choice of g, Jones et al. (1991) showed that

$$n^{1/2}(\widehat{h}_{\mathrm{SCV}}/h_{\mathrm{MISE}} - 1) \xrightarrow{d} N(0, \sigma^2_{\mathrm{SCV}}).$$

2.6.7 Bandwidth selection in practice

The asymptotic results of the previous section need to be viewed with some caution. Apart from requiring that the sample size be sufficiently large they also have the defect of often masking the choice of various auxiliary parameters, such as the choice of scale estimate for a normal scale rule, or the number of stages of a plugin strategy. These parameters can have a significant effect on the performance of a bandwidth selector in practice. The main tool for assessing the practical performance of a bandwidth selector is simulation. Figure2.23 show that important insight into the effect of the number of stages of a plug-in rule can be obtained from simulation. Figure 2.24 provides a similar comparison of the selectors $\widehat{h}_{\mathrm{LSCV}}$, $\widehat{h}_{\mathrm{BCV}}$ (with K equal to the standard normal kernel) and the versions of $\widehat{h}_{\mathrm{DPI},2}$ and $\widehat{h}_{\mathrm{SCV},2}$. The sample size was taken to be $n = 100$. For this particular setting we see that $\widehat{h}_{\mathrm{DPI},2}$ provides the best compromise between bias and variance among these four selectors. It is difficult to give a concise summary of simulation results in general since the rankings of the selectors change for different densities (see Wand and Jones (1995)).

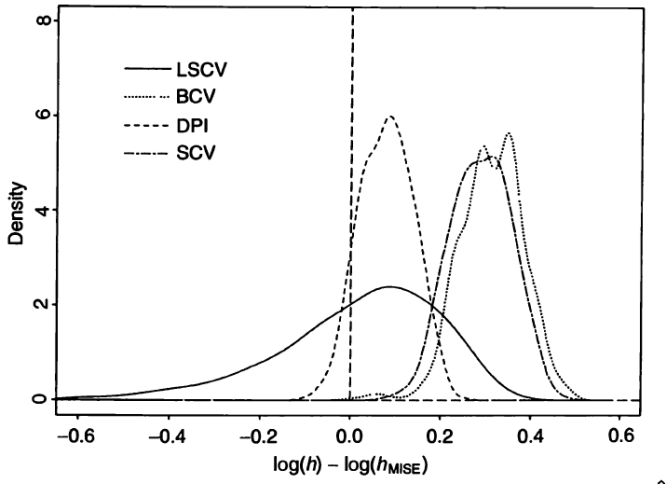


Figure 2.24: Density estimates based on values of $\log_{10}(\widehat{h}) - \log_{10}(h_{\mathrm{MISE}})$ for several bandwidth selectors. Selected bandwidths are based on 500 simulated samples of size $n = 100$ from a normal mixture density. Taken from Wand and Jones (1995).

In summary, while considerable recent progress has been made in the development towards high-performance bandwidth selectors, no rule comes with a guarantee that it will work satisfactorily in all cases. A sensible data analytic strategy is that of obtaining estimates for a variety of bandwidths, perhaps obtained from a variety of bandwidth selectors and choices of auxiliary parameters. If a single objective bandwidth selector is required then the general recommendation, based on simulation evidence, is to use a version of $\widehat{h}_{\mathrm{DPI}}$, $\widehat{h}_{\mathrm{STE}}$ or $\widehat{h}_{\mathrm{SCV}}$, rather than $\widehat{h}_{\mathrm{LSCV}}$ or $\widehat{h}_{\mathrm{BCV}}$

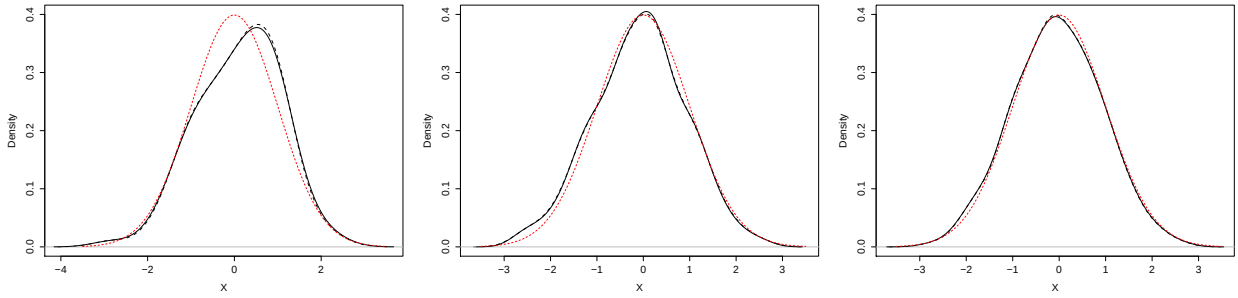
2.7 Kernel density estimation in R

By means of several examples we will demonstrate the use of kernel density estimation in R.

2.7.1 Toy examples

First, consider the data set $X \sim N(0,1)$ for various sample sizes $n = 100, 500, 1000$. The bandwidth is obtained by the normal reference rule (2.14) (dashed bold line) and least squares cross-validation (2.16) (bold line). For comparison the real density is shown as a dashed line (see Figure 2.25). On the normal density, there is almost no difference between the two bandwidth selectors. The code is given below:

```
> n <- 1000
> x <- rnorm(n,0,1)
> a <- seq(-3.5, 3.5, length.out=200)
> Xr <- dnorm(a,0,1)

# Normal reference rule
> plot(density(x, bw = "nrd"), main = "", ylim = c(0,0.4), xlab = "X", lty = 2, lwd = 2)

# least squares cross-validation
> lines(density(x, bw = "ucv"), lwd = 2)
> lines(a, Xr ,col = 2, lwd = 1, lty = 2)
```



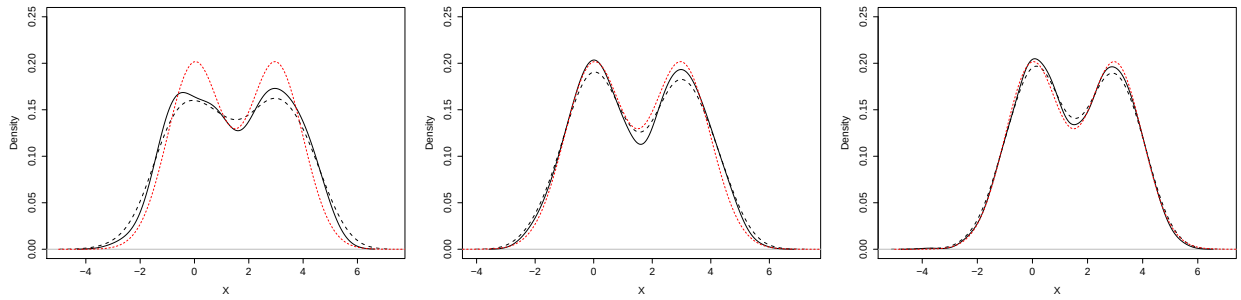
(a) $n = 100$ **(b)** $n = 500$ **(c)** $n = 1000$

Figure 2.25: Kernel density estimates (Gaussian kernel) for the data set $X \sim N(0,1)$ for various sample sizes $n = 100, 500, 1000$. The bandwidth is obtained by the normal reference rule (2.14) (dashed bold line) and least squares cross-validation (2.16) (bold line). For comparison the real density is shown as a dashed line.

Second, let's take a bimodal density. Although the normal reference rule is not designed for mixtures it still performs decent in this example. Figure 2.26 shows the result and the code is given below

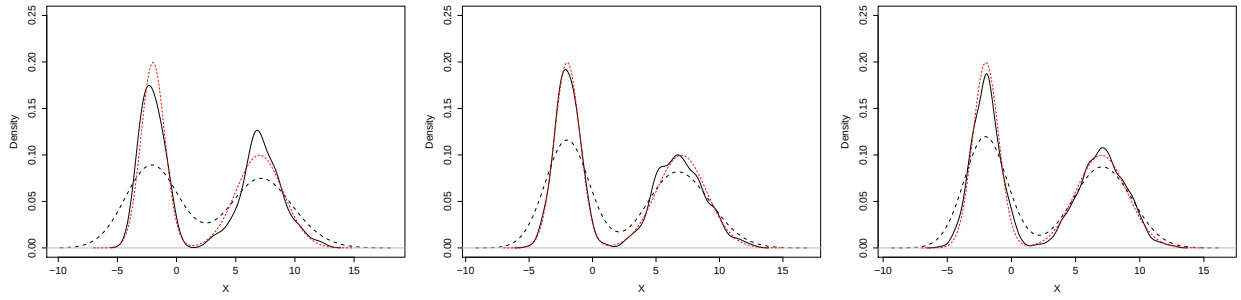
```
> n <- 1000
> x <- c(rnorm(n/2, 0, 1), rnorm(n/2 , 3, 1))
> a <- seq(-5, 10, length.out=200)
> Xr <- 0.5*dnorm(a, 0, 1) + 0.5*dnorm(a, 3, 1)

# Normal reference rule
> plot(density(x, bw = "nrd"), main = "", ylim = c(0,0.25), xlab = "X", lty = 2, lwd = 2)

# least squares cross-validation
> lines(density(x, bw = "ucv"), lwd = 2)
> lines(a, Xr ,col = 2, lwd = 1, lty = 2)
```

Of course, as the mixture becomes more strongly bimodal (2.14) produces an estimate which is less good than the one based on LSCV, see Figure 2.27. It is clear that the normal reference cannot capture both peaks very well.



(a) $n = 100$ **(b)** $n = 500$ **(c)** $n = 1000$

Figure 2.26: Kernel density estimates (Gaussian kernel) for the data set generated by the density $0.5N(0,1) + 0.5N(3,1)$ for various sample sizes $n = 100, 500, 1000$. The bandwidth is obtained by the normal reference rule (2.14) (dashed bold line) and least squares cross-validation (2.16) (bold line). For comparison the real density is shown as a dashed line.

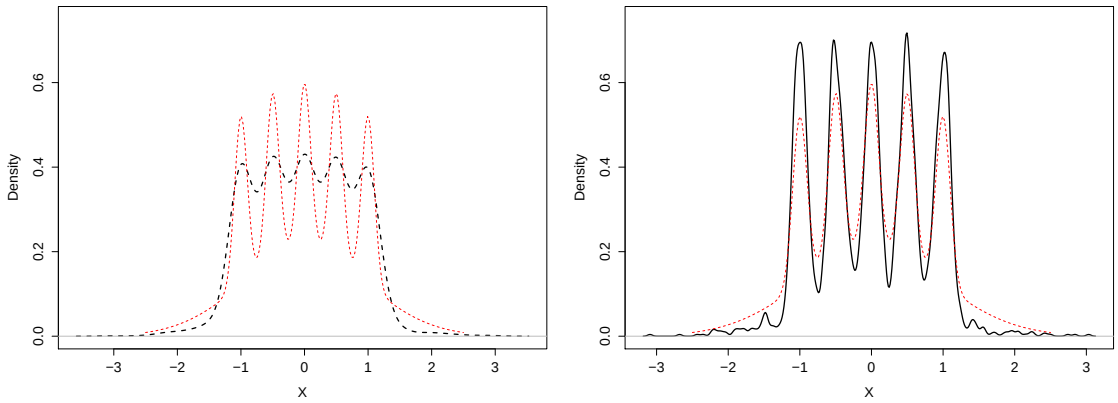


(a) $n = 100$ **(b)** $n = 500$ **(c)** $n = 1000$

Figure 2.27: Kernel density estimates (Gaussian kernel) for the data set generated by the density $0.5N(-2,1) + 0.5N(7,2)$ for various sample sizes $n = 100, 500, 1000$. The bandwidth is obtained by the normal reference rule (2.14) (dashed bold line) and least squares cross-validation (2.16) (bold line). For comparison the real density is shown as a dashed line.

Third, consider the claw density given by $0.5N(0,1) + 0.1 \sum_{i=1}^{4} N(i/2 - 1, 0.1)$. This density is quite challenging since it contains many modes. As shown in Figure 2.28, both bandwidth selectors are unable to capture the density very well even though $n = 3000$. More sophisticated bandwidth selection rules are necessary here i.e, direct plug-ins, solve-the-equation rules or the double kernel method.

```
> n <- 500
> sigma <- 0.1
> x1 <- rnorm(n,0,1)
> x2 <- rnorm(n,-1,sigma)
> x3 <- rnorm(n,-0.5,sigma)
> x4 <- rnorm(n,0,sigma)
> x5 <- rnorm(n,0.5,sigma)
> x6 <- rnorm(n,1,sigma)
> x <- c(x1,x2,x3,x4,x5,x6)
> a <- seq(-2.5, 2.5, length.out=200)
> Xr <- 0.5*dnorm(a,0,1)+ 0.1*dnorm(a,-1,sigma)+0.1*dnorm(a,-0.5,sigma)+0.1*dnorm(a,0,sigma)+...
        0.1*dnorm(a,0.5,sigma)+0.1*dnorm(a,1,sigma)

> plot(density(x,bw="nrd"),main="",ylim=c(0,0.65),xlab="X",lty=2,lwd=2)
> lines(density(x,bw="ucv"),lwd=2,main="",ylim=c(0,0.75),xlab="X")
> lines(a,Xr,col=2,lwd=1,lty=2)
```



(a) Normal reference rule **(b)** LSCV

Figure 2.28: Kernel density estimates (Gaussian kernel) for the data set generated by the density $0.5N(0,1) + 0.1\sum_{i=1}^{4} N(i/2 - 1,0.1)$ for a sample size $n = 3000$. (a) The bandwidth is obtained by the normal reference rule (2.14) (dashed bold line) and (b) least squares cross-validation (2.16) (bold line). For comparison the real density is shown as a dashed line.

2.7.2 Real data examples

Consider the famous Old Faithful geyser data set with $n = 272$. This data set represents the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Figure 2.29 shows the kernel density estimate based on the reference rule (dashed) and least squares cross-validation method (full). The R code is given below:

```
> data(faithful)
> attach(faithful)
> x <- faithful$eruptions
> plot(density(x,bw = "nrd"), main="", ylim=c(0,0.65), xlab="Eruption time (min)", lty=2, lwd=2)
> lines(density(x,bw = "ucv"), lwd=2)
```

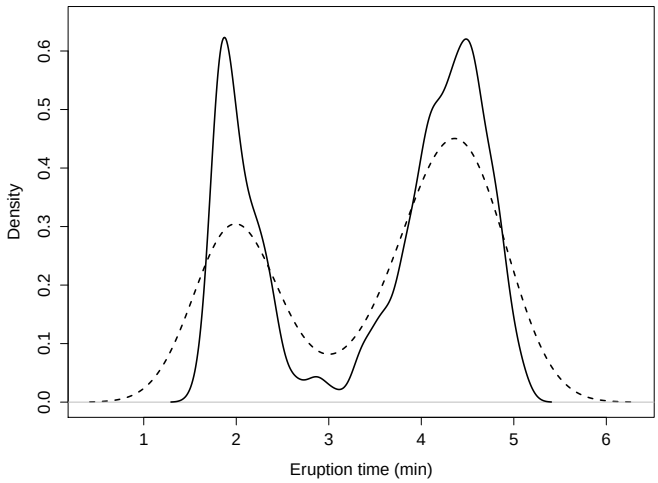


Figure 2.29: Kernel density estimate of the duration of the eruption for the Old Faithful geyser in Yellowstone National Park for two bandwidth selection methods i.e., normal reference rule (dashed) and least squares cross-validation (full).

A second example is the Suicide data set which comprises the lengths of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks reported by Copas and Fryer (1980). Figure 2.30 shows the result for the two bandwidth selection methods. Do you think this is a good density estimate?

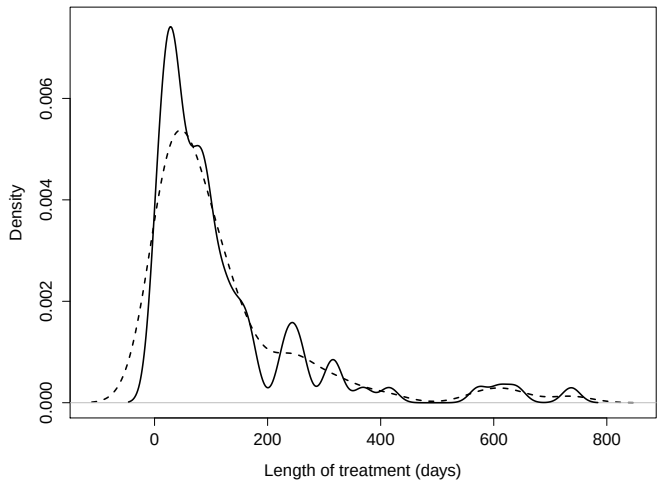


Figure 2.30: Kernel density estimate of the lengths of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks for two bandwidth selection methods i.e., normal reference rule (dashed) and least squares cross-validation (full).

Chapter 3

Multivariate density estimation

3.1 Multivariate histograms

Given a sample from $f_X(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^d$, the histogram is determined by a partition of the space. Consider a regular partition by hyper-rectangles of size $h_1 \times h_2 \times \cdots \times h_d$. Choosing hypercubes as bins would be sufficient if the data were properly scaled, but in general that will not be the case. Further improvements may be obtained by considering nonregular or rotated bins. See the book of Scott (1992) for an overview of multivariate density estimation.

Consider a generic hyper-rectangular bin labeled B_k containing ν_k points with $\sum_{k=1}^{n} \nu_k = n$. The multivariate histogram is defined as

$$\widehat{f}_X(\mathbf{x}) = \frac{\nu_k}{n h_1 h_2 \cdots h_d} \qquad \text{for } \mathbf{x} \in B_k.$$

Similar to the univariate case, the bias and variance are (for $\mathbf{x} \in B_k$)

$$\mathbf{E}[\widehat{f}_X(\mathbf{x})] - f_X(\mathbf{x}) = \frac{p_k}{h_1 h_2 \cdots h_d} - f_X(\mathbf{x}) \qquad \text{with } p_k = \int_{B_k} f_X(\mathbf{t})\, d\,\mathbf{t}$$

and

$$\mathbf{Var}[\widehat{f}_X(\mathbf{x})] = \frac{p_k(1-p_k)}{n(h_1 h_2 \cdots h_d)^2}.$$

As in the univariate case, approximations of the exact bias and variance expressions can be obtained, yielding an asymptotic expression for the MISE

$$\text{AMISE}(\widehat{f}_X) = \frac{1}{n h_1 h_2 \cdots h_d} + \frac{1}{12} \sum_{i=1}^{d} h_i^2 R(f_X^{(i)}),$$

where $f_X^{(i)}$ denotes the partial derivative of f_X with respect to x_i, the i^{th} component of the vector $\mathbf{x}$. Minimizing $\text{AMISE}(\widehat{f}_X)$ over $h_k, k = 1, \ldots, d$ yields

$$h_{k,\,\text{AMISE}} = \{R(f_X^{(k)})\}^{-1/2} \left\{ 6 \prod_{i=1}^{d} \{R(f_X^{(i)})\}^{1/2} \right\}^{1/(2+d)} n^{-1/(2+d)}.$$

Plugging the asymptotically optimal bin width in the AMISE expression gives

$$\inf_{h_{k,\,\text{AMISE}}} \text{AMISE}(\widehat{f}_X) = \frac{6^{2/(2+d)}}{4} \left\{ \prod_{i=1}^{d} R(f_X^{(i)}) \right\}^{1/(2+d)} n^{-2/(2+d)} = O(n^{-2/(2+d)}).$$

A quick and simple bin width $(h_1, \ldots, h_d)$ selection rule is obtained by referring to a multivariate normal density $N(0,\Sigma)$ with $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ yielding

$$h_{k,\,\text{AMISE}} = 2 \cdot 3^{1/(2+d)} \pi^{d/(4+2d)} \sigma_k n^{-1/(2+d)}.$$

This leads to the practical bin width selection rule

$$\widehat{h}_{k,\,\text{AMISE}} \approx 3.5 \widehat{\sigma}_k n^{-1/(2+d)}.$$

3.2 Multivariate kernel density estimation

In its most general form, the d-dimensional kernel density estimator is given by

$$\widehat{f}_X(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

where $\mathbf{H}$ is a symmetric positive definite $d \times d$ matrix called the bandwidth matrix and

$$K_{\mathbf{H}}(\mathbf{x}) = |\,\mathbf{H}\,|^{-1/2} K(\mathbf{H}^{-1/2}\,\mathbf{x}),$$

with K a d-variate kernel function satisfying

$$\int K(\mathbf{x})\, d\,\mathbf{x} = 1.$$

The kernel function is often taken to be a d-variate probability density function. There are two common techniques for generating multivariate kernels from a symmetric univariate kernel κ

1. *product kernel* $K^p(\mathbf{x}) = \prod_{i=1}^{d} \kappa(x_i)$
2. *spherically or radially symmetric kernel* $K^S(\mathbf{x}) = c_{k,d}\kappa\big\{(\mathbf{x}^T\,\mathbf{x})^{1/2}\big\}$ with $c_{k,d}^{-1} = \int \kappa\big\{(\mathbf{x}^T\,\mathbf{x})^{1/2}\big\}\, d\,\mathbf{x}$.

A popular choice for K is the standard d-variate normal density

$$K(\mathbf{x}) = (2\pi)^{-d/2} \exp\!\left(-\frac{1}{2}\,\mathbf{x}^T\,\mathbf{x}\right).$$

A simpler form of the multivariate kernel density estimator is obtained by choosing $\mathbf{H}$ to be of diagonal form, i.e.

$$\mathbf{H} = \text{diag}(h_1^2, \ldots, h_d^2),$$

in which case the multivariate kernel density estimator can be re-written as

$$\widehat{f}_X(\mathbf{x}) = n^{-1}\left(\prod_{l=1}^{d} h_l\right)^{-1} \sum_{i=1}^{n} K\!\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right).$$

A further simplification follows from the restriction to take $\mathbf{H} = h^2\mathbf{I}$ with $\mathbf{I}$ the $d \times d$ identity matrix

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

Asymptotic expressions for the MISE can be worked out, as in the univariate case. In the simplest case where $\mathbf{H} = h^2\mathbf{I}$, this leads to the asymptotic MISE

$$\text{AMISE}(\widehat{f}_X) = \frac{R(K)}{nh^d} + \frac{h^4\mu_2^2}{4}\int \big\{\nabla^2 f_X(\mathbf{x})\big\}^2\, d\,\mathbf{x},$$

where

$$\nabla^2 f_X(\mathbf{x}) = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} f_X(\mathbf{x}).$$

Minimizing the AMISE yields the following optimally constant bandwidth

$$h_{\text{AMISE}} = \left\{\frac{dR(K)}{\mu_2^2 \int \big\{\nabla^2 f_X(\mathbf{x})\big\}^2\, d\,\mathbf{x}}\right\}^{1/(d+1)} n^{-1/(d+4)}$$

leading to

$$\inf_{h_{k,\,\text{AMISE}}} \text{AMISE}(\widehat{f}_X) = O(n^{-4/(d+4)}).$$

Chapter 4

Nonparametric Regression

4.1 Introduction

Consider (X,Y) a bivariate random variable. Y will be called the response variable and X the covariate. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be an independent identically distributed sample from (X,Y). Consider the model

$$Y_i = m(X_i) + e_i, \qquad i = 1,\ldots,n,$$

with design variable X_i and error term e_i. Examples of parametric regression models are:

- linear regression model
$$m(X_i) = \beta_0 + \beta_1 X_i$$
where β_0 and β_1 are unknown parameters
- quadratic regression model
$$m(X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$
where β_0, β_1 and β_2 are unknown parameters
- nonlinear regression model
$$m(X_i) = \frac{\beta_0}{1 + \beta_1 \exp(-\beta_2 X_i)}$$
where β_0, β_1 and β_2 are unknown parameters.

