Towards Semiparametric Bandwidth Selectors for Kernel Density Estimators

Nils Lid Hjort

University of Oslo

May 1999

ABSTRACT. There is an intense and partly recent literature focussing on the problem of selecting the bandwidth parameter for kernel density estimators. Available methods are largely 'very nonparametric', in the sense of not requiring any knowledge about the underlying density, or 'very parametric', like the normality-based reference rule. This report aims at widening the scope towards the inclusion of many semiparametric bandwidth selectors, via Hermite type expansions around the normal distribution. The resulting bandwidths may be seen as carrying out suitable corrections on the normal reference rule, requiring a low number of extra coefficients to be estimated from data.

The present report introduces and discusses some basic ideas and develops the necessary initial theory, but modestly chooses to stop short of giving precise recommendations for specific procedures among the many possible constructions. This will require some further analysis and some simulation-based exploration. Future work in this direction is planned with esteemed colleagues I. Gijbels and M.C. Jones.

KEY WORDS: *bandwidth selection, exact* MISE, *integrated squared derivatives, Hermite approximations, kernel density estimation, smoothed cross validation*

1. Background, motivation and summary

Suppose that $X_1, \ldots, X_n$ are independent and that it is required to estimate its density f. The classic kernel estimator is of the form $\widehat{f}(x) = n^{-1}\sum_{i=1}^n K_h(X_i - x)$, in which $K_h(u) = h^{-1}K(h^{-1}u)$ is the h-scaled version of K, the kernel function, which here is taken to be a symmetric probability density function. This paper is about the already well-researched problem of selecting a good bandwidth h from data; consult for example Scott (1992), Wand and Jones (1995), Simonoff (1996), Jones, Marron and Sheather (1996) or Chiu (1996) for recent and quite comprehensive overviews.

There is a strong tradition favouring mean integrated squared error as performance criterion for density estimation. The familiar decomposition into variance and squared bias gives

$$\text{MISE}(h) = \mathrm{E}\int \{\widehat{f}(x) - f(x)\}^2 \,\mathrm{d}x = \int \mathrm{Var}\widehat{f}(x)\,\mathrm{d}x + \int \{e_h(x) - f(x)\}^2\,\mathrm{d}x, \tag{1.1}$$

where $e_h(x)$ is the mean of $\widehat{f}(x)$. Accordingly many efforts have been exuded on in some way or another estimating

$$h_n = h_n(f), \quad \text{the minimiser of MISE}(h), \tag{1.2}$$

the best possible bandwidth value in this sense.

Several of the best known approaches work with natural large-sample approximations to bias and variance. These lead for example to

$$\text{MISE}(h) \approx (nh)^{-1}R(K) + \tfrac{1}{4}k_2^2h^4R(f'') - n^{-1}R(f), \tag{1.3}$$

where $k_2 = \int u^2K(u)\,\mathrm{d}u$ while $R(p) = \int p(x)^2\,\mathrm{d}x$ is generic notation for functions p. One of the few methods that go directly for the exact MISE is the so-called unbiased cross validation method that goes back to Rudemo (1982) and Bowman (1984). It consists in minimising

$$\text{UCV}(h) = \int \widehat{f}(x)^2\,\mathrm{d}x - 2n^{-1}\sum_{i=1}^{n}\widehat{f}_{(i)}(X_i), \tag{1.4}$$

where $\widehat{f}_{(i)}$ is the kernel estimator constructed from the data set of size $n-1$ where X_i is deleted; note the dependence of $\widehat{f}$ and $\widehat{f}_{(i)}$ on h. The method derives its name from the fact that $\text{UCV}(h)$ is an unbiased estimate of $\text{MISE}(h)$ minus the constant $R(f)$.

The present paper is also concerned with methods that at least at the outset are strictly non-asymptotic in spirit, for estimating the exact MISE curve and hence its minimiser. Its point of departure is a simple identity for MISE entirely in terms of the density function g for a difference of two data points. This identity, presented in Section 2, is neither complicated nor entirely new, and is at least indirectly present in other literature, see for example Kim, Park and Marron (1994, Section 2) and Wand and Jones (1995, Ch. 2.6). We are able to exploit this identity in several novel ways, however, and in the process also shed some fresh light on some well-established h selection methods, including cross validation and the normal reference rule, as well as smoothed cross validation and on traditional large-sample theory. This is discussed in Section 3.

