7

Choosing the Smoothing Parameter

1. Introduction

We have now come to just about the most important aspect of nonparametric density estimation: choosing the smoothing parameter in kernel estimation that will give near-optimal results for large classes of densities. The same problem arises for maximum penalized likelihood estimation, or any other method for that matter. Actually, in kernel estimation, both the kernel and the smoothing parameter need to be chosen appropriately, whereas in maximum penalized likelihood estimation, the roughness penalization functional and the smoothing parameter are subject to choice. Choosing the roughness penalization is essentially uncharted territory. The authors do not even know whether it is an important question, and we shall not explore it. As discussed in § 4.7, in kernel estimation, the choice of the smoothing parameter is much more critical than the choice of the kernel. Why this should be so is not *a priori* clear, although deep statistical insights could be quoted.

In this chapter, we study some current methods for choosing the smoothing parameter h in the kernel estimator

$$f^{nh}(x) = \frac{1}{n} \sum_{i=1}^{n} A_h(x - X_i) \;, \quad x \in \mathbb{R} \;, \tag{1.1}$$

to wit, least-squares cross-validation and least-squares plug-in methods, the double kernel method, various L^1 plug-in methods, and a method based on a discrepancy principle. The L^1 plug-in methods require pilot estimators, of which we single out the double kernel method. We also discuss variational analogues of the plug-in methods, in which there is no need for pilot estimators. There are many more methods for smoothing parameter selection, of which methods based on the bootstrap and on spacings should be mentioned. However, these are not considered here. Instead, the reader is referred to the surveys and (simulation) comparisons in Park and Turlach (1992), Berlinet and Devroye (1994), Cao, Cuevas and González-Manteiga (1994), and Devroye (1997). We also dis-

cuss a discrepancy principle for selecting the smoothing parameter for the GOOD estimator of § 5.2.

In the remainder of this introduction, we make some general observations regarding smoothing parameter selection procedures. Three questions are addressed. The first one addresses the issue of quantifying what the "optimal" smoothing parameter is supposed to achieve. In Chapter 8, this provides the basis for the objective comparison of various selection procedures with each other. The second question concerns the kind of densities one is likely to encounter in practice, and with the desired asymptotic behavior of the selected smoothing parameter under these circumstances. The third question deals with the fact that the smoothing parameter should dependent on the data only, and not on the intuition or the deep statistical insight of the experimenter. One requirement is that the selected smoothing parameter should be scaling and translation invariant.

What is the purpose of selecting the smoothing parameter? From the L^1 point of view of this text, the "optimal" method would select h for each sample so as to

$$\text{minimize} \quad \| f^{nh} - f_o \|_1 \quad \text{over } h > 0 \ . \tag{1.2}$$

This may be called *finite sample optimality*, which in practice must be deemed unattainable, even if we restrict f_o to "reasonable" classes of densities. A perhaps more accessible goal is to minimize the "risk", that is,

$$\text{minimize} \quad \mathbb{E}\big[\| f^{nh} - f_o \|_1 \big] \quad \text{over } h > 0 \ , \tag{1.3}$$

but even this is much too hard to achieve. The next choice is to strive for *asymptotic* optimality, that is, construct any kernel estimator $f_{n,ANY}$ that satisfies for *every* density f_o,

$$\limsup_{n\to\infty} \ \frac{\| f_{n,ANY} - f_o \|_1}{\min_h \ \| f^{nh} - f_o \|_1} =_{\text{as}} 1 \ , \tag{1.4}$$

but perhaps with expected values in both the numerator and the denominator. This is still hard to achieve. The best result until now is by DEVROYE and LUGOSI (1996), (1997): For every $\varepsilon > 0$, they can construct methods for which the limsup is $\leqslant_{\text{as}} 3 + \varepsilon$, for every density. For densities f_o satisfying the usual nonparametric assumptions, the kernel method satisfies the expected value version of (1.4), see DEVROYE (1989) and § 3 below. With L^2 norms and for bounded densities (1.4) is the famous result of STONE (1984) for least-squares cross-validation, see § 2. The practical significance of these *universal* asymptotically optimal selection procedures is limited, since typically the density to be estimated is known (assumed) to be smooth or to satisfy certain shape constraints. Morover, one is dealing with the small sample case, and small sample adjustments have to be made. This explains the plethora of techniques in the literature, of which we cover only a few.

What kind of densities are we likely to encounter in nonparametric density estimation? Naturally, we make the usual nonparametric assumptions regarding smoothness and light tails. The smoothness assumption appears to be controversial, but the authors find the following justification convincing. With small sample sizes, one can reasonably hope to recover only the global features of a density, and one must consider the small-scale features to be inaccessible. Alternatively, for small sample sizes, one cannot hope to distinguish between a very rough density and a smoothed version of it, cf. Exercise (8.1.1) in the next chapter. This is tantamount to saying that only a smoothed version of the unknown density can be estimated well. So, exaggerating a bit, in the small sample case, all densities are smooth. Regarding the tail conditions, we note that evidence regarding the (alleged) light tails is embodied in the sample and, thus, is open to inspection. However, the existence of a finite moment of order > 1 allows (roughly) the tail behavior

$$f_o(x) = \mathcal{O}\big(\,|\,x\,|^{-\alpha}\,\big)\ , \quad |\,x\,| \to \infty\ , \tag{1.5}$$

for some $\alpha > 2$. This should be contrasted with the two-sided exponential density, which in practice still has quite heavy tails. (Finite samples contain many outliers, see § 2.5.) Thus, the nonparametric moment condition is not very stringent.

We next discuss the desired asymptotic behavior of the smoothing parameter, when considering densities that satisfy the usual nonparametric assumptions. Recall from Chapter 4 that for these densities, the asymptotically optimal smoothing parameter satisfies $h_{\text{asymp}} \asymp n^{-1/5}$, with the corresponding L^1 error of order $n^{-2/5}$. So, as a minimal requirement, it is reasonable to insist that H_n, the smoothing parameter chosen, satisfies

$$H_n \asymp_{\text{as}} n^{-1/5} \quad \text{for } n \to \infty\ , \tag{1.6}$$

by which we mean that

$$0 < \liminf_{n\to\infty}\ n^{1/5}\, H_n \leqslant \limsup_{n\to\infty}\ n^{1/5}\, H_n < \infty \quad \text{almost surely}\ , \tag{1.7}$$

and that

$$\|\, f^{nH_n} - f_o\,\|_1 =_{\text{as}} \mathcal{O}\big(\, n^{-2/5}\,\big)\ . \tag{1.8}$$

In statements like (1.6), we usually drop the qualification $n \to \infty$, but it is intended nevertheless. Equation (1.8) is a case in point. As discussed before, one would like to achieve asymptotic or even small sample optimality, but that is outside of the scope of this text.

For a few selection procedures to be discussed later, we prove (1.6), under the usual assumptions. We also show, in § 3, that (1.6) implies (1.8), provided A is the Gaussian or two-sided exponential kernel. For general kernels, the proof does not work. Then, our only alternative is to use fractional integration by parts, see § 4.3, but this yields only the suboptimal rate of $n^{-2/5+\varepsilon}$, for arbitrary $\varepsilon > 0$. However, we venture to guess

that here too (1.6) implies (1.8). For general kernels, the expected value version is treated in the next exercise.

(1.9) EXERCISE. Suppose H_n satisfies (1.6), i.e., for deterministic constants $0 < c < C < \infty$, assume that $H_n \in I_n$ almost surely, where

$$I_n = [\, h_n \,,\, h^n \,] \quad \text{with} \quad h_n = c\, n^{-1/5} \quad , \quad h^n = C\, n^{-1/5} \ .$$

Show that the bound (1.8) holds, under the usual nonparametric assumptions on f_o, and suitable conditions on A, as follows.
(a) Show that $\| A_h * (dF_n - dF_o) \|_1$ is a.e. differentiable with respect to h, and that

$$\left| \frac{d}{dh} \| A_h * (dF_n - dF_o) \|_1 \right| \leqslant h^{-1} \| B_h * (dF_n - dF_o) \|_1 \ ,$$

where $B(x) = -\dfrac{d}{dx}\Big\{ x\, A(x) \Big\}$ and, as usual, $B_h(x) = h^{-1}\, B(h^{-1}\, x)$.
(b) Show that

$$\| A_{H_n} * (dF_n - dF_o) \|_1 \leqslant \sup_{h \in I_n} \| A_h * (dF_n - dF_o) \|_1 \ ,$$

(c) and that

$$\sup_{h \in I_n} \| A_h * (dF_n - dF_o) \|_1 \leqslant$$

$$\| A_{h_n} * (dF_n - dF_o) \|_1 + \int_{h_n}^{h^n} h^{-1} \| B_h * (dF_n - dF_o) \|_1 \, dh \ .$$

(d) Now, take expectations in (b) and (c), and take care of the difference between $A_{H_n} * (dF_n - dF_o)$ and $A_{H_n} * dF_n - f_o$.

Finally, how should the selected h depend on the data? It goes almost without saying that the selected h should be a statistic. Equivalently, the procedure should be "rational", and not require input from the user, whatever that might mean exactly. However, how complicated a function of the data must it be? To partly answer this question, we investigate the scaling invariance of the smoothing parameter. By way of example, whether the Buffalo snowfall data are presented in inches or centimeters, one should insist that the selected h change accordingly, so that the two estimators based on the data in inches or in centimeters are "the same". A precise way of saying this is as follows. Suppose that the random variable X has density f_o. Then, for any (deterministic) $t > 0$, the random variable $t\,X$ has density f_t, with $f_t(x) = t^{-1} f_o(t^{-1}x)$, $t > 0$ (so $f_1 = f_o$). Let $X_1, X_2, \cdots, X_n$ be an iid sample with common pdf f. To simplify notation, let

$$\mathbb{X}_n = (X_1, X_2, \cdots, X_n) \in \mathbb{R}^{1\times n} \ , \tag{1.10}$$

so then $t\,\mathbb{X}_n = (\, t\,X_1, t\,X_2, \cdots, t\,X_n\,)$ is an iid sample with common density f_t. Now, consider a kernel estimator of f based on the sample $\mathbb{X}_n$,

written as

$$f^{nh}(x\,;\,\mathbb{X}_n\,) = \tfrac{1}{nh}\,\sum_{i=1}^{n} A\big(h^{-1}(x - X_i)\big)\ , \quad -\infty < x < \infty\ . \tag{1.11}$$

Then, the corresponding kernel estimator of f_t based on $t\,\mathbb{X}_n$ is

$$f^{nh}(x\,;\,t\,\mathbb{X}_n\,) = \tfrac{1}{nh\,t}\,\sum_{i=1}^{n} A\big(h^{-1}(x - t\,X_i)\big)\ , \tag{1.12}$$

and so,

$$f^{nh}(x\,;\,t\,\mathbb{X}_n\,) = t^{\,-1}\,f^{n,h\,t}(\,t^{\,-1}\,x\,;\,\mathbb{X}_n\,)\ , \quad -\infty < x < \infty\ . \tag{1.13}$$

Now, suppose a hypothetical selection procedure applied to $\mathbb{X}_n$ selects the smoothing parameter $H_{n,HYP} = H_{n,HYP}(\,\mathbb{X}_n\,)$. Denoting the corresponding kernel density estimator (1.11) as

$$f^{n,H_{n,HYP}}(x\,;\,\mathbb{X}_n\,) = f_{n,HYP}(x\,;\,\mathbb{X}_n\,)\ , \tag{1.14}$$

the two estimators are "the same" if

$$f_{n,HYP}(x\,;\,\mathbb{X}_n\,) = t^{\,-1}\,f_{n,HYP}(\,t^{\,-1}x\,;\,t\,\mathbb{X}_n\,)\ , \quad -\infty < x < \infty\ . \tag{1.15}$$

This occurs if the selected $H_{n,HYP}$ is scaling invariant in the sense that

$$H_{n,HYP}(\,\mathbb{X}_n\,) = t^{\,-1}\,H_{n,HYP}(\,t\,\mathbb{X}_n\,)\ . \tag{1.16}$$

(1.17) Exercise. (a) Verify (1.13) and that (1.16) implies (1.15).
(b) Show that for all $t > 0$,

$$\|\,f_t - f^{n,h\,t}(\,\cdot\,;\,t\,\mathbb{X}_n)\,\|_1 = \|\,f_o - f^{n,h}(\,\cdot\,;\mathbb{X}_n)\,\|_1\ .$$

It is easy to construct scaling-invariant smoothing parameters that even satisfy the asymptotic size information of (1.6), e.g., take H_n as

$$H_n = c\,n^{-1/5}\,\Big\{\,\tfrac{1}{n}\,\sum_{i=1}^{n}\,|\,X_i - \overline{X}\,|^{\,2}\,\Big\}^{1/2}\ , \tag{1.18}$$

where c is a universal constant and $\overline{X}$ is the sample mean. Unfortunately, this hardly solves the problem. To illustrate this, consider the problem of estimating a normal density versus estimating a mixture of two normals, see Figure 1.1. The two densities in question are the normal $\phi_\sigma(x)$ with $\sigma = 0.714$ and the mixture

$$\tfrac{9}{10}\,\phi_{1/2}(x-5) + \tfrac{1}{10}\,\phi_{1/2}(x-7)\ ,$$

which have the same standard deviations. Here, as usual, $\phi_\sigma = \sigma^{-1}\,\phi(\sigma^{-1}x)$. Graphs of these densities are shown in the left diagram of Figure 1.1. In the right diagram of Figure 1.1, we show the L^1 errors of the kernel estimators as functions of h, for two "typical" samples of size 100.

Two conclusions may be drawn. First, the mixture of normals being "rougher" than the normal, the (asymptotically) optimal values of the

Figure 1.1. On the left, graphs of the normal density $\phi_\sigma(x-5)$ with $\sigma = 0.714$ and the mixture $\frac{9}{10}\phi_{1/2}(x-5) + \frac{1}{10}\phi_{1/2}(x-7)$, which have the same variance. On the right, graphs of the L_1 errors (as functions of h) of kernel estimators using the normal kernel, for two "typical" iid samples of size 100 from each density. The locations of both minima are indicated on both curves.

smoothing parameters are quite different in each case. Secondly, it is clear that a lot would be lost if a single value of h were used in both cases. The conclusion is that a data-driven smoothing parameter like (1.18) is not fully satisfactory in practice.

The remainder of this chapter is put together as follows. We discuss L^2 cross validation and L^2 plug-in methods in § 2. The remainder of the chapter deals with L^1 errors: For kernel estimation, the double kernel method and a compelling modification are discussed in § 3, and small sample and asymptotic plug-in methods in §§ 4 and 5. A discrepancy principle for kernel estimators and the GOOD estimator is discussed in §§ 6 and 7. Heuristic justifications of the various methods are given, as well as some proofs regarding asymptotic rates of convergence, but only when it can be done by the methods of Chapters 4 and 5.

EXERCISES: (1.9), (1.17).

2. Least-squares cross-validation and plug-in methods

In this section, we take the L^2 point of view, and study chosing h so as to

$$\text{minimize} \quad \| f^{nh} - f_o \|_2^2 \quad \text{subject to} \quad h > 0 \,. \tag{2.1}$$

The methods discussed are L^2 cross validation and plug-in methods. In the least-squares cross-validation method, one minimizes an *unbiased* estimator

of $\| f^{nh} - f_o \|_2^2$. The idea was first considered by RUDEMO (1982) and BOWMAN (1984), and furthered by HALL (1983) and STONE (1984). See also WAHBA (1981). The second type of methods to be considered are plug-in methods, based on minimizing asymptotic expressions for the expected squared L^2 error. As discussed earlier, the drawback of these approaches is that they deal with the L^2 error, which has no obvious interpretation in the context of estimating densities.

In the cross-validation approach, one first derives an unbiased estimator of $\| f^{nh} - f_o \|_2^2$ by observing that

$$\| f^{nh} - f_o \|_2^2 = \| f^{nh} \|_2^2 + \| f_o \|_2^2 \ - 2 \int_{\mathbb{R}} f^{nh}(x)\, dF_o(x) \ . \tag{2.2}$$

Now, the second term on the right is independent of h, and since ultimately we wish to minimize over h, only the last term needs to be estimated. It was written in a rather suggestive manner: A "natural" estimator for it is

$$\int_{\mathbb{R}} f^{nh}(x)\, dF_o(x) \approx \int_{\mathbb{R}} f^{nh}(x)\, dF_n(x) = \tfrac{1}{n} \sum_{i=1}^{n} f^{nh}(X_i) \ . \tag{2.3}$$

However, this turns out to be a *biased* estimator of $\int_{\mathbb{R}} f^{nh} dF_o$. This may be traced to the fact that

$$f^{nh}(X_i) = (nh)^{-1} A(0) + \tfrac{1}{n} \sum_{j \neq i} A_h(X_i - X_j) \ , \tag{2.4}$$

and it is clear (?) that the first term does not "belong". Thus, the biasedness may be fixed by using the approximation

$$\int_{\mathbb{R}} f^{nh}(x)\, dF_o(x) \approx \tfrac{1}{n} \sum_{i=1}^{n} f^{nh}_{(i)}(X_i) \ , \tag{2.5}$$

where

$$f^{nh}_{(i)}(x) = \tfrac{1}{n-1} \sum_{j \neq i} A_h(x - X_j) \ , \quad x \in \mathbb{R} \ . \tag{2.6}$$

One interpretation of (2.6) is that we are estimating $f_o(X_i)$ by a kernel estimator based on the data with X_i omitted. For this reason, this method goes by the name of the "leave-one-out method", but "cross validation method" is the standard designation.