In a nonparametric regression model no assumptions are made on the form of the function m.

Example 4.1 (Motorcycle data) *The data concern 133 observations of a variable X which is the time (in milliseconds) after a simulated impact with motorcycles, and a variable Y which is the head acceleration (in g) of a PMTO (post mortem human test object). Figure 4.1a shows the local cubic kernel estimate on the Motorcycle data set.*

Example 4.2 (Old Faithful geyser data) *A version of the eruptions data from the Old Faithful geyser in Yellowstone National Park, Wyoming. This version comes from Azzalini and Bowman (1990) and is of continuous measurement from August 1 to August 15, 1985. Figure 4.1b shows a local linear fit to the data.*

In nonparametric regression one makes the distinction between a fixed design model and a random design model.

- In the fixed design case the response variables are assumed to satisfy
$$Y_i = m(x_i) + e_i, \qquad i = 1,\ldots,n,$$
where the $x_1,\ldots,x_n$ are **nonrandom** numbers and $e_1,\ldots,e_n$ are independent random variables with
$$\mathbf{E}[e_i] = 0 \qquad \text{and} \qquad \mathbf{Var}[e_i] = \sigma^2(x_i). \tag{4.1}$$

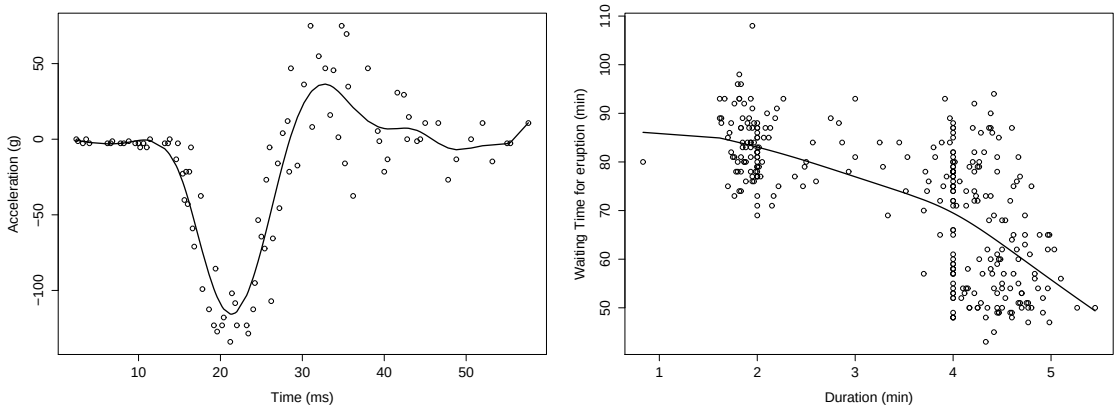


Figure 4.1: (a) Scatterplot of the Motorcycle data set with local cubic kernel estimate. (b) Scatterplot of the Old Faithful geyser data set with local linear kernel estimate.

We call m the mean regression function, or simply the regression function, since from model (4.1)

$$\mathbf{E}[Y_i] = m(x_i)$$

while

$$\mathbf{Var}[Y_i] = \sigma^2(x_i)$$

is called the variance function. The following distinction can now be made:

- $\sigma^2(x_i) = \sigma^2$: homoscedasticity
- heteroscedasticity: $\sigma^2(x_i)$

Hence in the fixed design context the nonrandom numbers are chosen by the experimentor. Some special cases of fixed designs are regular design and equally spaced design.

- The random design regression model is given by

$$Y_i = m(X_i) + \sigma(X_i)e_i, \qquad i = 1, \ldots, n,$$

where, conditional on $X_1, \ldots, X_n$, the e_i are independent random variables with

$$\mathbf{E}[e_i|X = x] = 0 \qquad \text{and} \qquad \mathbf{Var}[e_i|X = x] = 1.$$

In this random design context we have that

$$\begin{aligned} \mathbf{E}[Y|X = x] &= \mathbf{E}[m(X) + \sigma(X)e|X = x] \\ &= m(x) + \sigma(x)\,\mathbf{E}[e|X = x] \\ &= m(x), \end{aligned}$$

and

$$\begin{aligned} \mathbf{Var}[Y|X = x] &= \mathbf{E}[Y^2|X = x] - m^2(X) \\ &= \mathbf{E}[m^2(X) + \sigma^2(X)e^2 + 2m(X)\sigma(X)e|X = x] - m^2(X) \\ &= \sigma^2(x). \end{aligned}$$

Hence $m(x)$ is the conditional mean of Y given $X = x$ and $\sigma^2(x)$ is the conditional variance of Y given $X = x$.

4.2 Nadaraya-Watson regression estimator

We have that

$$m(x) = \mathbf{E}[Y|X = x] = \int y f_{Y|X}(x,y)\, dy = \frac{1}{f_X(x)} \int y f_{XY}(x,y)\, dy. \tag{4.2}$$

The unknown quantity $f_{XY}(x,y)$ can be estimated by a bivariate kernel density estimator with product kernel

$$\widehat{f}_{XY}(x,y) = \frac{1}{nhh^\star} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h^\star}\right),$$

with bandwidth h and $h^\star$ in the X-direction and Y-direction respectively. We can estimate $\int y f_{XY}(x,y)\, dy$ by

$$\begin{aligned}
\int y \widehat{f}_{XY}(x,y)\, dy &= \frac{1}{nhh^\star} \sum_{i=1}^{n} \int y K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h^\star}\right) dy \\
&= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{1}{h^\star} \int y K\left(\frac{y - Y_i}{h^\star}\right) dy \\
&= \frac{h^\star}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \frac{1}{h^\star} \int (Y_i + uh^\star) K(u)\, du \\
&= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i,
\end{aligned}$$

if $\int K(u)\, du = 1$ and $\int uK(u)\, du = 0$. By replacing $f_X(x)$ by its kernel density estimator in (4.2) we obtain the Nadaraya-Watson kernel regression estimator (independently introduced by Nadaraya (1964) and Watson (1964))

$$\widehat{m}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x - X_i}{h}\right)}. \tag{4.3}$$

Remark 4.1 *Note that this estimator can be written in the following form*

$$\widehat{m}(x) = \sum_{i=1}^{n} \left(\frac{w_i}{\sum_{j=1}^{n} w_j}\right) Y_i \quad \textit{with} \quad w_i = K\left(\frac{x - X_i}{h}\right).$$

This is a linear combination of the Y_i's. In general, an estimator of the form

$$\sum_{i=1}^{n} W_i(x; X_1, \ldots, X_n) Y_i$$

is called a linear smoother.

Next, we will show that the Nadaraya-Watson estimator (4.3) is a consistent estimator for $m(x)$. Let

$$\widehat{m}(x) = \sum_{i=1}^{n} \frac{\frac{1}{nh} K\left(\frac{x - X_i}{h}\right) Y_i}{\frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_i}{h}\right)} \equiv \frac{\widehat{r}(x)}{\widehat{f}_X(x)}$$

and $m(x) \equiv \dfrac{r(x)}{f_X(x)}$. From Chapter 2 we already know that if $f_X(\cdot)$ is continuous at x, $h \to 0$, $nh \to \infty$ as $n \to \infty$ and some conditions on the kernel function (see Theorem 2.3) then

$$\widehat{f}_X(x) \xrightarrow{p} f_X(x)$$

and (if f_X has at least two derivatives)

$$\begin{aligned}\mathbf{E}[\widehat{f}_X(x)] &= f_X(x) + \frac{h^2}{2}\mu_2 f_X''(x) + o(h^2) \\ \mathbf{Var}[\widehat{f}_X(x)] &= \frac{R(K)}{nh} f_X(x) + o\left(\frac{1}{nh}\right).\end{aligned}$$

For the bias of $\widehat{r}(x)$ we have

$$\begin{aligned}\mathbf{E}[\widehat{r}(x)] &= \frac{1}{h}\mathbf{E}\left[K\left(\frac{x-X}{h}\right)Y\right] \\ &= \frac{1}{h}\mathbf{E}\left\{\mathbf{E}\left[K\left(\frac{x-X}{h}\right)Y|X\right]\right\} \\ &= \frac{1}{h}\mathbf{E}\left\{K\left(\frac{x-X}{h}\right)\mathbf{E}[Y|X]\right\} \\ &= \frac{1}{h}\mathbf{E}\left\{K\left(\frac{x-X}{h}\right)m(X)\right\} \\ &= \frac{1}{h}\int K\left(\frac{x-y}{h}\right)m(y)f_X(y)\,dy \\ &= \frac{1}{h}\int K\left(\frac{x-y}{h}\right)r(y)\,dy \\ &= \frac{1}{h}\int K\left(\frac{u}{h}\right)r(x-u)\,du.\end{aligned}$$

Applying Bochner's lemma (Lemma 2.1) yields

$$\lim_{n\to\infty}\mathbf{E}[\widehat{r}(x)] = r(x)$$

provided that $m(\cdot)$ and $f(\cdot)$ are continuous at the point x, $h \to 0$ as $n \to \infty$ and under the necessary assumptions on K. Also, using a Taylor expansion yields (if $m(\cdot)$ and $f(\cdot)$ have at least two derivatives)

$$\mathbf{E}[\widehat{r}(x)] = r(x) + \frac{h^2}{2}\mu_2 r''(x) + o(h^2). \tag{4.4}$$

For the variance of $\widehat{r}(x)$ we have

$$\begin{aligned}\mathbf{Var}[\widehat{r}(x)] &= \mathbf{Var}\left[\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)Y_i\right] \\ &= \frac{1}{n}\mathbf{Var}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)Y\right] \\ &= \frac{1}{n}\left\{\mathbf{E}\left[\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)Y^2\right] - \mathbf{E}^2\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)Y\right]\right\}.\end{aligned}$$

For the first term we have

$$\begin{aligned}\mathbf{E}\left[\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)Y^2\right] &= \mathbf{E}\left\{\mathbf{E}\left[\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)Y^2|X\right]\right\} \\ &= \mathbf{E}\left\{\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)\mathbf{E}[Y^2|X]\right\} \\ &= \mathbf{E}\left\{\frac{1}{h^2}K^2\left(\frac{x-X}{h}\right)\left(\mathbf{Var}[Y|X] + m^2(X)\right)\right\} \\ &= \frac{1}{h^2}\mathbf{E}\left\{K^2\left(\frac{x-X}{h}\right)\sigma^2(X)\right\} + \frac{1}{h^2}\mathbf{E}\left\{K^2\left(\frac{x-X}{h}\right)m^2(X)\right\}\end{aligned}$$

Putting everything together we have

$$nh\,\mathbf{Var}[\widehat{r}(x)] = \frac{1}{h}\,\mathbf{E}\left\{K^2\left(\frac{x-X}{h}\right)\sigma^2(X)\right\} + \frac{1}{h}\,\mathbf{E}\left\{K^2\left(\frac{x-X}{h}\right)m^2(X)\right\} - h\big(\mathbf{E}[\widehat{r}(x)]\big)^2.$$

Taking the limit for $n \to \infty$ gives (and by applying Bochner's lemma (Lemma 2.1))

$$\lim_{n\to\infty} nh\,\mathbf{Var}[\widehat{r}(x)] = \sigma^2(x)f_X(x)R(K) + m^2(x)f_X(x)R(K) + 0$$

for $\sigma(\cdot), m(\cdot)$ and $f_X(\cdot)$ continuous at the point x. For $h \to\ 0$ and $nh \to \infty$ as $n \to \infty$, the MSE of $\widehat{r}(x)$ is

$$\begin{aligned}\mathbf{E}[\widehat{r}(x) - r(x)]^2 &= \frac{h^4}{4}\mu_2^2\{r''(x)\}^2 + \frac{1}{nh}\{\sigma^2(x) + m^2(x)\}f_X(x)R(K) + o\left(h^4 + \frac{1}{nh}\right)\\ &\to\ 0, \text{ as } n \to \infty.\end{aligned}$$

Hence, by Chebyshev's inequality we have that for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}[|\widehat{r}(x) - r(x)| \geq \epsilon] = 0 \quad \text{or} \quad \widehat{r}(x) \xrightarrow{\text{p}} r(x)$$

and from Chapter 2

$$\lim_{n\to\infty} \mathbf{P}[|\widehat{f}_X(x) - f_X(x)| \geq \epsilon] = 0 \quad \text{or} \quad \widehat{f}_X(x) \xrightarrow{\text{p}} f_X(x).$$

Then by Slutsky's theorem we have

$$\widehat{m}(x) = \frac{\widehat{r}(x)}{\widehat{f}_X(x)} \xrightarrow{\text{p}} \frac{r(x)}{f_X(x)} = m(x),$$

provided that $f_X(x) > 0$. Next, we can show the bias and variance expressions for the Nadaraya-Watson estimator. First, consider the following expansion

$$\begin{aligned}\widehat{m}(x) - m(x) &= \left(\frac{\widehat{r}(x)}{\widehat{f}_X(x)} - m(x)\right)\left\{\frac{\widehat{f}_X(x)}{f_X(x)} + \left(1 - \frac{\widehat{f}_X(x)}{f_X(x)}\right)\right\}\\ &= \frac{\widehat{r}(x) - m(x)\widehat{f}_X(x)}{f_X(x)} + \frac{1}{f_X(x)}(\widehat{m}(x) - m(x))(f_X(x) - \widehat{f}_X(x)).\end{aligned} \tag{4.5}$$

For the bias part we have

$$\begin{aligned}\mathbf{E}[\widehat{m}(x) - m(x)] &= \mathbf{E}\left[\frac{\widehat{r}(x) - m(x)\widehat{f}_X(x)}{f_X(x)}\right] + \frac{1}{f_X(x)}\,\mathbf{E}[(\widehat{m}(x) - m(x))(f_X(x) - \widehat{f}_X(x))]\\ &= B_1 + B_2.\end{aligned}$$

For the first term and using (4.4),

$$\begin{aligned}B_1 &= \frac{1}{f_X(x)}\left\{r(x) + \frac{h^2}{2}\mu_2 r''(x) - m(x)\big[f(x) + \frac{h^2}{2}f''(x)\mu_2\big]\right\} + o(h^2)\\ &= m(x) + \frac{h^2}{2}\mu_2\frac{r''(x)}{f_X(x)} - m(x) - m(x)\frac{h^2}{2}\mu_2\frac{f_X''(x)}{f_X(x)} + o(h^2)\\ &= \frac{h^2}{2}\mu_2\frac{m''(x)f_X(x) + 2m'(x)f_X'(x) + m(x)f_X''(x) - m(x)f_X''(x)}{f_X(x)} + o(h^2)\\ &= \frac{h^2}{2}\mu_2\left(2m'(x)\frac{f_X'(x)}{f_X(x)} + m''(x)\right) + o(h^2)\end{aligned} \tag{4.6}$$

and the second term gives

$$\begin{aligned}B_2 &\leq \frac{1}{f_X(x)}\big|\mathbf{E}[(\widehat{m}(x) - m(x))(f_X(x) - \widehat{f}_X(x))]\big|\\ &\overset{\text{Cauchy-Schwartz}}{\leq} \frac{1}{f_X(x)}\sqrt{\mathbf{E}[\widehat{m}(x) - m(x)]^2}\sqrt{\mathbf{E}[f_X(x) - \widehat{f}_X(x)]^2}.\end{aligned}$$

The asymptotic order of MSE$(\widehat{m}(x))$ will be the same as MSE$(\widehat{r}(x))$ i.e., for a bandwidth $h = O(n^{-1/5})$ we have MSE$(\widehat{m}(x)) = O(n^{-4/5})$. Also, for a bandwidth $h = O(n^{-1/5})$ we know from Chapter 2 that MSE$(\widehat{f}_X(x)) = O(h^4 + 1/(nh)) = O(n^{-4/5})$. Then, for the term B_2 we have

$$B_2 = O(\sqrt{n^{-4/5}})O(\sqrt{n^{-4/5}}) = O(n^{-4/5}). \tag{4.7}$$

Combining (4.6) and (4.7) gives the bias of the Nadaraya-Watson estimator

$$\text{bias}[\widehat{m}(x)] = \frac{h^2}{2}\mu_2\bigg(2m'(x)\frac{f'_X(x)}{f_X(x)} + m''(x)\bigg) + o(h^2) + O(n^{-4/5}) = \frac{h^2}{2}\mu_2\bigg(2m'(x)\frac{f'_X(x)}{f_X(x)} + m''(x)\bigg) + o(h^2).$$

Next, by using (4.5)

$$\begin{aligned}
\textbf{Var}[\widehat{m}(x)] &= \textbf{Var}\bigg[\frac{\widehat{r}(x) - m(x)\widehat{f}_X(x)}{f_X(x)} + \frac{1}{f_X(x)}\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}\bigg] \\
&= \frac{1}{f_X^2(x)}\textbf{Var}\big[\widehat{r}(x) - m(x)\widehat{f}_X(x)\big] + \frac{1}{f_X^2(x)}\textbf{Var}\big[\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}\big] \\
&+ \frac{2}{f_X^2(x)}\textbf{Cov}[\widehat{r}(x) - m(x)\widehat{f}_X(x), \{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}] \\
&= V_1 + V_2 + V_3.
\end{aligned} \tag{4.8}$$

For each of the three terms we have and using independence

$$\begin{aligned}
V_1 &= \frac{1}{f_X^2(x)}\textbf{Var}\big[\widehat{r}(x) - m(x)\widehat{f}_X(x)\big] \\
&= \frac{1}{f_X^2(x)}\textbf{Var}\bigg[\frac{1}{nh}\sum_{i=1}^n K\bigg(\frac{x - X_i}{h}\bigg)Y_i - \frac{1}{nh}\sum_{i=1}^n K\bigg(\frac{x - X_i}{h}\bigg)m(x)\bigg] \\
&= \frac{1}{nh^2 f_X^2(x)}\textbf{Var}\bigg[K\bigg(\frac{x - X}{h}\bigg)\{Y - m(x)\}\bigg] \\
&= \frac{1}{nh^2 f_X^2(x)}\mathbf{E}\bigg[K^2\bigg(\frac{x - X}{h}\bigg)\{Y - m(x)\}^2\bigg] - \frac{1}{nh^2 f_X^2(x)}\mathbf{E}^2\bigg[K\bigg(\frac{x - X}{h}\bigg)\{Y - m(x)\}\bigg] \\
&= V_{11} + V_{12}.
\end{aligned}$$

Analyzing each term separate, using Taylor series and $h \to 0$ yields

$$\begin{aligned}
V_{11} &= \frac{1}{nh^2 f_X^2(x)}\mathbf{E}\bigg[K^2\bigg(\frac{x - X}{h}\bigg)\mathbf{E}[\{Y^2 - 2Ym(x) + m^2(x)\}|X]\bigg] \\
&= \frac{1}{nh^2 f_X^2(x)}\mathbf{E}\bigg[K^2\bigg(\frac{x - X}{h}\bigg)\{\sigma^2(X) + m^2(X) - 2m(X)m(x) + m^2(x)\}\bigg] \\
&= \frac{1}{nhf_X^2(x)}\int K^2(u)\sigma^2(x - uh)f_X(x - uh)\,du + \frac{1}{nhf_X^2(x)}\int K^2(u)m^2(x - uh)f_X(x - uh)\,du \\
&\quad - \frac{2m(x)}{nhf_X^2(x)}\int K^2(u)m(x - uh)f_X(x - uh)\,du + \frac{m^2(x)}{nhf_X^2(x)}\int K^2(u)f_X(x - uh)\,du \\
&= \frac{1}{nhf_X^2(x)}\int K^2(u)\{\sigma^2(x) + o(1)\}\{f_X(x) + o(1)\}\,du + \frac{1}{nhf_X^2(x)}\int K^2(u)\{m^2(x) + o(1)\}\{f_X(x) + o(1)\}\,du \\
&\quad - \frac{2m(x)}{nhf_X^2(x)}\int K^2(u)\{m(x) + o(1)\}f_X(x - uh)\,du + \frac{m^2(x)}{nhf_X^2(x)}\int K^2(u)\{f_X(x) + o(1)\}\,du \\
&= \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + \frac{m^2(x)}{nhf_X(x)}\int K^2(u)\,du - \frac{2m^2(x)}{nhf_X(x)}\int K^2(u)\,du + \frac{m^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\bigg(\frac{1}{nh}\bigg) \\
&= \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\bigg(\frac{1}{nh}\bigg)
\end{aligned}$$

and

$$\begin{aligned} V_{12} &= \frac{1}{nh^2 f_X^2(x)}\,\mathbf{E}^2\left[K\left(\frac{x-X}{h}\right)\mathbf{E}[\{Y-m(x)\}|X]\right] \\ &= \frac{1}{nf_X^2(x)}\left[\int K(u)\{m(x-uh)-m(x)\}f_X(x-uh)\,du\right]^2 \\ &= \frac{1}{nf_X^2(x)}\left[\int K(u)\{m(x)+o(1)-m(x)\}\{f_X(x)+o(1)\}\,du\right]^2 = o\left(\frac{1}{n}\right). \end{aligned}$$

Hence, for the first term V_1 we have

$$V_1 = \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right) + o\left(\frac{1}{n}\right) = \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right).$$

To obtain the order of the next two terms in (4.8) we need the following two technical lemmas.