One conspicuous aspect of the identity is that only the g density for $X_i - X_j$ matters, other aspects of the f object are irrelevant. In particular g is symmetric and often much more well-behaved than the original f, and consequently more amenable to statistical modelling. This viewpoint invites several natural semiparametric h selection rules. These consist in minimising an estimate of the MISE curve constructed by insertion of a suitable semiparametric density estimate for g. This is illustrated in Section 4 for a family of Hermite expansion based models. These h rules are easy to use and can be seen as 'extended rules of thumb' that in various ways correct for non-normal features of g. In practice these rules will typically be less variable than most of the purely nonparametric ones, but will perhaps aim at a slightly sub-optimal h. Such a rule might nevertheless often win over sophisticated nonparametric ones, in that a small bias and smaller variance often beats zero bias and bigger variance. Theory is developed in Section 5, with clarification of the basic bias and variance issues involved in the process. An initial and not yet conclusive discussion is given in Section 6. Finally some extraneous results and comments are offered in Section 7.

In bandwidth selection literature the big space between parametric and nonparametric methods has been left quite un-explored. This paper can be viewed as reporting on initial discoveries and consequent problems related to attempts of opening up this box of possibilities. It is admitted that many of the questions emerging from the box, partly due to the large number of sensible new constructions, need further attention before they can be conclusively answered. In particular, further work and simulation explorations are needed to determine which of the indicated proposals of Section 6 are more fruitful than others.

2. The exact MISE

The aim of this section is to work out a fruitful exact expression for MISE in terms of

$$g(y) = \int f(x) f(x+y)\,\mathrm{d}x, \tag{2.1}$$

the density of a difference $Y_{i,j} = X_j - X_i$ between two different data points. Let similarly $g_K(y) = \int K(u)K(u+y)\,\mathrm{d}u$ for the kernel function K.

PROPOSITION. *The exact* MISE *for the kernel estimator can be expressed as*

$$\begin{aligned} \mathrm{MISE}(h) &= (nh)^{-1} R(K) + \int A_K(v) g(hv)\,\mathrm{d}v + R(f) \\ &= (nh)^{-1} R(K) + \int h^{-1} A_K(h^{-1}y) g(y)\,\mathrm{d}y + R(f), \end{aligned} \tag{2.2}$$

where $A_K(v) = (1 - n^{-1}) g_K(v) - 2K(v)$.

To prove this, note first that the mean of $\widehat{f}(x)$ can be expressed as

$$e_h(x) = \mathrm{E} K_h(X_i - x) = \int K_h(y - x) f(y)\,\mathrm{d}y = \int K(u) f(x + hu)\,\mathrm{d}u.$$

Similarly its variance can be written $(nh)^{-1} a_h(x) - n^{-1} e_h(x)^2$, in which $a_h(x) = \int K(u)^2 f(x + hu)\,\mathrm{d}u$. Next observe that $\int a_h(x)\,dx = \int K(u)^2\,\mathrm{d}u = R(K)$, after interchanging order of integrals, so that the integrated variance becomes equal to $(nh)^{-1}R(K) - n^{-1}\int e_h^2\,\mathrm{d}x$; thus there is nothing asymptotic about the familiar $(nh)^{-1}R(K)$ term that starts out the MISE approximations of type (1.3). Next two similar Fubini operations lead to $\int e_h^2\,\mathrm{d}x = \int g_K(v) g(hv)\,\mathrm{d}v$. Also, $\int e_h f\,\mathrm{d}x$ is seen to be the same as $\int K(v) g(hv)\,\mathrm{d}v$. This gives the required result via (1.1).

We note that the A_K function is almost independent of n, and that the $R(f) = g(0)$ term of course is irrelevant for the minimisation of MISE. In other words, the direct, non-asymptotic function to minimise is

$$\mathrm{DNA}(h) = (nh)^{-1} R(K) + q(h) = \mathrm{MISE}(h) - R(f), \tag{2.3}$$

where

$$
\begin{aligned}
q(h) &= \mathrm{E}\, h^{-1} A_K(h^{-1} Y) \\
&= (1-n^{-1}) \int h^{-1} g_K(h^{-1} y) g(y)\,\mathrm{d}y - 2 \int h^{-1} K(h^{-1} y) g(y)\,\mathrm{d}y,
\end{aligned}
\tag{2.4}
$$

writing Y for a generic difference $X_j - X_i$ between two different data points. Again, note that $q(h)$ is almost independent of n.