To summarize, if we set

$$CV(h) = \| f^{nh} \|_2^2 - \tfrac{2}{n} \sum_{i=1}^{n} f^{nh}_{(i)}(X_i) \ , \tag{2.7}$$

then

$$\mathbb{E}\big[\, CV(h) \,\big] = \| f^{nh} - f_o \|_2^2 - \| f_o \|_2^2 \ . \tag{2.8}$$

(2.9) EXERCISE. (a) Show that for $i \neq j$,

$$\mathbb{E}[\, A_h(X_i - X_j) \,] = \int_{\mathbb{R}} f_o(x)\, [\, A_h * f_o \,](x)\, dx \ .$$

(b) Verify that the estimator of (2.3) is a biased estimator of $\int_{\mathbb{R}} f^{nh}\, dF_o$, in general.
(c) Verify (2.8).

(2.10) EXERCISE. Verify that

$$CV(h) = (nh)^{-1} \| A \|_2^2 \; - \; [\, n(n-1)\,]^{-1} \sum_{i \neq j} B_h(X_i - X_j) \; - \\ [\, n^2(n-1)\,]^{-1} \sum_{i \neq j} [\, A_h * A_h\,](X_i - X_j) \; ,$$

where $B_h(x) = h^{-1} B(h^{-1} x)$ and $B = 2\,A - A * A$. Here, the summation over $i \neq j$ is over all i, j, with $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, n$, but $i \neq j$.

We are thus lead to the least-squares cross-validation method.

(2.11) In the least-squares cross-validation method, the smoothing parameter is chosen so as to

$$\text{minimize} \quad CV(h) \quad \text{over } h > 0 \; .$$

The h so chosen is denoted by $H_{n,CV}$ and the corresponding kernel estimator by $f_{n,CV}$.

We shall not attempt to analyze this method and merely state its asymptotic optimality for L^2 errors.

(2.12) THEOREM. [STONE (1984)] *Let A be a symmetric, Hölder continuous kernel with compact support and integral equal to one. If f_o is a bounded density, then*

$$\limsup_{n \to \infty} \; \frac{\| f_{n,CV} - f_o \|_2}{\inf_h \; \| f^{nh} - f_o \|_2} =_{\text{as}} 1 \; .$$

An amazing feature of this theorem is the almost complete lack of conditions on the density f_o. It even holds in the multivariate case, if in addition all the one-dimensional marginals of f_o are bounded. Note also that higher order kernels are allowed. On the negative side, least-squares cross validation seems to have practical drawbacks. The smoothing parameter $H_{n,CV}$ selected seems to show too much variability and too large a negative correlation with the optimal smoothing parameter $h_{n,OPT}$. The various fixes based on slight modifications of $CV(h)$ seem to have their own drawbacks, see SCOTT and TERRELL (1987), HALL, MARRON and PARK (1992), and HALL and JOHNSTONE (1992). Another issue is the scaling invariance of $H_{n,CV}$, prompted by the lack of scaling invariance of the L^2

distance. However, the above treatment can be repeated for

$$\frac{\| f^{nh} - f_o \|_2}{\| f_o \|_2} \ ,$$

and this expression is scaling invariant. Because the denominator is independent of h, minimizing the numerator over $h > 0$ is the same as minimizing the quotient. So the estimator is indeed scaling invariant.

(2.13) EXERCISE. Verify that $H_{n,CV}$ is scaling invariant in the sense of (1.16).

We now consider plug-in methods. In the asymptotic L^2 plug-in methods, the squared error $\| f^{nh} - f_o \|_2^2$ is estimated by first replacing it by the asymptotic expansion of its expected value

$$\mathbb{E}[\, \| f^{nh} - f_o \|_2^2 \,] = \tfrac{1}{4}\, h^4 \sigma^4(A)\, \| (f_o)'' \|_2^2 + (nh)^{-1}\, \| A \|_2^2 + o(h^4 + (nh)^{-1}) \ , \tag{2.14}$$

for $nh \to \infty$, $h \to 0$. Naturally, the assumptions required are that

$$f_o \in W^{2,2}(\mathbb{R}) \ , \tag{2.15}$$

i.e., f_o and its second derivative both belong to $L^2(\mathbb{R})$, and that

(2.16) the kernel A is a symmetric, square integrable pdf, with

$$\sigma(A) = \Big\{ \int_{\mathbb{R}} x^2\, A(x)\, dx \Big\}^{1/2} < \infty \ .$$

Thus, a theoretically interesting choice for the smoothing parameter h is the one that minimizes

$$\tfrac{1}{4}\, h^4 \sigma^4(A)\, \| (f_o)'' \|_2^2 + (nh)^{-1}\, \| A \|_2^2$$

over $h > 0$, and this is given by

$$h_{\text{asymp}} = n^{-1/5}\ r(A)\, \varrho(f_o) \ , \tag{2.17}$$

where

$$r(A) = \left\{ \frac{\| A \|_2}{\sigma^2(A)} \right\}^{2/5} \ , \quad \varrho(f_o) = \left\{ \frac{1}{\| (f_o)'' \|_2} \right\}^{2/5} \ . \tag{2.18}$$

Of course, this hardly solves the problem: The factor $\varrho(f_o)$ or $\| (f_o)'' \|_2^2$ must be estimated from the data.

One of the first such estimators was proposed by DEHEUVELS (1977). His idea was to use a *parametric* estimator for f_o to estimate $\| (f_o)'' \|_2$ by pretending that f_o may be approximated well by some element from a specific parametric family. This is somewhat at odds with the nonparametric approach to density estimation, but never mind. Thus, consider a scaling family of densities $\gamma(\,\cdot\,;\theta) = \theta^{-1}\, g(\theta^{-1}\, x\,)$ for some known density g, and

let θ_n be an estimator of the "true" scale parameter θ_o. The estimator of h_{asymp} is then

$$H_{n,Deh} \stackrel{\text{def}}{=} n^{-1/5}\, r(A)\, \varrho(\gamma(\,\cdot\,;\,\theta_n))\;, \tag{2.19}$$

and it is an exercise to show that

$$H_{n,Deh} = \theta_n\, r(A)\, \varrho(g)\; n^{-1/5}\;. \tag{2.20}$$

Moreover, if θ_n is scaling invariant, then so is $H_{n,Deh}$. Precisely, as usual, let $\mathbb{X}_n = (X_1, X_2, \cdots, X_n)$. Then, for all $t > 0$,

$$\begin{aligned} &\text{If} \quad \theta_n(\,\mathbb{X}_n\,) = t^{-1}\,\theta_n(\,t\,\mathbb{X}_n\,), \quad \text{then} \\ &H_{n,Deh}(\,\mathbb{X}_n\,) = t^{-1}\,H_{n,Deh}(\,t\,\mathbb{X}_n\,)\;. \end{aligned} \tag{2.21}$$

(2.22) Exercise. Verify (2.20) and (2.21).

If θ_n is selected properly, then typically, something more can be said. That is, whether $\gamma(\,\cdot\,;\,\theta)$ is the correct parametric family or not, usually there exists a $\theta_o = \theta(f_o)$ such that

$$\theta_n - \theta_o =_{\text{as}} \mathcal{O}\big(\,(n^{-1}\log\log n)^{1/2}\,\big)\;, \tag{2.23}$$

and hence,

$$H_{n,Deh} =_{\text{as}} \theta_n\, r(A)\, \varrho(g)\, n^{-1/5} + \mathcal{O}\big(\,n^{-7/10}(\log\log n)^{1/2}\,\big)\;. \tag{2.24}$$

Thus, as long as (2.23) holds, $H_{n,Deh}$ passes the minimum requirements imposed on a smoothing parameter. Deheuvels (1977) proposed this with A the Epanechnikov kernel, g the standard normal, and $\theta_n = s_n$, the sample standard deviation. With these choices, (2.20)–(2.23) give the asymptotic plug-in-normal $H_{n,Deh}$ as

$$H_{n,Deh} = 0.7443\; s_n\; n^{-1/5}\;. \tag{2.25}$$

(2.26) Exercise. Verify (2.25).

It is of course not surprising that we were able to pinpoint what is involved in establishing the a.s. asymptotic behavior of $H_{n,Deh}$, because everything is based on asymptotic considerations. But what about its small sample behavior? Then, everything depends on the appropriateness of the parametric family $\gamma(\,\cdot\,;\theta)$ and the scale estimator θ_n. Thus, in the present context, it seems more natural to use a nonparametric estimator for $\|\,(f_o)''\,\|_2^2$. Such a procedure was first proposed by Woodroofe (1970), see also Deheuvels and Hominal (1980), and has resulted is a sizable literature. The development here is based on a long series of papers culminating in Sheather and Jones (1991). An obvious estimator for $\|\,(f_o)''\,\|_2^2$ is $\|\,(\varphi^{n\lambda})''\,\|_2^2$, where $\varphi^{n\lambda} = B_\lambda * dF_n$ is a kernel estimator for f_o based on a smooth symmetric kernel B, and we must select the new

smoothing parameter λ! The precise assumptions on B are that it satisfies (2.16) and that $B'' \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Unfortunately, the above estimator is biased, as we now show. With

$$\varphi^{n\lambda}(x) = \tfrac{1}{n} \sum_{i=1}^{n} B_\lambda(x - X_i) \ , \quad x \in \mathbb{R} \ , \tag{2.27}$$

one has that

$$\| (\varphi^{n\lambda})'' \|_2^2 = \frac{1}{n^2 \lambda^4} \sum_{i,j=1}^{n} C_\lambda(X_i - X_j) \ , \tag{2.28}$$

in which $C = (B * B)^{(4)}$ (fourth derivative).

(2.29) Exercise. Derive (2.28).

One verifies that

$$\mathbb{E}[\, \| (\varphi^{n\lambda})'' \|_2^2 \,] = n^{-1} \lambda^{-5} \| B'' \|_2^2 \ + \ \frac{n-1}{n} \| B_\lambda * (f_o)'' \|_2^2 \ , \tag{2.30}$$

provided $f_o \in W^{4,2}(\mathbb{R})$. Now, one could argue that the first term on the right does not belong and view $\| (\varphi^{n\lambda})'' \|_2^2 - n^{-1} \lambda^{-5} \| B'' \|_2^2$ as an estimator for $\| (f_o)'' \|_2^2$, but this is not what is done. Instead, one observes that

$$\| B_\lambda * (f_o)'' \|_2^2 = \| (f_o)'' \|_2^2 - \tfrac{1}{2} \lambda^2 \sigma^2(B) \| (f_o)^{(3)} \|_2^2 + \mathcal{O}(\lambda^4) \ , \tag{2.31}$$

and so

$$\mathbb{E}[\, \frac{n}{n-1} \| (\varphi^{n\lambda})'' \|_2^2 - \| (f_o)'' \|_2^2 \,] = (n-1)^{-1} \lambda^{-5} \| B'' \|_2^2 \ - \ \tfrac{1}{2} \lambda^2 \sigma^2(B) \| (f_o)^{(3)} \|_2^2 \ + \ \mathcal{O}(\lambda^4) \ . \tag{2.32}$$

So, λ should/could be chosen so as to set the bias equal to 0. Omitting the $\mathcal{O}(\lambda^4)$ term in (2.32) and ignoring the difference between $n-1$ and n, the bias vanishes for $\lambda = \lambda_{\text{asymp}}$, given by

$$\lambda_{\text{asymp}} = n^{-1/7} \ \left\{ \frac{2 \| B'' \|_2^2}{\sigma^2(B) \| (f_o)^{(3)} \|_2^2} \right\}^{1/7} \ . \tag{2.33}$$

Since the natural question now is how to estimate $\| (f_o)^{(3)} \|_2^2$, it seems that little progress has been made. However, at this point, Sheather and Jones (1991) observe that as functions of n, the smoothing parameters h_{asymp} of (2.17) and λ_{asymp} are asymptotically related by

$$\lambda_{\text{asymp}} \asymp \left(h_{\text{asymp}} \right)^{5/7} \ , \quad n \to \infty \ .$$

In particular, taking $A = B$ for simplicity,

$$\lambda_{\text{asymp}} = c(B) \ \left\{ \frac{\| (f_o)'' \|_2}{\| (f_o)^{(3)} \|_2} \right\}^{2/7} \ \left(h_{\text{asymp}} \right)^{5/7} \ , \tag{2.34}$$

with

$$c(B) = \left\{ \frac{\sqrt{2}\ \sigma(B)\, \| B'' \|_2}{\| B \|_2} \right\}^{2/7} . \tag{2.35}$$

Now, to make (2.34) practicable, the expressions $\| (f_o)'' \|_2^2$ and $\| (f_o)^{(3)} \|_2^2$ are estimated by the double sums

$$S_4(\mu) = \frac{1}{n(n-1)\mu^5} \sum_{i,j=1}^{n} \phi^{(4)}\big(\mu^{-1}(X_i - X_j) \big) \tag{2.36}$$

and

$$S_6(\nu) = \frac{1}{n(n-1)\nu^7} \sum_{i,j=1}^{n} \phi^{(6)}\big(\nu^{-1}(X_i - X_j) \big) , \tag{2.37}$$

with ϕ the normal kernel. The smoothing parameters μ and ν are chosen so as to obtain optimal estimators if f_o is a normal density. The hope is that the normal parametric model is sufficiently accurate for the purpose. The net result is that

$$\mu = 0.920\ q_n\ n^{-1/7} \quad , \quad \nu = 0.912\ q_n\ n^{-1/9} , \tag{2.38}$$

where q_n is the sample interquartile range

$$\begin{aligned} q_n &= \big(X_{[\,3n/4],n} - X_{[\,n/4],n} \big) \,/\, \big(\Phi^{\text{inv}}(\tfrac{3}{4}) - \Phi^{\text{inv}}(\tfrac{1}{4}) \big) \\ &\doteq \big(X_{[\,3n/4],n} - X_{[\,n/4],n} \big) \,/\, 1.35 . \end{aligned} \tag{2.39}$$

(Here, Φ is the distribution of the standard Gaussian density.)

So the asymptotic relationship (2.34) is replaced by the concrete one

$$\lambda(h) = c(B) \left\{ \frac{S_4(\mu)}{S_6(\nu)} \right\}^{1/7} h^{5/7} . \tag{2.40}$$

Now, Sheather and Jones (1991) obtain their estimator $H_{n,SJ}$ of the smoothing parameter as the solution to

$$h = n^{-1/5}\ r(B)\, \varrho\big(\varphi^{n,\lambda(h)} \big) , \tag{2.41}$$

with $\varphi^{n\lambda}$ as in (2.27). We refer to the solution $H_{n,SJ}$ of (2.40)–(2.41) as *the* Sheather-Jones estimator, although they have a number of estimators to their credit.

It is not at all obvious, but in practice the equations (2.40)–(2.41) have a unique solution, which is easily found using safe versions of the Secant method. It is possible to derive (quite amazing) bounds on $H_{n,SJ} - h_{\text{asymp}}$, but we shall not do so. Wand and Jones (1995) contains all of the details.

(2.42) EXERCISE. Verify (2.31). [Hint: Use

$$\int_{\mathbb{R}} f^{(2)}(x)\, f^{(4)}(x)\, dx = -\,\| (f_o)^{(3)} \|_2^2 \,.]$$

EXERCISES: (2.9), (2.10), (2.13), (2.22), (2.26), (2.29), (2.42).

3. The double kernel method

We now return to the L^1 point of view. It seems to make eminent sense to choose the smoothing parameter as the solution to

$$\text{minimize} \quad \| f^{nh} - f_o \|_1 \quad \text{over } h > 0 \,, \tag{3.1}$$

but, of course, the loss $\| f^{nh} - f_o \|_1$ must be estimated first. The niceties associated with the L^2 norm, especially (2.2), do not apply to the L^1 norm, so the goal of getting an *unbiased* estimator of $\| f^{nh} - f_o \|_1$ seems unattainable. There appears to be no other choice but to use another estimator of f_o. Thus, let B be some other kernel, and consider

$$\varphi^{nh}(x) = B_h * dF_n(x) = \tfrac{1}{n} \sum_{i=1}^{n} B_h(x - X_i) \,, \quad x \in \mathbb{R} \,, \tag{3.2}$$

and suppose that φ^{nh} is much more accurate than f^{nh}. In view of § 4.7 on optimal kernels, if f_o is very smooth, one could take

$$B = 2\,A - A * A \,. \tag{3.3}$$

So, assuming that

$$\| \varphi^{nh} - f_o \|_1 \ll \| f^{nh} - f_o \|_1 \,, \tag{3.4}$$

then

$$\| f^{nh} - \varphi^{nh} \|_1 \approx \| f^{nh} - f_o \|_1 \,, \tag{3.5}$$

and one would expect to do well by minimizing $\| f^{nh} - \varphi^{nh} \|_1$. This is the "double kernel method" for choosing the smoothing parameter, due to DEVROYE (1989).

So, for a suitable pair of kernels A and B,

(3.6) in the double kernel method, the smoothing parameter is chosen so as to

$$\text{minimize} \quad DBL(h) \stackrel{\text{def}}{=} \| (A_h - B_h) * dF_n \|_1 \quad \text{over } h > 0 \,.$$

The resulting h is denoted as $H_{n,DBL}$ and the associated kernel estimator $f^{n,H_{n,DBL}}$ as $f_{n,DBL}$.