Lemma 4.1 *For any two random variables X and Y with finite variances*

$$\mathbf{Var}[X \pm Y] \le 2\,\mathbf{Var}\,X + 2\,\mathbf{Var}\,Y.$$

PROOF. From the variance of the sum of two random variables we have

$$0 \le \mathbf{Var}[X \pm Y] = \mathbf{Var}\,X + \mathbf{Var}\,Y \pm 2\,\mathbf{Cov}[X,Y]$$

and

$$|2\,\mathbf{Cov}[X,Y]| \le \mathbf{Var}\,X + \mathbf{Var}\,Y.$$

Substituting the latter equation into the first gives the result. ■

Lemma 4.2 *Let X and Y be any two random variables with $\mathbf{E}[X] < \infty$, $\mathbf{Var}\,X < \infty$, $\mathbf{Var}\,Y < \infty$. Further, assume there exists a $B \ge 0$ such that $\mathbf{P}[|Y| \le B] = 1$, then*

$$\mathbf{Var}[XY] \le 2\|Y\|_\infty^2\,\mathbf{Var}\,X + 2(\mathbf{E}[X])^2\,\mathbf{Var}\,Y$$

where $\|Y\|_\infty = \inf\{B \ge 0 : \mathbf{P}[|Y| \le B] = 1\}$.

PROOF. Using Lemma 4.1 with $X = (X - \mathbf{E}[X])Y$ and $Y = \mathbf{E}[X]Y$ yields

$$\begin{aligned} \mathbf{Var}[XY] &\le 2\,\mathbf{Var}[(X-\mathbf{E}[X])Y] + 2\,\mathbf{Var}[\mathbf{E}[X]Y] \\ &= 2\,\mathbf{Var}[(X-\mathbf{E}[X])Y] + 2(\mathbf{E}[X])^2\,\mathbf{Var}[Y]. \end{aligned}$$

The first term is (with the 2 omitted)

$$\begin{aligned} \mathbf{Var}[(X-\mathbf{E}[X])Y] &= \mathbf{E}\{[(X-\mathbf{E}[X])Y]^2\} - \{\mathbf{E}[(X-\mathbf{E}[X])Y]\}^2 \\ &\le \mathbf{E}\{[(X-\mathbf{E}[X])Y]^2\} \\ &\le \mathbf{E}\{|X-\mathbf{E}[X]|^2|Y|^2\} \\ &\le \|Y\|_\infty^2\,\mathbf{Var}\,X. \end{aligned}$$

■

Next, our goal is to show that the order of the two last terms in (4.8) are of lower order than the first term. Using Lemma 4.2, the order second term V_2 is

$$\begin{aligned} V_2 &= \frac{1}{f_X^2(x)}\,\mathbf{Var}\big[(\widehat{m}(x)-m(x))(f_X(x)-\widehat{f}_X(x))\big] \\ &\le \frac{\big[2(\inf\{B \ge 0 : \mathbf{P}[|f_X(x)-\widehat{f}_X(x)| \le B] = 1\})^2\,\mathbf{Var}[\widehat{m}(x)] + 2\{\mathbf{E}[\widehat{m}(x)-m(x)]\}^2\,\mathbf{Var}[\widehat{f}_X(x)]\big]}{f_X^2(x)}. \end{aligned}$$

By Theorem 2.6 we know that there exist a B such that $\inf\{B \geq 0 : \mathbf{P}[|f_X(x) - \widehat{f}_X(x)| \leq B] = 1\})$. In fact, Theorem 2.6 states that B goes to zero almost surely and hence the first term in the inequality is $o(1)$. Consequently

$$V_2 = o(1)O\Big(\frac{1}{nh}\Big) + O(h^4)O\Big(\frac{1}{nh}\Big) = o\Big(\frac{1}{nh}\Big).$$

Finally, for the third term (using the covariance inequality)

$$\begin{aligned} V_3 &= \frac{2}{f_X^2(x)}\,\mathbf{Cov}[\widehat{r}(x) - m(x)\widehat{f}_X(x), \{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}] \\ &= O\Big(\sqrt{\mathbf{Var}[\widehat{r}(x) - m(x)\widehat{f}_X(x)]}\sqrt{\mathbf{Var}[\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}]}\Big) \\ &= O\Big(\frac{1}{\sqrt{nh}}\Big)o\Big(\frac{1}{\sqrt{nh}}\Big) = o\Big(\frac{1}{nh}\Big). \end{aligned}$$

Combing the three terms together gives the variance of the Nadaraya-Watson estimator

$$\mathbf{Var}[\widehat{m}(x)] = \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\Big(\frac{1}{nh}\Big).$$

4.3 Local polynomial regression

There exist a vast number of methods to construct nonparametric regression estimates e.g., splines, wavelets, support vector machines, local polynomial regression, Nadaraya-Watson regression, Gasser-Müller estimator, Priestly-Chao estimator, orthogonal series estimator, k nearest neighbors, etc. A good overview can be found in Wasserman (2006). A thorough theoretical study of nonparametric regression is given in Györfi et al. (2002). In what follows we will focus on local polynomial regression which is a very popular method in statistics. Most of this material is based on Fan and Gijbels (1996).

4.3.1 Local polynomial regression framework

Consider the bivariate data $(X_1,Y_1),\ldots,(X_n,Y_n)$, which form an independent and identically distributed (i.i.d.) sample from a population (X,Y). Our interest is to estimate the regression function $m(x_0) = \mathbf{E}[Y|X = x_0]$ and its derivatives $m'(x_0), m''(x_0),\ldots,m^{(p)}(x_0)$. To understand the estimation methodology, we can regard the data as being generated from the model

$$Y = m(X) + \sigma(X)e, \tag{4.9}$$

where $\mathbf{E}[e] = 0, \mathbf{Var}[e] = 1$ and X and e are independent. We always denote the conditional variance of Y given $X = x_0$ by $\sigma^2(x_0)$ and the marginal density of X i.e., the *design density*, by f_X.

Suppose that the $(p+1)$th derivative of the regression function m at the point x_0 exists. We then approximate the unknown regression function m locally by a polynomial of order p. Figure 4.2 illustrates the idea of local constant ($p = 0$) and linear kernel regression ($p = 1$).
A Taylor expansion gives, for x in the neighborhood of x_0,

$$\begin{aligned} m(x) &= m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p + o\big(|x - x_0|^p\big) \\ &= \sum_{j=0}^{p}\frac{m^{(j)}(x_0)}{j!}(x - x_0)^j + o\big(|x - x_0|^p\big) \\ &=: \sum_{j=0}^{p}\beta_j(x - x_0)^j + o\big(|x - x_0|^p\big). \end{aligned} \tag{4.10}$$

This polynomial is fitted locally by the following weighted least squares regression problem:

$$\min_{\beta\in\mathbb{R}^{p+1}}\sum_{i=1}^{n}\big\{Y_i - \sum_{j=0}^{p}\beta_j\,(X_i - x_0)^j\big\}^2 K_h(X_i - x_0), \tag{4.11}$$

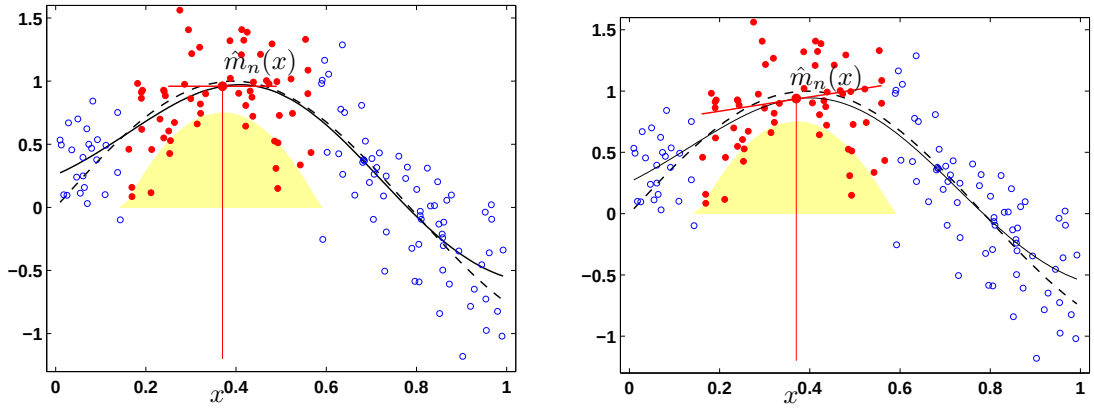


Figure 4.2: 100 pairs (X_i,Y_i) are generated at random from $Y = \sin(4X)$ (dashed line) with Gaussian errors $e \sim \mathcal{N}(0,1/3)$ and $X \sim \mathcal{U}[0,1]$. The dot around 0.38 (vertical line) is the fitted constant $\hat{m}_n(x)$, and the full circles indicate those observations contributing to the fit at x. The solid region indicates the weights assigned to observations according the Epanechnikov kernel. (a) The full NW estimate is shown by the full line. (b) The full local linear estimate is shown by the full line.

where β_j are the solutions to the weighted least squares problem, $\beta = (\beta_0,\beta_1,\ldots,\beta_p)^T$, $h > 0$ is the bandwidth controlling the size of the local neighborhood and $K_h(\cdot) = K(\cdot/h)/h$ with K a kernel function assigning weights to each point. From the Taylor expansion (4.10) it is clear that $\widehat{m}^{(\nu)}(x_0) = \nu!\widehat{\beta}_\nu$ is an estimator for the νth order derivative $m^{(\nu)}(x_0)$, $\nu = 0,1,\ldots,p$.

It is often more convenient to work with matrix notation. Denote by $\mathbf{X}$ the design matrix of problem (4.11):

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

and put

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{and} \qquad \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix}.$$

Further, $\mathbf{W}$ is the $n \times n$ diagonal matrix of weights

$$\mathbf{W} = \text{diag}\{K_h(X_i - x_0)\}.$$

The weighted least squares problem (4.11) can be written as

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\,\beta)^T\, \mathbf{W} (\mathbf{Y} - \mathbf{X}\,\beta),$$

with $\beta = (\beta_0,\ldots,\beta_p)^T$. The solution vector is provided by weighted least squares theory and is given by

$$\widehat{\beta} = (\mathbf{X}^T\, \mathbf{W}\, \mathbf{X})^{-1}\, \mathbf{X}^T\, \mathbf{W}\, \mathbf{Y}\,. \tag{4.12}$$

There are several important issues which have to be discussed. First of all there is the choice of the bandwidth h, which plays a very crucial role (see also the binwidth for histograms). A too large bandwidth under-parametrizes the regression function, causing a large modelling bias, while a too small bandwidth over-parametrizes the unknown function and results in noisy estimates. In what follows, we will show how to obtain ideal theoretical bandwidth choices. As we will see, this theoretical choice is not directly usable since it depends on unknown quantities. Finding a practical procedure for selecting the bandwidth parameter is one of the most important tasks.

Another issue in local polynomial fitting is the choice of the order of the local polynomial. Since the modelling bias is primarily controlled by the bandwidth, this issue is less crucial however. For a given bandwidth h, a large value of p would expectedly reduce the modelling bias, but would cause a large variance and considerable computational cost.

How good are the local polynomial estimators compared to other estimators? An answer to this question is provided by studying the efficiency of the local polynomial fit. It is beyond the scope of the course to prove the efficiency of local polynomial estimators, but it can be shown that local polynomial fitting is nearly optimal in an asymptotic minimax sense (Fan and Gijbels, 1996, Chapter 3).

From a computational point of view local polynomial estimators are attractive, due to their simplicity. It might be desirable however to speed up the computations especially when computing intensive procedures e.g., bandwidth selection, are to be implemented (Fan and Marron, 1994).

4.4 Advantages of local polynomial fitting

Local polynomial fitting is an attractive method both from theoretical and practical point of view. Other commonly used kernel estimators, such as the Nadaraya-Watson (NW) estimator and the Gasser-Müller (GM) estimator suffer from some drawbacks. In summary, the NW estimator leads to an undesirable form of the bias, while the GM estimator has to pay a price in variance when dealing with a random design model. Local polynomial fitting also has other advantages. The method adopts to various types of designs such as random and fixed designs, highly clustered and nearly uniform designs. Furthermore, there is an absence of boundary effects: the bias at the boundary stays automatically of the same order as in the interior, without the use of specific boundary kernels! The local polynomial approximation method is appealing on general scientific grounds: the least squares principle to be applied opens the way to a wealth of statistical knowledge and thus easy generalizations.

4.4.1 Bias and variance of local polynomial fitting

When dealing with the bandwidth selection problem, a key issue is to have a good insight into bias and variance of the estimators, since a trade-off between these two quantities forms the core of many bandwidth selection criteria. The conditional bias and variance of the estimator $\widehat{\beta}$ are derived immediately from its definition in (4.12)

$$\begin{aligned}
\mathbf{E}[\widehat{\beta}\,|\mathbb{X}] &= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{m} \\
&= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}(\mathbf{m}-\mathbf{X}\,\beta+\mathbf{X}\,\beta) \\
&= \beta+(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}(\mathbf{m}-\mathbf{X}\,\beta) \\
&= \beta+(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{r} \qquad (4.13)
\end{aligned}$$

$$\begin{aligned}
\mathbf{Var}[\widehat{\beta}\,|\mathbb{X}] &= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathrm{diag}\{\sigma^2(X_i)\}\,\mathbf{W}^T\,\mathbf{X}(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}(\mathbf{X}^T\,\Sigma\,\mathbf{X})(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}, \qquad (4.14)
\end{aligned}$$

where $\mathbb{X}=(X_1,\ldots,X_n)$, $\mathbf{m}=(m(X_1),\ldots,m(X_n))^T$, $\beta=(m(x_0),\ldots,m^{(p)}(x_0)/p!)^T$, $\mathbf{r}=\mathbf{m}-\mathbf{X}\,\beta$, the vector of residuals of the local polynomial approximation and $\Sigma=\mathrm{diag}\{K_h^2(X_i-x_0)\sigma^2(X_i)\}$.

These exact bias and variance expressions are not directly usable, since they depend on unknown quantities: the residual **r**and the diagonal matrix Σ. Hence, there is a need for approximating the bias and variance. We first of all show how to derive the asymptotic expression for the conditional variance given in (4.21). Denote by $S_n=\mathbf{X}^T\,\mathbf{W}\,\mathbf{X}$ and $S_n^\star=\mathbf{X}^T\,\Sigma\,\mathbf{X}$ the $(p+1)\times(p+1)$ matrix $(S^\star_{n,j+l})_{0\le j,l\le p}$ with $S^\star_{n,j}=\sum_{i=1}^n(X_i-x_0)^jK_h^2(X_i-x_0)\sigma^2(X_i)$. Then, the conditional variance in (4.21) can be re-expressed as

$$S_n^{-1}S_n^\star S_n^{-1}, \qquad (4.15)$$

and the task is now to find the approximations for the two matrices S_n and $S_n^\star$. Since $S_{n,j}=\sum_{i=1}^n K_h(X_i-x_0)(X_i-x_0)^j$, we have that

$$\begin{aligned}
S_{n,j} &= \mathbf{E}[S_{n,j}]+\frac{S_{n,j}-\mathbf{E}[S_{n,j}]}{\sqrt{\mathbf{Var}[S_{n,j}]}}\sqrt{\mathbf{Var}[S_{n,j}]} \\
&= \mathbf{E}[S_{n,j}]+O_p\big(\sqrt{\mathbf{Var}[S_{n,j}]}\big). \qquad (4.16)
\end{aligned}$$

Because the data is i.i.d., we have

$$\begin{aligned}
\mathbf{E}[S_{n,j}] &= n\,\mathbf{E}[K_h(X-x_0)(X-x_0)^j] \\
&= \frac{n}{h}\int K\left(\frac{x-x_0}{h}\right)(x-x_0)^j f_X(x)\,dx \\
&= nh^j\int u^j K(u) f_X(x_0+uh)\,du \\
&= nh^j\int u^j K(u)(f_X(x_0)+o(1))\,du \\
&= nh^j\left[f_X(x_0)\int u^j K(u)\,du + o(1)\right] \\
&= nh^j f_X(x_0)\mu_j\left[1+o(1)\right],
\end{aligned}$$

with $\mu_j = \int u^j K(u)\,du$. Substituting the above expression into (4.16) and using the Cauchy-Schwartz inequality yields

$$S_{n,j} = nh^j f_X(x_0)\mu_j\left[1+o(1)\right] + O_p\big(\sqrt{n\,\mathbf{E}[(X-x_0)^{2j}K_h^2(X-x_0)]}\big). \tag{4.17}$$

Next we need to find the order of the last term. Similarly, we have that

$$\begin{aligned}
n\,\mathbf{E}[(X-x_0)^{2j}K_h^2(X-x_0)] &= \frac{n}{h^2}\int K^2\left(\frac{x-x_0}{h}\right)(x-x_0)^{2j} f_X(x)\,dx \\
&= nh^{2j-1}\int K^2(u)u^{2j} f_X(x_0+uh)\,du.
\end{aligned}$$

It immediately follows that

$$\begin{aligned}
S_{n,j} &= nh^j f_X(x_0)\mu_j\left[1+o(1)\right] + O_p\big(\sqrt{nh^{2j-1}}\big) \\
&= nh^j f_X(x_0)\mu_j\left[1+o(1)+O_p\big(1/\sqrt{nh}\big)\right] \\
&= nh^j f_X(x_0)\mu_j\left[1+o_p(1)\right],
\end{aligned} \tag{4.18}$$

provided that $h\to 0$ and $nh\to\infty$. Since

$$S_n = \mathbf{X}^T\mathbf{W}\mathbf{X} = \begin{pmatrix} S_{n,0} & S_{n,1} & \cdots & S_{n,p} \\ S_{n,1} & S_{n,2} & \cdots & S_{n,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p} & S_{n,p+1} & \cdots & S_{n,2p} \end{pmatrix}$$

it follows that

$$\begin{aligned}
S_n &= \begin{pmatrix} nh^0 f_X(x_0)\mu_0\left[1+o_p(1)\right] & nhf_X(x_0)\mu_1\left[1+o_p(1)\right] & \cdots & nh^p f_X(x_0)\mu_p\left[1+o_p(1)\right] \\ nh^1 f_X(x_0)\mu_1\left[1+o_p(1)\right] & nh^2 f_X(x_0)\mu_2\left[1+o_p(1)\right] & \cdots & nh^{p+1} f_X(x_0)\mu_{p+1}\left[1+o_p(1)\right] \\ \vdots & \vdots & \ddots & \vdots \\ nh^p f_X(x_0)\mu_p\left[1+o_p(1)\right] & nh^{p+1} f_X(x_0)\mu_{p+1}\left[1+o_p(1)\right] & \cdots & nh^{2p} f_X(x_0)\mu_{2p}\left[1+o_p(1)\right] \end{pmatrix} \\
&= nf_X(x_0)\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & h & 0 & \cdots & 0 \\ 0 & 0 & h^2 & & \vdots \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & h^p \end{pmatrix}\begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & h & 0 & \cdots & 0 \\ 0 & 0 & h^2 & & \vdots \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & h^p \end{pmatrix}\left[1+o_p(1)\right] \\
&= nf_X(x_0)HSH\left[1+o_p(1)\right],
\end{aligned} \tag{4.19}$$

where $H = \text{diag}\{1,h,\ldots,h^p\}$. Using similar arguments, we have that

$$\begin{aligned} S^\star_{n,j} &= nh^{j-1} f_X(x_0)\sigma^2(x_0) \int u^j K^2(u)\, du \,\big[1+o_p(1)\big] \\ &= nh^{j-1} f_X(x_0)\sigma^2(x_0)\nu_j \,\big[1+o_p(1)\big] \end{aligned}$$

and hence

$$S^\star_n = nh^{-1} f_X(x_0)\sigma^2(x_0) H S^\star H \,\big[1+o_p(1)\big] \tag{4.20}$$

with

$$S^\star = \begin{pmatrix} \nu_0 & \nu_1 & \cdots & \nu_p \\ \nu_1 & \nu_2 & \cdots & \nu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_p & \nu_{p+1} & \cdots & \nu_{2p} \end{pmatrix}.$$

Now, starting from (4.15) and using (4.19) and (4.20) we find that

$$\mathbf{Var}[\widehat{\beta}\,|\mathbb{X}] = \frac{\sigma^2(x_0)}{f_X(x_0)nh} H^{-1} S^{-1} S^\star S^{-1} H^{-1} \,\big[1+o_p(1)\big]\,,$$

and since $\widehat{m}_\nu(x_0) = \nu!\varepsilon^T_{\nu+1}\widehat{\beta}$, with $\varepsilon_{\nu+1} = (0,\ldots,0,1,0,\ldots,0)^T$ the unit vector with 1 on the $(\nu+1)$th place, the following theorem follows readily.

Theorem 4.1 (Variance of the local polynomial regression estimator) *Assume that $f_X(x_0) > 0$, $f_X(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x_0. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\widehat{m}_\nu(x_0)$ is given by*

$$\mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \varepsilon^T_{\nu+1} S^{-1} S^\star S^{-1} \varepsilon_{\nu+1} \frac{\nu!^2\sigma^2(x_0)}{f_X(x_0)nh^{1+2\nu}} + o_p\left(\frac{1}{nh^{1+2\nu}}\right). \tag{4.21}$$

Second, we derive the asymptotic expression for the bias. Here, we have to distinguish between the case that $p-\nu$ is odd and $p-\nu$ is even. Let's consider the case $p-\nu$ first. By using a Taylor expansion the conditional bias, see (4.24), $S_n^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{r}$ of $\widehat{\beta}$ can be written as

$$\begin{aligned} \text{bias}[\widehat{\beta}\,|\mathbb{X}] &= S_n^{-1}\,\mathbf{X}^T\,\mathbf{W} \left\{ \beta_{p+1} \begin{pmatrix} (X_1-x_0)^{p+1} \\ \vdots \\ (X_n-x_0)^{p+1} \end{pmatrix} + o_p\Big(\begin{pmatrix} (X_1-x_0)^{p+1} \\ \vdots \\ (X_n-x_0)^{p+1} \end{pmatrix} \Big) \right\} \\ &= S_n^{-1} \left\{ \beta_{p+1} \begin{pmatrix} S_{n,p+1} \\ \vdots \\ S_{n,2p+1} \end{pmatrix} + o_p\Big(\begin{pmatrix} nh^{p+1} \\ \vdots \\ nh^{2p+1} \end{pmatrix} \Big) \right\} = S_n^{-1} \left\{ \beta_{p+1} c_n + o_p\Big(\begin{pmatrix} nh^{p+1} \\ \vdots \\ nh^{2p+1} \end{pmatrix} \Big) \right\}, \end{aligned} \tag{4.22}$$

with $c_n = (S_{n,p+1},\ldots,S_{n,2p+1})^T$ and $\beta_{p+1} = m^{(p+1)}(x_0)/(p+1)!$. Applying (4.18) and (4.19), we obtain from (4.22)

$$\begin{aligned} \text{bias}[\widehat{\beta}\,|\mathbb{X}] &= \frac{1}{nf_X(x_0)} H^{-1} S^{-1} H^{-1} \left\{ \beta_{p+1} \begin{pmatrix} nh^{p+1} f_X(x_0)\mu_{p+1}[1+o_p(1)] \\ \vdots \\ nh^{2p+1} f_X(x_0)\mu_{2p+1}[1+o_p(1)] \end{pmatrix} + o_p\Big(\begin{pmatrix} nh^{p+1} \\ \vdots \\ nh^{2p+1} \end{pmatrix} \Big) \right\} \big[1+o_p(1)\big] \\ &= H^{-1} S^{-1} \beta_{p+1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1/h & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 1/h^p \end{pmatrix} \left\{ \begin{pmatrix} h^{p+1} & 0 & \cdots & 0 \\ 0 & h^{p+2} & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & h^{2p+1} \end{pmatrix} \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix} [1+o_p(1)] \right\} \big[1+o_p(1)\big] \\ &= H^{-1} S^{-1} c_p \beta_{p+1} h^{p+1} [1+o_p(1)]. \end{aligned} \tag{4.23}$$

The above derivation of course holds for any value of $p-\nu$, but the problem is that for $p-\nu$ even the $(\nu+1)$th element of the vector $S^{-1}c_p$ is zero. This can be easily seen by writing out the structure of the matrix S and the vector c_p, and recalling that odd order moments of a symmetric kernel are zero. Hence the main term of (4.23) is zero, and one clearly has to proceed to higher order expansions. This essentially means that in all derivations we derived to obtain (4.23) some extra terms have to be taken along. In what follows, we derive the case $p-\nu$ odd. Using (4.23), the conditional bias of the local polynomial estimator is then given by

$$\begin{aligned}
\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] &= \text{bias}[\nu!\varepsilon_{\nu+1}^T\widehat{\beta}\,|\mathbb{X}] = \nu!\varepsilon_{\nu+1}^T H^{-1}S^{-1}c_p\beta_{p+1}h^{p+1}[1+o_p(1)] \\
&= \frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)\varepsilon_{\nu+1}^T\begin{pmatrix} h^{p+1} & 0 & \cdots & 0 \\ 0 & h^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & h \end{pmatrix}S^{-1}c_p[1+o_p(1)] \\
&= \varepsilon_{\nu+1}^T S^{-1}c_p\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu}+o_p\big(h^{p+1-\nu}\big).
\end{aligned}$$

We can now finalize the following theorem.