We shall take special interest in two popular kernel functions. The first is the the standard normal $K = \phi$, for which g_K is the $N(0,2)$ density. The second is the one that manages to minimise both the pointwise and integrated mean squared error for large samples, namely the Yepanechnikov kernel $K(u) = \frac{3}{2}(1-4u^2)$ for $|u| \le \frac{1}{2}$. For this kernel function somewhat long calculations give

$$
g_K(y) = \begin{cases} (6/5)(1-5y^2+5y^3-y^5) & \text{for } 0 \le y \le 1, \\ (6/5)(1-5y^2-5y^3+y^5) & \text{for } -1 \le y \le 0, \end{cases}
$$

see Byholt and Hjort (1999).

3. Normal reference, UCV and SCV rules and asymptotic approximations revisited

In view of the result above a group of natural h selection procedures emerges: for each well-motivated proposal $\widehat{q}(h)$ for estimating $q(h)$, form

$$
\widehat{\mathrm{DNA}}(h) = (nh)^{-1} R(K) + \widehat{q}(h), \tag{3.1}
$$

and let $\widehat{h}$ be its minimiser. This and the following section consider several appealing special cases. The present section in particular sheds some fresh light on several well-established bandwidth selection strategies, while several new rules emerge naturally in the following sections.

3.1. Normal difference density: the normal reference rule in new light. At one extreme is the hard parametric assumption that Y is normal. Let us write this as $Y \sim N(0, 2\sigma^2)$, in terms of the standard deviation σ for X_i (Y is necessarily symmetric about zero). With $g(u) = (\sqrt{2}\sigma)^{-1}\phi((\sqrt{2}\sigma)^{-1}u)$ function (2.2) can be minimised over h. The result is of the form $h_n = b_n\sigma/n^{1/5}$, where b_n is the well-defined minimiser of the function $\mathrm{MISE}_0(b/n^{1/5})$. Here $\mathrm{MISE}_0(h)$ is as in (2.2), calculated under $g = N(0,2)$, that is, with $\sigma = 1$. This leads to the proposal

$$
\widehat{h}_n = b_n \widehat{\sigma}/n^{1/5}, \tag{3.2}
$$

inserting any reasonable estimator for σ.

With a standard normal kernel one finds from (2.4) that $q(h) = \sigma^{-1} q_0(\sigma^{-1} h)$, where $q_0(\cdot)$ is computed under $\sigma = 1$, giving

$$
q_0(h) = (2\sqrt{\pi})^{-1}\{(1-n^{-1})(1+h^2)^{-1/2} - 2(1+\tfrac{1}{2}h^2)^{-1/2}\}.
$$

The method is as in (3.2) with b_n chosen to minimise

$$\text{MISE}_0(b/n^{1/5}) = \frac{1}{2\sqrt{\pi}}\Big\{\frac{1}{n^{4/5}b} + \Big(1-\frac{1}{n}\Big)\frac{1}{(1+b^2/n^{2/5})^{1/2}} - 2\frac{1}{(1+\frac{1}{2}b^2/n^{2/5})^{1/2}}\Big\}.$$

This is a more precise finite-sample version of the classical $(4/3)^{1/5}\widehat{\sigma}/n^{1/5}$ rule which stems from minimising the approximate MISE as in (1.3) under a normal reference assumption for f, cf. Silverman (1986, Ch. 3.4) and Scott (1992, Ch. 6.2). The table below illustrates that with finite n one should stretch h a little more, to gain a couple of percent in MISE reduction.

For the Yepanechnikov kernel similar but much longer calculations are provided in Byholt and Hjort (1999), with the g_K function found above as one ingredient. Again the result is an h_n of the form $c_n\sigma/n^{1/5}$, where c_n minimises a well-defined function of c for the given n, and the normal reference rule becomes $\widehat{h}_n = c_n\widehat{\sigma}/n^{1/5}$. The simplistic version of this, using (1.3) with a normal reference for f, is $4.6898\,\widehat{\sigma}/n^{1/5}$. A small table of b_n and c_n is given here.