For the double kernel method to work, and to be able to say something about it, the kernels A and B must satisfy some minimal conditions. The various assumptions needed at one point or another are as follows.

(3.7) A and B are bounded, symmetric about 0, have compact support, and

$$\int_{\mathbb{R}} A(x)\,dx = \int_{\mathbb{R}} B(x)\,dx = 1 \ .$$

Moreover, it is assumed that A and B are distinct, in the sense that there exists an $\omega_o > 0$ such that

$$\widehat{A}(\omega) \neq \widehat{B}(\omega) \ , \quad 0 < |\omega| \leqslant \omega_o \ , \tag{3.8}$$

where $\widehat{A}$ and $\widehat{B}$ are the Fourier transforms of A and B. Finally, there should exist a constant c such that for all $h > 1$,

$$\| A - A_h \|_1 \leqslant c\,|\,1-h\,| \ , \quad \| B - B_h \|_1 \leqslant c\,|\,1-h\,| \ . \tag{3.9}$$

Some examples of pairs of kernels satisfying these conditions are A the uniform kernel or the Epanechnikov kernel, and $B = 2\,A - A * A$ (without proof).

What can we say about the double kernel method? The first concern is whether $H_{n,DBL}$ exists and is scaling invariant. Existence would be no problem, if we were to allow $H_{n,DBL} = 0$ or $= +\infty$. But, in fact, things are much nicer than that.

(3.10) Exercise. Let A and B satisfy (3.7). Show that for every realization of $X_1, X_2, \cdots, X_n$,
(a) $\| (A_h - B_h) * dF_n \|_1 \leqslant \| A - B \|_1$,
(b) $\| (A_h - B_h) * dF_n \|_1 \longrightarrow \| A - B \|_1$ for $h \to 0$, as well as for $h \to \infty$,
(c) $\| (A_h - B_h) * dF_n \|_1$ is continuous in h, and
(d) conclude that $H_{n,DBL}$ exists.

(3.11) Exercise. Show that $H_{n,DBL}$ is scale invariant, in the sense of (1.16).

The second worry is whether the double kernel method gives consistent estimators. Amazingly, it does so without any assumptions on the density.

(3.12) Theorem. [Devroye (1989)] *Under the assumptions* (3.7), (3.8), *and* (3.9) *on the kernels, for every density* f_o,

$$\lim_{n\to\infty} \| f_{n,DBL} - f_o \|_1 =_{\text{as}} 0 \ .$$

The assumptions on the kernel can be relaxed somewhat. Moreover, the theorem holds for suitable higher order kernels. We shall not go into the details.

As the lack of smoothness and tail assumptions on f_o indicates, we are in no position to prove this. The next theorem states that under the usual smoothness and tail conditions on f_o, one gets the optimal rate of convergence, but even this is out of our reach.

(3.13) THEOREM. [DEVROYE (1989)] *Under the assumptions* (3.7), (3.8), *and* (3.9) *on the kernels, if* $f_o \in W^{2,1}(\mathbb{R})$ *and has a moment of order* > 1, *and*

$$\varepsilon \stackrel{\text{def}}{=} \{4 \,\|\, B \,\|_2 / \|\, A \,\|_2 \}^{1/2} < 1 \ ,$$

then

$$\limsup_{n\to\infty} \ \frac{\mathbb{E}[\,\|\, f_{n,DBL} - f_o \,\|_1\,]}{\inf_h \ \mathbb{E}[\,\|\, f^{nh} - f_o \,\|_1\,]} \leqslant_{\text{as}} \ \frac{1+\varepsilon}{1-\varepsilon} \ .$$

Note that one can make ε arbitrarily close to 0 by choosing the kernels A and B appropriately.

As lamented above, we cannot prove either theorem with the methods that were explored in Chapter 4. We are not even able to prove that

$$H_{n,DBL} \asymp_{\text{as}} n^{-1/5} \ . \tag{3.14}$$

However, one side is easy, sort of.

(3.15) EXERCISE. Let $\varepsilon > 0$ be arbitrary. Show that if $f_o \in W^{2,1}(\mathbb{R})$ and $\sqrt{f_o} \in L^1(\mathbb{R})$, then $H_{n,DBL} =_{\text{as}} \mathcal{O}(\, n^{-1/5+\varepsilon}\,)$.

So, in view of Theorems (3.12) and (3.13), the double kernel method asymptotically is all peaches. Unfortunately, for small sample sizes, things are not as clear. By way of example, note that for the choice (3.3),

$$f^{nh} - \varphi^{nh} = A_h * A_h * dF_n - A_h * dF_n \ , \tag{3.16}$$

which is just the difference between "any" two kernel estimators, presumably of comparable accuracy: One would not expect either one to be much more accurate than the other. So the original motivation (3.4)–(3.5) may not be quite relevant. See also Exercise (3.48)(a) below. Actually, under the usual nonparametric conditions, asymptotically the optimal h for $A_h * dF_n$ satisfies $h \asymp n^{-1/5}$, and the optimal h for $B_h * dF_n$ satisfies $h \asymp n^{-1/9}$, provided $f_o \in W^{4,1}(\mathbb{R})$. Thus, it seems that drastically different scales are required for A and B. In fact, it suggests that B_h should be replaced by

$$B_h = 2\, A_\lambda - A_\lambda * A_\lambda \ , \tag{3.17}$$

with $\lambda \asymp h^{5/9}$, if only one knew how to do this in a data-driven way. Although changing the scale of the second kernel just a little is enough to get the $(1+\varepsilon)/(1-\varepsilon)$ bound of Theorem (3.13), the above seems to explain

the careful tweaking of the scales of A and B necessary to make the double kernel method work well in the small sample case, see Chapter 8.

Following Devroye (1989), in the simulations of Chapter 8, the following double kernel methods are considered. With A the Epanechnikov kernel, the second kernel is taken to be

(3.18) $B = L_s$, with $s \in \{ 2,\ 2.4,\ 2.88,\ 3 \}$ (the tweaking parameter) ,

where L is the Berlinet-Devroye kernel given by

$$L(x) = \begin{cases} \frac{1}{4}\,(\,7 - 31\,x^2\,) & , \quad |\,x\,| \leqslant \frac{1}{2}\ , \\ \frac{1}{4}\,(\,x^2 - 1\,) & , \quad \frac{1}{2} < |\,x\,| \leqslant 1\ , \end{cases} \tag{3.19}$$

and $= 0$ otherwise. There are good reasons for chosing this kernel L, which we shall not discuss. However, it is an exercise to show that L is a fourth-order kernel.

(3.20) Exercise. Investigate the theoretical and practical virtues, if any, of the following method for choosing h:

$$\text{minimize} \quad \| f^{nh} - \varphi^{n,h^{5/9}} \|_1 \quad \text{over } h > 0\ .$$

Here, φ^{nh} is given by (3.2), and B by (3.3). [Hint: Use the methods below.]

It is perhaps worthwhile to pinpoint the difficulties in establishing (3.14). Following the martingale results of § 4.4, we have

$$\| (\,A_h - B_h\,) * dF_n \|_1 =_{\text{as}} \mathbb{E}[\,\| (\,A_h - B_h\,) * dF_n \|_1\,] + \mathcal{O}\big((n^{-1}\,\log n)^{1/2} \big)\ . \tag{3.21}$$

The first difficulty is that the above holds for deterministic h only, but we pretend it holds for random h also. Now, by the triangle inequality,

$$\begin{aligned} \| (\,A_h - B_h\,) * dF_n \|_1 &\geqslant \\ \| (\,A_h - B_h\,) * f_o \|_1 &- \| (\,A_h - B_h\,) * (\,dF_n - dF_o\,) \|_1 \\ &\geqslant_{\text{as}} c_1\,h^2 - c_2\,(nh)^{-1/2}\ . \end{aligned} \tag{3.22}$$

This statement too holds for deterministic h only. Pretending that it holds for $h = H_{n,DBL}$ would give

$$c_1\,(H_{n,DBL})^2 \leqslant_{\text{as}} c_3\,n^{-2/5} + c_2\,(nH_{n,DBL})^{-1/2}\ ,$$

which implies that there exists a constant c_4 such that

$$H_{n,DBL} \leqslant_{\text{as}} c_4\,n^{-1/5}\ . \tag{3.23}$$

For the lower bound on $H_{n,DBL}$, we have similar to (3.22)

$$\begin{aligned} \| (\,A_h - B_h\,) * dF_n \|_1 &\geqslant \\ \| (\,A_h - B_h\,) * (\,dF_n - dF_o\,) \|_1 &- \| (\,A_h - B_h\,) * f_o \|_1\ , \end{aligned} \tag{3.24}$$

and we would be in great shape if

$$\| (A_h - B_h) * (dF_n - dF_o) \|_1 \geqslant_{\text{as}} c_5 \, (1 + nh)^{-1/2} \tag{3.25}$$

for all $h \leqslant h_n \asymp n^{-1/5}$. For deterministically, smoothly varying h, this definitely holds, but it needs to hold for random h. Proceeding fearlessly, (3.25) gives the inequality

$$c_3 \, n^{-2/5} + c_1 \, (H_{n,DBL})^2 \leqslant_{\text{as}} c_5 \, (n H_{n,DBL})^{-1/2} \ ,$$

and this indeed implies that there exists a constant $c_6 > 0$ such that

$$H_{n,DBL} \geqslant_{\text{as}} c_6 \, n^{-1/5} \ . \tag{3.26}$$

Thus, (3.14) would hold.

In the above, it is possible to finesse one's way around the randomness of $H_{n,DBL}$, but showing (3.25) is amazingly hard. In fact, it is not obvious that it is true. However, following the theme of this text (when in doubt, penalize), if we add a roughness penalization to the $DBL(h)$ function, then we ought to be able to show (the analogue) of (3.14). So,

(3.27) in the perverted double kernel method, the smoothing parameter is taken to be the smallest solution to

$$\text{minimize} \quad PER(h) \stackrel{\text{def}}{=} \| (A_h - B_h) * dF_n \|_1 + (nh)^{-1/2} \, \text{var}_n(B; h) \ ,$$

where $\text{var}_n(B; h)$ is defined as

$$\text{var}_n(B; h) = \| \sqrt{(B^2)_h * dF_n - h \, (B_h * dF_n)^2} \, \|_1 \ .$$

The h so selected is denoted by $H_{n,PER}$ and the corresponding kernel estimator by $f_{n,PER}$. Recall that $B = 2A - A * A$.

It should be noted that $(B_h)^2 = h^{-1} \, (B^2)_h$, so that $\text{var}_n(B; h)$ is more like a (normalized) standard deviation than a variance. Notationally, we are safe by calling it a "variation" term.

A simple, but honest, motivation for this method is that it allows a proof of the analogue of (3.14) along the lines (3.21)–(3.26). In particular, there is no need to show (3.25), since there is already a term $(nh)^{-1/2}$ present. A more ambitious motivation is as follows. Note that

$$\begin{aligned} \| f^{nh} - f_o \|_1 \leqslant \| \, (A_h - B_h \,) * dF_n \, \|_1 + \\ \| \, B_h * f_o - f_o \, \|_1 + \| \, B_h * (\, dF_n - dF_o \,) \, \|_1 \ . \end{aligned} \tag{3.28}$$

Since B is a fourth-order kernel, if f_o is smooth enough, then

$$\| \, B_h * f_o - f_o \, \|_1 = \mathcal{O}(h^4) \, ,$$

which is much smaller than the bias in $A_h * dF_n$, and hence, we may ignore this term. Continuing, the third term on the right of (3.28) behaves pretty much like its expectation, see § 4.4, and

$$\mathbb{E}[\, \| \, B_h * (dF_n - dF_o) \, \|_1 \,] \leqslant (nh)^{-1/2} \, \text{var}_o(B; h) \ , \tag{3.29}$$

with

$$\mathrm{var}_o(B;h) = \| \sqrt{(B^2)_h * dF_o - h\,(B_h * dF_o)^2}\,\|_1 \quad . \tag{3.30}$$

Of course, $\mathrm{var}_n(B;h)$ is the obvious estimator of $\mathrm{var}_o(B;h)$. Now, if the above inequalities are just about equalities, then *PER*(h) should be a good estimator of $\| f^{nh} - f_o \|_1$, and so minimizing *PER*(h) should yield a good smoothing parameter.

So much for heuristics. We now prove that the modified kernel method works in the sense of (1.6) and (1.8).

(3.31) THEOREM. *Suppose $f_o \in W^{2,1}(\mathbb{R})$ and has a finite moment of order $\lambda > 1$. If A is the standard Gaussian density and $B = 2A - A * A$, then $H_{n,PER} \asymp_{\mathrm{as}} n^{-1/5}$ and $\| f_{n,PER} - f_o \|_1 =_{\mathrm{as}} \mathcal{O}\big(n^{-2/5}\big)$.*

The crucial property to be used is that the normal density satisfies

$$A_h * A_\lambda = A_\sigma \ , \quad \text{with} \quad \sigma^2 = h^2 + \lambda^2 \ , \tag{3.32}$$

which implies that expressions like

$$\| A_h * (dF_n - dF_o) \|_1 \quad \text{and} \quad \| \sqrt{(A^2)_h * dF_n}\,\|_1 \tag{3.33}$$

are monotone functions of h, see Exercises (3.48) and (3.51). For arbitrary kernels, this does not work, and we are up the creek without a paddle.

The first step in the study of $H_{n,PER}$ is to determine an upper bound for the minimum of *PER*(h) over $h > 0$.

(3.34) LEMMA. *Under the assumptions of Theorem* (3.31)*, there exists a constant K depending on f_o only such that*

$$\limsup_{n\to\infty}\ n^{2/5} \inf_h PER(h) \leqslant_{\mathrm{as}} K \ .$$

Lower and upper bounds on $H_{n,PER}$ are then provided by the following two lemmas, which, combined, say that the infimum of *PER*(h) occurs a.s. on an interval $\delta\, n^{-1/5} \leqslant h \leqslant \gamma\, n^{-1/5}$, for suitable (deterministic) δ and γ.

(3.35) LEMMA. *Under the assumptions of Theorem* (3.31)*, for a large enough constant γ, depending on f_o only,*

$$\liminf_{n\to\infty}\ n^{2/5} \inf \big\{\, PER(h) \,:\, h \geqslant \gamma\, n^{-1/5} \,\big\} >_{\mathrm{as}} K \ .$$

(3.36) LEMMA. *Under the assumptions of Theorem* (3.31)*, for a small enough positive constant δ, depending on f_o only,*

$$\liminf_{n\to\infty}\ n^{2/5} \inf \big\{\, PER(h) \,:\, h \leqslant \delta\, n^{-1/5} \,\big\} >_{\mathrm{as}} K \ .$$

As warning to the reader, in the proofs to follow, a careful distinction must be made between $(A_h)^2$ and $(A^2)_h$, which refer to the kernels

$$\big(\, h^{-1}A(h^{-1}x)\,\big)^2 \quad \text{and} \quad h^{-1}\big(\, A(h^{-1}x)\,\big)^2 \;, \quad \text{respectively} \;.$$

Proof of Lemma (3.34). Of the three, this proof is the simplest, since it suffices to stick to a deterministic choice of h. With $C = A - B$, we have

$$\| C_h * dF_n \|_1 \leqslant \| C_h * dF_o \|_1 + \| C_h * (dF_n - dF_o) \|_1$$

and

$$\begin{aligned} \| C_h{*}(dF_n - dF_o) \|_1 \leqslant \\ &\leqslant \| A_h * A_h * (dF_n - dF_o) \|_1 + \| A_h * (dF_n - dF_o) \|_1 \\ &\leqslant 2 \| A_h * (dF_n - dF_o) \|_1 \;, \end{aligned}$$

where we used Young's inequality in the form

$$\| A_h * A_h * (dF_n - dF_o) \|_1 \leqslant \| A_h \|_1 \| A_h * (dF_n - dF_o) \|_1 \;.$$

It follows that

$$PER(h) \leqslant \| C_h{*}f_o \|_1 + 2 \| A_h{*}(dF_n - dF_o) \|_1 + (nh)^{-1/2} \operatorname{var}_n(B;h) \;. \tag{3.37}$$

An appeal to Exercise (3.49) gives the bound

$$\| C_h * f_o \|_1 \leqslant c\, h^2 \;,$$

for a suitable constant c, depending on f_o.