Theorem 4.2 (bias of the local polynomial regression estimator) *Assume that $f_X(x_0)>0$, $f_X(\cdot)$ and $m^{(p+1)}(\cdot)$ are continuous in a neighborhood of x_0. Further assume that $h\to 0$ and $nh\to\infty$. Then the asymptotic conditional bias of $\widehat{m}_\nu(x_0)$ for $p-\nu$ odd is given by*

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \varepsilon_{\nu+1}^T S^{-1}c_p\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu}+o_p\big(h^{p+1-\nu}\big). \tag{4.24}$$

Further, for $p-\nu$ even the asymptotic conditional bias of $\widehat{m}_\nu(x_0)$ is given by

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \varepsilon_{\nu+1}^T S^{-1}\tilde{c}_p\frac{\nu!}{(p+2)!}\left\{m^{(p+2)}(x_0)+(p+2)m^{(p+1)}(x_0)\frac{f_X'(x_0)}{f_X(x_0)}\right\}h^{p+2-\nu}+o_p\big(h^{p+2-\nu}\big),$$

provided that $f_X'(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3\to\infty$ and $\tilde{c}_p=(\mu_{p+2},\ldots,\mu_{2p+2})^T$.

A deeper result than the previous theorem, specifying higher order terms in the asymptotic bias and variance expressions, can be found in Fan et al. (1996). From the previous theorem it is already clear there is a theoretical difference between the cases $p-\nu$ odd and $p-\nu$ even. For $p-\nu$ even, the leading term $O_p\big(h^{p+1}\big)$ in the bias expression is zero due to symmetry of the kernel K and hence the second order term is represented in the theorem. For $p-\nu$ odd, the asymptotic bias has a simpler structure and does not involve $f_X'(x_0)$, a factor appearing in the asymptotic bias when $p-\nu$ is even. This theorem is in fact a generalization of what has already been observed for the special case of the local constant fit ($p=0$) used for estimating the regression function ($\nu=0$). The estimator resulting from such a fit, the Nadaraya-Watson (NW) estimator (Nadaraya, 1964; Watson, 1964) has an additional term in the asymptotic bias expression. The NW estimator is defined as

$$\widehat{m}_0(x_0)=\sum_{i=1}^n\frac{K\left(\frac{X_i-x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)}Y_i. \tag{4.25}$$

It can be shown that polynomial fits with $p-\nu$ odd outperform those with $p-\nu$ even.

4.4.2 Equivalent kernels

Next, we will show how the local polynomial approximation method assigns weights to each point. Note that (4.25) can rewritten as

$$\widehat{m}_0(x_0)=\sum_{i=1}^n W_0^n\left(\frac{X_i-x_0}{h}\right)Y_i \qquad \text{with} \quad W_0^n\left(\frac{X_i-x_0}{h}\right)=\frac{K\left(\frac{X_i-x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)}.$$

Is it possible to write a similar expression for the more general local polynomial estimator? If possible, this will provide further insight into the method and serves as a technical tool for understanding and deriving its asymptotic properties. The answer is YES!

Recall the notation

$$S_{n,j} = \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^j$$

and let $S_n = \mathbf{X}^T \mathbf{W} \mathbf{X}$ denote the $(p+1) \times (p+1)$ matrix $(S_{n,j+l})_{0 \le i,j \le p}$. Then, the estimator $\widehat{\beta}_\nu$ can be written as

$$\begin{aligned} \widehat{\beta}_\nu = \varepsilon_{\nu+1}^T \widehat{\beta} &= \varepsilon_{\nu+1}^T S_n^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \\ &= \sum_{i=1}^{n} W_\nu^n \left(\frac{X_i - x_0}{h} \right) Y_i, \end{aligned}$$

with $W_\nu^n(t) = \varepsilon_{\nu+1}^T S_n^{-1}(1,th,\ldots,(th)^p)^T K(t)/h$. The above expression reveals that the estimator $\widehat{\beta}_\nu$ is very much like a conventional kernel estimator except that the "kernel" W_ν^n depends on the design points AND locations. This explains why the local polynomial fit can adapt automatically to various designs and to boundary estimation.

Substituting (4.19) into the definition of W_ν^n yields

$$W_\nu^n(t) = \frac{1}{nh^{\nu+1} f_X(x_0)} \varepsilon_{\nu+1}^T S^{-1}(1,t,\ldots,t^p)^T K(t)[1 + o_p(1)]$$

and therefore

$$\widehat{\beta}_\nu = \frac{1}{nh^{\nu+1} f_X(x_0)} \sum_{i=1}^{n} K_\nu^\star \left(\frac{X_i - x_0}{h} \right) Y_i [1 + o_p(1)] \tag{4.26}$$

with

$$K_{\nu,p}^\star(t) = \varepsilon_{\nu+1}^T S^{-1}(1,t,\ldots,t^p)^T K(t) = \left(\sum_{l=0}^{p} S^{\nu l} t^l \right) K(t), \tag{4.27}$$

with $S^{-1} = \left(S^{jl}\right)_{0 \le j,l \le p}$. This kernel satisfies the following moment conditions:

$$\int u^q K_{\nu,p}^\star(u)\, du = \delta_{\nu,q} \quad 0 \le \nu,q \le p. \tag{4.28}$$

Table 4.1 gives the forms of some equivalent kernel functions. To emphasize the dependence of p, we use $K_{\nu,p}^\star$ to denote the equivalent kernel given by (4.27). As an illustration we plot in Figure 4.3 the Epanechnikov kernel

ν	p	Equivalent kernel function $K_{\nu,p}^\star(t)$
0	1	$K(t)$
0	3	$(\mu_4 - \mu_2^2)^{-1}(\mu_4 - \mu_2 t^2)K(t)$
1	2	$\mu_2^{-1} t K(t)$
2	3	$(\mu_4 - \mu_2^2)^{-1}(t^2 - \mu_2)K(t)$

Table 4.1: Equivalent kernel functions $K_{\nu,p}^\star$. Taken from Fan and Gijbels (1996, p. 66)

$K(u) = \frac{3}{4}(1-u^2)_+$ as well as the equivalent kernel $K_{\nu,p}^\star$ for some values of ν and p.

The conditional bias and variance of the estimator $\widehat{m}_\nu(x_0)$, given in (4.21) and (4.24) respectively, can equally well be re-expressed in terms of the equivalent kernel $K_{\nu,p}^\star$, leading to the asymptotic expression (for $p - \nu$ odd)

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \left(\int t^{p+1} K_{\nu,p}^\star(t)\, dt \right) \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} + o_p\left(h^{p+1-\nu}\right), \tag{4.29}$$

and its variance equals

$$\mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \left(\int K_{\nu,p}^{\star 2}(t)\, dt \right) \frac{\nu!^2 \sigma^2(x_0)}{f_X(x_0) n h^{1+2\nu}} + o_p\left(\frac{1}{nh^{1+2\nu}} \right). \tag{4.30}$$

These expressions can be obtained from (4.26) and (4.28).

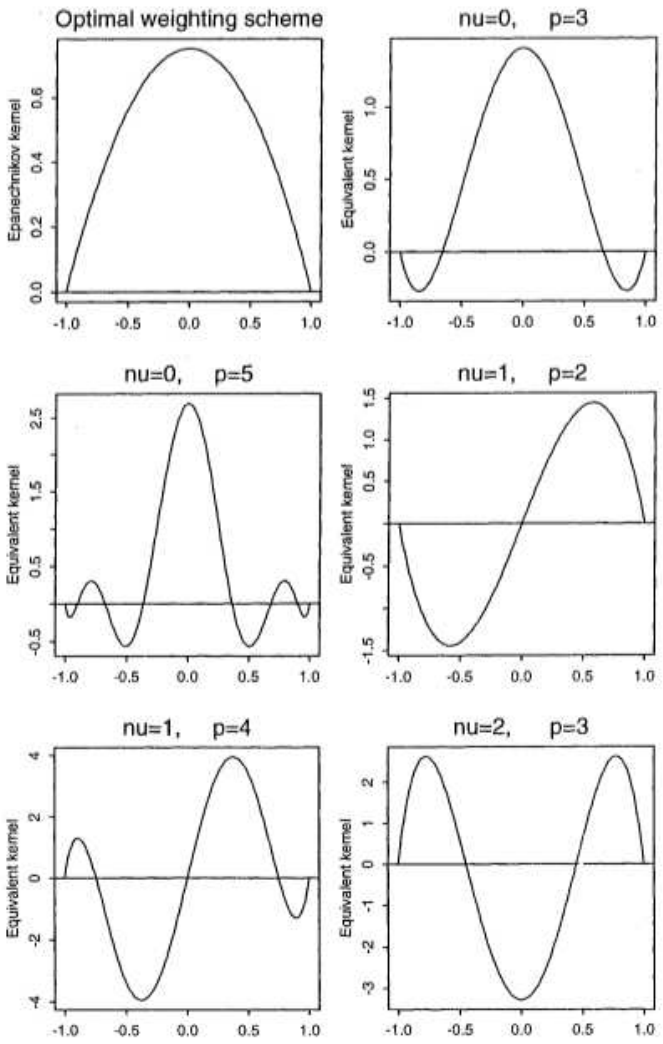


Figure 4.3: Epanechnikov kernel and its equivalent kernel for some values of p and ν. Taken from Fan and Gijbels (1996).

4.4.3 Ideal choice of bandwidth

The choice of the bandwidth parameter is rather crucial and hence should be done with a lot of care. As was the case for density estimation, we can make a distinction between a global or local varying bandwidth. A theoretical optimal local bandwidth for estimating $m^{(\nu)}(x_0)$ is obtained by minimizing the conditional mean squared error given by

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}]^2 + \mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}]. \tag{4.31}$$

The ideal choice of bandwidth can be approximated by the asymptotically optimal local bandwidth, i.e. the bandwidth which minimizes the asymptotic MSE. It is easy to show that minimizing (4.31) using (4.29) and (4.30), leads to (for $p-\nu$ odd)

$$\begin{aligned} h_{opt}(x_0) &= \left[\frac{(p+1)!^2(2\nu+1)\int K^{\star 2}_{\nu,p}(t)\,dt}{2(p+1-\nu)\left(\int t^{p+1}K^{\star}_{\nu,p}(t)\,dt\right)^2}\right]^{1/(2p+3)} \left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f_X(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)} \\ &= C_{\nu,p}(K)\left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f_X(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)}. \end{aligned}$$

The constant $C_{\nu,p}(K)$ is easy to calculate and Table 4.2 lists some of them for different ν and p. If we want a global

ν	p	Gaussian	Uniform	Epanechnikov
0	1	0.776	1.351	1.719
0	3	1.161	2.813	3.243
1	2	0.884	1.963	2.275
2	3	1.006	2.604	2.893

Table 4.2: Constant $C_{\nu,p}(K)$ for different kernel functions

measure of error, we could opt for a weighted MISE given by

$$\int (\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}]^2 + \textbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}])w(x)\,dx,$$

with $w \geq 0$ some weight function, leads to a theoretical optimal constant bandwidth. Usually w is taken to be the design density f_X. It can be shown that an asymptotically optimal constant bandwidth is given by (with $w = f_X$)

$$h_{opt} = C_{\nu,p}(K) \left[\frac{\int \sigma^2(x)\,dx}{\int \{m^{(p+1)}(x)\}^2 f_X(x)\,dx} \right]^{1/(2p+3)} n^{-1/(2p+3)}. \tag{4.32}$$

In the latter it is assumed that the integrals are finite and the that the denominator does not vanish.

In practice (4.32) is not usable since it depends on several unknown quantities. In what follows we present a simple way of estimating these quantities. The most simple way to do this is by fitting a polynomial of order $p+3$ **globally** to $m(x)$ (via ordinary least squares), leading to the parametric fit

$$\widetilde{m}(x) = \widetilde{\alpha}_0 + \cdots + \widetilde{\alpha}_{p+3} x^{p+3}.$$

The choice of a global fit results results in a derivative function $\tilde{m}^{(p+1)}(x)$ which is of quadratic form, allowing for certain flexibility in estimating the curvature. Assuming a constant error variance $\sigma^2(x) = \sigma^2$, then we can estimate this e.g., in a model-free way (Hall et al., 1990)

$$\widetilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n-2} (0.809 Y_{[i]} - 0.5 Y_{[i+1]} - 0.309 Y_{[i+2]})^2,$$

where $Y_{[j]}$ denotes the jth order observation corresponding to the ordered $X_{[j]}$. Other model-free error variance estimators, not necessarily restricted to the one dimensional case, can be found in Devroye et al. (2013) and De Brabanter et al. (2014). Of course, one could also use a model based estimator. Further assume that $x \in [a,b]$, then (4.32) can be estimated by

$$\widetilde{h}_{opt} = C_{\nu,p}(K) \left[\frac{\widetilde{\sigma}^2 \int_a^b dx}{\int \{\widetilde{m}^{(p+1)}(x)\}^2 f_X(x)\,dx} \right]^{1/(2p+3)} n^{-1/(2p+3)}.$$

Using the strong law of large numbers, the final rule of thumb bandwidth selector $\widetilde{h}_{\text{ROT}}$ (for $p-\nu$ odd) is given by

$$\boxed{\widetilde{h}_{\text{ROT}} = C_{\nu,p}(K) \left[\frac{\widetilde{\sigma}^2 (b-a)}{\sum_{i=1}^n \widetilde{m}^{(p+1)}(X_i)^2} \right]^{1/(2p+3)}} \tag{4.33}$$

Although (4.33) is derived under certain conditions, it can be applied in situations where these conditions are not strictly fulfilled.

4.4.4 Design adaptation property

The bias and variance expressions in (4.21) and (4.24) are obtained under the random design model, but remain valid for fixed designs. Hence, local polynomial estimators adapt to both random and fixed designs. This is in contrast with the Gasser-Müller estimator which cannot adapt to random designs: the unconditional variance is higher by a factor 1.5 for random designs. More explanation about this statement can be found in Mack and Müller (1989).

Recall that for $p-\nu$ even, additional terms arise in the asymptotic conditional bias. For example, when estimating the regression function $m(x_0)(\nu = 0)$, an extra term $m'(x_0)f'(x_0)/f(x_0)$ appears in the asymptotic bias of the Nadaraya-Watson estimator (4.25). The bias of this estimator depends on the intrinsic part $m''(x_0)$ interplaying with the artifact $m'(x_0)f'(x_0)/f(x_0)$. Keeping $m''(x_0)$ fixed, we first remark that in the highly clustered (asymmetric) design where $|f'(x_0)/f(x_0)|$ is large, the bias of the Nadaraya-Watson estimator can be large. Thus this estimator cannot adapt to highly clustered designs. Similar artifacts hold true for polynomial fits of an even order $p-\nu$. Local polynomial fitting with $p-\nu$ odd however rules out such artifacts and results in design-adaptive estimators.

4.4.5 Automatic boundary carpentry

In applications design points always have a bounded support. For estimating $m^{(\nu)}(x_0)$, with x_0 a point close to the boundary, the local neighborhood $x_0 \pm h$ can lie outside the design region. Hence, certain symmetric moment conditions, valid for all interior points, are no longer valid for x_0 in a boundary region, causing a large boundary bias for most of the smoothing techniques. If a bandwidth is chosen to be 25% of the data range, then for about 50% of the data range the local neighborhood will lie partly outside the design region. Hence the boundary region is about 50% of the whole data range. In higher dimensions these figures are even more striking, reflecting even more severe problems. Since many of the smoothing techniques show the aforementioned bias problem at the boundary, considerable efforts have been devoted to methods for correcting this boundary bias. Two popular approaches are boundary kernel methods and reflection methods. But none of these methods are as simple and as efficient as the automatic boundary correction when using local polynomial fitting. Without loss of generality we assume that the design density has a bounded support [0,1]. A left boundary point is thought of as being of the form $x := ch$, with $c > 0$, whereas a right boundary point is of the form $x = 1 - ch$.

The behavior of the estimator $\widehat{m}^{(\nu)}(x_0)$ for points x_0 at the interior of the support has been studied in the previous sections. In this section we address the question of how local polynomial estimators behave at boundary points. For most regression smoothers the rate of convergence at boundary points is slower than that at points in the interior. In the literature one refers to this problem as boundary effects or edge effects. These effects are visually very disturbing in practice, and in addition they can play a dominant role in theoretical analysis. Hence, in the case of boundary effects there is a strong request for boundary modifications, in order to overcome the problem.

The aforementioned automatic boundary carpentry can be easily seen from the representation of the local polynomial estimator in terms of an equivalent kernel (4.27). Consider a left boundary point $x = ch$. Similar as before, the finite moments are

$$S_{n,j} = nh^j f(0+)\mu_{j,c}\{1 + o_p(1)\}$$

where $\mu_{j,c} = \int_{-c}^{+\infty} u^j K(u)\, du$. This leads to the following equivalent kernel at the boundary

$$K^{\star}_{\nu,c,p}(t) = \varepsilon_{\nu+1}^T S_c^{-1}(1,t,\ldots,t^p)^T K(t) \quad \text{with} \quad S_c = (\mu_{j+l,c})_{0\le j,l\le p}.$$

This equivalent kernel differs from $K^{\star}_{\nu,p}$ only in the matrix S. This reflects the automatic adaptation to the boundary. Figure 4.4 shows the Epanechnikov kernel and some of its equivalent kernels for some boundary points and for various values of p and ν.

Theorem 4.3 (Fan and Gijbels (1996)) *Assume that $f(0+) > 0$ and that $f(\cdot), m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are right continuous at the point 0. Then, the conditional MSE of the estimator $m^{(\nu)}(x)$ at the left boundary point $x = ch$ is given by*

$$\left[\left\{\int_{-c}^{+\infty} t^{p+1} K^{\star}_{\nu,c,p}(t)\, dt\right\}^2 \left\{\nu!\frac{m^{(p+1)}(0+)}{(p+1)!}\right\}^2 h^{2(p+1-\nu)} + \int_{-c}^{+\infty} K^{\star 2}_{\nu,c,p}(t)\, dt \frac{\nu!^2\sigma^2(0+)}{f(0+)nh^{1+2\nu}}\right]\{1 + o_p(1)\}.$$

4.4.6 Which order of polynomial fit?

Fitting polynomials of higher order leads to a possible reduction of the bias, but on the other hand also to an increase of the variability, caused by introducing more local parameters. Intuitively it is clear that in a flat non-sloped region a local constant or linear fit is recommendable, whereas at peaks and valleys local quadratic and cubic fits are preferable. Thus, for a very spatially inhomogeneous curve, the order of the polynomial approximation should be adjusted to the curvature of the unknown regression function. We would like to mention, however, that for many applications

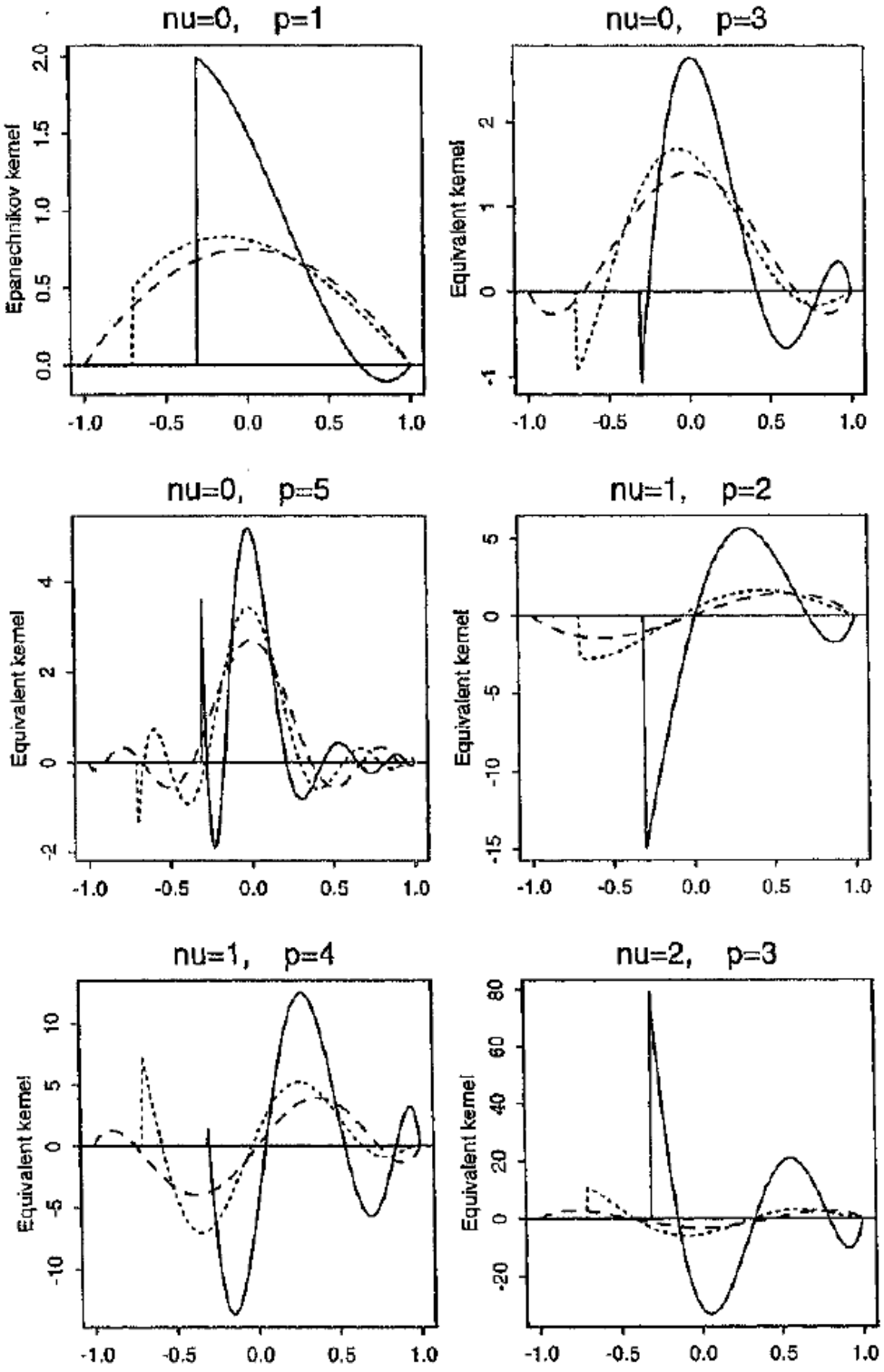


Figure 4.4: The Epanechnikov kernel and its equivalent kernels at the boundary points $c = 0.3$ (solid line) and $c = 0.7$ (dotted line) and interior points $c \geq 1$ (dashed line) for various values of p and ν. Taken from Fan and Gijbels (1996).

the choice $p = \nu + 1$ suffices. A variable order selection carries a possible price including the stochastic element introduced by the selection procedure and computational costs. Such an order selection procedure is mainly proposed for recovering spatially inhomogeneous curves.