TABLE 1. *Optimal reference rules for the normal and for the Yepanechnikov kernels: these are respectively $b_n\widehat{\sigma}/n^{1/5}$ and $c_n\widehat{\sigma}/n^{1/5}$, where b_n and c_n are given here for some sample sizes.*

3	1.2871	5.2821	14	1.1849	5.0198
4	1.2628	5.2177	15	1.1816	5.0117
5	1.2458	5.1737	16	1.1786	5.0043
6	1.2331	5.1411	17	1.1759	4.9975
7	1.2230	5.1156	18	1.1734	4.9913
8	1.2148	5.0949	19	1.1711	4.9855
9	1.2080	5.0776	20	1.1689	4.9801
10	1.2021	5.0628	50	1.1368	4.8996
11	1.1970	5.0500	100	1.1190	4.8540
12	1.1925	5.0388	1000	1.0842	4.7617
13	1.1885	5.0288	∞	1.0592	4.6898

3.2. Why does the normal reference rule work so well? Note that assuming a normal density for $Y = X_1 - X_2$ is not only implied by but actually also equivalent to having a normal density for X. This is a non-trivial mathematical characterisation theorem for the normal distribution. However, even though the two statements 'f is normal' and 'g is normal' are mathematically equivalent, it is fair to say that g is often much more normal than f. The f to g operation averages and symmetrises at the same time, and the result is smoother and indeed 'more normal'. One way of appreciating closeness to normality is to look at closeness to zero of all cumulants from order three on. Here the odd ones vanish while the even ones decrease; the $2j$th cumulant coefficient for g is equal to $(\frac{1}{2})^{j-1}$ times that of f, for $j \geq 2$.

3.3. Direct nonparametric: the UCV *revisited.* The simplest straightforward nonparametric proposal estimates the theoretical mean by the empirical mean over all differences, giving

$$\widehat{\mathrm{DNA}}(h) = (nh)^{-1}R(K) + \frac{1}{n(n-1)}\sum_{i\neq j} h^{-1}A_K(h^{-1}Y_{i,j}); \tag{3.3}$$

the A_K function is symmetric so we only have to take the mean of the summands with $i < j$ in practice. This is the natural assumption-free unbiased estimate of $\mathrm{DNA}(h)$. But it turns out via some algebraic manipulations that this is exactly the same as the UCV formula (1.4).

One might argue that (3.3) has an even more intuitive motivation than the the UCV in the form (1.4). In particular the unbiasedness property, which with the (1.4) approach takes a little bit of algebra to demonstrate, is very direct here.

3.4. From estimation of g to smoothed cross validation. There is also an interesting link from the present approach, estimating DNA curves, to the so-called smoothed cross validation method, as studied by Müller (1985), Staniswalis (1989), Hall, Marron and Park (1992) and Wand and Jones (1995, Section 3.7), among others. This link will not be discussed further here, but rather in future work with Irène Gijbels and Chris Jones.

3.5. Links to large-sample theory. Here we indicate how the familiar approximations to MISE of the type (1.3) can be derived via our g-based identity (2.2). Some finite-sample corrections to these are also suggested by the following simple calculations. This will also help us later in understanding the behaviour of the Hermite selectors.

The trick is to start with the first formula in (2.2), and then Taylor expanding $g(hv)$ for small h. Hence the $q(h) = \int A_K(v)g(hv)\,dv$ term can be approximated with

$$\begin{aligned} g(0)\{(1-n^{-1})\cdot 1 - 2\} &+ \tfrac{1}{2}h^2 g''(0)\{(1-n^{-1})2k_2 - 2k_2\} \\ &+ \tfrac{1}{24}h^4 g^{(4)}(0)\{(1-n^{-1})(2k_4 + 6k_2^2) - 2k_4\} \\ &+ \tfrac{1}{720}h^6 g^{(6)}(0)\{(1-n^{-1})(2k_6 + 30k_2k_4) - 2k_6\} + \cdots, \end{aligned}$$

using k_j for the jth moment of K. Thus

$$\begin{aligned} q(h) \doteq -(1+n^{-1})g(0) &+ \tfrac{1}{4}k_2^2 h^4 g^{(4)}(0)(1-n^{-1}) \\ &+ \tfrac{1}{24}k_2k_4h^6 g^{(6)}(0)(1-n^{-1}) - k_2h^2n^{-1}g''(0), \end{aligned}$$