Next, for suitable constants c_1 and c_2,

$$\operatorname{var}_n(B;h) \leqslant \| \sqrt{(B^2)_h * dF_n} \|_1 \leqslant_{\text{as}} c_1 + c_2\, h^{1/(2\lambda)} \;, \tag{3.38}$$

the last inequality by Exercise (4.2.26), since f_o has a finite moment of order $\lambda > 1$ and the density $B^2 / \| B \|_2^2$ has finite *exponential* moments. Finally, the term $\| A_h * (dF_n - dF_o) \|_1$ has been adequately treated in § 4.4: For $h \asymp n^{-\beta}$ (deterministic), with $0 < \beta < 1$,

$$\begin{aligned} \| A_h * (dF_n - dF_o) \|_1 =_{\text{as}} \\ \mathbb{E}[\, \| A_h * (dF_n - dF_o) \|_1 \,] + \mathcal{O}\big((n^{-1} \log n)^{1/2} \big) \;. \end{aligned} \tag{3.39}$$

Moreover,

$$\begin{aligned} \mathbb{E}[\, \| A_h * (dF_n - dF_o) \|_1 \,] &\leqslant (nh)^{-1/2} \operatorname{var}_o(A;h) \\ &\leqslant c_4\, (nh)^{-1/2} + c_5\, n^{-1/2} \;. \end{aligned} \tag{3.40}$$

Putting everything together gives for $h \asymp n^{-\beta}$ and (other) suitable constants c_i,

$$PER(h) \leqslant_{\text{as}} c_1 h^2 + c_2 (nh)^{-1/2} (\, 1 + c_3\, h^{\mu}\,) + c_4 n^{-1/2} \;, \quad n \to \infty \;.$$

Here, $\mu = \frac{1}{2} + \frac{1}{2\lambda}$. For $\beta = \frac{1}{5}$, this proves the lemma. Q.e.d.

PROOF OF LEMMA (3.35). Let $h_n = \gamma\, n^{-1/5}$ for a suitable constant γ to be determined later. The starting point is the inequality

$$PER(h) \geqslant \| C_h * f_o \|_1 - \| A_h * (dF_n - dF_o) \|_1 \ ,$$

which implies that

$$\inf_{h \geqslant h_n} PER(h) \geqslant \inf_{h \geqslant h_n} \| C_h * f_o \|_1 - \sup_{h \geqslant h_n} \| A_h * (dF_n - dF_o) \|_1 \ . \tag{3.41}$$

By Exercise (3.49), we have the lower bound

$$\inf_{h \geqslant h_n} \| C_h * f_o \|_1 \geqslant c_1 \ \min(\, (h_n)^2 \, , \, 1 \,) \ , \quad h > 0 \ ,$$

for some positive constant c_1.

For the second term on the right of (3.41), we have by Exercise (3.48)(b) that $\| A_h * (dF_n - dF_o) \|_1$ is decreasing in h, which implies the bound

$$\sup_{h \geqslant h_n} \| A_h * (dF_n - dF_o) \|_1 = \| A_{h_n} * (dF_n - dF_o) \|_1 \leqslant c_2 \, (nh_n)^{-1/2} \ ,$$

the last bound by (3.39)–(3.40), for a suitable positive constant c_2. Combining these lower bounds gives

$$\inf_{h \geqslant h_n} PER(h) \geqslant_{\text{as}} c_1 \ \min((h_n)^2 \, , \, 1 \,) - c_2 \, (nh_n)^{-1/2} \ .$$

It follows that

$$\liminf_{n \to \infty} \ n^{2/5} \inf_{h \geqslant h_n} PER(h) \geqslant_{\text{as}} c_1 \, \gamma^2 - c_2 \, \gamma^{-1/2} \ ,$$

and this dominates K for large enough γ. Q.e.d.

PROOF OF LEMMA (3.36). Let $h_n = \delta\, n^{-1/5}$, for a small enough positive constant δ. The starting point is the inequality

$$\begin{aligned} PER(h) &\geqslant (nh)^{-1/2} \mathrm{var}_n(B; h) \\ &\geqslant (nh)^{-1/2} \| \sqrt{(B^2)_h * dF_n} \, \|_1 - n^{-1/2} \, \| B \|_1 \ , \end{aligned}$$

the last inequality by Exercise (3.50). Recall that

$$B = 2\, A - A * A = 2\, A - A_{\sqrt{2}} \ .$$

We work temporarily with $(B_h)^2$ rather than with $(B^2)_h$. The triangle inequality for the Euclidean norm on $\mathbb{R}^n$ gives

$$\sqrt{(B_h)^2 * dF_n(x)} \ \geqslant 2 \sqrt{(A_h)^2 * dF_n(x)} \ - \sqrt{(\, A_{h\sqrt{2}})^2 * dF_n(x)} \ .$$

Upon integration, we get

$$\begin{aligned} \| \sqrt{(B_h)^2 * dF_n} \, \|_1 &\geqslant 2 \| \sqrt{(A_h)^2 * dF_n} \, \|_1 - \| \sqrt{(A_{h\sqrt{2}})^2 * dF_n} \, \|_1 \\ &\geqslant \| \sqrt{(A_h)^2 * dF_n} \, \|_1 \ , \end{aligned}$$

the last inequality by Exercise (3.51)(d). Thus, by translating back in terms of $(B^2)_h$ and $(A^2)_h$,

$$\| \sqrt{(B^2)_h * dF_n(x)} \|_1 \geqslant \| \sqrt{(A^2)_h * dF_n(x)} \|_1 .$$

Now, Exercise (3.51)(d) below implies that $\| \sqrt{(A_h)^2 * dF_n} \|_1$ is a decreasing function of h. Thus,

$$\begin{aligned} \inf_{h \leqslant h_n} PER(h) &\geqslant \inf_{h \leqslant h_n} (nh)^{-1/2} \| \sqrt{(A^2)_h * dF_n} \|_1 - n^{-1/2} \| B \|_2 \\ &\geqslant (nh_n)^{-1/2} \| \sqrt{(A^2)_{h_n} * dF_n} \|_1 - n^{-1/2} \| B \|_2 . \end{aligned}$$

In the next lemma, we show that there exists a positive constant c_3 such that for deterministic h, with $h \log n \to 0$,

$$\| \sqrt{(A^2)_h * dF_n} \|_1 \geqslant_{\text{as}} c_3 (nh)^{1/2} (1 + nh)^{-1/2} ,$$

and so

$$\liminf_{n \to \infty} n^{2/5} \inf_{h \leqslant h_n} PER(h) \geqslant_{\text{as}} c_3 \delta^{-1/2} .$$

For δ small enough, this dominates K. Q.e.d.

To wrap up the above proof, we need to provide a.s. lower bounds on $\| \sqrt{(A^2)_h * dF_n} \|_1$. Since A^2 is nonnegative and integrable, it suffices to do this for an arbitrary nonnegative kernel K. Lucky for us, the material of § 4.4 on martingales and exponential inequalities applies here as well and seems to give sharp bounds. We formulate it as a lemma.

(3.42) Lemma. *Let K be a nonnegative kernel, with a finite exponential moment. There exists a positive constant c such that for deterministic h satisfying $h \log n \to 0$,*

$$(nh)^{-1/2} \| \sqrt{K_h * dF_n} \|_1 \geqslant_{\text{as}} c (1 + nh)^{-1/2} , \quad n \to \infty .$$

Proof. This is an application of the Devroye (1991) approach to obtaining exponential inequalities for kernel estimators, see § 4.4. Using the abbreviation $\mathbb{X}_n = (X_1, X_2, \cdots, X_n)$, let

$$\psi_n(\mathbb{X}_n) = \| \sqrt{K_h * dF_n} \|_1 . \tag{3.43}$$

We leave it as an exercise to verify that, with $(x)_{n-1} = (x_1, x_2, \cdots, x_{n-1})$ and $\big((x)_{n-1}, u \big) = (x_1, x_2, \cdots, x_{n-1}, u)$,

$$\sup_{u,w} | \psi_n\big((x)_{n-1}, u \big) - \psi_n\big((x)_{n-1}, w \big) | \leqslant 2 \big(h/n \big)^{1/2} . \tag{3.44}$$

Thus, Theorem (4.4.20) implies the exponential inequality

$$\mathbb{P}\big[\, \big| \psi_n(\mathbb{X}_n) - \mathbb{E}[\psi_n(\mathbb{X}_n)] \big| > 2t\sqrt{h} \, \big] \leqslant 2 \exp(-2t^2) , \tag{3.45}$$

which in turn implies the a.s. bound

$$\psi_n(\,\mathbb{X}_n\,) =_{\text{as}} \mathbb{E}[\,\psi_n(\,\mathbb{X}_n\,)\,] + \mathcal{O}\big(\,(\,h\,\log n\,)^{1/2}\,\big)\ . \tag{3.46}$$

To obtain a lower bound on the expected value, note that Hölder's inequality implies for all x,

$$\mathbb{E}[\,K_h * dF_n(x)\,] \leqslant \big(\,\mathbb{E}\big[\,\sqrt{K_h * dF_n(x)}\,\big]\,\big)^{2/3}\,\big(\,\mathbb{E}[\,\big(\,K_h * dF_n(x)\,\big)^2\,]\,\big)^{1/3}\ ,$$

and so

$$\mathbb{E}[\,\|\,\sqrt{K_h * dF_n}\,\|_1\,] \geqslant \int_{\mathbb{R}} \frac{(K_h * f_o(x)\,)^{3/2}}{\big(\,Lf_o(x)\,\big)^{1/2}}\,dx\ , \tag{3.47a}$$

where $Lf_o(x) = \mathbb{E}[\,(K_h * dF_n(x)\,\big)^2\,]$, or

$$Lf_o(x) = (nh)^{-1}\,[\,(K^2)_h * f_o\,](x) + (K_h * f_o(x))^2\ . \tag{3.47b}$$

Now, once more using Hölder's inequality gives

$$\int_{\mathbb{R}} K_h * f_o(x)\,dx \leqslant \left\{\,\int_{\mathbb{R}} \frac{\big(\,K_h * f_o(x)\,\big)^{3/2}}{\big(\,Lf_o(x)\,\big)^{1/2}}\,dx\,\right\}^{2/3}\ \left\{\,\int_{\mathbb{R}} Lf_o(x)\,dx\,\right\}^{1/3}\ .$$

The integral on the left equals $\|\,K\,\|_1$, and obviously,

$$\int_{\mathbb{R}} Lf_o(x)\,dx = (nh)^{-1}\,\|\,K\,\|_2^2 + \|\,K\,\|_1\ .$$

It follows that

$$\int_{\mathbb{R}} \frac{\big(\,K_h * f_o(x)\,\big)^{3/2}}{\big(\,Lf_o(x)\,\big)^{1/2}}\,dx\ \geqslant \|\,K\,\|_1^{3/2}\,\big\{\,(nh)^{-1}\,\|\,K\,\|_2^2 + \|\,K\|_1\,\big\}^{-1/2}\ ,$$

and (3.47a) clinches the argument. Q.e.d.

In the above proofs, we referred to a number of results, which we formulate as exercises.

(3.48) Exercise. For the kernels of Theorem (3.31), show that

(a) $\|\,A_h * (dF_n - dF_o)\,\|_1 \leqslant \|\,B_h * (dF_n - dF_o)\,\|_1 \leqslant 3\,\|\,A_h * (dF_n - dF_o)\,\|_1$,

(b) $\|\,A_h * (dF_n - dF_o)\,\|_1$ is a decreasing function of h.

Likewise for the double exponential kernel. [Hint: See (3.32) for Part (b).]

(3.49) Exercise. Let the kernels A and B be as in Theorem (3.31), let $C = A - B$, and suppose the density f_o belongs to $W^{2,1}(\mathbb{R})$.

(a) Show that $\|\,C_h * f_o\,\|_1 = \frac{1}{2}\,\sigma^2(A)\,h^2\,\|\,f''\,\|_1 + o(h^2)\ ,\quad h \to 0$.

(b) Show that $\lim_{h\to\infty} \|\,C_h * f_o\,\|_1 = \|\,C\,\|_1$.

(c) Show that $\|\,C_h * f_o\,\|_1$ is a continuous function of h, and that it is positive for all $h > 0$.

(d) Show that there exist constants $0 < c_1 \leqslant c_2 < \infty$ such that

$$c_1 \min(\, h^2\,,\, 1\,) \leqslant \| \,(A_h - B_h) * f_o\,\|_1 \leqslant c_2 \,\min(\, h^2\,, 1\,) \quad \text{for all } h > 0\,.$$

(e) Do the same for the double exponential kernel.

(3.50) Exercise. (a) Show that $\sqrt{x} - \sqrt{y} \leqslant \sqrt{x-y}$ for all $x \geqslant y \geqslant 0$.
(b) For any kernel K, show that

$$\big|\, \mathrm{var}_n(K;h) - \| \,\sqrt{(K^2)_h * dF_n}\,\|_1 \,\big| \leqslant \| \,K\,\|_1 \,\sqrt{h}\ .$$

(3.51) Exercise. Let A be the Gaussian kernel. For all $X_1, X_2, \cdots, X_n$, show that for every x,
(a) the map $\varphi \longmapsto \sqrt{[\,(\varphi^2) * dF_n\,](x)}$ is convex in φ,
(b) the map $\varphi \longmapsto \sqrt{[\,\varphi * dF_n\,](x)}$ is concave in φ (nonnegative),
(c) $\| \,\sqrt{(A^2)_h * dF_n}\,\|_1$ is an increasing function of h, and
(d) $\| \,\sqrt{(A_h)^2 * dF_n}\,\|_1$ is a decreasing function of h.

(3.52) Exercise. Complete the proof of Lemma (3.42) by verifying the statements (3.44)–(3.47).

(3.53) Exercise. Wrap up the proof of Theorem (3.31) by showing that

$$\| \,f_{n,PER} - f_o\,\|_1 =_{\text{as}} \mathcal{O}(\, n^{-2/5}\,)\ ,$$

using the fact that $H_{n,PER} \asymp_{\text{as}} n^{-1/5}$. [Hint: Use the monotonicity of $\| \,A_h * (dF_n - dF_o)\,\|_1$ as function of h.]

(3.54) Exercise. Use the integration by parts trick of § 4.3 to prove the following (weakened) version of Theorem (3.31):
If A and B satisfy (3.7) *and (3.8), and if $f \in W^{2,1}(\mathbb{R})$ and $\sqrt{f_o} \in L^1(\mathbb{R})$, then for all $\varepsilon > 0$ there exist positive constants c_1, c_2 such that*

$$c_1\, n^{-1/5-\varepsilon} \leqslant_{\text{as}} H_{n,PER} \leqslant_{\text{as}} c_2\, n^{-1/5+\varepsilon}$$

and

$$\| \,f_{n,PER} - f_o\,\|_1 =_{\text{as}} \mathcal{O}(\, n^{-2/5+\varepsilon}\,)\ .$$

(3.55) Exercise. For deterministic h, under the usual assumptions, provide a lower bound for

$$\mathbb{E}[\,\big|\,[A_h * (dF_n - dF_o)](x)\big|\,]\,,$$

which will show that

$$\mathbb{E}[\,\| \,A_h * (dF_n - dF_o)\,\|_1\,] \geqslant c\,(\,1 + nh\,)^{-1/2}\ .$$

(3.56) Exercise. You thought we would forget about it, didn't you? Prove Theorem (4.1.53) for the normal and two-sided exponential kernel.

Solution to Exercise (3.56). We prove the "only if" part of Theorem (4.1.53). The proof is by way of contradiction.

So, suppose that the statement "$H \longrightarrow_{\rm as} 0$, $nH \longrightarrow_{\rm as} \infty$" is not true. First, assume that "$H \longrightarrow_{\rm as} 0$" does not hold. Thus, pick a subsequence $\{ H_{n_i} \}_i$ for which

$$\lim_{i \to \infty} H_{n_i} = 2\,\delta > 0\,,$$

and consider the sequence $\{ \widetilde{H}_k \}_k$ with $\widetilde{H}_k = H_{n_i}$ for $n_i \leqslant k < n_{i+1}$. We denote $\widetilde{H}$ by just H. It now suffices to show that

$$\liminf_{n \to \infty} \| A_H * dF_n - f_o \|_1 > 0 \ .$$

First, the triangle inequality gives

$$\| A_H * dF_n - f_o \|_1 \geqslant \| A_H * f_o - f_o \|_1 - \| A_H * (dF_n - dF_o) \|_1 \ . \tag{3.57}$$

Since $H_n \geqslant \delta > 0$ for all n large enough (depending on the sample $X_1, X_2, \cdots, X_n$), then

$$\| A_H * (dF_n - dF_o) \|_1 \leqslant_{\rm as} \| A_\delta * (dF_n - dF_o) \|_1 =_{\rm as} \mathcal{O}\big((n^{-1} \log n)^{1/2} \big) \ ,$$

by Theorem (4.4.22). Also, because f_o is a density,

$$\| A_H * f_o - f_o \|_1 \geqslant_{\rm as} \inf_{h > \delta} \| A_h * f_o - f_o \|_1 = \eta > 0 \ .$$

From (3.57), it follows that $\liminf_n \| A_H * dF_n - f_o \|_1 >_{\rm as} 0$, and the same holds for the original sequence $\{ H_n \}_n$.

Now, suppose that $H \longrightarrow_{\rm as} 0$, but $\limsup_n nH \leqslant_{\rm as} C < \infty$. Take a subsequence for which $\liminf_n nH_n \leqslant_{\rm as} C$, and replace the whole sequence by a sequence for which $\lim_n nH_n \leqslant_{\rm as} C$, similar to the first part. The triangle inequality gives

$$\| A_H * dF_n - f_o \|_1 \geqslant \| A_H * (dF_n - dF_o) \|_1 - \| A_H * f_o - f_o \|_1 \ .$$

The last term converges to 0 a.s., since $H \longrightarrow_{\rm as} 0$. Now, since $nH \leqslant 2\,C$ for all n large enough (depending on $X_1, X_2, \cdots, X_n$), by monotonicity,

$$\begin{aligned} \| A_H * (dF_n - dF_o) \|_1 &\geqslant_{\rm as} \inf_{h \leqslant 2\,C/n} \| A_h * (dF_n - dF_o) \|_1 \\ &\geqslant_{\rm as} \| A_\lambda * (dF_n - dF_o) \|_1 \ , \end{aligned}$$

with $\lambda = 2\,C\,n^{-1}$. Now, by Theorem (4.4.22),

$$\| A_\lambda * (dF_n - dF_o) \|_1 \geqslant_{\rm as} \mathbb{E}[\, \| A_\lambda * (dF_n - dF_o) \|_1 \,] - \mathcal{O}\big((n^{-1} \log n)^{1/2} \big) \ .$$

By Exercise (3.55), the expected value exceeds $c\,(1 + n\lambda)^{-1/2}$, which is bounded away from 0. Thus, $\| A_\lambda * (dF_n - dF_o) \|_1$ will not tend to 0.