Increase in variability

Suppose we fit a local polynomial of order p in order to estimate the derivative $m^{(\nu)}(x_0)$. The bias of such a fit will be of order $h^{p+1-\nu}$ (for $p-\nu$ odd) or of order $h^{p+1-\nu}$ (for $p-\nu$ even) as can be seen from Theorem 4.2. So, higher order polynomial approximations result in a smaller order of the bias. But let's see what happens to the variance if we increase the order of the approximation. The asymptotic variance of the estimator $\widehat{m}_\nu(x_0)$ is given by (see Theorem 4.1)

$$\begin{aligned} \mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}] &= \varepsilon_{\nu+1}^T S^{-1} S^\star S^{-1} \varepsilon_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f_X(x_0) n h^{1+2\nu}} \{1 + o_p(1)\} \\ &= \left(\int K_{\nu,p}^{\star 2}(t)\, dt \right) \frac{\nu!^2 \sigma^2(x_0)}{f_X(x_0) n h^{1+2\nu}} \{1 + o_p(1)\} \end{aligned}$$

which is of order $n^{-1}h^{-(1+2\nu)}$ and hence not affected by the order of the polynomial fit. But let's take a look at the constant terms. To simplify, take $\nu = 0$. This is by no means a restriction; conclusions drawn for the case of estimating the regression function carry over to the estimation of its derivative functions. The asymptotic variance of the estimator for the regression function is of the form

$$V_p \frac{\sigma^2(x_0)}{f_X(x_0) n h},$$

where V_p is the $(1,1)^{\text{th}}$ element of the matrix $S^{-1}S^\star S^{-1}$. Table 4.3 shows how much the variance increases with the order of the approximation for several kernel functions, relative to the variance of the Nadaraya-Watson estimator (local constant fit, $p = 0$). Table 4.3 summarizes the values for V_p/V_0 for various commonly used kernel functions. Note that there is no loss in terms of asymptotic variance by doing a local linear instead of a local constant fit. This remark applies to the comparison of any even order approximation with its consecutive odd order approximation. However, the asymptotic variance increases when moving from an odd order approximation to its consecutive even order approximation. For example in the case of the Epanechnikov kernel, the variance increases by a factor of 2.0833 when a local quadratic instead of a local linear fit is used. The increase in variability is the most pronounced for the uniform kernel. A graphical representation of Table 4.3 is given in Figure 4.5.

p	Gaussian	Uniform	Epanechnikov	Biweight	Triweight
1	1	1	1	1	1
2	1.6876	2.25	2.0833	1.9703	1.9059
3	1.6876	2.25	2.0833	1.9703	1.9059
4	2.2152	3.5156	3.1550	2.8997	2.7499
5	2.2152	3.5156	3.1550	2.8997	2.7499
6	2.6762	4.7852	4.2222	3.8133	3.5689
7	2.6762	4.7852	4.2222	3.8133	3.5689
8	3.1224	6.0562	5.2872	4.7193	4.3753
9	3.1224	6.0562	5.2872	4.7193	4.3753
10	3.5704	7.3281	6.3509	5.6210	5.1744

Table 4.3: Increase of the variability with the order of the polyno1nial approximation p. Results taken from Fan and Gijbels (1995).

It's an odd world

It now becomes clear that odd order fits are preferable. A fit of odd order $2p+1$ introduces an extra parameter in comparison with a fit of even order $2p$, but there is no increase of variability caused by this. With this extra parameter

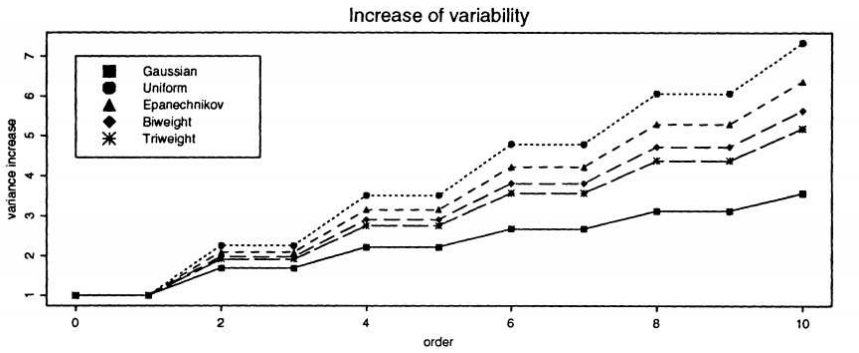


Figure 4.5: Increase of the variability with the order of the polynomial approximation p. Taken from Fan and Gijbels (1995).

an opportunity is created for a significant bias reduction especially in the boundary regions and in highly clustered design regions. Moreover, even order fits suffer from low efficiency, as was established by Fan (1993) for the local constant fit. In addition serious boundary effects appear when using even order fits; this contrasts with odd order fits which have the nice boundary adaptive property. The above asymptotic considerations demonstrate that it is "an odd world": odd order polynomial fits are preferable to even order polynomial fits. An odd world indeed, and that is why we say that $p - \nu$ odd is natural.

4.5 Data driven bandwidth choices: Cross-validation

4.5.1 Leave-one-out cross-validation (LOO-CV)

Consider the integrated squared error (ISE) as a measure of accuracy for the estimator $\hat{m}_h(x)$, let h denote the bandwidth of the kernel, and given a data set $\mathcal{D}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. Note that h should be strictly positive. The main idea is to construct an estimate $\widehat{m}_h$ such that the ISE is small. Let F_X denote the distribution over the input space, then

$$\int |\widehat{m}_h(x) - m(x)|^2 \, dF_X(x) = \int m^2(x) \, dF_X(x) + \int \widehat{m}_h^2(x) \, dF_X(x) - 2 \int \widehat{m}_h(x) m(x) \, dF_X(x). \tag{4.34}$$

Since the first term in (4.34) is independent of h, minimizing (4.34) is equivalent to minimizing

$$\int \widehat{m}_h^2(x) \, dF_X(x) - 2 \int \widehat{m}_h(x) m(x) \, dF_X(x). \tag{4.35}$$

In practice this would be impossible to compute since this quantity depends on the unknown real-valued (true) function m and the density f. The first term of (4.35) can be entirely computed from the data $\mathcal{D}_n$ and the second term can be written as

$$\int \widehat{m}_h(x) m(x) \, dF_X(x) = \mathbf{E}[\widehat{m}_h(X) m(X) | \mathcal{D}_n]. \tag{4.36}$$

If one estimates (4.36) by its empirical version $n^{-1} \sum_{i=1}^n Y_i \widehat{m}_h(X_i)$ the selection will be a biased estimator of the ISE. The bias is due to the fact that the observation Y_i is used in $\widehat{m}_h(X_i)$ to predict itself. However, there exist several methods to find an unbiased estimate of the ISE e.g. plug-in methods, leave-one-out (LOO) technique and a modification so that bias cancels out asymptotically. Here we will use the LOO technique in which one observation is left out. Therefore, a better estimator for (4.36) instead of its straight empirical version is

$$\frac{1}{n} \sum_{i=1}^n Y_i \widehat{m}_h^{(-i)}(X_i), \tag{4.37}$$

where $\widehat{m}_h^{(-i)}(X_i)$ denotes the LOO estimator with point i left out from the training. Similarly, the first term of (4.35) can be written as

$$\frac{1}{n} \sum_{i=1}^n \left| \widehat{m}_h^{(-i)}(X_i) \right|^2. \tag{4.38}$$

From (4.37) and (4.38), the LOO-CV function is given by

$$\text{LOO-CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{m}_h^{(-i)}(X_i)\right|^2.$$

The LOO cross-validated selection of h is

$$\boxed{\widehat{h}_{\text{LOO-CV}} = \arg\min_h \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{m}_h^{(-i)}(X_i)\right|^2}$$

It is interesting to know that this LOO-CV criterion actually estimates the following quantity (under independent errors)

$$\frac{1}{n}\sum_{i=1}^{n}|m(X_i) - \widehat{m}_h(X_i)|^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma^2(X_i).$$

4.5.2 v-fold Cross-Validation

In general there is no reason that training sets should be of size $n-1$ as in the LOO-CV case. There is the possibility that small perturbations, when single observations are left out, make LOO-CV(h) too variable, if the fitted values $\hat{m}_h(x)$ do not depend smoothly on the empirical distribution $\hat{F}_n$ or if the loss function $L(Y, \widehat{m}_h(X))$ is not continuous. These potential problems can be avoided, to a large extent, by leaving out groups of observations, rather than single observations. Also, it offers a computational advantage since we do not have to compute n estimates but only v in v-fold CV. The latter plays an important role for large data sets. Note that v-fold CV with $v = n$ is LOO-CV.

The use of groups have the desired effect of reducing variance, but at the cost of increasing bias. According to Beran (1984) and Burman (1989) the bias of v-fold CV yields

$$a_0\left[(v-1)^{-1}n^{-1}\right].$$

For LOO-CV the bias is of order $O(n^{-2})$, but when v is small the bias term is not necessarily small. Therefore, the use of 2-fold CV is never recommended. The term a_0, depending on the loss function L used in the CV procedure and the empirical distribution $\hat{F}_n$, is of the order of the number of effective parameters being estimated. As a result, if the number of effective parameters is not small, the v-fold CV is a poor estimate of the prediction error. However, there are adjustments possible to reduce the bias in v-fold CV, see e.g. Burman (1989, 1990); Tibshirani and Tibshirani (2009); Arlot and Celisse (2010). These adjustments to the v-fold CV procedure reduce the bias to

$$a_1\left[(v-1)^{-1}n^{-2}\right],$$

for some constant a_1 depending on the loss function L used in the CV procedure and the empirical distribution $\hat{F}_n$.

Precise understanding of how **Var**$[v$-fold CV$]$ depends on the splitting scheme is rather complex since the number of splits (folds) v is linked with the number of points used as validation. Furthermore, the variance of CV strongly depends on the framework and on the stability of the algorithm. Therefore, radically different results have been obtained in different frameworks, in particular on the value of v for which the v-fold CV estimator has a minimal variance, see e.g. Burman (1989) and Hastie et al. (2009, Chapter 7).

What is a suitable value for v? Davison and Hinkley (2003) have suggested the following rule of thumb. Take $v = \min(\sqrt{n}, 10)$, because taking $v > 10$ maybe computationally too expensive while taking groups of size at least $\sqrt{n}$ should perturb the data sufficiently to give a small variance of the estimate.

4.6 Local polynomial regression in R

4.6.1 Toy example

Create a toy example data set (200 points and normal random noise with $\sigma = 0.2$) and load the appropriate library:

```
> set.seed(729)
> x <- seq(0, 1, length.out = 200)
> y <- (sin(2*pi*(x-0.5)))^2 + rnorm(200, 0, 0.2)
> d <- data.frame(x,y)
> library(locpol)
```

Suppose we want to fit a local linear regression ($p = 1$) with Gaussian kernel to the data. We choose the bandwidth with the rule of thumb selector and via cross-validation:

```
> p <- 1
# Rule of Thumb
> hrot <- thumbBw(d$x, d$y, deg = p, kernel = gaussK)

0.03095715

# Cross-validation
> hcv <- regCVBwSelC(d$x,d$y,deg=p,kernel=gaussK,interval=c(seq(0.001,0.5,length.out=1000)))

0.03165582
```

In this case both bandwidths are not that different. This is of course not always the case. In practice it is recommended to use cross-validation. The problem with cross-validation is that it also can have multiple local minima. Therefore, it is recommended to specify a range or interval to look for a (local) minimum. Both methods assume that the data is i.i.d! Finally, we can fit the local linear regression estimate (for both bandwidths) and evaluate in each point of the data set by

```
> r_rot <- locpol(y~x, d, deg = p, bw = hrot, kernel = gaussK, xeval = d$x)
> r_cv <- locpol(y~x, d, deg = p, bw = hcv, kernel = gaussK, xeval = d$x)
```

Figure 4.6 shows the local linear estimate based on bandwidth $\widehat{h}_{\text{ROT}}$ (full line) and the true regression function (dashed line).

```
> plot(x, y)
> lines(x, fitted(r_rot), lwd = 2, col = "red")
> lines(x, (sin(2*pi*(x-0.5)))^2, lwd = 2, col = "blue", lty = 2)
```

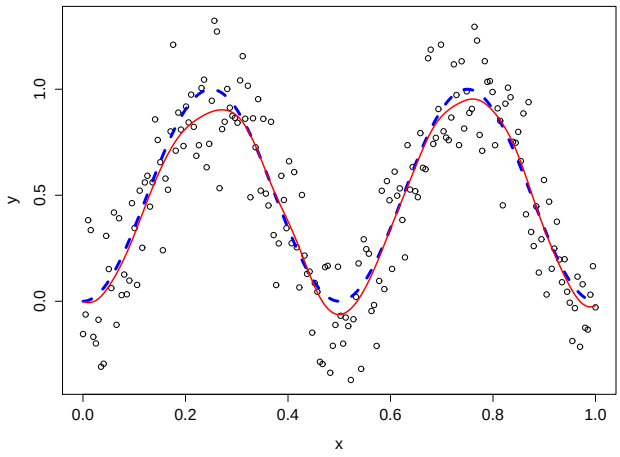


Figure 4.6: Local linear estimate based on bandwidth $\widehat{h}_{\text{ROT}}$ (full line) and the true regression function $\sin^2[2\pi(x-0.5)]$ (dashed line).

4.6.2 LIDAR data example

The lidar data frame has 221 observations from a light detection and ranging (LIDAR) experiment and can be found in the R package *SemiPar*. Loading the data set and finding the bandwidth for a local cubic fit ($p = 3$) based on a Gaussian kernel is done as follows:

```
> library(SemiPar)
> data(lidar)
> d <- data.frame(x = lidar$range, y = lidar$logratio)
> p <- 3
> hrot <- thumbBw(lidar$range, lidar$logratio, deg = p, kernel = gaussK)

10.89228

> hcv <- regCVBwSelC(lidar$range, lidar$logratio, deg = p, kernel = gaussK,...
                   interval=c(seq(22, 26, length.out = 10000)))

24.97158

> r_rot <- locpol(y~x, d, deg = p, bw = hrot, kernel = gaussK, xeval = d$x)
> r_cv <- locpol(y~x, d, deg = p, bw = hcv, kernel = gaussK, xeval = d$x)
> plot(d$x, d$y, xlab = "Range", ylab = "Logratio")
> lines(d$x, fitted(r_rot), lwd=4, col = "red", lty = 2)
> lines(d$x, fitted(r_cv), lwd=2, col = "blue")
```

The result is given in Figure 4.7. It is clear that the estimate based on $\widehat{h}_{\text{ROT}} = 10.89228$ is slightly more wiggly than the one based on $\widehat{h}_{\text{LOO-CV}} = 24.97158$. Do you know why? *Hint:* Check the assumptions on which the ROT is based and compare with LOO-CV!

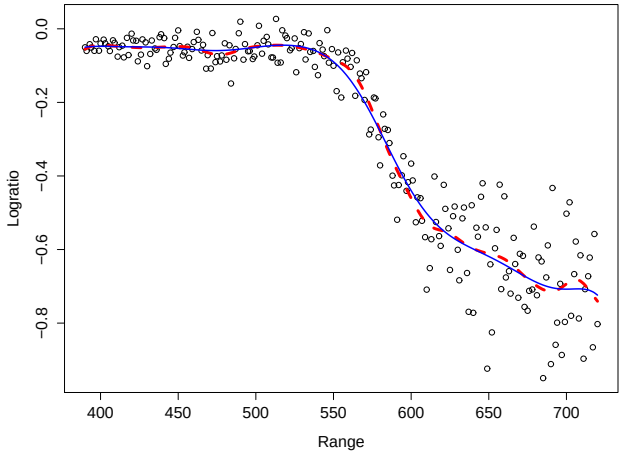


Figure 4.7: Local cubic estimate based on bandwidth $\widehat{h}_{\text{ROT}}$ (dashed line) and $\widehat{h}_{\text{LOO-CV}}$ (full line).

Chapter 6

Resampling methods: The jackknife and bootstrap

6.1 Introduction

This section is mainly based on the monograph of Efron and Tibshirani (1993).

A study was done to see if small aspirin doses would prevent heart attacks in healthy middle-aged men. The data for the aspirin study were collected in a particularly efficient way; by a controlled, randomized, double-blind study. One half of the subjects received aspirin and the other half received a control substance, or placebo, with no active ingredients. The subjects were randomly assigned to the aspirin or placebo groups. Both the subjects and the supervising physicians were blinded to the assignments, with the statisticians keeping a secret code of who received which substance. Scientists, like everyone else, want the project they are working on to succeed. The elaborate precautions of a controlled, randomized, blinded experiment guard against seeing benefits that do not exist, while maximizing the chance of detecting a genuine positive effect. The summary statistics in the newspaper article are very simple:

	heart attacks (fatal + non-fatal)	♯ subjects
aspirin group	104	11037
placebo group	189	11034

What strikes the eye here is the lower rate of heart attacks in the aspirin group. The ratio of the two rates is

$$\widehat{\theta} = \frac{104/11037}{189/11034} = 0.55.$$

If this study can be believed, and its solid design makes it very believable, the aspirin-takers only have 55% as many heart attacks as placebo-takers. Of course we are not really interested in $\widehat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio, that is the ratio we would see if we could treat all subjects, and not just a sample of them. The value $\widehat{\theta} = 0.55$ is only an estimate of θ. The sample seems large here, 22071 subjects in all, but the conclusion that aspirin works is really based on a smaller number, the 293 observed heart attacks. How do we know that $\widehat{\theta}$ might not come out much less favorably if the experiment were run again?

This is were statistical inference comes into play. It would be helpful to make the following claim: the true value of θ lies in the interval

$$0.143 < \theta < 0.70$$

with 95% confidence. The latter is a classical confidence interval. We almost certainly would not decide that θ exceeded 1, that is that aspirin was actually harmful. It is really rather amazing that the same data that give us an estimated value, $\widehat{\theta} = 0.55$ in this case, also can give us a good idea of the estimate's accuracy.

The aspirin study tracked strokes as well as heart attacks, with the following results:

	strokes	♯ subjects
aspirin group	119	11037
placebo group	89	11034

For strokes, the ratio of rates is

$$\widehat{\theta} = \frac{119/11037}{89/11034} = 1.21.$$

It now looks like taking aspirin is actually harmful. However the interval for the true stroke ratio θ turns out to be

$$0.93 < \theta < 1.59$$

with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo in relation to strokes. In the language of statistical hypothesis testing, *aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes*.

6.2 The jackknife

The jackknife is a technique for estimating the bias and standard error of an estimator. The jackknife method was invented in 1956 by Quenouille and further developed by Tukey in 1957. The word "jackknife" refers to the handy "knife" that one should always have.