omitting terms of lesser importance. Note now that in addition to $g(0) = R(f)$ we have $g''(0) = \int ff'' = -R(f')$, $g^{(4)}(0) = \int ff^{(4)} = R(f'')$, and so on, under sufficient regularity assumptions on the density. In particular, using the (2.2) identity again, we recognise the familiar (1.3) approximation. It also follows from the above that the optimal bandwidth h_n of (1.2) for moderate to large n can be approximated with the explicit minimiser, say $h_{n,a}$, of the asymptotic MISE expression (1.3), in that

$$h_n = h_{n,a} + O(n^{-3/5}) = \{R(K)/k_2^2\}^{1/5}\{g^{(4)}(0)\}^{-1/5}n^{-1/5} + O(n^{-3/5}). \tag{3.4}$$

See for example Wand and Jones (1995, Ch. 3) or Fan and Marron (1992).

Several of our h selection schemes rely on estimating $q(h)$ by plugging in a nonparametric or partly parametric estimate $\widehat{g}(y)$ for $g(y)$. An exact parallel to the calculation above shows that in that case,

$$\widehat{q}(h) = -(1+n^{-1})\widehat{g}(0) + \tfrac{1}{4}k_2^2h^4\widehat{g}^{(4)}(0)(1-n^{-1}) + \tfrac{1}{24}k_2k_4h^6\widehat{g}^{(6)}(0)(1-n^{-1})$$

plus terms that are either smaller in size or independent of h. Several matters become clearer in this light.

Note first that we do not really need an ambitiously complete model for the g density; what matters is reasonable estimation of the quantity $\int A_K(v)g(hv)\,\mathrm{d}v$ for h of size $n^{-1/5}$. The crux, at least for large n, is that the fourth derivative of $\widehat{g}$ of zero should be close to the real $g^{(4)}(0) = R(f'')$ with high probability. Next note that this sometimes can be achieved by using a partly or fully parametric model, say $g(y,\theta)$, where there could be modelling bias but smaller sampling variability in comparison with the fully nonparametric schemes. For large n the modelling bias would be measured by the ratio $\rho = \{g^{(4)}(0,\theta_0)\}^{-1/5}/\{g^{(4)}(0)\}^{-1/5}$, where θ_0 is the limit in probability of $\widehat{\theta}$. The variance of $\widehat{h}_n/h_n$ would be of size $O(n^{-1})$ under traditional parametric assumptions, and this is smaller than for many of the nonparametric h rules. Thus h rules of this type are good whenever the ρ ratio is close to 1 and the number of parameters used is small. For smaller n the more precise quantity to consider would be $\rho_n = h_{n,\theta_0}/h_n$, where h_n is the exact MISE-minimiser and $h_{n,\theta}$ minimises the appropriate $(nh)^{-1}R(K) + q(h,\theta)$.

The following section introduces extended h rules via Hermite extensions. How successful can they be? In view of remarks above this is essentially determined by two issues, at least for large n. The first is closeness of the approximation of the model's $g^{(4)}(0)^{1/5}$, say $R_{2m}^{1/5}$ in an expansion with $2m$ terms, to the real quantity $R(f'')^{1/5}$. The second issue is that of small enough sampling variability of the estimate of this quantity, say $\widehat{R}_{2m}^{1/5}$. These issues are investigated in Section 5.

4. New rules using short Hermite expansions

This section considers a new class of natural selection rules based on the (3.1) minimisation recipe, through models for the symmetric g density that are far less restricted than the normal but also not as wide than the utterly nonparametric. The tools will be those of approximations to densities of the form a normal times short Hermite polynomial expansions, which we review first.

4.1. Approximations using Hermite expansions. Let H_j be the jth Hermite polynomial, defined via $\phi^{(j)}(x) = (-1)^j\phi(x)H_j(x)$; the first few are $H_0 = 1$, $H_1 = x$, $H_2 = x^2 - 1$, $H_3 = x^3 - 3x$, and $H_4 = x^4 - 6x^2 + 3$. These have the property that $\int H_j^2\phi\,\mathrm{d}x = j!$ while $\int H_jH_k\phi\,\mathrm{d}x = 0$ for $j \neq k$.