This proves the "only if" part of the Theorem. The "if" part goes along the same lines. Q.e.d.

Exercises: (3.10), (3.11), (3.15), (3.20), (3.48), (3.49), (3.50), (3.51), (3.52), (3.53), (3.54), (3.55), (3.56).

4. Asymptotic plug-in methods

In the remainder of this chapter, we are interested in smooth densities for which kernel estimators can achieve the $n^{-2/5}$ rate of convergence for the L^1 error. With this in mind, the kernel A is assumed to be a symmetric, square integrable pdf, with finite variance. See (2.16).

We recall that the goal of smoothing parameter selection in kernel density estimation is to minimize $\| f^{nh} - f_o \|_1$. The starting point of plug-in methods is the realization of § 4.4 that there is not much of a difference between $\| f^{nh} - f_o \|_1$ and its expectation. This is followed by the decomposition of the expected L^1 error into bias and variance components,

$$\mathbb{E}\big[\, \| f^{nh} - f_o \|_1 \,\big] \leqslant \mathrm{bias}(h) + (nh)^{-1/2}\, \mathrm{var}_o(A_h) \;, \tag{4.1}$$

where

$$\mathrm{bias}(h) = \| A_h * f_o - f_o \|_1 \;, \tag{4.2}$$

$$\mathrm{var}_o(A;h) = \| \sqrt{(A^2)_h * dF_o - h\,(A_h * dF_o)^2} \,\|_1 \;, \tag{4.3}$$

with the assumption that there is just about equality in (4.1). One additional step is taken, viz. the bias and variance terms in the bound (4.1) are replaced by the leading terms of their asymptotic expansions

$$\mathbb{E}\big[\, \| f^{nh} - f_o \|_1 \,\big] \leqslant \tfrac{1}{2}\, \sigma^2(A)\, \| f_o'' \|_1\, h^2 + (nh)^{-1/2}\, \| A \|_2\, \| \sqrt{f_o} \,\|_1 + \cdots \;. \tag{4.4}$$

Thus, a theoretically interesting choice of h is the minimizer of the right-hand side of (4.4), that is,

$$h_{n,\mathit{API}} = r_1(A)\, \varrho_1(f_o)\, n^{-1/5} \;, \tag{4.5}$$

where

$$r_1(A) = \big\{\, \tfrac{1}{2}\, \| A \|_2 \,/\, \sigma^2(A) \,\big\}^{2/5} \;, \tag{4.6}$$

$$\varrho_1(f_o) = \big\{\, \| \sqrt{f_o} \,\|_1 \,/\, \| f_o'' \|_1 \,\big\}^{2/5} \;. \tag{4.7}$$

Note that $r_1(A)$ and $\varrho_1(f_o)$ differ from the corresponding expressions in the asymptotic L^2 error, see § 2.

In the above development, the inequalities in (4.1) and (4.4) must be considered blemishes. It may be corrected by the following interesting device connected with the Central Limit Theorem, see Chapter 5 of Devroye and Györfi (1985) and also Hall and Wand (1988). The starting point is the *exact* decomposition

$$f^{nh}(x) - f_o(x) = \mathrm{bias}(x;h) + [\, A_h * (dF_n - dF_o) \,](x) \;, \tag{4.8}$$

where

$$\mathrm{bias}(x;h) = A_h * f_o(x) - f_o(x) \;. \tag{4.9}$$

Now, for fixed x and deterministically varying h,

$$n\,[\,A_h * (dF_n - dF_o)\,](x) = \sum_{i=1}^{n} Z_i$$

is the sum of the iid random variables

$$Z_i = A_h(x - X_i) - A_h * f_o(x)\ , \quad i = 1, 2, \cdots, n\ .$$

So by the Central Limit Theorem,

$$\sqrt{n}\,[A_h * (dF_n - dF_o)\,](x) \longrightarrow_{\mathrm{d}} h^{-1/2}\,\mathrm{var}_o(A; h; x)\,Y\ , \tag{4.10}$$

where $Y \sim N(0,1)$ and

$$\mathrm{var}_o(A; h; x) = \sqrt{(A^2)_h * f_o(x) - h\,(\,A_h * f_o(x)\,)^2}\ . \tag{4.11}$$

It is then reasonable to conclude that

$$\mathbb{E}[\,|\,f^{nh}(x) - f_o(x)\,|\,] \longrightarrow [\,\Psi_{nh} f_o\,](x)\ , \tag{4.12}$$

with

$$[\,\Psi_{nh} f_o\,](x) = \mathbb{E}[\,|\,\mathrm{bias}(x; h) + (nh)^{-1/2}\,\mathrm{var}_o(A; h; x)\,Y\,|\,]\ . \tag{4.13}$$

Therefore, upon integration with respect to $x \in \mathbb{R}$, we find an asymptotic expression for $\mathbb{E}[\,\|\,f^{nh} - f_o\,\|_1\,]$, but its correctness is somewhat suspect.

The above may be made rigorous based on the following precise result, taken lock, stock, and barrel from Devroye and Györfi (1985).

(4.14) Lemma. [Devroye and Györfi (1985)] *Let $X_1, X_2, \cdots, X_n$ be iid random variables with $\mathbb{E}[\,X_1\,] = 0$, $\mathbb{E}[\,|\,X_1\,|^2\,] = 1$, and $\mathbb{E}[\,|\,X_1\,|^3\,] < \infty$. Let*

$$S_n = n^{-1/2} \sum_{i=1}^{n} X_i\ ,$$

and let $Y \sim N(0,1)$ be independent of the X_i. Then,

$$\sup_{a \in \mathbb{R}} \left|\,\mathbb{E}[\,|\,a + S_n\,|\,] - \mathbb{E}[\,|\,a + Y\,|\,]\,\right| \leqslant c\,n^{-1/2}\,\mathbb{E}[\,|\,X_1\,|^3\,]\ ,$$

for some universal constant c.

Proof. Let $P_n(x) = \mathbb{P}[\,S_n \leqslant x\,]$, and let $\Phi(x)$ be the standard normal distribution. Then,

$$\begin{aligned}\mathbb{E}[\,|\,a + S_n\,|\,] - \mathbb{E}[\,|\,a + Y\,|\,] &= \int_{\mathbb{R}} |\,a + x\,|\,\{\,dP_n(x) - d\Phi(x)\,\} \\ &= \int_{\mathbb{R}} \{\,P_n(x) - \Phi(x)\,\}\,\mathrm{sign}(\,a + x\,)\,dx\ ,\end{aligned}$$

and so

$$\left|\,\mathbb{E}[\,|\,a + S_n\,|\,] - \mathbb{E}[\,|\,a + Y\,|\,]\,\right| \leqslant \int_{\mathbb{R}} \left|\,P_n(x) - \Phi(x)\,\right| dx\,.$$

Now, using the bound of Berry-Esseen type,

$$\big|\, P_n(x) - \Phi(x) \,\big| \leqslant c\; n^{-1/2}\, \mathbb{E}[\,|\, X_1 \,|^3\,]\, (\,1 + |\, x \,|\,)^{-3}\;, \quad x \in \mathbb{R}\;,$$

proves the lemma. Q.e.d.

The lemma implies that

$$\mathbb{E}[\,|\, f^{nh}(x) - f_o(x) \,|\,] = \mathbb{E}[\,|\, \mathrm{bias}(x;h) + (nh)^{-1/2}\, \mathrm{var}_o(A;h;x)\, Y \,|\,] + \varepsilon_{nh}(x)\;,$$

with

$$|\, \varepsilon_{nh}(x) \,| \leqslant c\, n^{-1}\, \frac{\mathbb{E}[\,|\, A_h(x - X_1) - A_h * f_o(x) \,|^3\,]}{\mathbb{E}[\,|\, A_h(x - X_1) - A_h * f_o(x) \,|^2\,]} \leqslant c_1\, (nh)^{-1}\;, \tag{4.15}$$

as $h \to 0$, for universal constants c, c_1.

Unfortunately, it is not permissible to integrate (4.15) with respect to $x \in \mathbb{R}$, but for any $T > 0$, we have

$$\mathbb{E}\Big[\int_{|\, x \,| < T} |\, f^{nh}(x) - f_o(x) \,|\, dx\,\Big] = \int_{|\, x \,| < T} \mathbb{E}[\,|\, \mathrm{bias}(x;h) + (nh)^{-1/2}\, \mathrm{var}_o(A;h;x)\, Y \,|\,]\, dx \;+\; \delta_{nhT}\;,$$

where

$$|\, \delta_{nhT} \,| \leqslant 2\, c_1\, T\, (nh)^{-1}\;.$$

On $|\, x \,| > T$, we use that

$$\mathbb{E}[\,|\, f^{nh}(x) - f_o(x) \,|\,] \leqslant \mathrm{bias}(x;h) + (nh)^{-1/2}\, \mathrm{var}_o(A;h;x)\;,$$

which implies that

$$\mathbb{E}[\int_{|\, x \,| > T} |\, f^{nh}(x) - f_o(x) \,|\, dx\,] \leqslant C\,(\,h^2 + (nh)^{-1/2}\,)\, \eta(T)\;,$$

for some function η, with $\eta(T) \longrightarrow 0$ for $T \to \infty$. Consequently,

$$\big|\, \mathbb{E}[\,\|\, f^{nh} - f_o \,\|_1\,] - \|\, \Psi_{nh} f_o \,\|_1 \,\big| \leqslant 2\, c_1\, T\, (nh)^{-1} + C\, \eta(T)\, (h^2 + (nh)^{-1/2})\;.$$

Now, take $T = o(\,\{\, h^2 + (nh)^{-1/2} \,\}^{-1}\,)$, and we have proven the following theorem.

(4.16) Theorem. [Devroye and Györfi (1985)] *If f_o satisfies the usual nonparametric assumptions, as well as $\mathbb{E}[\,|\, X_1 \,|^3\,] < \infty$, then for $h \to 0$, $nh \to \infty$,*

$$\mathbb{E}[\,\|\, f^{nh} - f_o \,\|_1\,] = \|\, \Psi_{nh} f_o \,\|_1 + o(\, h^2 + (nh)^{-1/2}\,)\;.$$

(4.17) Corollary. [Devroye and Györfi (1985)]

$$\mathbb{E}[\,\|\, f^{nh} - f_o \,\|_1\,] \leqslant \mathrm{bias}(h) + \sqrt{2/(\pi nh)}\;\, \mathrm{var}_o(A;h) + o(\, h^2 + (nh)^{-1/2}\,)\;.$$

Ignoring the $o(\cdots)$ terms in these two results then gives

$$\mathbb{E}[\,\|\,f^{nh} - f_o\,\|_1\,] = \int_{\mathbb{R}} \mathbb{E}[\,\Big|\,\tfrac{1}{2}h^2\sigma^2(A)\,(f_o)''(x) + (nh)^{-1/2}\,\|\,A\,\|_2\,\sqrt{f_o(x)}\;Y\,\Big|\,]\,dx \tag{4.18}$$

and

$$\mathbb{E}[\,\|\,f^{nh} - f_o\,\|_1\,] \leqslant \tfrac{1}{2}h^2\sigma^2(A)\,\|\,(f_o)''\,\|_1 + (2/\pi nh)^{1/2}\,\|\,A\,\|_2\,\|\,\sqrt{f_o}\,\|_1\;. \tag{4.19}$$

The first expression is the basis of L^1 plug-in methods, see Hall and Wand (1988) and the next section. Minimizing the second (asymptotic) upper bound gives the minimizer

$$h_{n,\,API} = (2/\pi)^{1/5}\,r_1(A)\,\varrho_1(f_o)\;n^{-1/5}\;. \tag{4.20}$$

A useful upper bound on $h_{n,\,API}$ is given in the next exercise.

(4.21) Exercise. [Devroye and Györfi (1985), p. 113] Show that for A the Epanechnikov kernel $h_{n,\,API} \leqslant h_{n,\,UP}$, where

$$h_{n,\,UP} = \sigma\,\left(\frac{98415}{65536}\,\frac{\pi^4}{n}\right)^{1/5} \doteq 2.71042\,\sigma\,n^{-1/5}\;,$$

in which σ is the standard deviation of X_1. Estimating σ, e.g., by the interquartile estimator q_n, see (2.39), then gives an estimator for this "upper bound", denoted by $H_{n,\,UP\text{–}E}$ for the Epanechnikov kernel. There is a similar bound for the upper bound $H_{n,\,UP\text{–}N}$ for the the normal kernel. [Hint: The framework of § 11.2 seems to fit.]

As in § 2, the expression (4.20) may be used as the starting point for estimation procedures. The factor $r_1(A)$ can be computed exactly, but $\varrho_1(f_o)$ is unknown and must be estimated. In the asymptotic plug-in methods, $\varrho_1(f_o)$ is estimated based on kernel estimators of f_o, with smoothing parameters chosen by some (other) method.

The literature on L^1 plug-in methods is not large. Deheuvels and Hominal (1980) suggest a normal pilot estimator, similar to Deheuvels (1977) in the L^2 context, see (2.25). The simulations of Berlinet and Devroye (1994) and others suggest that the success of any plug-in method depends mainly on the performance of the pilot estimator. So the pilot estimator must be selected carefully. According to Berlinet and Devroye (1994) a carefully tweaked double kernel method works the best. Their L^1 plug-in method with the pilot double kernel method is as follows. Let A be the Epanechnikov kernel, let $B = L_3$, as in (3.18)–(3.19), so that B is a fourth-order kernel, and let $H_{n,\,DBL}$ be the corresponding double kernel smoothing parameter: It solves

$$\text{minimize} \quad \|\,(A_h - B_h) * dF_n\,\|_1 \quad \text{over } h > 0\;.$$

Now, $\| \sqrt{f_o} \|_1$ is estimated by $\| \sqrt{f_{n,DBL}} \|_1$. For the term $\| f_o'' \|_1$, the starting point is (4.4), with the variance term omitted, thus,

$$\| (A_h - B_h) * dF_n \|_1 = \tfrac{1}{2} h^2 \sigma^2(A) \| f_o'' \|_1 + \cdots \ ,$$

and so $\| (f_o)'' \|_1$ is estimated by

$$D_n = 2 \left\{ H_{n,DBL} \, \sigma(A) \right\}^{-2} \| (A_{H_{n,DBL}} - B_{H_{n,DBL}}) * dF_n \|_1 \ . \tag{4.22}$$

Actually, this is not it completely. To get the plug-in method to work for small sample sizes, a few more tweaks are required. First, to make sure that the smoothing parameter is large enough that the bias term in (4.4) is (much) larger than the variance term, in (4.22), the smoothing parameter $H_{n,DBL}$ is replaced by

$$H_{n,DBL} \cdot \max(\, 1 \,,\, 10 \, R \,), \tag{4.23}$$

where

$$R = \frac{\| \sqrt{f_{n,DBL}} \|_1 \, \| A - B \|_2}{(n H_{n,DBL})^{1/2} \, \| (A_{H_{n,DBL}} - B_{H_{n,DBL}}) * dF_n \|_1} \ . \tag{4.24}$$

Here, R is an estimator for the ratio of the bias and the variance term. Secondly, the initial choice of the smoothing parameter is taken to be

$$\widetilde{H}_{\, n,API-DBL} = \left(\| (f_{n,DBL})^{1/2} \|_1 \, D_n^{-1} \right)^{-2/5} \left(15/2\pi n \right)^{-1/5} \ , \tag{4.25}$$

with D_n as in (4.22). Finally, in view of Exercise (4.21), the final choice in the asymptotic plug-in method for the smoothing parameter is

$$H_{n,API-DBL} = \min(\, \widetilde{H}_{\, n,API-DBL} \, , \, 2.71042 \, q_n \, n^{-1/5} \,) \ , \tag{4.26}$$

with q_n the inner quartile estimator of the scale, see (2.39).

As a closing comment, we note that this method is relatively simple and contains essentially only one tuning parameter, to wit, the stretch parameter in the underlying double kernel method. Therefore, its unparalleled practical performance, especially for "weird" densities, is quite remarkable. See Berlinet and Devroye (1994) and Chapter 8.

Exercise: (4.21).

5. Away with pilot estimators!?

There is something unsatisfying about the use of pilot estimators: If the pilot estimators are that good, then why bother plugging them into asymptotic formulas? The adjective "asymptotic" actually refers to two types of asymptotic behavior. On the one hand, the sample size n tends to ∞, and on the other, the window parameter h tends to 0. The former is stochastic in nature, the latter is mostly deterministic. In this section, we concentrate on the stochastic aspects, and let the window parameter be. It will

transpire that this way one can eliminate the pilot estimators, at least for smooth densities. All of the methods turn out to be variational in nature, i.e., the window parameter is chosen by minimizing a certain estimator of the "error". We begin with a simple-minded approach suggested by the previous section, after which, we are fully armed to deal with the method of Devroye and Györfi (1985) and Hall and Wand (1988).