6.2.1 The jackknife method

The basic idea behind the jackknife estimator lies in systematically recomputing the estimate leaving out one observation at a time from the sample set. From this new set of "observations" for the statistic an estimate for the bias can be calculated as well as an estimate for the variance of the statistic.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the population X with cumulative distribution function F. Let $\theta(F)$ be the quantity of interest. Denote by $T_n = T(X_1, \ldots, X_n)$ an estimator of $\theta(F)$. Consider

$$T_n = \theta(F_n)$$

where F_n is the empirical cumulative distribution function. Define the leave-one-out estimator of $\theta(F)$ by

$$T_{n-1}^{(i)} = T(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n). \tag{6.1}$$

We can look at the jackknife from two different ways, of course, leading to the same results. A first way is to define

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)}.$$

If T_n is unbiased

$$\mathbf{E}(\overline{T}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(T_{n-1}^{(i)}) = \theta.$$

If $\mathbf{E}(T_n) = \theta + a(\theta)n^{-1} + b(\theta)n^{-2} + O(n^{-3})$ we have that

$$\begin{aligned}
\mathbf{E}(\overline{T}_n - T_n) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(T_{n-1}^{(i)}) - \theta - a(\theta)n^{-1} - b(\theta)n^{-2} - O(n^{-3}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\theta + \frac{a(\theta)}{n-1} + \frac{b(\theta)}{(n-1)^2} + O(n^{-3})\right) - \theta - a(\theta)n^{-1} - b(\theta)n^{-2} - O(n^{-3}) \\
&= \frac{a(\theta)}{n(n-1)} + O(n^{-3})
\end{aligned}$$

and therefore

$$\widehat{\text{bias}}_{\text{jack}}(T_n) = (n-1)(\overline{T}_n - T_n) = (n-1)\left[\frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)} - T_n\right].$$

This means that the jackknife estimate of the bias is correct up to second order. Consequently, the bias-corrected jackknife estimate is

$$\widehat{\theta}_{\text{jack}} = \widehat{T}_n = T_n - \widehat{\text{bias}}_{\text{jack}}(T_n).$$

A second way is to define "pseudo-values" $T_{n,i}$'s by

$$T_{n,i} = nT_n - (n-1)T_{n-1}^{(i)}, \qquad i = 1,\ldots,n. \tag{6.2}$$

We now explain the motivation behind the "pseudo-values" $T_{n,i}$'s. Suppose T_n admits the following bias expansion

$$\mathbf{E}(T_n) = \theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \cdots.$$

Then we have for $\widehat{T}_n = n^{-1}\sum_{i=1}^{n} T_{n,i}$

$$\begin{aligned}
\mathbf{E}(\widehat{T}_n) &= \mathbf{E}\left(nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)}\right) \\
&= n\,\mathbf{E}(T_n) - \frac{n-1}{n}\sum_{i=1}^{n}\mathbf{E}(T_{n-1}^{(i)}) \\
&= n\,\mathbf{E}(T_n) - (n-1)\,\mathbf{E}(T_{n-1}^{(1)}) \\
&= n\left(\theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \cdots\right) - (n-1)\left(\theta + \frac{a(\theta)}{n-1} + \frac{b(\theta)}{(n-1)^2} + \cdots\right) \\
&= \theta + \frac{b(\theta)}{n} - \frac{b(\theta)}{n-1} + \cdots \\
&= \theta - \frac{b(\theta)}{n(n-1)} + \cdots
\end{aligned}$$

and consequently the estimator $\widehat{T}_n$ should be a more accurate estimator for θ than T_n. The jackknife estimator of the mean of T_n is given by

$$\begin{aligned}
\widehat{\mathbf{E}}_{\text{jack}}(T_n) = \widehat{T}_n &= \frac{1}{n}\sum_{i=1}^{n} T_{n,i} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[nT_n - (n-1)T_{n-1}^{(i)}\right] \\
&= nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)} \\
&= T_n + (n-1)\left[T_n - \frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)}\right] \\
&= T_n - \widehat{\text{bias}}_{\text{jack}}(T_n).
\end{aligned}$$

The jackknife estimator for the mean of T_n is thus equal to the mean of the pseudo-values. The jackknife estimator for the bias of T_n can also be found as

$$\begin{aligned}
\widehat{\text{bias}}_{\text{jack}}(T_n) &= T_n - \widehat{\mathbf{E}}_{\text{jack}}(T_n) \\
&= T_n - nT_n + \frac{n-1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)}
\end{aligned}$$

and consequently we have a similar result as before

$$\widehat{\text{bias}}_{\text{jack}}(T_n) = (n-1)\left[\frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)} - T_n\right]. \tag{6.3}$$

Tukey's jackknife estimator for the variance of T_n is given by the sample variance of the pseudo-values $T_{n,i}$

$$\widehat{\mathbf{Var}}_{\text{jack}}(T_n) = \frac{1}{n}\left\{\frac{1}{n-1}\sum_{i=1}^{n}\big(T_{n,i} - \widehat{\mathbf{E}}_{\text{jack}}(T_n)\big)^2\right\}. \tag{6.4}$$

6.2.2 Some examples

Mean of a random variable

Suppose that

$$\theta(F) = \mathbf{E}(X)$$

and let

$$T_n = \overline{X}_n = \text{sample mean.}$$

First calculate the pseudo-values $T_{n,i}$

$$\begin{aligned} T_{n,i} &= nT_n - (n-1)T_{n-1}^{(i)} \\ &= n\overline{X}_n - (n-1)\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} X_j \\ &= \sum_{j=1}^{n} X_j - \sum_{\substack{j=1\\ j\neq i}}^{n} X_j \\ &= X_i. \end{aligned}$$

The jackknife estimator for the sample mean is

$$\widehat{E}_{\text{jack}}(T_n) = \widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{n,i} = \overline{X}_n = T_n.$$

The jackknife estimator for the variance of T_n is

$$\begin{aligned} \widehat{\mathbf{Var}}_{\text{jack}}(T_n) &= \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\big(T_{n,i} - \widehat{\mathbf{E}}_{\text{jack}}(T_n)\big)^2\right) \\ &= \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}_n\big)^2\right) \\ &= \frac{S_n^2}{n} \end{aligned}$$

with S_n^2 the sample variance of the sample $X_1, \ldots, X_n$.

Variance of a random variable

Let

$$\theta(F) = \mathbf{Var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2.$$

Consider

$$T_n = \theta(F_n) = \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}_n\big)^2$$

the empirical variance. The pseudo-values are

$$\begin{aligned} T_{n,i} &= nT_n - (n-1)T_n^{(i)} \\ &= n\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}_n\big)^2 - (n-1)\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\big(X_j - \overline{X}_{n-1}^{(i)}\big)^2. \end{aligned}$$

Since

$$\overline{X}_{n-1}^{(i)} = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} X_j = \frac{1}{n-1}(n\overline{X}_n - X_i),$$

we have

$$\begin{aligned} T_{n,i} &= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2 - \sum_{\substack{j=1\\ j\neq i}}^{n} X_j^2 + 2(n-1)\big(\overline{X}_{n-1}^{(i)}\big)^2 - (n-1)\big(\overline{X}_{n-1}^{(i)}\big)^2 \\ &= X_i^2 - n\overline{X}_n^2 + (n-1)\big(\overline{X}_{n-1}^{(i)}\big)^2 \\ &= X_i^2 - n\overline{X}_n^2 + (n-1)\left(\frac{1}{n-1}(n\overline{X}_n - X_i)\right)^2 \\ &= X_i^2 - n\overline{X}_n^2 + \frac{n^2}{n-1}\overline{X}_n^2 - 2\frac{n}{n-1}X_i\overline{X}_n + \frac{1}{n-1}X_i^2 \\ &= \frac{n}{n-1}X_i^2 + \frac{n^2 - n(n-1)}{n-1}\overline{X}_n^2 - 2\frac{n}{n-1}X_i\overline{X}_n \\ &= \frac{n}{n-1}X_i^2 + \frac{n}{n-1}\overline{X}_n^2 - 2\frac{n}{n-1}X_i\overline{X}_n \\ &= \frac{n}{n-1}(X_i - \overline{X}_n)^2. \end{aligned}$$

Finally,

$$\widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{n,i} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = S_n^2,$$

the sample variance of the sample $X_1,\ldots,X_n$. Note that $\mathbf{E}(T_n) = \frac{n-1}{n}\theta$ and $\mathbf{E}(\widehat{T}_n) = \theta$. Therefore, $\widehat{T}_n$ is a better estimator than T_n.

6.2.3 Failure of the jackknife

The statistics to be analyzed need to be "smooth" in the data, meaning that small changes in the data set cause only small changes in the statistic. Consider the following classical example to illustrate this effect.

Let (the median)

$$\theta(F) = F^{-1}(0.5) = \inf\{x : F(x) \geq 0.5\}.$$

Further, consider an i.i.d. sample $X_1,\ldots,X_n$ from X having distribution F and assume the case $n = 2m$ (even sample size). Then the sample median is

$$T_n = \frac{1}{2}(X_{(m)} + X_{(m+1)}),$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ denote the order statistics of $X_1,\ldots,X_n$. In order to calculate the pseudo-values, let's first take a look at $T_{n-1}^{(i)}$

$$T_{n-1}^{(i)} = \begin{cases} X_{(m+1)}, & i = 1,\ldots,m; \\ X_{(m)}, & i = m+1,\ldots,n. \end{cases}$$

The pseudo-values are

$$T_{n,i} = \begin{cases} nT_n - (n-1)X_{(m+1)}, & i = 1,\ldots,m; \\ nT_n - (n-1)X_{(m)}, & i = m+1,\ldots,n, \end{cases}$$

and

$$\begin{aligned}\widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{n,i} &= nT_n - \frac{n-1}{n} m X_{(m+1)} - \frac{n-1}{n}(n-m)X_{(m)} \\ &= nT_n - \frac{n-1}{n} m(X_{(m+1)} - X_{(m)}) \\ &= nT_n - (n-1)T_n \\ &= T_n.\end{aligned}$$

The jackknife estimator of the bias is

$$T_n - \widehat{E}_{\text{jack}}(T_n) = T_n - \widehat{T}_n = 0.$$

However, the jackknife estimator for the variance does not work. Note that

$$\begin{aligned}T_{n,i} - \widehat{T}_n &= T_{n,i} - T_n \\ &= \begin{cases} (n-1)T_n - (n-1)X_{(m+1)}, & i = 1,\ldots,m; \\ (n-1)T_n - (n-1)X_{(m)}, & i = m+1,\ldots,n, \end{cases} \\ &= \begin{cases} (n-1)(T_n - X_{(m+1)}), & i = 1,\ldots,m; \\ (n-1(T_n - X_{(m)}), & i = m+1,\ldots,n, \end{cases}\end{aligned}$$

and it follows that

$$\begin{aligned}\widehat{\mathbf{Var}}_{\text{jack}}(T_n) &= \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}\big(T_{n,i} - \widehat{T}_n\big)^2\right) \\ &= \frac{1}{n}\left(\frac{1}{n-1} m(n-1)^2 (T_n - X_{(m+1)})^2 + \frac{1}{n-1}(n-m)(n-1)^2 (T_n - X_{(m)})^2\right) \\ &= \frac{m(n-1)}{n}\big((T_n - X_{(m+1)})^2 + (T_n - X_{(m)})^2\big) \\ &= \frac{n-1}{2}\left[\left(\frac{1}{2}X_{(m)} - \frac{1}{2}X_{(m+1)}\right)^2 + \left(\frac{1}{2}X_{(m+1)} - \frac{1}{2}X_{(m)}\right)^2\right] \\ &= \frac{n-1}{2}\left[\frac{1}{2}X_{(m)}^2 - \frac{1}{2}X_{(m)}X_{(m+1)} + \frac{1}{2}X_{(m+1)}^2 - \frac{1}{2}X_{(m)}X_{(m+1)}\right] \\ &= \frac{n-1}{4}\big(X_{(m+1)} - X_{(m)}\big)^2.\end{aligned}$$

In order to verify whether the jackknife estimator of the variance is consistent, we derive the asymptotic distribution of sample quantiles.

Suppose $X_1,\ldots,X_n$ are i.i.d. continuous random variables from a distribution with CDF F. Let $F_n(x)$ (empirical CDF) be a random variable defined for fixed $x \in \mathbb{R}$ by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}.$$

It immediately follows that $\mathbf{E}(I\{X_i \le x\}) = F(x)$, $\mathbf{Var}(I\{X_i \le x\}) = F(x)\{1-F(x)\}$ and by the central limit theorem

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)\{1-F(x)\}).$$

Now consider the transformation through function $g(t)$ with $0 < t < 1$ by $g(t) = F^{-1}(t)$. The first order derivative of $g(t)$ is

$$g'(t) = \frac{1}{f(F^{-1}(t))}.$$

By the Delta method we have

$$\sqrt{n}\big(F^{-1}(F_n(x)) - F^{-1}(F(x))\big) \xrightarrow{d} N\left(0, \frac{F(x)\{1-F(x)\}}{\{f(F^{-1}(F(x)))\}^2}\right)$$

and letting $p = F(x)$

$$\sqrt{n}\big(F^{-1}(F_n(x)) - x\big) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f(x)^2}\right).$$

$F^{-1}(F_n(x))$ is a random variable lying between the $(p-1)$st and pth sample quantile and can be written using via order statistic notation as $X_{(np)}$. Moreover,

$$|X_{(np)} - F^{-1}(F_n(x))| \xrightarrow{a.s.} 0$$

and hence

$$\sqrt{n}\big(X_{(np)} - x\big) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f(x)^2}\right).$$

For the median ($p = \frac{1}{2}$) of a symmetric (around 0) density we have that the asymptotic variance is given by $\dfrac{1}{4nf^2(0)}$. It is clear that

$$\widehat{\mathbf{Var}}_{\text{jack}}(T_n) - \frac{1}{4nf^2(0)} \nrightarrow 0, n \to \infty$$

and therefore the jackknife estimator for the variance of the median is not consistent as it fails to converge to the true one.

Jackknife in R

We can use the jackknife to calculate the bias and variance of the sample mean and variance. Let $X_1, \ldots, X_n \sim U(0,1), i = 1, \ldots, 1000$. The following R code calculates the jackknife estimate for the bias and variance of the sample mean.

```
> library(bootstrap)
> b <- 1
> a <- 0
> X <- runif(1000, a, b)
> jackres <- jackknife(X,mean)
> jackres$jack.bias
[1] 0

> jackres$jack.se
[1] 0.009056706
```

The real standard deviation of the sample mean is

$$\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\frac{1}{12}}{1000}} = 0.009128709.$$

Similarly, using the jackknife to estimate the bias for the MLE for the variance i.e., $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

```
> theta <- function(x){(1/length(x)) * sum((x-mean(x))^2)}
jackres <- jackknife(x,theta)
> jackres$jack.bias
[1] -0.00008450318

> jackres$jack.se
[1] 0.002417544
```

The true bias and variance of the MLE sample variance are given by

$$\text{bias}(\widehat{\sigma}^2) = -\frac{\sigma^2}{n} = -\frac{1}{12 \times 1000} = -0.000083333$$

and

$$\mathbf{Var}(\widehat{\sigma}^2) = \frac{(n-1)^2}{n^3}\,\mathbf{E}(X - \mathbf{E}\,X)^4 - \frac{(n-1)(n-3)}{n^3}\sigma^4 = 5.558325 \times 10^{-6} \quad \text{and} \quad \sqrt{\mathbf{Var}(\widehat{\sigma}^2)} = 0.00235761$$

respectively.

6.3 The bootstrap

6.3.1 Principle of the bootstrap

The bootstrap is a resampling mechanism designed to provide approximations to the sampling distribution of a functional $T(X_1,X_2,\ldots,X_n,F)$ where F is a CDF, typically on some Euclidean space, and $X_1,X_2,\ldots,X_n$ are independent sample observations from F. For example, F could be some continuous CDF on the real line, and $T(X_1,X_2,\ldots,X_n,F)$ could be $\sqrt{n}(\overline{X}-\mu)$, where $\mu = \mathbf{E}_F(X_1) = \mathbf{E}(X_1)$. The problem of approximating the distribution is important, because even when the statistic $T(X_1,X_2,\ldots,X_n,F)$ is a simple one, such as the sample mean, we usually cannot find the distribution of $T(X_1,X_2,\ldots,X_n,F)$ exactly for given n. Sometimes, there may be a suitable asymptotic normality result known about the statistic T, which may be used to form an approximation to the distribution of T . A remarkable fact about the bootstrap is that even if such an asymptotic normality result is available, the bootstrap often provides a better approximation to the true distribution of T than does the normal approximation.

The bootstrap is not limited to the i.i.d. situation. It has been studied for various kinds of dependent data and highly complex situations. In fact, this versatility of the bootstrap is the principal reason for its huge popularity and the impact that it has had on practice.

Suppose $X_1,X_2,\ldots,X_n \overset{i.i.d}{\sim} F$ and $T(X_1,X_2,\ldots,X_n,F)$ is a functional, for example, $T(X_1,X_2,\ldots,X_n,F) = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ where $\mu = \mu(F) = \mathbf{E}(X_1)$, and $\sigma^2 = \sigma^2(F) = \mathbf{Var}_F(X_1) = \mathbf{Var}(X_1)$, assumed to be finite. In statistical problems, we frequently need to know something about the sampling distribution of T , for example, $\mathbf{P}(T(X_1,X_2,\ldots,X_n,F) \le t)$. If we had replicated samples from the population, resulting in a series of values for the statistic T , then we could form estimates of $\mathbf{P}_F(T \le t) = \mathbf{P}(T \le t)$ by counting how many of the T_i's are $\le t$. But statistical sampling is not done that way. Usually, we do not obtain replicated samples; we obtain just one set of data values of some size n. The intuition of the canonical bootstrap is that by the Glivenko–Cantelli theorem, the empirical CDF F_n should be very close to the true underlying CDF F, and so, sampling from F_n, which amounts to simply resampling n values with replacement from the already available data $(X_1,X_2,\ldots,X_n)$, should produce new sets of values that act like samples from F itself. So, although we did not have replicated datasets to start with, it is as if by resampling from the available dataset we now have the desired replications. There is a certain element of faith in this idea, unless we have demonstrable proofs that this simple idea will in fact work, that is, that these resamples lead us to accurate approximations to the true distribution of T . It turns out that such theorems are available, and have led to the credibility and popularity of the bootstrap as a distribution approximation tool. To implement the bootstrap, we only need to be able to generate enough resamples from the original dataset. So, in a sense, the bootstrap replaces a hard mathematical calculation in probability theory by an omnibus and almost automated computing exercise. It is the automatic nature of the bootstrap that makes it so appealing. However, it is also frequently misused in situations where it should not be used, because it is theoretically unjustifiable in those problems, and will in fact give incorrect and inaccurate answers.

Suppose for some number B, we draw B resamples of size n from the original sample. Denoting the resamples from the original sample as

$$(X^\star_{11},X^\star_{12},\ldots X^\star_{1n}),(X^\star_{21},X^\star_{22},\ldots X^\star_{2n}),\ldots,(X^\star_{B1},X^\star_{B2},\ldots X^\star_{Bn})$$

with corresponding values $T^\star_1, T^\star_2,\ldots,T^\star_B$ for the functional T. One can use simple frequency based estimates such as $\frac{\#\{j:T^\star_j \le t\}}{B}$ to estimate $\mathbf{P}(T \le t)$. This is the basic idea of the bootstrap.

Applying the bootstrap in practice goes as follows. Suppose $T(X_1,X_2,\ldots,X_n,F) = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$. In the canonical bootstrap scheme, we take i.i.d. samples from F_n. By a simple calculation, the mean and the variance of the empirical distribution F_n are $\overline{X}$ and $s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. The bootstrap is a device for estimating

$$\mathbf{P}_F\left(\frac{\sqrt{n}(\overline{X}-\mu(F))}{\sigma} \le x\right) \quad \text{by} \quad \mathbf{P}_{F_n}\left(\frac{\sqrt{n}(\overline{X}^\star_n - \overline{X})}{s} \le x\right).$$

We further approximate $\mathbf{P}_{F_n}\left(\frac{\sqrt{n}(\overline{X}_n^\star-\overline{X})}{s}\leq x\right)$ by resampling only B times from the original sample set $(X_1,X_2,\ldots,X_n)$. Finally, we report as our estimate for $\mathbf{P}_F\left(\frac{\sqrt{n}(\overline{X}-\mu(F))}{\sigma}\leq x\right)$ the number

$$\frac{\#\left\{j:\frac{\sqrt{n}(\overline{X}^\star_{n,j}-\overline{X})}{s}\leq x\right\}}{B}.$$

This number depends on the original sample set $(X_1,X_2,\ldots,X_n)$, the particular resampled sets $(X^\star_{i1},X^\star_{i2},\ldots,X^\star_{in})$ and the bootstrap Monte Carlo sample size B. If the bootstrap Monte Carlo is repeated, then for the same B, and of course, the same original sample set $(X_1,X_2,\ldots,X_n)$, the bootstrap estimate will be a different number. We would like the bootstrap estimate to be close to the true value $\mathbf{P}_F\left(\frac{\sqrt{n}(\overline{X}-\mu(F))}{\sigma}\leq x\right)$; consistency of the bootstrap is about our ability to guarantee that for large n, and an implicit unspoken assumption of a large B.

6.3.2 Consistency of the bootstrap

Consider the sample mean of i.i.d random variables. If $X_1,X_2,\ldots,X_n\overset{i.i.d}{\sim}F$ and if $\mathbf{Var}_F(X)<\infty$ then $\sqrt{n}(\overline{X}-\mu)$ has a limiting normal distribution by the central limit theorem (CLT). A probability $\mathbf{P}_F(\sqrt{n}(\overline{X}-\mu(F))\leq x)$ could be approximated by, for example, $\Phi(x/s)$, with s the sample standard deviation. An interesting property of the bootstrap approximation is that even when the CLT approximation $\Phi(x/s)$ is available, the bootstrap approximation may be more accurate. Such results are generally known as *higher-order accuracy* of the bootstrap.

Before proving consistency of the bootstrap we need the following definition.

Definition 6.1 *Let $\rho(F,G)$ be a metric space of CDFs on $\mathcal{X}\in\mathbb{R}$. For a given functional $T(X_1,X_2,\ldots,X_n,F)$, let*

$$\begin{aligned}H_n(x) &= \mathbf{P}_F\big(T(X_1,X_2,\ldots,X_n,F)\leq x\big)\\ H_{boot}(x) &= \mathbf{P}_\star\big(T(X_1^\star,X_2^\star,\ldots,X_n^\star,F_n)\leq x\big).\end{aligned}$$

The the bootstrap is weakly consistent under ρ for T if $\rho(H_n,H_{boot})\overset{p}{\to}0$ as $n\to\infty$. The bootstrap is strongly consistent under ρ for T is $\rho(H_n,H_{boot})\overset{a.s.}{\to}0$ as $n\to\infty$.

Remark 6.1 *It is common to use the notation $\mathbf{P}_\star$ to denote probabilities under the bootstrap distribution. $\mathbf{P}_{F_n}(\cdot)$ corresponds to probability statements corresponding to all the n^n possible with replacement resamples from the original sample $X_1,X_2,\ldots,X_n$. Recalculating T from all n^n resamples is basically impossible unless n is very small. Therefore one uses a smaller number of B resamples and recalculates T only B times. Thus $H_{boot}(\cdot)$ is itself estimated by Monte Carlo, known as bootstrap Monte Carlo. So the final estimate for $\mathbf{P}_F\big(T(X_1,X_2,\ldots,X_n,F)\leq x\big)$ absorbs errors from two sources: (i) pretending that $(X^\star_{i1},X^\star_{i2},\ldots,X^\star_{in})$ are bona fide samples from F; (ii) estimating the true $H_{boot}(\cdot)$ by a Monte Carlo.*

Note that the need for mentioning convergence to zero in probability or a.s. in this definition is due to the fact that the bootstrap distribution H_{boot} is a random CDF. It is a random CDF because as a function it depends on the original sample $(X_1,X_2,\ldots,X_n)$. Thus, the bootstrap uses a random CDF to approximate a deterministic but unknown CDF, namely the true CDF H_n of the functional T. In principle, a sequence of random CDFs could very well converge to another random CDF, or not converge at all! It is remarkable that under certain minimal conditions, those disasters do not happen, and H_{boot} and H_n get close as $n\to\infty$.

Consider the following metric (Kolmogorov metric)

$$K(F,G)=\sup_{-\infty<x<\infty}|F(x)-G(x)|.$$

Of course, other metrics can also be considered e.g. the Mallows-Wasserstein metric (Bickel and Freedman, 1981). We can formulate the consistency of the bootstrap.