Now consider in general terms the possibility of approximating a given density $p(x)$ with a standard normal times an additive expansion. For any positive h_H the functions $H_j(h_H^{-1}x)$ are orthogonal with respect to $h_H^{-1}\phi(h_H^{-1}x)$. To form an expansion, minimise

$$\int\Big\{\frac{p(x)}{\phi(x)}-\sum_{j=0}^m a_jH_j(h_H^{-1}x)\Big\}^2 h_H^{-1}\phi(h_H^{-1}x)\,\mathrm{d}x$$
$$=\int\Big\{p(x)-\phi(x)\sum_{j=0}^m a_jH_j(h_H^{-1}x)\Big\}^2\frac{h_H^{-1}\phi(h_H^{-1}x)}{\phi(x)^2}\,\mathrm{d}x$$

with respect to the coefficients $a_0,\ldots,a_m$. This brings forwards the best approximation of order m, for given h_H, namely

$$p_m(x)=\phi(x)\sum_{j=0}^m\frac{\alpha_j}{j!}H_j(h_H^{-1}x)\quad\text{with}\quad\alpha_j=\mathrm{E}_pH_j(h_H^{-1}X)\frac{h_H^{-1}\phi(h_H^{-1}X)}{\phi(X)}.\tag{4.1}$$

In our applications we shall have occasion to consider several possible values of h_H. The standard choice is actually $h_H=1$, leading to the direct Hermite expansion $p_m(x)=\phi(x)\sum_{j=0}^m(\gamma_j/j!)\,H_j(x)$ which uses $\gamma_j=\mathrm{E}H_j(X)$. For a standardised X with mean zero and variance one this gives $\gamma_0=1$, $\gamma_1=\gamma_2=0$, and then the familiar skewness, kurtosis and so on for γ_3, γ_4 and so on. While well-known and possibly popular, also due to its connection to cumulants, it is easy to see that this direct expansion has serious drawbacks; the moments of p must be finite and the natural sample estimates become quite noisy and non-robust. There are also cases in which all coefficients exist but where the expansion simply does not converge; see Fenstad and Hjort (1997). Part of the reason is visible from the arguments above; this expansion corresponds to using weighted L_2 as criterion function, with weight function $1/\phi(x)$, or $\exp(\frac{1}{2}x^2)$, which means extreme weight for rather non-central x values. A more natural choice, in this light, is $h_H=1/\sqrt{2}$, corresponding to canonical, non-weighted L_2 distance. The result is what is called the robust Hermite expansion in Fenstad and Hjort (1997), $p_m(x)=\phi(x)\sum_{j=0}^m(\delta_j/j!)\,H_j(\sqrt{2}x)$ with $\delta_j=\sqrt{2}\mathrm{E}_pH_j(\sqrt{2}X)\exp(-\frac{1}{2}X^2)$.

However, in the situation at hand we are interested in applying this machinery to approximating and then estimating the difference density $g(y)$ with a view towards achieving best precision around zero, cf. the $\int A_K(v)g(hv)\,\mathrm{d}v$ formula for $q(h)$. This wish could be reflected in choosing a weight function for the L_2 criterion, whose form in the setting above is $\exp\{-\frac{1}{2}(1/h_H^2-2)x^2\}$, with most weight around zero. This corresponds to choosing a smaller value for h_H. That this is a good idea will be elaborated on in Section 5. There is a balance to be struck in that too small h_H values leads to higher estimation variability.

4.2. Expressions for MISE and its estimation. Since the g density is symmetric with variance $2\sigma^2$, the theory above naturally lends itself to the approximation of the density for $Y/(\sqrt{2}\sigma)$ around the standard normal. Hence we study the expansion model

$$g_{2m}(y) = \phi\Big(\frac{y}{\sqrt{2}\sigma}\Big)\frac{1}{\sqrt{2}\sigma}\sum_{j=0}^{m}\frac{\alpha_{2j}}{(2j)!}H_{2j}\Big(\frac{h_H^{-1}y}{\sqrt{2}\sigma}\Big). \tag{4.2}$$