As in § 4, we consider smooth densities, and assume that A is a symmetric, square integrable pdf with finite variance.

We are interested in estimating $\| f^{nh} - f_o \|_1$, but § 4.5 tells us we may just as well consider its expected value. One simple-minded message of the previous section is that

$$\mathbb{E}[\, \| f^{nh} - f_o \|_1 \,] \approx \mathrm{bias}(h) + (2/\pi nh)^{-1/2} \, \mathrm{var}_o(A;h) \ , \tag{5.1}$$

with $\mathrm{bias}(h)$ and $\mathrm{var}_o(A;h)$ as in (4.2)–(4.3). Now, the question arises of whether it is possible to estimate the right-hand side of (5.1).

As in § 3, we estimate $\mathrm{var}_o(A;h)$ by $\mathrm{var}_n(A;h)$,

$$\mathrm{var}_n(A;h) = \| \sqrt{(A_h)^2 * dF_n - (A_h * dF_n)^2} \, \|_1 \ . \tag{5.2}$$

This should work reasonably well, for h not too small.

The bias term causes more difficulty. We do not know of a (good) method for estimating it over a wide range of h and, hence, must resort to small h asymptotics after all. For small h, we have the asymptotic expansion

$$E[\, \| A_h * f_o - f_o \|_1 \,] = \tfrac{1}{2} \, \sigma^2(A) \, h^2 \, \| \, (f_o)'' \, \|_1 + o(h^2) \ , \quad h \to 0 \ , \tag{5.3}$$

cf. Theorem (4.7.2), and this may be used to our advantage. A similar asymptotic behavior is obtained for $\| \, (A_h - B_h) * f_o \, \|_1$, provided that the kernel B is symmetric, with finite $\sigma^2(B)$, viz.

$$\| \, (A_h - B_h) * f_o \, \|_1 = \tfrac{1}{2} \, h^2 \, |\sigma^2(A - B)| \, \| \, (f_o)'' \, \|_1 + o(h^2) \ . \tag{5.4}$$

To simplify the presentation, we henceforth assume that

$$B \text{ is a fourth-order kernel} \ , \tag{5.5}$$

beacuse then $\sigma^2(B) = 0$, and asymptotically, the right-hand sides of (5.3) and (5.4) are the same.

(5.6) Exercise. Prove (5.4).

It is useful to introduce the shorthand notations

$$C = A - B \tag{5.7}$$

and $C_h = A_h - B_h$. Continuing, (5.3) and (5.4) imply that

$$\| \, C_h * dF_o \, \|_1 = \| \, A_h * f_o - f_o \, \|_1 + o(h^2) \ ,$$

and the left-hand side may be estimated by $\| C_h * dF_n \|_1$, except that this introduces another variance term. Employing the analogue of (5.1), we obtain

$$\mathbb{E}[\, \| C_h * dF_n \|_1 \,] \approx \| C_h * f_o \| + (2/\pi nh)^{-1/2} \operatorname{var}_o(C; h) \;, \tag{5.8}$$

with $\operatorname{var}_o(C; h)$ as in (4.3). Again, we assume that there is just about equality in (5.8). Because we know how to estimate the variance term in (5.8), we are thus lead to estimating the error $\| A_h * dF_n - f_o \|_1$ by

$$\begin{aligned} \mathit{SPI}(h) \stackrel{\text{def}}{=} \; & \| (A_h - B_h) * dF_n \|_1 + \\ & (2/\pi nh)^{-1/2} \left\{ \operatorname{var}_n(A; h) - \operatorname{var}_n(A - B; h) \right\} . \end{aligned} \tag{5.9}$$

A good way to select the smoothing parameter h then ought to be:

(5.10) in the *SPI* method, the smoothing parameter selected is the solution to

$$\text{minimize} \quad \mathit{SPI}(h) \quad \text{subject to} \quad h > 0 \;.$$

The h so selected is denoted by $H_{n,\mathit{SPI}}$ and the corresponding kernel estimator by $f_{n,\mathit{SPI}}$.

Before continuing, some comments regarding the *SPI* method are in order. First, "*SPI*" stands for Small sample Plug-In, which seems apt enough. Second, there is a striking similarity with the perverted double kernel method. Third, the estimator of the bias term in the small sample plug-in method is copied from the double kernel method, although the motivation is somewhat different. Fourth, all of the considerations that went into the definition of $\mathit{SPI}(h)$ hinge on approximate equality in the inequalities (5.1) and (5.8). Although this is not true for all values of h (e.g., for $h \to \infty$), it seems to hold for all relevant h.

The above motivation of the *SPI* method is a very rough version of the asymptotic considerations of § 4. We may repeat those considerations in the present context, without (explicitly) letting $h \to 0$. Thus, from Theorem (4.16),

$$\mathbb{E}[\, \| f^{nh} - f_o \|_1 \,] = \| \Psi_{nh} f_o \|_1 + o(\, (nh)^{-1/2} + h^2 \,) \;, \tag{5.11}$$

where, see (4.13),

$$[\, \Psi_{nh} f_o \,](x) = \mathbb{E}[\, | \operatorname{bias}_o(x) + (nh)^{-1/2} \operatorname{var}_o(A; h; x)\, Y \, | \,] \;, \tag{5.12}$$

in which $Y \sim N(0, 1)$. As before, $\operatorname{var}_o(A; h; x)$ is estimated (accurately) by $\operatorname{var}_n(A; h; x)$. As usual, the bias term causes problems, but it is clear that its estimation should involve $C_h * dF_n = (A_h - B_h) * dF_n(x)$. Hindsight suggests one consider

$$\mathbb{E}_Z[\, | \, C_h * dF_n + (nh)^{-1/2} \operatorname{stuff}(x) Z \, | \,] \;,$$

where $Z \sim N(0, 1)$ is independent of $X_1, X_2, \cdots, X_n$, and $\mathbb{E}_Z$ denotes expectation with respect to Z. Here, "$\operatorname{stuff}(x)$" is independent of Z and is

to be determined such that

$$\mathbb{E}_Z[\,|\,C_h * dF_n + (nh)^{-1/2}\,\mathrm{stuff}\,(x)Z\,|\,] \approx [\,\Psi_{nh} f_o\,](x)\ . \tag{5.13}$$

Now, similar to what went on in § 4,

$$\begin{aligned}\mathbb{E}_Z[\,|\,C_h * dF_n + (nh)^{-1/2}\,\mathrm{stuff}\,(x)Z\,|\,] \approx \\ \mathbb{E}_{Z,Y}[\,|\,C_h * dF_o + (nh)^{-1/2}\,\{\,\mathrm{stuff}\,(x)Z\ + \mathrm{var}_o(C;h;x)\,Y\,\}\,|\,]\ ,\end{aligned} \tag{5.14}$$

where $Y \sim N(0,1)$ is independent of Z. Thus, since we know how to add independent normal random variables, the last expectation equals

$$\mathbb{E}_Z[\,|\,C_h * dF_o + (nh)^{-1/2}\,\sqrt{\mathrm{stuff}\,(x)^2 + (\mathrm{var}_o(C;h;x))^2}\ Z\,|\,]\ ,$$

where Z is another $N(0,1)$ random variable, and this expression equals $[\,\Psi_{nh} f_o\,](x)$ if

$$\big(\,\mathrm{stuff}(x)\,\big)^2 = (\mathrm{var}_o(A;h;x))^2 - (\mathrm{var}_o(C;h;x))^2\ .$$

Beacuse the right-hand side could well be negative, we choose

$$\mathrm{stuff}\,(x) = \mathrm{var}_o(A,C;h;x)\ ,$$

where

$$\mathrm{var}_o(A,C;h;x) \stackrel{\mathrm{def}}{=} \sqrt{0 \vee \big\{\,(\mathrm{var}_o(A;h;x))^2 - (\mathrm{var}_o(C;h;x))^2\,\big\}}\ . \tag{5.15}$$

This may be estimated in the usual manner by $\mathrm{var}_n(A,C;h;x)$, defined similarly to (5.15). Thus, we define the empirical functional, freely after Devroye and Györfi (1985) and Hall and Wand (1988),

$$\mathit{DGHW}(h) \stackrel{\mathrm{def}}{=} \big\|\,\mathbb{E}_Z[\,|\,C_h * dF_n + (nh)^{-1/2}\,\mathrm{var}_n(A,C;h;\,\cdot\,)\,Z\,|\,]\,\big\|_1\ , \tag{5.16}$$

where $C_h = A_h - B_h$, and $Y \sim N(0,1)$, which leads to the following method for choosing the smoothing parameter:

(5.17) in the *DGHW* method, the smoothing parameter is the solution to

$$\text{minimize}\quad \mathit{DGHW}(h)\quad \text{subject to}\quad h > 0\ .$$

The h so selected is denoted by $H_{n,\mathit{DGHW}}$ and the corresponding kernel estimator by $f_{n,\mathit{DGHW}}$.

Another method immediately suggests itself. The triangle inequality gives the upper bound for $\mathit{DGHW}(h)$

$$\mathit{MOD}(h) \stackrel{\mathrm{def}}{=} \|\,(A_h - B_h) * dF_n\,\|_1 + \frac{2}{\sqrt{\pi n h}}\,\|\,\mathrm{var}_n(A,C;h;\,\cdot\,)\,\|_1\ , \tag{5.18}$$

and this gives yet another method for choosing the window parameter:

(5.19) in the *MOD* method, the smoothing parameter is the solution to

$$\text{minimize} \quad MOD(h) \quad \text{subject to} \quad h > 0 \ .$$

The h so selected is denoted by $H_{n,MOD}$, and the corresponding kernel estimator by $f_{n,MOD}$.

The rather limp designation *MOD* stands for MODified, as in modified double kernel method.

We shall not prove that the methods *DGHW* and *MOD* work, either in our sense or in the sense of asymptotic optimality. One would certainly expect under the usual nonparametric smoothness and tail assumptions that the *DGHW* method is asymptotically optimal, and that for the *MOD* method, one can prove with "our" methods that

$$H_{n,MOD} \asymp_{\text{as}} n^{-1/5} \ , \tag{5.20}$$

but *we* shall not do so.

At this point, something should be said about the choice of the kernels A and B. We mention two possibilities based on A being the normal kernel. The first choice for B is to take $B = 2\,A - A * A = 2\,A - A_{\sqrt{2}}$ or, admitting a stretch (shrink) parameter,

$$B = 2\,A_\lambda - A_{\lambda\sqrt{2}} \ , \tag{5.21}$$

for which $\sigma^2(B) = 0$. Simulations (unreported) suggest that $\lambda \approx 0.83$ gives the best results. The other choice for the kernel B is

$$B = A_\lambda \ , \tag{5.22}$$

with stretch parameter λ. Simulations (again unreported) seem to indicate that $\lambda \approx 0.995$ is the most preferable choice. This also suggests the limiting case $\lambda \to 1$, i.e., with the kernel C of (5.7) (recall A is the normal kernel) given by

$$C(x) = \frac{1}{\sqrt{8\pi}}\,(x^2 - 1)\,\mathrm{e}^{-x^2/2} \ , \tag{5.23}$$

which performs just about the same as the previous method, except for the uniform density where it is noticeably worse. This also indicates that we are skating on thin ice with $\lambda = 0.995$. In Chapter 8, we show some simulations for these methods.

We finish this section by proving that the *SPI* method works, in the sense that

$$H_{n,SPI} \asymp_{\text{as}} n^{-1/5} \quad , \quad n \to \infty \ . \tag{5.24}$$

The conditions required are that A and B are as in (5.21) with $\lambda = 1$ (no stretching), and that f_o satisfies the usual nonparametric assumptions, see

(4.1.47)–(4.1.48). The general approach to the proof of (5.24) is as in § 3, where most of the hard work was done. In particular, we leave as exercises the following three lemmas.

(5.25) Lemma. *There exists a constant K depending on f_o only such that*

$$\limsup_{n\to\infty}\ n^{2/5}\ \inf_h SPI(h) \leqslant_{\text{as}} K\ .$$

(5.26) Lemma. *For a large enough constant γ, depending on f_o only,*

$$\liminf_{n\to\infty}\ n^{2/5}\ \inf\left\{\, SPI(h)\ :\ h \geqslant \gamma\, n^{-1/5}\,\right\} >_{\text{as}} K\ .$$

(5.27) Lemma. *For a small enough positive constant δ, depending on f_o only,*

$$\liminf_{n\to\infty}\ n^{2/5}\ \inf\left\{\, SPI(h)\ :\ h \leqslant \delta\, n^{-1/5}\,\right\} >_{\text{as}} K\ .$$

In comparison with § 3, an extra ingredient is required for the proof of Lemma (5.27). It would seem obvious that to prove this lemma, we need a suitable lower bound on the quantity $\text{var}_n(A;h) - \text{var}_n(A-B;h)$. The first step is to consider instead lower bounds for

$$\|\,\sqrt{(A^2)_h * dF_n}\,\|_1 - \|\,\sqrt{((A-B)^2)_h * dF_n}\,\|_1\ .$$

A nice bound would follow if for some $0 < r < 1$,

$$A(x) \geqslant r\,|\,A(x) - B(x)\,|\ ,\quad x \in \mathbb{R}\ , \tag{5.28}$$

but this fails for $x \to \pm\infty$. (Recall that A is the standard normal, and $B = 2\,A - A * A$.) Of course, where it matters, $|\,A - B\,|$ is much smaller than A. This may be expressed is as follows.

(5.29) Exercise. For A the normal density and $B = 2\,A - A * A$ show that $\|\,A\,\|_2 = 0.531\cdots$, and

$$\|\,A\,\|_2 - \|\,A - B\,\|_2 = \sqrt{\frac{1}{\sqrt{4\,\pi}}} - \sqrt{\frac{1}{\sqrt{4\,\pi}} - \frac{2}{\sqrt{6\,\pi}} + \frac{1}{\sqrt{8\,\pi}}} = 0.38653\cdots\ .$$

However, instead of (5.28), we have the following inequality, which we elevate to the status of lemma, although no formal proof is given.

(5.30) Lemma. *For A the normal kernel,*

$$|\,A(x) - A_{\sqrt{2}}(x)\,| \leqslant r\,A_{\sqrt{3}}(x)\ ,\quad x \in \mathbb{R}\ ,$$

where $r = 0.515\cdots$.

PROOF BY GRAPHICS. Plot the ratio of the left- and right-hand side of the inequality, and read off the value of r. Q.e.d.

(5.31) EXERCISE. As the graph of $|A - A_{\sqrt{2}}|/A_{\sqrt{3}}$ indicates, there are three local maxima, the middle one of which is a tad below the global maximum. This suggests that the comparison with $A_{\sqrt{3}}$ is not optimal. To get the optimal comparison, consider the problem

$$\text{minimize} \max_{x\in\mathbb{R}} \frac{|A(x) - A_{\sqrt{2}}(x)|}{A_\lambda(x)} \quad \text{subject to } \lambda > 0 \ .$$

($\lambda = \sqrt{3}$ is very close to the optimum.)

With the previous lemma in hand, we now prove the following lemma.

(5.32) LEMMA. *For all realizations of $X_1, X_2, \cdots, X_n$,*

$$\big\|\sqrt{(A_h)^2 * dF_n}\,\big\|_1 - \big\|\sqrt{(A_h - A_{h\sqrt{2}})^2 * dF_n}\,\big\|_1 \geqslant (1-r)\,\big\|\sqrt{(A_h)^2 * dF_n}\,\big\|_1 \ .$$

PROOF. By Lemma (5.30) and Exercise (3.51)(d), we have the inequalities

$$\begin{aligned}\big\|\sqrt{(A_h - A_{h\sqrt{2}})^2 * dF_n}\,\big\|_1 &\leqslant r\,\big\|\sqrt{(A_{h\sqrt{3}})^2 * dF_n}\,\big\|_1 \\ &\leqslant r\,\big\|\sqrt{(A_h)^2 * dF_n}\,\big\|_1 \ .\end{aligned}$$

The lemma follows. Q.e.d.

(5.33) EXERCISE. (a) Prove Lemmas (5.25), (5.26), and (5.27), and that the *SPI* method works in the sense of (5.24).
(b) Prove that the *MOD* method works in the sense of (5.20).

(5.34) EXERCISE. Prove the asymptotic optimality of the *DGHW* method, and send the authors a copy: They are not actually sure they know how to do this!

(5.35) EXERCISE. For the actual computation of *DGHW*(h), the following identity is useful (see DEVROYE and GYÖRFI (1985), Chapter 5). Let $Y \sim N(0,1)$ and let a, $b \in \mathbb{R}$, with $b > 0$. Then,

$$\mathbb{E}[\,|\,a + b\,Y\,|\,] = a\ \mathrm{erf}(x) + b\,(2/\pi)^{1/2}\ \exp(-x^2) \ ,$$

where $x = a/(b\sqrt{2})$ and $\mathrm{erf}(x)$ is the error function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2}\,dt \ .$$

EXERCISES: (5.6), (5.29), (5.31), (5.33), (5.34), (5.35).