Theorem 6.1 *Suppose $X_1,X_2,\ldots,X_n\overset{i.i.d.}{\sim}F$ and suppose $\mathbf{E}_F(X_1^2)<\infty$. Let $T(X_1,X_2,\ldots,X_n,F)=\sqrt{n}(\overline{X}-\mu)$ then $K(H_n,H_{boot})\overset{a.s.}{\to}0$ as $n\to\infty$.*

Note that $\mathbf{E}_F(X_1^2) < \infty$ guarantees that $\sqrt{n}(\overline{X} - \mu)$ admits a CLT. Theorem 6.1 states that the bootstrap is strongly consistent, w.r.t. K, under that assumption. A good rule of thumb: if a functional $T(X_1,X_2,\ldots,X_n;F)$ admits a CLT, then the bootstrap would be at least weakly consistent for T. Strong consistency might require more assumptions.

In order to prove the consistency of the bootstrap we need the following three theorems.

Theorem 6.2 (Zygmund-Marcinkiewicz SLLN) *Let $Y_1,Y_2,\ldots,Y_n$ be i.i.d. random variables with CDF F and suppose, for some $0 < \delta < 1$, $\mathbf{E}_F\,|Y_1|^\delta < \infty$. Then*

$$n^{-1/\delta}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} 0.$$

Theorem 6.3 (Berry-Esseen Theorem) *Let X_i, $i \geq 1$ be i.i.d. with $\mathbf{E}(X_1) = \mu$, $\mathbf{Var}(X_1) = \sigma^2$ and $\mathbf{E}\,|X_1 - \mu|^3 < \infty$. Then there exists a universal constant C not depending on n or the distribution F of the X_i such that*

$$\sup_{-\infty<x<\infty}\left|\mathbf{P}\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq x\right) - \Phi(x)\right| \leq C\frac{\mathbf{E}\,|X_1-\mu|^3}{\sigma^3\sqrt{n}}.$$

Remark 6.2 *Calculated values of the constant C have decreased markedly over the years, from the original value of 7.59 by Esseen (1942), to 0.7882 by van Beek (1972), then 0.7655 by Shiganov (1986), then 0.7056 by Shevtsova (2007), then 0.7005 by Shevtsova (2008), then 0.5894 by Tyurin (2009), then 0.5129, then 0.4785 by Tyurin (2010). The best estimate as of 2017, $C < 0.4748$ due to Shevtsova (2011). Further, it can also be shown (Esseen, 1956) that $C \geq \frac{\sqrt{10}+3}{6\sqrt{2\pi}} \approx 0.40973$.*

Theorem 6.4 (Pólya's Theorem) *Let X_n, $n \geq 1$ have CDF F_n and let X have CDF F. If F is everywhere continuous and if $X_n \xrightarrow{d} X$ then*

$$\sup_{-\infty<x<\infty}|F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

as $n \to \infty$.

PROOF. [Theorem 6.1] Using the definition of the Kolmogorov metric

$$\begin{aligned}
K(H_n,H_{\text{boot}}) &= \sup_{-\infty<x<\infty}\left|\mathbf{P}_F(T_n \leq x) - \mathbf{P}_\star(T_n^\star \leq x)\right| \\
&= \sup_{-\infty<x<\infty}\left|\mathbf{P}_F\left(\frac{T_n}{\sigma} \leq \frac{x}{\sigma}\right) - \mathbf{P}_\star\left(\frac{T_n^\star}{s} \leq \frac{x}{s}\right)\right| \\
&= \sup_{-\infty<x<\infty}\left|\mathbf{P}_F\left(\frac{T_n}{\sigma} \leq \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) + \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) + \Phi\left(\frac{x}{s}\right) - \mathbf{P}_\star\left(\frac{T_n^\star}{s} \leq \frac{x}{s}\right)\right| \\
&\leq \sup_{-\infty<x<\infty}\left|\mathbf{P}_F\left(\frac{T_n}{\sigma} \leq \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right)\right| + \sup_{-\infty<x<\infty}\left|\Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right)\right| \\
&\quad + \sup_{-\infty<x<\infty}\left|\Phi\left(\frac{x}{s}\right) - \mathbf{P}_\star\left(\frac{T_n^\star}{s} \leq \frac{x}{s}\right)\right| \\
&= A_n + B_n + C_n.
\end{aligned}$$

By Pólya's theorem $A_n \to 0$ almost surely. Also s^2 converges almost surely to σ^2 and by the continuous mapping theorem s converges almost surely to σ. Then $B_n \to 0$ almost surely since $\Phi(\cdot)$ is uniform continuous. Finally, by applying Berry-Esseen's theorem on the last term with $0.40973 \leq C < 0.4748$ yields

$$\begin{aligned}
C_n &\leq \frac{C}{\sqrt{n}}\frac{\mathbf{E}_{F_n}\,|X_1^\star - \overline{X}|^3}{[\mathbf{Var}_{F_n}(X_1^\star)]^{3/2}} \\
&= \frac{C}{\sqrt{n}}\frac{\sum_{i=1}^n |X_i - \mu + \mu - \overline{X}|^3}{ns^3} \\
&\leq \frac{C}{n^{3/2}s^3}2^{(3-1)}\left[\sum_{i=1}^n |X_i-\mu|^3 + n|\mu - \overline{X}|^3\right] \\
&= \frac{C_1}{s^3}\left[\frac{1}{n^{3/2}}\sum_{i=1}^n |X_i-\mu|^3 + \frac{|\overline{X}-\mu|^3}{\sqrt{n}}\right],
\end{aligned}$$

where the last inequality follows from the C_r inequality and $C_1 = 2^{(3-1)}C$.

By the strong law of large numbers $\overline{X} \stackrel{a.s.}{\to} \mu$ and $s \stackrel{a.s.}{\to} \sigma > 0$, it follows that

$$\frac{|\overline{X} - \mu|^3}{\sqrt{n}} \stackrel{a.s.}{\to} 0.$$

For the other term, let $Y_i = |X_i - \mu|^3$ and $\delta = 2/3$. Then the Y_i are i.i.d. and

$$\mathbf{E}\,|Y_i|^\delta = \mathbf{E}_F\,|X_i - \mu|^{3\delta} = \mathbf{Var}_F(X_1) < \infty.$$

From the Zygmund-Marcinkiewicz SLLN we have

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} |X_i - \mu|^3 = n^{-1/\delta} \sum_{i=1}^{n} Y_i \stackrel{a.s.}{\to} 0, \quad \text{as } n \to \infty.$$

Hence, $A_n + B_n + C_n \to 0$ almost surely and consequently $K(H_n,H_{\text{boot}}) \stackrel{a.s.}{\to} 0$ as $n \to \infty$. ■

It is natural to ask if the bootstrap is consistent for $\sqrt{n}(\overline{X} - \mu)$ even when $\mathbf{E}_F(X_1^2) = \infty$. If we insist on strong consistency, then the answer is negative. The point is that the sequence of bootstrap distributions is a sequence of random CDFs and so it can converge to a random CDF, depending on the particular realization $X_1,X_2,\ldots,X_n$ if $\mathbf{E}_F(X_1^2) = \infty$.

Example 6.1 (Practical accuracy of the bootstrap) *How does the bootstrap compare with the CLT approximation in actual applications? The question can only be answered by case-by-case simulation. The results are mixed in the following numerical table (Table 6.1). The X_i are i.i.d. $Exp(1)$ in this example and $T = \sqrt{n}(\overline{X} - 1)$ with $n = 100$. For the bootstrap approximation we used $B = 10000$. The exact distribution can be easily derived in this case: Since $X_i \sim Exp(1)$, it follows that $\sum_{i=1}^n X_i \sim \Gamma_{n,1}$. The density of T, f_T, is given by*

$$f_T(t) = \frac{1}{\sqrt{n}} f\big(1 + \frac{t}{\sqrt{n}} \mid n\big), \quad -\sqrt{n} \le t < +\infty,$$

with

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

t	$F_T(t)$	CLT	$H_{\text{boot}}(t)$
-2	0.0171	0.0228	0.0150
-1	0.1582	0.1587	0.1467
0	0.5133	0.5	0.5103
1	0.8417	0.8413	0.8506
2	0.9721	0.9772	0.9783

Table 6.1: Accuracy of the bootstrap.

Example 6.2 (Bootstrap failure) *In spite of the many consistency theorems for the bootstrap, there are instances where the ordinary bootstrap with sampling with replacement from F_n actually does not work. Typically, these are instances where the functional T_n fails to admit a CLT. Here is a simple example where the ordinary bootstrap fails to consistently estimate the true distribution of a statistic. Let $X_1,X_2,\ldots,X_n$ be i.i.d. uniform(0,θ), with $\theta \ge 0$ and let $T_n = n(\theta - X_{(n)})$ and $T_n^\star = n(X_{(n)} - X_{(n)}^\star)$. The ordinary bootstrap will fail in the sense that the conditional distribution of $T_n^\star$ given $X_{(n)}$ will not converge to the true one almost surely. Let's first find the exact distribution of T_n. The distribution (under F) of $M = X_{(n)}$, the maximum of the sample is*

$$\mathbf{P}(M \le m) = \mathbf{P}(X_1 \le m,\ldots,X_n \le m) \stackrel{i.i.d.}{=} [F_X(m)]^n.$$

Then for X_1,X_2,...,$X_n \sim U(0,\theta)$ we have that

$$\mathbf{P}(M \le m) = \begin{cases} 1, & m \ge \theta; \\ \dfrac{m^n}{\theta^n}, & 0 \le m \le \theta; \\ 0, & \textit{else.} \end{cases}$$

Then

$$\begin{aligned} F_{T_n}(t) &= \mathbf{P}\big(n(\theta - X_{(n)}) \le t\big) \\ &= 1 - F_M\left(\theta - \frac{t}{n}\right) \\ &\overset{n\to\infty}{\to} 1 - e^{-t/\theta}, \quad t \ge 0. \end{aligned}$$

Without loss of generality, assume $\theta = 1$. Then for $t \ge 0$,

$$\begin{aligned} \mathbf{P}_{F_n}(T_n^\star \le t) &\ge \mathbf{P}_{F_n}(T_n^\star = 0) \\ &= \mathbf{P}_{F_n}\big(X_{(n)}^\star = X_{(n)}\big) \\ &= 1 - \left(1 - \frac{1}{n}\right)^n \\ &\overset{n\to\infty}{\to} 1 - e^{-1}. \end{aligned}$$

Let, for example, $t = 0.0001$; then $\lim_{n\to\infty} \mathbf{P}_{F_n}(T_n^\star \le t) \ge 1 - e^{-1}$, whereas $\lim_{n\to\infty} \mathbf{P}_F(T_n \le t) = 1 - e^{-t} = 1 - e^{-0.0001} \approx 0$. Therefore $\mathbf{P}_{F_n}(T_n^\star \le t) \nrightarrow \mathbf{P}_F(T_n \le t)$. The phenomenon of this example can be generalized essentially to any CDF F with a compact support with some conditions on F , such as existence of a smooth and positive density.

6.3.3 Bootstrap bias and variance estimates

The bootstrap is used in practice for a variety of purposes. It is used to estimate a CDF, or a percentile, or the bias or variance of a statistic T_n. For example, if T_n is an estimate for some parameter θ and if $\mathbf{E}_F(T_n - \theta)$ is the bias of T_n, the bootstrap estimate $E_{F_n}(T_n^\star - T_n)$ can be used to estimate the bias. Likewise, variance estimates can be obtained by estimating $\mathbf{Var}_F(T_n)$ by $\mathbf{Var}_{F_n}(T_n^\star)$. To estimate $\mathbf{Var}_F(T_n)$, we sample B sets of samples of size n with replacement from the original sample set, say

$$(X_{11}^\star,\ldots,X_{1n}^\star),\ldots,(X_{B1}^\star,\ldots,X_{Bn}^\star).$$

We compute $T_i^\star = T(X_{i1}^\star,\ldots,X_{in}^\star)$,$i = 1,\ldots,B$ and their mean $\overline{T}_n^\star$. The bootstrap estimate for the variance is given by $\frac{1}{B}\sum_{i=1}^n (T_i^\star - \overline{T}_n^\star)^2$. One wants to know how accurate the bootstrap-based estimates are in reality. This can only be answered on the basis of case-by-case investigation. Some overall qualitative phenomena have emerged from these investigations. For instance

- The bootstrap distribution estimate captures information about skewness that the CLT will miss.
- The bootstrap tends to underestimate the variance of a statistic T.

That the bootstrap tends to underestimate the variance of a statistic T can be seen as follows. Let $(X_1^\star,\ldots,X_n^\star)$ be an i.i.d. sample drawn for the empirical CDF F_n. Then we have "in the real world"

$$X \sim F \qquad E_F(X) = \mu \qquad \mathbf{Var}_F(X) = \sigma^2,$$

"in the bootstrap world" however,

$$\begin{aligned} &X^\star \mid (X_1,\ldots,X_n) \sim F_n \\ &\mathbf{E}^\star(X^\star) = \mathbf{E}_{F_n}(X^\star) = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n \\ &\mathbf{Var}^\star(X^\star) = \mathbf{Var}_{F_n}(X^\star) = \mathbf{E}_{F_n}(X^\star - \mathbf{E}_{F_n}(X^\star))^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{n-1}{n} S_n^2 \end{aligned}$$

and therefore, regarding the sample properties of $\overline{X}_n$ we have "in the real world"

$$\mathbf{E}(\overline{X}_n) = \mathbf{E}_F(\overline{X}_n) = \mu \quad \text{and} \quad \text{bias}(\overline{X}_n) = 0$$
$$\mathbf{Var}(\overline{X}_n) = \mathbf{Var}_F(\overline{X}_n) = \frac{\mathbf{Var}_F(X)}{n},$$

while "in the bootstrap world"

$$\mathbf{E}^\star(\overline{X}_n) = \mathbf{E}_{F_n}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{F_n}(X_i^\star) = \overline{X}_n$$
$$\mathbf{Var}^\star(\overline{X}_n) = \mathbf{Var}_{F_n}(\overline{X}_n) = \frac{1}{n}\,\mathbf{Var}_{F_n}(X^\star) = \frac{n-1}{n}\frac{S_n^2}{n}.$$

From this we can conclude that the bootstrap estimates of the bias and variance of $T_n = \overline{X}_n$ are zero and $\frac{n-1}{n}\frac{S_n^2}{n}$ respectively.

6.3.4 Bootstrap in R

Let's look at our earlier example of the MLE of the variance. Let $X_1,\ldots,X_n \sim U(0,1), i = 1,\ldots,1000$. The following R code calculates the bootstrap estimate for the bias and variance of the MLE of the variance.

```
> library(bootstrap)
> theta <- function(x){(1/length(x)) * sum((x-mean(x))^2)}
> B <- 10000
> bootres <- bootstrap(x,B,theta)

# Bootstrap estimate for the bias
> mean(bootres$thetastar)-theta(x)
[1] -0.0000893451

# Bootstrap estimate for the standard deviation
> sqrt((1/B) * sum((bootres$thetastar -mean(bootres$thetastar))^2))
[1] 0.002429728
```

Chapter 7

Introduction to nonparametric deconvolution problems

7.1 Introduction

Deconvolution problems occur in many fields of nonparametric statistics, for example, density estimation based on contaminated data (Delaigle and Gijbels, 2006), nonparametric regression with errors-in-variables (Delaigle and Meister, 2011), image and signal deblurring (Qiu, 2005). During the last decades, these topics have received considerable attention. As applications of deconvolution procedures concern many real-life problems in econometrics, biometrics, medical statistics and image reconstruction. On the other hand, some rigorous results from Fourier analysis, functional analysis and probability theory are required to understand the construction of deconvolution techniques and their properties.

The general problem of deconvolution in statistics can be described as follows: Our goal is to estimate a function f while an empirical access is restricted to some quantity

$$z = f * G = \int f(x-y)\, dG(y), \tag{7.1}$$

that is, the convolution of f and some probability distribution G. Hence, the function f can be estimated from some observations only indirectly. The strategy is estimating z first; this means producing an empirical version $\hat{z}$ of z followed by a deconvolution procedure to $\hat{z}$ to estimate f. Therefore, we have to invert the convolution operator with G where some regularization is required to guarantee that $\hat{z}$ is contained in the invertibility domain of the deconvolution operator. The estimator $\hat{z}$ has to be chosen with respect to the statistical experiment. Obviously, to ensure that the specific convolution operator is known, we have to assume that the distribution G is known. Although not realistic in many practical problems, full knowledge of G is assumed in the classical deconvolution approaches (Devroye, 1989). Then, G may be used in the construction of deconvolution estimators. Recent approaches relaxing the exact knowledge of G can be found in Delaigle (2008) and Meister (2009). Nevertheless, as one faces troubles of identifiability in problems with unknown G, either more restrictive conditions on f or additional data (Efromovich, 1997) or repeated measurements are required (Meister, 2009). While there are discrete deconvolution problems (Hall and Qiu, 2005) where all probability mass of G is concentrated on a finite set, the large majority of problems discussed here deals with continuous convolution models, where G has a density function g in the Lebesgue sense. Then, g is called error density or blurring density, according to the corresponding model

$$z = f * g = \int f(x-y)g(y)\, dy.$$

Then, the integral is to be understood in the Lebesgue sense; and f and g are real-valued functions mapping into $\mathbb{R}$.

Let us roughly explain why Fourier methods are very popular in deconvolution problems. The Fourier transform of a distribution G, defined by

$$\mathcal{G}(t) = \int \exp(itx)\, dG(x), \quad t \in \mathbb{R}$$

and also the Fourier transform of a function f (not necessarily a density function), defined in the same way when $dG(x)$ is replaced by $f(x)dx$, are utilized. Throughout this tutorial paper, the Fourier transform is denoted by $\mathcal{G}$ and $\mathcal{F}$, respectively. Using the Fourier transform is motivated by the fact that it changes convolution into simple multiplication. More concretely, (7.1) is equivalent with

$$\mathcal{Z} = \mathcal{F} \cdot \mathcal{G}.$$

It is now clear that the construction of f from z just becomes dividing $\mathcal{Z}$ (empirically accessible) by $\mathcal{G}$ in the Fourier domain. As rough guidelines we give the following scheme for the construction of deconvolution estimators:

1. Estimate $\mathcal{Z}$ based on empirical information, denoted by $\hat{\mathcal{Z}}$.
2. Calculate $\hat{\mathcal{Z}}(t)$ and divide it by $\mathcal{G}(t)$, leading to $\hat{\mathcal{F}}(t)$.
3. Regularize $\hat{\mathcal{F}}(t)$ so that the inverse Fourier transform $\hat{f}$ exists.

However, this scheme seems straight forward but the mathematical effort for the regularization must not be underestimated.

7.2 Density deconvolution

7.2.1 Assumptions and general estimation procedure

In many real-life situations, direct data are not available since measurement error occurs. Then, we observe the contaminated data $Y_1, \ldots, Y_n$ instead of the true data $X_1, \ldots, X_n$. The elementary model of noisy data is the additive measurement error, that is, any empirical access is restricted to the data $Y_1, \ldots, Y_n$ with

$$Y_j = X_j + \varepsilon_j, \quad j \in \{1, \ldots, n\}$$

instead of the incorrupted independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$. Nevertheless, our goal is still to estimate the density f of the incorrupted, but unobserved random variable X. The i.i.d. random variables $\varepsilon_1, \ldots, \varepsilon_n$ represent the error or the contamination of the data; the density of the random variables ε, consequently called error density, is denoted by g. Further, we assume that X_j and ε_j are real valued and independent and that $\mathbf{E}(\varepsilon_j | X_j) = 0$ and $\mathbf{Var}(\varepsilon_j | X_j) < \infty$.

Following the classical approach to this model, g is assumed to be exactly known; although in many real-life situations this condition cannot be justified. However, in most practical applications, we are able to estimate the error density g from replicated measurements. Discussion on the case of unknown g is a current topic of research in the area of nonparametric statistics and therefore we will give a brief summary in Section 7.4 so, in what follows, we assume that g is perfectly known so that we can fully concentrate on the deconvolution step itself. It is an elementary result of probability theory that the density of the sum of two independent random variables is equal to the convolution of the densities of both addends. Hence,

$$z = f * g = \int f(x - y) g(y)\, dy,$$

where z denotes the density of the observation Y. Since any direct empirical access is restricted to z, the problem is included in the general deconvolution framework.

Following the general deconvolution scheme in the introduction, the first step is estimating the density z of the observations Y_j. But we see that the main intention at this stage is making the Fourier transform $\mathcal{Z}$ empirically accessible. As the characteristic function Ψ_Y of a random variable Y is just the Fourier transform of the density of Y, we have

$$\mathcal{Z}(t) = \int \exp(itx) z(x)\, dx = \mathbf{E} \exp(itY) = \Psi_Y(t).$$

By replacing the expectation by averaging with respect to the i.i.d. data yields

$$\hat{\Psi}_Y(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(itY_j).$$

There is a simple multiplicative link between $\mathcal{Z}$ and $\mathcal{F}$ because

$$\begin{aligned}\mathcal{Z}(t) &= \Psi_Y(t) = \mathbf{E}\exp(it(X+\varepsilon)) = \mathbf{E}[\exp(itX)\exp(it\varepsilon)] \\ &= \mathbf{E}\exp(itX)\cdot\mathbf{E}\exp(it\varepsilon) = \Psi_X(t)\cdot\Psi_\varepsilon(t) = \mathcal{F}(t)\cdot\mathcal{G}(t),\end{aligned}$$

where the independence of X and ε implies the independence of $\exp(itX)$ and $\exp(it\varepsilon)$. It would be reasonable to consider

$$\hat{\Psi}_X(t) = \frac{1}{n}\frac{\sum_{j=1}^n \exp(itY_j)}{\mathcal{G}(t)}$$

as an estimator of $\mathcal{F}(t)$ assuming $\mathcal{G}$ is bounded away from zero. It is easy to show that this estimator is unbiased and consistent (by the strong law of large numbers). Hence, by taking the inverse Fourier transform, a naive estimator of f is

$$\hat{f}_{\text{naive}}(x) = \frac{1}{2\pi}\int \exp(-itx)\hat{\Psi}_X(t)\,dt,$$

where the integral is taken over the whole real line. However, the estimator $\hat{f}_{\text{naive}}$ is not well-defined as $\hat{\Psi}_X$ is neither integrable nor square integrable over $\mathbb{R}$. Unlike its true counterpart $\mathcal{F}$ to be estimated, which is square-integrable whenever f is square-integrable, due to Parseval's identity. Apparently, for large $|t|$, $\hat{\Psi}_X(t)$ is no good estimator for $\mathcal{F}(t)$ as the tail behavior is significantly different. Therefore, there is some necessity to regularize $\hat{\Psi}_X$ before the Fourier inversion is employed.

7.2.2 Rozenblatt-Parzen kernel density deconvolution estimator

One of the most well-known method for density estimation based on direct data (i.e., the error-free case) is the Rosenblatt-Parzen kernel density estimator (Parzen, 1962), defined by

$$\hat{z}(x) = \frac{1}{nh}\sum_{j=1}^n K\left(\frac{x-Y_j}{h}\right), \tag{7.2}$$

with kernel function $K:\mathbb{R}\to\mathbb{R}^+$ and a bandwidth parameter $h>0$. If $K\in L_1(\mathbb{R})\cap L_2(\mathbb{R})$, that is, the intersection of the sets of all absolutely or square integrable functions over the whole real line, respectively, in Lebesgue sense, the estimator $\hat{z}$ also lies $L_1(\mathbb{R})\cap L_2(\mathbb{R})$ almost surely so that its Fourier transform exist. It is given by

$$\begin{aligned}\mathcal{Z}(t) &= \frac{1}{nh}\int\sum_{j=1}^n \exp(itx)K\left(\frac{x-Y_j}{h}\right)dx = \frac{1}{nh}\sum_{j=1}^n\int \exp(itx)K\left(\frac{x-Y_j}{h}\right)dx \\ &= \frac{1}{n}\sum_{j=1}^n\int\exp(it(uh+Y_j))K(u)du = \frac{1}{n}\sum_{j=1}^n\exp(itY_j)\int\exp(ithu)K(u)du \\ &= \hat{\Psi}_Y(t)\cdot\mathcal{K}(th),\end{aligned}$$

by using $u=(x-Y_j)/h$ and let $\mathcal{K}$ denote the Fourier transform of the kernel K. Then,

$$\hat{\Psi}_X(t) = \frac{\hat{\Psi}_Y(t)\mathcal{K}(th)}{\mathcal{G}(t)} \tag{7.3}$$

is a second empirical version for $\mathcal{F}$. There are kernel functions whose Fourier transforms are bounded and compactly supported, for example, the de la Vallée Poussin kernel $K(x) = (1/2\pi)[\sin(\frac{1}{2}x)/\frac{1}{2}x]^2$ having Fourier transform $\mathcal{K}(t) = \max(1-|t|,0)$. It can be shown that both integrability and square-integrability of (7.3) hold. Hence, the inverse Fourier transform can be applied to (7.3) leading to the estimator

$$\hat{f}(x) = \frac{1}{2\pi}\int\exp(-itx)\mathcal{K}(th)\frac{\frac{1}{n}\sum_{j=1}^n\exp(itY_j)}{\mathcal{G}(t)}\,dt, \tag{7.4}$$

which is well-defined for any nonvanishing $\mathcal{G}$ whenever $\mathcal{K}$ is compactly supported. This in contrast to the naive estimator $\hat{f}_{\text{naive}}$. The estimator (7.3) has become known as the standard deconvolution kernel density estimator (see Carroll and Hall (1988); Devroye (1989); Stefanski and Carroll (1990) for a rigorous study of this estimator).

Example 7.1 *To illustrate the effect of a contaminated sample on density estimation we consider the following example for 1000 data points. We applied the classical Rosenblatt-Parzen kernel density estimator and the estimator* (7.4) *to simulated examples from two densities* f_X*: (1)* $X \sim 0.5N(-3,1^2) + 0.5N(2,1^2)$ *and (2)* $0.5N(0,1^2) + 0.5N(3,(1/2)^2)$*. The error density* g *is Laplace distributed,* $\mathcal{L}(\mu,b)$*, with location parameter* $\mu = 0$ *and scale parameter* $b = 0.5$*. The Fourier transform of the error density is* $\mathcal{G}(t) = 4/(4+t^2)$*. For the two examples we have chosen de la Vallée Poussin kernel.*

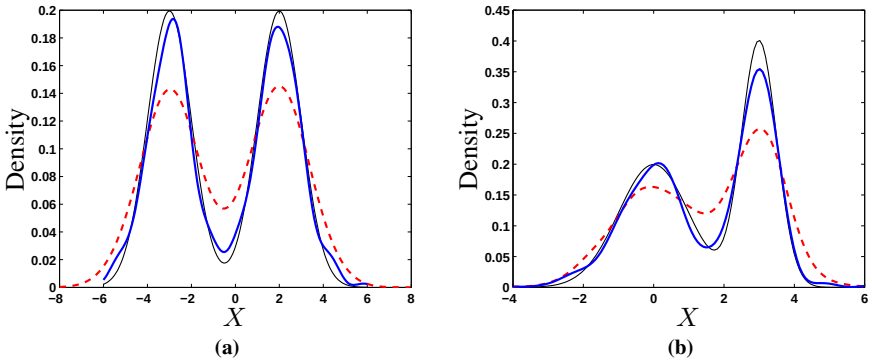


Figure 7.1: (a), (b) Effect of a contaminated sample on density estimation for two normal mixtures. The thine line is the true density, bold line is the estimated density based on (7.4) and bold dashed line represents the estimate based on the Rosenblatt-Parzen kernel density estimator.

7.3 Nonparametric regression with errors-in-variables

7.3.1 Errors-in-variables problem formulation

As a broad field in statistics in general, the investigation of the link or the dependence between some quantity, which is affected by random noise, and some circumstances, under which the quantity is observed, is referred to as regression. We assume that those circumstances may be represented by a real number X, which is called the covariate or the independent variable. In the standard nonparametric measurement error model, we assume that the covariates X can only be observed with some additive independent noise. Therefore, we change the observation scheme into the i.i.d. dataset $(W_1,Y_1),\ldots,(W_n,Y_n)$, where

$$W_j = X_j + \delta_j \qquad \text{and} \qquad Y_j = m(X_j) + \varepsilon_j \quad \text{for } j = 1,\ldots,n \tag{7.5}$$

and m is the regression function. The covariate errors δ_j are i.i.d. unobservable random variables having error density g. Note that they are different from the regression errors ε_j. The δ_j are stochastically independent of the X_j and the Y_j. As in the previous section on density deconvolution, we assume that g is known in the standard setting, while the distribution of the ε_j need not be known.

7.3.2 Kernel regression with errors-in-variables

In case the covariates are not affected by contamination, the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) is a well-known kernel regression estimator. It is defined as follows

$$\hat{m}(x) = \sum_{j=1}^{n} \frac{K\left(\frac{x-X_j}{h}\right) Y_j}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)},$$

with $K : \mathbb{R} \to \mathbb{R}$ and bandwidth $h > 0$. As we may equivalently multiply the numerator and the denominator by $1/(nh)$, we realize the close relation to the kernel density estimator (7.2). Indeed, the kernel density estimator of f_X based on the i.i.d. data $X_1,\ldots,X_n$ occurs as the denominator of $\hat{m}$.

We can now focus on extending the Nadaraya-Watson estimator to our contaminated data $(W_1,Y_1),\ldots,(W_n,Y_n)$. The denominator of the Nadaraya-Watson estimator may be replaced by the deconvolution kernel density estimator (7.4) using the data $W_1,\ldots,W_n$, which are additively corrupted by unobservable random variables with density g. Then the denominator is an empirical version of the density f_X as in the error-free setting. We must also alter the numerator of the Nadaraya-Watson estimator so it does not require knowledge of the unobservable data $X_1,\ldots,X_n$ but only uses the data $W_1,\ldots,W_n$. This can be done as follows. The kernel deconvolution estimator (7.4), based on the data $W_1,\ldots,W_n$, can be written as

$$\frac{\sum_{j=1}^{n}\int \exp(-itx)\mathcal{K}(th)\frac{\exp(itW_j)}{\mathcal{G}(t)}dt}{2n\pi} = \frac{\sum_{j=1}^{n}\int \exp\left[-i\left(\frac{x-W_j}{h}\right)u\right]\frac{\mathcal{K}(u)}{\mathcal{G}(\frac{u}{h})}du}{2\pi nh},$$

by using the substitution $u=th$. Then, the latter can be written as

$$\hat{f}(x)=\frac{1}{nh}\sum_{j=1}^{n}H\left(\frac{x-W_j}{h}\right),$$

where

$$H(x)=\frac{1}{2\pi}\int \exp(-ixu)\frac{\mathcal{K}(u)}{\mathcal{G}(\frac{u}{h})}\,du. \tag{7.6}$$

By appealing to (7.4), (4.25) and (7.6), the following kernel regression estimator involving errors-in-variables is given by Fan and Truong (1993)

$$\hat{m}(x)=\frac{\frac{1}{nh}\sum_{j=1}^{n}H\left(\frac{x-W_j}{h}\right)Y_j}{\frac{1}{2n\pi}\sum_{j=1}^{n}\int \exp(-itx)\mathcal{K}(th)\frac{\exp(itW_j)}{\mathcal{G}(t)}\,dt}. \tag{7.7}$$

The steps in deriving the Nadaraya-Watson kernel regression estimator involving errors-in-variables boil down to replacing the unobserved $K\left(\frac{x-X_j}{h}\right)$ by an observable quantity $H\left(\frac{x-W_j}{h}\right)$ satisfying (see also Delaigle et al. (2009))

$$\mathbf{E}\left[H\left(\frac{x-W_j}{h}\right)|X_j\right]=K\left(\frac{x-X_j}{h}\right).$$

In the usual nomenclature of measurement error models, this simply means that $H\left((x-W_j)/h\right)$ is an unbiased score for the kernel function $K\left((x-X_j)/h\right)$.

Example 7.2 *To illustrate the effect of a contaminated sample on regression estimation via model* (7.5)*, we consider the following example for 350 equispaced data points. We consider the following regression function*

$$m(x)=2x\exp(-10x^4/81) \quad \textit{with} \quad x\in[-2,2]$$

with $\varepsilon\sim N(0,0.2^2)$, $\delta\sim\mathcal{L}(0,0.2)$, $\mathbf{E}[\varepsilon|X]=0$ *and the* δ_j *are independent of the* (X_j,Y_j)*. Figure 7.2 shows the result for the artificial data set and illustrate the difference between the observed and uncontaminated data.*

Remark 7.1 (Berkson regression) *There is another errors-in-variables problem in nonparametric regression, which is closely related to the model* (7.5)*, but not identical. In literature, it is usually referred to as the Berkson regression model. It was first mentioned in the paper of Berkson (1950) which had been published before conventional nonparametric techniques such as kernel smoothing were introduced. The main difference between the Berkson model and* (7.5) *concerns the fact that, in the Berkson context, the covariate is affected by additive noise after it was measured. In the Berkson model, we observe the i.i.d. data* $(X_1,Y_1),\ldots,(X_n,Y_n)$ *where*

$$Y_j=m(X_j+\delta_j)+\varepsilon_j, \quad j=1,\ldots,n.$$

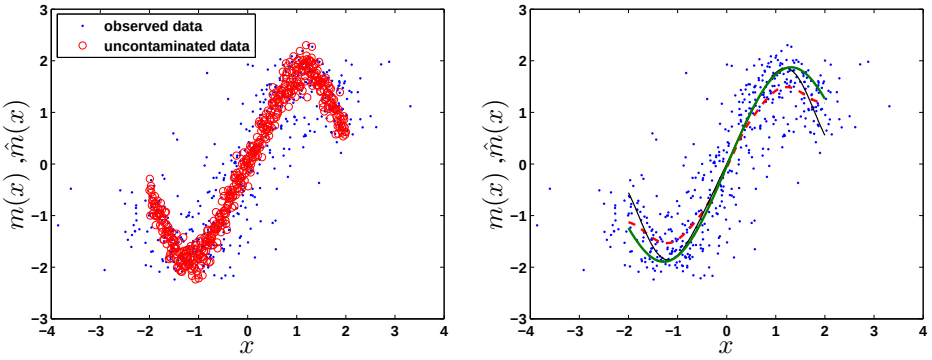


Figure 7.2: (a) Difference between the observed and uncontaminated data;(b) Effect of a contaminated sample on regression estimation. The thine line is the true regression function, the bold line is the regression estimate based on (7.7) and the bold dashed line represents the regression estimation using the standard Nadaraya-Watson estimator assuming the error-free case.

7.4 Current state-of-the-art

The current state-of-the-art regarding deconvolution methods is to relax the assumption that the error density g has to be known. Several approaches exist to estimate this density. A first approach is based on additional data. This means that the error density g is unknown but can be estimated directly from i.i.d. data $\varepsilon'_1, \ldots, \varepsilon'_n$, which are collected in a separate independent experiment. This model was studied in Efromovich (1997); Neumann (1997) and in the book Efromovich (1999). Of course, its applicability is restricted to cases where the system of measurement can be calibrated somehow. In particular, the model should be considered when, in some cases, the same individual or quantity can be observed both in an error-free way (call this measurement $X'_{j,1}$) and by a measurement procedure, which is affected by nonnegligible noise (denote this observation by $X'_{j,2}$). Then put $\varepsilon'_j = X'_{j,2} - X'_{j,1}$ where ε'_j is indeed a direct observation from the error distribution. The previously estimated error density may be employed in the deconvolution step for that latter data set.

A second approach is based on replicated measurements. Here, the same uncorrupted but unobserved random variable X_j is independently measured for several times, but each measurement is affected by error. Suppose we observe data $Y_{j,k}$, $j = 1, \ldots, n$ and $k = 1, \ldots, m_j$ with $m_j \neq 2$ defined by

$$Y_{j,k} = X_j + \varepsilon_{j,k}.$$

Then each $\varepsilon_{j,k}$ has error density g. Consider the accessible differences (set $m_j = 2$)

$$\Delta Y_j = Y_{j,1} - Y_{j,2} = \varepsilon_{j,1} - \varepsilon_{j,2}, \quad j \in \{1, \ldots, n\}.$$

It can be shown that characteristic function of ΔY_j yields

$$\psi_{\Delta Y_j}(t) = \mathbf{E} \exp(it\Delta Y_j) = |\mathcal{G}(t)|^2.$$

Hence, under some conditions on $\mathcal{G}$, a reasonable estimate would be

$$\hat{\mathcal{G}}(t) = \left| \frac{1}{n} \sum_{j=1}^{n} \exp(it\Delta Y_j) \right|^{1/2}$$

as an estimator for $\mathcal{G}(t)$. Such approaches have been studied in Horowitz and Markatou (1996) and Delaigle et al. (2009). Delaigle et al. (2009) also extended local polynomial regression to the error-in-variables framework.

Other topics within this area such as model selection methods Delaigle and Gijbels (2006), optimal kernels for deconvolution Delaigle and Hall (2006) and establishing all its statistical properties is an active research area. But perhaps one of the most challenging part is to implement all these methods in a numerically stable way, see e.g. Hall and Qiu (2005) and Meister (2009).

References

I.S. Abramson. On bandwidth variation in kernel estimates–a square root law. *The Annals of Statistics*, 4(400): 168–176, 1982.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

A. Azzalini and A.W. Bowman. A look at some data on the old faithful geyser. *Applied Statistics*, 39(3):357–365, 1990.

R.J. Beran. Jackknife approximation to bootstrap estimates. *The Annals of Statistics*, 12(1):101–118, 1984.

J. Berkson. Are there two regression problems? *Journal of the American Statistical Association*, 45:164–180, 1950.

P. Bickel and D. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(1196–1217), 1981.

S. Bochner. *Harmonic Analysis and the Theory of Probability*. University of California Press, 1955.

A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2): 353–360, 1984.

P. Burman. A comparative study of ordinary cross-validation, v-fold crossvalidation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

P. Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā: The Indian Journal of Statistics, Series A*, 52(3):314–345, 1990.

R.J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

D.B.H. Cline and J.D. Hart. Kernel estimation of densities with discontinuous derivatives. *Statistics*, 22:69–84, 1991.

J.B. Copas and M.J. Fryer. Density estimation and suicide risks in psychiatric treatment. *Journal of the Royal Statistical Society. Series A*, 143(2):167–176, 1980.

A. Cowling and P. Hall. On pseudodata methods for removing boundary effects in kernel density estimation. *J. Royal. Stat. Soc. B*, 58:551–563, 1996.

A.C. Davison and D.V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 2003. (reprinted with corrections).

K. De Brabanter, P. G. Ferrario, and L. Györfi. Detecting ineffective features for nonparametric regression. In J. A. K. Suykens, M. Signoretto, and A. Argyriou, editors, *Regularization, Optimization, Kernels, and Support Vector Machines*, Machine Learning & Pattern Recognition, To Appear. Chapman & Hall, 2014.

A. Delaigle. An alternative view of the deconvolution problem. *Statistica Sinica*, 18(3):1025–1045, 2008.

A. Delaigle and I. Gijbels. Data-driven boundary estimation in deconvolution problems. *Computational Statistics & Data Analysis*, 50(8):1965–1994, 2006.

A. Delaigle and P. Hall. On optimal kernel choice for deconvolution. *Statistics & Probability Letters*, 76:1594–1602, 2006.

A. Delaigle and A. Meister. Rate-optimal nonparametric estimation in classical and Berkson errors-in-variables. *Journal of Statistical Planning and Inference*, 141(1):102–114, 2011.

A. Delaigle, J. Fan, and R.J. Carroll. A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*, 104(485):348–359, 2009.

L. Devroye. Consistent deconvolution in density estimation. *The Canadian Journal of Statistics*, 17(2):235–239, 1989.

L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 view*. Wiley, 1985.

L. Devroye, P. G. Ferrario, L. Györfi, and H. Walk. Strong universal consistent estimate of the minimum mean squared error. In B. Schölkopf, Z. Luo, and V. Vovk, editors, *Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik*, chapter 14, pages 143–160. Springer, 2013.

S. Efromovich. Density estimation for the case of supersmooth measurement error. *Journal of the American Statistical Association*, 92(438):526–535, 1997.

S. Efromovich. *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer-Verlag, 1999.

B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

V. A. Epanechnikov. Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Applications*, 13:153–158, 1969.

C-G. Esseen. On the liapunoff limit of error in the theory of probability. *Arkiv för matematik, astronomi och fysik*, A28:1–19, 1942.

C-G. Esseen. A moment inequality with an application to the central limit theorem. *Skand. Aktuarietidskr.*, 39: 160–170, 1956.

J. Fan. Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, 21:196–216, 1993.

J. Fan and I. Gijbels. Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *J. Comp. Graph. Statist.*, 4:213–227, 1995.

J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Applications*. Chapman & Hall, 1996.

J. Fan and J.S. Marron. Fast implementation of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3(1):35–56, 1994.

J. Fan and Y.K. Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, 21(4):1900–1925, 1993.

J. Fan, I. Gijbels, T.-C. Hu, and L.-S. Huang. An asymptotic study of variable bandwidth selection for local polynomial regression with application to density estimation. *Statistica Sinica*, 6(1):113–127, 1996.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

P. Hall and J.S. Marron. Extent to which least-squares cross-validation minimises integrated squared error in nonparametric density estimation. *Probab. Theory Rel. Fields*, 74:567–581, 1987.

P. Hall and P. Qiu. Discrete-transform approach to deconvolution problems. *Biometrika*, 92(1):135–148, 2005.

P. Hall, J. W. Kay, and D. M. Titterington. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

P. Hall, J.S. Marron, and B.U. Park. Smoothed cross-validation. *Probability Theory and Related Fields*, 92:1–20, 1992.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

N.L. Hjort and M.C. Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24(4): 1619–1647, 1996.

J.L. Horowitz and M. Markatou. Semiparametric estimation of regression models for panel data. *The Review of Economic Studies*, 63:145–168, 1996.

M.C. Jones. Simple boundary correction for kernel density estimation. *Statist. Comput.*, 3:135–146, 1993.

M.C. Jones and P.J. Foster. Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, 3 (1):81–94, 1993.

M.C. Jones and P.J. Foster. A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sinica*, 6:1005–1013, 1996.

M.C. Jones and R.F. Kappenman. On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, 19(4):337–349, 1991.

M.C. Jones and R.F. Kappenman. On a class of kernel density estimate bandwidth selectors. *Scand. J. Statistics*, 19 (4):337–349, 1992.

M.C. Jones, J.S. Marron, and B.U. Park. A simple root-n bandwidth selector. *Ann. Statist.*, 19:1919–1932, 1991.

D.O. Loftsgaarden and C.P. Quesenberry. A nonparametric estimate of a multivariate density. *Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Y.P. Mack and H.-G. Müller. Convolution type estimators for nonparametric regression estimation. *Statist. Prob. Lett.*, 1:229–239, 1989.

A. Meister. *Deconvolution Problems in Nonparametric Statistics*. Springer-Verlag, Berlin Heidelberg, 2009.

H.-G. Müller. Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statistics & Decisions*, Supplement 2:193–206, 1985.

E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

E.A. Nadaraya. On non-parametric estimates of a density function. *Theor. Probability Appl.*, 10:186–190, 1965.

M.H. Neumann. On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7:307–330, 1997.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3): 1065–1076, 1962.

K. Pearson. Contributions to the mathematical theory of evolution–II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186:343–414, 1895.

P. Qiu. *Image Processing and Jump Regression Analysis*. John Wiley & Sons, New York, 2005.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2): 64–78, 1982.

D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

H. Scheffé. A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18(3): 434–458, 1947.

E.F. Schuster. Estimation of a probability density and its derivatives. *Ann. Math. Statist.*, 40:1187–1196, 1969.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

D.W. Scott and G.R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987.

S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B*, 53(3):683–690, 1991.

I.G. Shevtsova. Sharpening of the upper bound of the absolute constant in the berry–esseen inequality. *Theory of Probability and its Applications*, 51(3):549–553, 2007.

I.G. Shevtsova. On the absolute constant in the berry–esseen inequality. *The Collection of Papers of Young Scientists of the Faculty of Computational Mathematics and Cybernetics*, 5:101–110, 2008.

I.G. Shevtsova. On the absolute constants in the berry esseen type inequalities for identically distributed summands. *arXiv:1111.6554*, 2011.

I.S. Shiganov. Refinement of the upper bound of a constant in the remainder term of the central limit theorem. *Journal of Soviet Mathematics*, 35(3):109–115, 1986.

B.W. Silverman. Weak and strong uniform consistency of the kernel estimate of density and its derivatives. *Ann. Stat.*, 6(2):177–184, 1978.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Springer, 1986.

J.S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

J.G. Staniswalis. Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, 84: 49–54, 1985.

L.A. Stefanski and R.J. Carroll. Deconvoluting kernel density estimators. *Statistics*, 21:169–184, 1990.

C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–645, 1977.

G.R. Terrell. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410):470–477, 1990.

R. Tibshirani and R. Tibshirani. A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics*, 3(2):822–829, 2009.

I.S. Tyurin. On the accuracy of the gaussian approximation. *Doklady Mathematics*, 80(3):840–843, 2009.

I.S. Tyurin. An improvement of upper estimates of the constants in the lyapunov theorem. *Russian Mathematical Surveys*, 65(3(393)):201–202, 2010.

P. van Beek. An application of fourier methods to the problem of sharpening the berry–esseen inequality. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 23(3):187–196, 1972.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

J. Van Ryzin. On strong consistency of density estimates. *Ann. Math. Statist.*, 40:1765–1772, 1969.

M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

G.S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.