6. A discrepancy principle

Until now, the procedures for selecting the smoothing parameter consisted of minimizing some estimate of $\| f^{nh} - f_o \|_1$ over h. A different kind of method may be based on discrepancy principles. The notion was introduced by applied mathematicians: by REINSCH (1967) in the context of smoothing splines, and by MOROZOV (1966) and ARCANGELI (1966) for the selection of the regularization parameter in ill-posed least-squares problems. See Volume II. Generally speaking, statisticians shudder at the thought, but it is a notion worth considering. We describe here one method for selecting the smoothing parameter for kernel density estimation, based on a discrepancy principle.

The assumptions on the kernel A are the usual ones, i.e., A is a symmetric, square-integrable pdf with finite variance, see (2.16). Regarding the density f_o, the usual nonparametric assumptions suffice.

To begin, we recall the Kolmogorov-Smirnov statistic $\| F_n - \Psi \|_\infty$ for testing whether an iid sample with empirical distribution F_n has been drawn from a distribution Ψ. Also recall that $\| F_n - F_o \|_\infty$ is distribution free, i.e., its distribution does not depend on F_o,

$$\| F_n - F_o \|_\infty =_{\mathrm{d}} \| U_n - U_o \|_\infty \ , \tag{6.1}$$

where $U_n(t)$ is the empirical distribution of an iid sample of size n from the uniform $(0,1)$ distribution and $U_o(t) = t$, $0 \leqslant t \leqslant 1$. Finally, recall the CHUNG (1949) law of the iterated logarithm

$$\| F_n - F_o \|_\infty =_{\mathrm{as}} \mathcal{O}\big((n^{-1} \log\log n)^{1/2} \big) \ . \tag{6.2}$$

Putting two and two together, the following would seem to make sense. For arbitrary $h > 0$, let F^{nh} be the distribution associated with f^{nh}, the kernel estimator under discussion (or any other estimator depending on h). Now, choose h such that

$$\| F_n - F^{nh} \|_\infty = c\, (n^{-1} \log\log n)^{1/2} \ , \tag{6.3}$$

for some (suitable) constant c. After all, we should not expect F^{nh} to be closer to F_n than the true F_o. Unfortunately, this does not quite work, apparently due to the ill-posedness of density estimation. As shown below, for all intents and purposes,

$$\| F_n - F^{nh} \|_\infty \approx c\, h^2 \ , \tag{6.4}$$

so (6.3) would imply that $h \asymp n^{-1/4}$, give or take a power of $\log\log n$. However, asymptotically, the optimal h satisfies $h \asymp n^{-1/5}$, so in (6.3), we are demanding that F^{nh} is too close to F_n. From this point of view, replacing the right-hand side of (6.3) by $n^{-2/5}$ could be right:

(6.5) In the *DP* method, the smoothing parameter h is selected as the smallest solution of

$$\| F_n - F^{nh} \|_\infty = c_{DP}\, n^{-2/5} \ ,$$

where $c_{DP} = 0.35$. The h so selected is denoted by $H_{n,DP}$ and the corresponding kernel estimator $f^{n,H_{n,DP}}$ by $f_{n,DP}$.

The numerical value of c_{DP} is of course rather mysterious, but works remarkably well for smooth densities with light tails. This will be appreciated more fully when we come to the simulations of Chapter 8.

We now show that $H_{n,DP}$ is a reasonable smoothing parameter.

(6.6) THEOREM. *Let $f_o \in W^{2,1}(\mathbb{R})$, $\sqrt{f_o} \in L^1(\mathbb{R})$. Then, $H_{n,DP}$ exists and satisfies*

$$H_{n,DP} =_{\text{as}} c\, n^{-1/5}\, (\, 1 + o(1)\,) \ ,$$

where $c > 0$ is given by

$$c^2 = 2\, c_{DP} \left\{ \sigma^2(A) \| f_o' \|_\infty \right\}^{-1} \ .$$

(6.7) COROLLARY. *If A is the normal or two-sided exponential density, then $\| f_{n,DP} - f_o \|_1 =_{\text{as}} \mathcal{O}\big(n^{-2/5} \big)$.*

(6.8) COROLLARY. *$H_{n,DP}$ is scale invariant in the sense of* (1.16).

The proof of Theorem (6.6) follows from a series of lemmas, which we leave as exercises. The first series deals with the existence of $H_{n,DP}$.

(6.9) LEMMA. *Let F_n be the empirical distribution of $X_1, X_2, \cdots, X_n$, and let $X_{1,n} \leqslant X_{2,n} \leqslant \cdots \leqslant X_{n,n}$ be the order statistics. If Ψ is a continuous distribution, then*

$$\| F_n - \Psi \|_\infty = \max_{1 \leqslant i \leqslant n} \left\{ \left| \tfrac{i-1}{n} - \Psi(X_{i,n}) \right| , \left| \tfrac{i}{n} - \Psi(X_{i,n}) \right| \right\} \ .$$

It is useful to introduce the distribution $\mathbb{A}$ corresponding to the density A,

$$\mathbb{A}(x) = \int_{-\infty}^{x} A(y)\, dy \ , \quad x \in \mathbb{R} \ . \tag{6.10}$$

Then, the distribution function for the density A_h is $\mathbb{A}_h(x) = \mathbb{A}(h^{-1}x)$ (note the missing factor h^{-1}), and

$$F^{nh}(x) = \tfrac{1}{n} \sum_{i=1}^{n} \mathbb{A}\big(h^{-1}(x - X_i) \big) \ , \quad x \in \mathbb{R} \ . \tag{6.11}$$

Now, observe that $\lim_{h \to 0}\ \mathbb{A}(h^{-1}z)$ equals 0 for $z < 0$ and equals 1 for $z > 0$, and that $\lim_{h \to \infty}\ \mathbb{A}(h^{-1}z) = \mathbb{A}(0) = \frac{1}{2}$ for all z. This gives:

(6.12) LEMMA. *For almost all realizations of $X_1, X_2, \cdots, X_n$, and for $1 \leqslant i \leqslant n$,*

(a) $$F^{nh}(X_{i,n}) \longrightarrow \tfrac{i-1/2}{n} \quad \text{for } h \to 0 \ , \ \text{and}$$

(b) $$F^{nh}(X_{i,n}) \longrightarrow \tfrac{1}{2} \quad \text{for } h \to \infty \ .$$

(6.13) LEMMA. *For almost all realizations of $X_1, X_2, \cdots, X_n$,*

(a) $$\lim_{h \to 0} \| F_n - F^{nh} \|_\infty = \tfrac{1}{2n} \ , \ \text{and}$$

(b) $$\lim_{h \to \infty} \| F_n - F^{nh} \|_\infty = \tfrac{1}{2} \ .$$

Together with the continuity of $\| F_n - F^{nh} \|_\infty$ as a function of h, the above lemmas suffice to show the existence of $H_{n,DP}$.

To prove the asymptotic behavior, we need one last ingredient. Let F_h be the distribution for the density $A_h * f_o$, so

(6.14) $$F_h = A_h * F_o = \mathbb{A}_h * dF_o \ .$$

(6.15) LEMMA. *If $f \in W^{2,1}(\mathbb{R})$, then*

(a) $$\| F_o - F_h \|_\infty = \tfrac{1}{2} h^2 \sigma^2(A) \| f_o' \|_\infty + o(h^2) \ , \ \text{and}$$

(b) $$\Big| \| F_n - F^{nh} \|_\infty - \| F_o - F_h \|_\infty \Big| \leqslant 2 \| F_o - F_n \|_\infty \ .$$

PROOF OF (a). Observe that

$$F_h(x) - F_o(x) = \int_{\mathbb{R}} A_h(x-y) \big(F_o(y) - F_o(x) \big) \, dy \ .$$

Now, Taylor's theorem with exact remainder gives

$$F_o(y) - F_o(x) = (y-x) f_o(x) + \tfrac{1}{2} (x-y)^2 f_o'(x) - \int_y^x (y-z)^2 f_o''(z) \, dz \ .$$

It follows that

$$F_h(x) - F_o(x) = \tfrac{1}{2} h^2 \sigma^2(A) f_o'(x) + R \ ,$$

with

$$R = \int_{\mathbb{R}} A_h(x-y) \int_y^x (y-z)^2 f_o''(z) \, dz \, dy \ .$$

It thus suffices to show that $R = o(h^2)$, uniformly in x. Note that

(6.16) $$| R | \leqslant \int_{\mathbb{R}} (x-y)^2 A_h(x-y) \Big| \int_y^x f_o''(z) \, dz \Big| \, dy \ .$$

To bound the right-hand side of (6.16), we split the integral into two parts. Let

$$R_1 = \int_{|x-y|>\sqrt{h}} (x-y)^2 \, A_h(x-y) \left| \int_y^x f_o''(z)\, dz \right| dy \ ,$$
$$R_2 = \int_{|x-y|<\sqrt{h}} (x-y)^2 \, A_h(x-y) \left| \int_y^x f_o''(z)\, dz \right| dy \ .$$

First, uniformly in x,

$$\begin{aligned} R_1 &\leqslant \| f_o'' \|_1 \int_{|y|>\sqrt{h}} y^2 \, A_h(y)\, dy \\ &\leqslant h^2 \, \| f_o'' \|_1 \int_{|y|>1/\sqrt{h}} y^2 \, A(y)\, dy = o(h^2) \ . \end{aligned}$$

Secondly, let

$$\omega(f\,;h) = \sup_{|x-y|<\sqrt{h}} \left| \int_y^x (y-z)^2 \, f_o''(z)\, dz \right| . \tag{6.17}$$

Since $f_o'' \in L^1(\mathbb{R})$, then $\lim_{h\to 0} \omega(f\,;h) = 0$. Consequently, uniformly in x,

$$\begin{aligned} R_2 &\leqslant \omega(f\,;h) \int_{|x-y|<\sqrt{h}} (x-y)^2 \, A_h(x-y)\, dy \\ &\leqslant h^2 \, \sigma^2(A)\, \omega(f\,;h) = o(h^2) \ . \end{aligned}$$

Part (a) follows. Q.e.d.

(6.18) Exercise. Prove the remaining results of these sections.

Exercise : (6.18).

7. The Good estimator

The last item in this chapter is choosing the smoothing parameter for maximum penalized likelihood estimation. We consider only the roughness penalization of Good (1971) and Good and Gaskins (1971), as elaborated in § 5.2. We discuss an old and very interesting proposal of Klonias (1984) and a method based on the discrepancy principle of § 6.

We recall that the Good (1971) estimator f^{nh} is given by $f^{nh} = (u^{nh})^2$, where u^{nh} solves

$$\begin{aligned} &\text{minimize} && -2 \int_{\mathbb{R}} \log u(x)\, dF_n(x) + \| u \|_2^2 + h^2 \, \| u' \|_2^2 \\ &\text{subject to} && u \in W^{2,1}(\mathbb{R}) \ , \ u \geqslant 0 \end{aligned} \tag{7.1}$$

($L^2(\mathbb{R})$ norms, and the prime denotes differentiation), and that

$$\| u^{nh} \|_2^2 = 1 - h^2 \| (u^{nh})' \|_2^2 \ . \tag{7.2}$$

Also, recall that u^{nh} is implicitly given by

$$u^{nh}(x) = \tfrac{1}{n} \sum_{i=1}^{n} \frac{\mathfrak{B}_h(x - X_i)}{u^{nh}(X_i)} \ , \quad x \in \mathbb{R} \ , \tag{7.3}$$

with $\mathfrak{B}_h(x) = (2\,h)^{-1} \exp(-h^{-1}\,|x|\,)$, the scaled, two-sided exponential pdf.

The first method for choosing h is a very interesting proposal of Klonias (1984), viz.

(7.4) in the *GK* method, the smoothing parameter is chosen as the (smallest) solution of

$$\text{minimize} \quad h^2 \| (u^{nh})' \|_2^2 \quad \text{over } h > 0 \ .$$

The h so chosen is denoted by $H_{n,GK}$ and the resulting estimator by $f_{n,GK}$.

Actually, the proposal of Klonias (1984) was to minimize the Lagrange multiplier when the pdf constraint $\| u^{nh} \|_2 = 1$ is incorporated into (7.1), but this is equivalent to minimizing $h^2 \| (u^{nh})' \|_2$ in our context. See Exercise (11.3.21). Also, note that as a consequence of the Comparison Lemma (5.2.22), see (5.2.23), we may write

$$h^2 \| (u^{nh})' \|_2^2 = \| f^{nh} - \mathfrak{B}_{h/\sqrt{2}} * dF_n \|_1 \ , \tag{7.5}$$

so that (7.4) may be thought of as a double kernel (or double estimator) method. It turns out that this method does not work, but it should be kept in mind that the proposal was made when little was known concerning rational ways of selecting the smoothing parameter. Be that as it may, in the course of its investigation, useful facts about the Good estimator have to be unearthed, which will be put to good use in the study of the discrepancy principle.

The first concern is the existence and scaling invariance of $H_{n,GK}$. The scaling invariance we leave to the end of this section. Establishing the existence involves some cute details, exhibited in the proofs of the following three lemmas.

(7.6) Lemma. *For every n and almost every realization of $X_1, X_2, \cdots, X_n$,*

(a) $$0 \leqslant h^2 \| (u^{nh})' \|_2^2 \leqslant \tfrac{1}{2} \ ,$$

(b) $h \| (u^{nh})' \|_2$ *is continuous for* $h > 0$, *and*

(c) $h^2 \| (u^{nh})' \|_2^2 \longrightarrow \frac{1}{2}$ *for* $h \to 0$ *and for* $h \to \infty$.

Proof. Part (b) follows from Exercise (5.2.54), and Part (a) from

$$h^2 \| (u^{nh})' \|_2^2 = 1 - \| u^{nh} \|_2^2 = 1 - \int_{\mathbb{R}} f^{nh}(x)\, dx \ ,$$

combined with the lower bound from the Comparison Lemma (5.2.22). In view of the equality above, to prove Part (c), it suffices to show that

$$\| u^{nh} \|_2^2 \longrightarrow \tfrac{1}{2} \ .$$

This we do in the following three lemmas. Q.e.d.

(7.7) Lemma. *For fixed n and almost all realizations of $X_1, X_2, \cdots, X_n$, for $i = 1, 2, \cdots, n$,*

$$(2\,nh)^{-1/2} \leqslant u^{nh}(X_i) = (2\,nh)^{-1/2}\,(\,1 + o(1)\,)\ , \quad h \to 0.$$

Proof. Let $u_i = u^{nh}(X_i)$, and write (7.3) as

$$u_i = (2\,nh\,u_i)^{-1} + (2\,nh)^{-1} \sum_{j \neq i} [\,u_j\,]^{-1} \exp(-h^{-1}(X_i - X_j)\,)\ , \tag{7.8}$$

so that $u_i \geqslant (2\,nh\,u_i)^{-1}$, and the lower bound follows:

$$u_i \geqslant (2\,nh)^{-1/2}\ , \quad i = 1, 2, \cdots, n\ . \tag{7.9}$$

(This holds for *all* $h > 0$, but never mind.) For the upper bound, consider $\delta = \min_{i \neq j} |\,X_i - X_j\,|$. Then, for fixed n, we have $\delta > 0$ almost surely. Consequently,

$$\varepsilon \stackrel{\text{def}}{=} \max_{i \neq j} (2h)^{-1} \exp(-h^{-1}\,|\,X_i - X_j\,|\,)$$

satisfies

$$\varepsilon = \mathcal{O}(\,(2h)^{-1} \exp(-h^{-1}\,\delta\,)\,) =_{\text{as}} o(\,1\,)\ .$$

So, from (7.8),

$$u_i \leqslant_{\text{as}} (2\,nh\,u_i)^{-1} + \frac{\varepsilon}{n} \sum_{j \neq i} [\,u_j\,]^{-1}\ ,$$

and then with (7.9),

$$u_i \leqslant_{\text{as}} (2\,nh)^{-1/2} + \varepsilon\,(2\,nh)^{1/2}\ ,$$

and the upper bound follows. Q.e.d.

(7.10) Lemma. *For fixed n and almost all realization of $X_1, X_2, \cdots, X_n$, for $i = 1, 2, \cdots, n$,*

$$u^{nh}(X_i) = (2\,h)^{-1/2}\,(\,1 + o(1)\,)\ , \quad h \to \infty.$$

Proof. Observe that for $h \to \infty$,

$$\mathfrak{B}_h(X_i - X_j) = (2\,h)^{-1} \exp(-h^{-1}\,|\,X_i - X_j\,|\,) \stackrel{\text{def}}{=} (2\,h)^{-1}(1 + \varepsilon_{ij}\,)\ .$$

Then, $\varepsilon_{ij} = \mathcal{O}(h^{-1}D)$, with

$$D = \max_{i \neq j} |\,X_i - X_j\,|\ .$$

So, for fixed n, we have $D < \infty$ almost surely. Then, from (7.3),

$$u_i = (2\,nh)^{-1} \sum_{j=1}^{n} [\,u_j\,]^{-1}(\,1 + \varepsilon_{ij}\,)\ ,$$

and so

$$(7.11)\quad (2\,nh)^{-1}\,(\,1-\varepsilon\,) \sum_{j=1}^{n} [\,u_j\,]^{-1} \leqslant u_i \leqslant (2\,nh)^{-1}\,(\,1+\varepsilon\,) \sum_{j=1}^{n} [\,u_j\,]^{-1}\ ,$$

with $\varepsilon = \max_{i\neq j}\,|\,\varepsilon_{ij}\,| = \mathcal{O}(\,h^{-1}\,D)$. Now, take reciprocals to obtain

$$[\,u_i\,]^{-1} \geqslant 2\,h\,(\,1+\varepsilon)^{-1}\,\big(\,\tfrac{1}{n} \sum_{j=1}^{n} [\,u_j\,]^{-1}\,\big)^{-1}\ ,$$

and sum over i. This yields

$$\tfrac{1}{n} \sum_{j=1}^{n} [\,u_j\,]^{-1} \geqslant (2\,h)^{1/2}\,(\,1+\varepsilon\,)^{-1/2}\ .$$

With the lower bound of (7.11), this gives that

$$u_i \geqslant (2\,h)^{-1/2}\,(\,1+\varepsilon\,)^{-1/2}\,(\,1-\varepsilon\,)\ ,$$

for each i. The upper bound follows similarly. Q.e.d.

(7.12) Lemma. *For fixed n, almost all realizations of $X_1, X_2, \cdots, X_n$, and all $h > 0$,*

$$\|\,u^{nh}\,\|_2^2 = n^{-2} \sum_{i,j=1}^{n} [\,u^{nh}(X_i)\,u^{nh}(X_j)\,]^{-1}\,a_{ij}(h)\ ,$$

with $\quad a_{ij}(h) = (4\,h)^{-1}\,(\,1 + h^{-1}\,|X_i - X_j|\,)\,\exp(-h^{-1}\,|X_i - X_j|\,)$.

In particular, with $\delta = \min_{i\neq j}\ |\,X_i - X_j\,| > 0$ almost surely,

$$\begin{aligned}
a_{ij}(h) &= (4\,h)^{-1}\,(\,1+o(1)\,) && ,\ h\to\infty\ , \quad \textit{all}\quad i\neq j\ ,\\
a_{ii}(h) &= (4\,h)^{-1} && ,\ \textit{all } h,\\
a_{ij}(h) &= \mathcal{O}(\,h^{-2}\,\exp(-h^{-1}\,\delta\,)\,) && ,\ \textit{for}\quad h\to 0,\ i\neq j\ .
\end{aligned}$$

Proof. From (7.3), one obtains

$$\|\,u^{nh}\,\|_2^2 = \frac{1}{n^2} \sum_{i,j} \frac{a_{ij}}{u^{nh}(X_i)\,u^{nh}(X_j)}\ ,$$

with

$$(7.13)\qquad a_{ij} = \int_{\mathbb{R}} \mathfrak{B}_h(x)\,\mathfrak{B}_h(x - X_i + X_j)\,dx = \mathfrak{B}_h * \mathfrak{B}_h(X_i - X_j)\ .$$

Now, from the identity $\mathfrak{B}_\lambda = (h/\lambda)^2\,\mathfrak{B}_h + (\,1-(h/\lambda)^2\,)\,\mathfrak{B}_\lambda * \mathfrak{B}_h$, see § 4.6, it follows that

$$\mathfrak{B}_\lambda * \mathfrak{B}_h = \frac{\lambda^2\,\mathfrak{B}_\lambda - h^2\,\mathfrak{B}_h}{\lambda^2 - h^2}\ .$$

So, upon taking the limit as $\lambda \to h$,

$$\mathfrak{B}_h * \mathfrak{B}_h(x) = (2\,h)^{-1}\,\frac{\partial}{\partial h}\,\big[\,h^2\,\mathfrak{B}_h(x)\,\big]\ ,$$

and the lemma follows. Q.e.d.

(7.14) Exercise. Prove Part (c) of Lemma (7.6).

The last concern is whether $H_{n,GK}$ has the correct asymptotic behavior for $n \to \infty$, under suitable conditions. First, we try to give sharp bounds on the minimum value of $h^2 \| \,(u^{nh})'\,\|_2^2$. Let w_h be the solution of the large sample asymptotic version of (7.1), and let $w_o = \sqrt{f_o}$. The triangle inequality implies that

$$h\,\|\,(u^{nh})'\,\|_2 = h\,\|\,w_o'\,\|_2 + \delta_{nh}\ , \tag{7.15}$$

where

$$|\,\delta_{nh}\,| \leqslant h\,\|\,(w_h - w_o)'\,\|_2 + h\,\|\,(u^{nh} - w_h)'\,\|_2\ . \tag{7.16}$$

From § 5.2, we have with $\lambda = h/\sqrt{2}$ that

$$h\,\|\,(u^{nh} - w_h)'\,\|_2 \leqslant c\,\|\,(T_\lambda * dF_n)^{1/2} - (T_\lambda * dF_o)^{1/2}\,\|_2\ , \tag{7.17}$$

and so, if $f_o \in W^{2,1}(\mathbb{R})$ and f_o has a finite exponential moment, then for deterministic or random h, for all $\varepsilon > 0$, there exists a constant c such that

$$|\,\delta_{nh}\,| \leqslant_{\rm as}\ c\,h^2 + c\,h^{-1/2-\varepsilon}\,n^{-1/2+\varepsilon}\ . \tag{7.18}$$

The second term in (7.18) comes from the integration by parts tricks of § 4.3 and the bounds from § 4.4. With (7.15), this gives

$$h\,\|\,(u^{nh})'\,\|_2 \leqslant_{\rm as}\ c\,h + c\,h^{-1/2-\varepsilon} n^{-1/2+\varepsilon}\ ,$$

and so, by taking $h \asymp n^{-1/3}$, for all $\varepsilon > 0$,

$$\min_{h>0}\ h\,\|\,(u^{nh})'\,\|_2 =_{\rm as}\ \mathcal{O}(\,n^{-1/3+\varepsilon}\,)\ . \tag{7.19}$$

For random h, similar arguments give

$$h\,\|\,(u^{nh})'\,\|_2 \geqslant h\,\|\,w_o'\,\|_2 - |\,\delta_{nh}\,| \geqslant_{\rm as}\ h\,\|\,w_o'\,\|_2 - c\,h^{-1/2-\varepsilon}\,n^{-1/2+\varepsilon}\ ,$$

and it follows that for all $\varepsilon > 0$, there exists a constant c such that

$$H_{n,GK}\|\,w_o'\,\|_2 \leqslant_{\rm as}\ c\,n^{-1/3+\varepsilon} + c\,(H_{n,GK})^{-1/2-\varepsilon}\,n^{-1/2+\varepsilon}\ . \tag{7.20}$$

Consequently, for all $\varepsilon > 0$,

$$H_{n,GK} =_{\rm as}\ \mathcal{O}(\,n^{-1/3+\varepsilon}\,)\ , \tag{7.21}$$

and we have proven the following result.

(7.22) LEMMA. *Let $\varepsilon > 0$ be arbitrary. If $f_o \in W^{2,1}(\mathbb{R})$, $\sqrt{f_o} \in W^{1,2}(\mathbb{R})$, and f_o has a finite exponential moment, then $H_{n,GK} =_{\text{as}} \mathcal{O}(\, n^{-1/3+\varepsilon}\,)$.*

So, this choice of the smoothing parameter will not work asymptotically since $H_{n,GK} \ll n^{-1/5}$, the optimal rate for h.

(7.23) EXERCISE. Show that (7.20) implies (7.21).

Are there *good* ways to select h? Nothing seems to be known about this! Of all the methods discussed for kernel estimation, only the discrepancy principle of § 6 is easily adapted to the present circumstances and easily analyzed. Thus, let F^{nh} be the distribution function with subdensity f^{nh}, the GOOD estimator. The discrepancy principle for selecting h may then be formulated as follows.

(7.24) In the *GDP* method, the smoothing parameter is chosen as the smallest solution of

$$\| F_n - F^{nh} \|_\infty = c_{GDP}\, n^{-2/5} \ ,$$

where $c_{GDP} = 0.35$. The h so selected is denoted by $H_{n,GDP}$ and the corresponding GOOD estimator $f^{n,H_{n,GDP}}$ by $f_{n,GDP}$.

Note that $c_{GDP} = c_{DP}$, as in § 6, but the authors have not experimented with other values of c_{GDP}.

We now address the usual concerns: existence, scaling invariance, and asymptotic behavior. The existence of $H_{n,GDP}$ is a consequence of the following lemma, whose proof we leave as an exercise.

(7.25) LEMMA. *For almost all realizations of $X_1, X_2, \cdots, X_n$,*

(a) $\| F_n - F^{nh} \|_\infty$ *is a continuous function of* h ,

(b) $\lim_{h\to 0} \| F_n - F^{nh} \|_\infty = \frac{3}{4\,n}$, *and*

(c) $\lim_{h\to\infty} \| F_n - F^{nh} \|_\infty = \frac{1}{2}$.

(7.26) EXERCISE. Prove Lemma (7.25). [Hint: The asymptotic behavior of the $u^{nh}(X_i)$ for $h \to 0$ and $h \to \infty$ comes in handy here.]

We next consider the asymptotic behavior of $H_{n,GDP}$.

(7.27) THEOREM. *If $\sqrt{f_o} \in W^{1,2,}(\mathbb{R})$ and f_o has a finite moment of order > 2, then*

$$H_{n,GDP} \asymp n^{-1/5} \quad \textit{almost surely} \ .$$

PROOF. In the following, let $\lambda = h/\sqrt{2}$. We start with the decomposition

$$F_n - F^{nh} = \{\, F_n - F_o - \mathfrak{B}_\lambda * (\, F_n - F_o\,)\,\} + F_o - \mathfrak{B}_\lambda * F_o + h^2\, \mathbb{B}_\lambda * |(u^{nh})'|^2 + \{\, \mathfrak{B}_\lambda * F_n - F^{nh} - h^2\, \mathbb{B}_\lambda * |(u^{nh})'|^2\,\}\ ,$$

where $\mathbb{B}_\lambda$ is the distribution corresponding to the density $\mathfrak{B}_\lambda$. The expression in curly brackets on the last line vanishes, cf. (5.2.18). The expression in curly brackets in the first line satisfies

$$\|\, F_n - F_o - \mathfrak{B}_\lambda * (F_n - F_o)\,\|_\infty \leqslant 2\,\|\, F_n - F_o\,\|_\infty =_{\rm as} \mathcal{O}\big(\,(n^{-1}\log\log n\,)^{1/2}\,\big)\ .$$

Finally, if $\sqrt{f_o} \in W^{1,2}(\mathbb{R})$, then

$$\begin{aligned} h^2\, \mathbb{B}_\lambda * |\,(u^{nh})'|^2 &= h^2\, \mathbb{B}_\lambda * |\,(f_o^{1/2})'|^2 + \mathcal{O}(\,h^3\,) + \varepsilon_{nh} \\ &= h^2\, \mathfrak{B}_\lambda * \Phi_o + \mathcal{O}(\,h^3\,) + \varepsilon_{nh}\ , \end{aligned}$$

with

$$\Phi_o(x) = \int_{-\infty}^{x} \big|\,[\,(f_o^{1/2})'\,](y)\,\big|^2\, dy$$

and

$$\varepsilon_{nh} = h^2\, \mathbb{B}_\lambda * \big(|\,(u^{nh} - w_h)'\,|^2\big)\ .$$

Here, w_h is the solution of the large sample asymptotic version of (7.1). With (7.17), we thus have

$$\begin{aligned} \|\, \varepsilon_{nh}\,\|_\infty &\leqslant c\, h^2\, \|\, \mathbb{B}_\lambda\,\|_\infty\, \|\, (u^{nh} - w_h)'\,\|_2^2 \\ &\leqslant c\, \|\, (T_\lambda * dF_n)^{1/2} - (T_\lambda * dF_0)^{1/2}\,\|_2^2 \\ &\leqslant c\, \mathrm{KL}(T_\lambda * dF_n\,,\, T_\lambda * dF_o)\ . \end{aligned}$$

The bounds now follow from § 4.5. The randomness of λ causes no problem, by the monotonicity in λ of $\mathrm{KL}(T_\lambda * dF_n\,,\, T_\lambda * dF_o)$.

Putting everything together gives

$$\Big|\,\|\, F_n - F^{nh}\,\|_\infty - \|\, F_o - \mathfrak{B}_\lambda * dF_o + h^2\, \mathfrak{B}_\lambda * \Phi_o\,\|_\infty\,\Big| \leqslant \|\, F_n - F_o - \mathfrak{B}_\lambda * (F_n - F_o)\,\|_\infty =_{\rm as} \mathcal{O}(\,(n^{-1}\log\log n)^{1/2}\,)\ .$$

It follows that for a suitable constant c,

$$\|\, F_n - F^{nh}\,\|_\infty =_{\rm as} c\, h^2 + \mathcal{O}(h^3\,) + \mathcal{O}(\,(n^{-1}\log\log n)^{1/2}\,) + o(\,(nh)^{-1/2}\,)\ .$$

Finally, the discrepancy principle says that this should equal $c_{GDP}\, n^{-2/5}$, and the conclusion follows. Q.e.d.

The last concern of this section is the scaling invariance of $H_{n,GK}$ and $H_{n,GDP}$. To phrase this properly, we need to exhibit their dependence on the sample $\mathbb{X}_n = (\, X_1, X_2, \cdots, X_n\,)$, and show that they satisfy (1.16) and (1.17). So, let $f^{nh}(x\,;\, \mathbb{X}_n\,)$ denote the GOOD estimator for a fixed value of

h and sample $\mathbb{X}_n$. Likewise, let $H_{n,GK}(\mathbb{X}_n)$ and $H_{n,GDP}(\mathbb{X}_n)$ denote $H_{n,GK}$ and $H_{n,GDP}$ for the given sample. The key to proving scaling invariance of the smoothing parameters selected is to see how scaling affects $f^{nh}(x\,;\,\mathbb{X}_n)$.

(7.28) LEMMA. *For all $h > 0$, $t > 0$ and all $X_1, X_2, \cdots, X_n$,*

$$t\, f^{n,\lambda}(t\,x\,;\; t\,\mathbb{X}_n) = f^{nh}(x\,;\,\mathbb{X}_n)\;, \quad x \in \mathbb{R}\;,$$

where $\lambda = t\,h$.

(7.29) EXERCISE. (a) For any smooth, nonnegative function f and $t > 0$, let f_t be defined by $f_t(x) = t^{-1} f(t^{-1}x)$. Show that

$$t^2 \,\|\,\{\sqrt{f_t}\,\}'\,\|_2^2 = \|\,\{\sqrt{f}\,\}'\,\|_2^2\;.$$

(b) Prove Lemma (7.28).

The interpretation of Lemma(7.28) is that the parameter h in (7.1) acts like the smoothing parameter h in kernel density estimation. This was more or less predicted by the Comparison with kernel density estimation lemma (5.2.22).

(7.30) LEMMA. *For all realizations of $\mathbb{X}_n$,*

(a) $$t^{-1}\, H_{n,GK}(t\,\mathbb{X}_n) = H_{n,GK}(\mathbb{X}_n)\;, \text{ and}$$

(b) $$t^{-1}\, H_{n,GDP}(t\,\mathbb{X}_n) = H_{n,GDP}(\mathbb{X}_n)\;.$$

(7.31) EXERCISE. Prove the lemma.

EXERCISES: (7.14), (7.23), (7.26), (7.29), (7.31).

8. Additional notes and comments

Ad § 3: The idea of adding a penalization term to the double kernel estimator of the L^1 error is similar to the idea of BARRON, BIRGÉ and MASSART (1999) in the model selection context.

Ad § 6: The development here follows EGGERMONT and LARICCIA (1996).

Ad § 8: Had there been a real § 8, it would have dealt with the smoothing parameter selection for the roughness penalization of the log-density. Some useful references are STONE (1990), STONE and KOO (1986), KOOPERBERG and STONE (1991), GU and QIU (1993), and GU (1993).

Ad § 8: Current interest in nonparametric density estimation is in *adaptive* estimation methods. A method is called adaptive if it is asymptotically

optimal over wide classes of densities. In a language similar to (1.4), one would like estimators f_n for which

$$\lim_{n\to\infty} \sup_{f_o\in\mathcal{F}} \frac{\| f_n - f_o \|_1}{\inf_{\varphi_n\in\Phi_n} \| \varphi_n - f_o \|_1} =_{\text{as}} 1 \ , \tag{8.1}$$

where Φ_n is the set of *all* estimators and $\mathcal{F}$ is the (large) class of densities one wishes to estimate. We have been concerned mostly with the class $PDF(C, C')$ of (1.5.1)–(1.5.2). In (8.1), the supremum should perhaps be outside the limit. Either way, shooting for (8.1) is quite ambitious, so typically one is satisfied if the above limit is finite. In the context of density estimation, one might want to estimate the (optimal) order of the kernel, as well as the smoothing parameter, see, e.g., Devroye, Lugosi and Udina (1998). In its own way, Watson and Leadbetter (1963) already investigated adaptive methods, by constructing optimal kernels based on *a priori* information on the unknown density. For more on adaptation in the density estimation and more general contexts, see Barron, Birgé and Massart (1999) and references therein.

http://www.springer.com/978-0-387-95268-0

Maximum Penalized Likelihood Estimation
Volume I: Density Estimation
Eggermont, P.P.B.; LaRiccia, V.N.
2001, XVIII, 512 p., Hardcover
ISBN: 978-0-387-95268-0