The odd moments vanish by symmetry, and

$$\begin{aligned}\alpha_{2j} &= \mathrm{E}\,H_{2j}\Big(\frac{h_H^{-1}Y}{\sqrt{2}\sigma}\Big)\frac{h_H^{-1}\phi(h_H^{-1}Y/(\sqrt{2}\sigma))}{\phi(Y/(\sqrt{2}\sigma))}\\ &= \int h_H^{-1}H_{2j}\Big(\frac{h_H^{-1}y}{\sqrt{2}\sigma}\Big)\exp\Big\{-\tfrac{1}{2}(1-h_H^2)\Big(\frac{h_H^{-1}y}{\sqrt{2}\sigma}\Big)^2\Big\}g(y)\,\mathrm{d}y.\end{aligned} \tag{4.3}$$

One way of estimating the α_{2j}s is presumably by first deriving expressions for these Y-related quantities in terms of related X-related ones. This seems difficult in any generality, however, and it is simpler to bypass the X_is and use the $Y_{i,j}$s directly. We use

$$\widehat{\alpha}_{2j} = \binom{n}{2}^{-1}\sum_{i<l}h_H^{-1}H_{2j}\Big(\frac{h_H^{-1}Y_{i,l}}{\sqrt{2}\sigma}\Big)\exp\Big\{-\tfrac{1}{2}(1-h_H^2)\Big(\frac{h_H^{-1}Y_{i,l}}{\sqrt{2}\sigma}\Big)^2\Big\} \tag{4.4}$$

(noting that all H_{2j} functions are symmetric). One would typically insert an estimate $\widehat{\sigma}$ for σ here, but it is useful to note that the expansion methods as such allow any positive value to be used here.

We now provide a formula for $q(h)$ of (2.4) in the case of a normal kernel. It is helpful to write this as $\sigma^{-1}q_0(\sigma^{-1}h)$ again, where it suffices to find the $q_0(\cdot)$ that comes from putting $\sigma = 1$. One finds $q_0(h)$ of the form $(1-n^{-1})S_0(h) - 2T_0(h)$, say. To find these terms, let $1/a = (1+1/h^2)^{1/2}$ and $1/b = (\frac{1}{2}+1/h^2)^{1/2}$, that is, $a = h/(1+h^2)^{1/2}$ and $b = h/(1+\frac{1}{2}h^2)^{1/2}$. Then

$$\begin{aligned}S_0(h) &= \int\phi\Big(\frac{y}{\sqrt{2}h}\Big)\frac{1}{\sqrt{2}h}\phi\Big(\frac{y}{\sqrt{2}}\Big)\frac{1}{\sqrt{2}}\sum_{j=0}^{m}\frac{\alpha_{2j}}{(2j)!}H_{2j}\Big(\frac{h_H^{-1}y}{\sqrt{2}}\Big)\,\mathrm{d}y\\ &= \frac{1}{2\sqrt{\pi}}\frac{a}{h}\sum_{j=0}^{m}\frac{\alpha_{2j}}{(2j)!}\mathrm{E}_N(h_H^{-1}aZ)\end{aligned}$$

and similarly

$$\begin{aligned}T_0(h) &= \int\phi\Big(\frac{y}{h}\Big)\frac{1}{h}\phi\Big(\frac{y}{\sqrt{2}}\Big)\frac{1}{\sqrt{2}}\sum_{j=0}^{m}\frac{\alpha_{2j}}{(2j)!}H_{2j}\Big(\frac{h_H^{-1}y}{\sqrt{2}}\Big)\,\mathrm{d}y\\ &= \frac{1}{2\sqrt{\pi}}\frac{b}{h}\sum_{j=0}^{m}\frac{\alpha_{2j}}{(2j)!}\mathrm{E}_N H_{2j}(h_H^{-1}bZ/\sqrt{2}),\end{aligned}$$

where the expectations are with respect to a Z being standard normal. These formulae can be worked with further through first detecting and then proving that $\mathrm{E}_N H_{2j}(cZ) = 1\cdot 3\cdots(2j-1)(c^2-1)^j$, which also can be written $\{(2j)!/(2^j j!)\}\,(c^2-1)^j$. Thus one finds

$$S_0(h) = \frac{1}{2\sqrt{\pi}}\frac{1}{(1+h^2)^{1/2}}\Big[\alpha_0 + \sum_{j=1}^m \frac{\alpha_{2j}}{2^j j!}(-1)^j\Big\{\frac{1-h^2(1-h_H^2)/h_H^2}{1+h^2}\Big\}^j\Big],$$

$$T_0(h) = \frac{1}{2\sqrt{\pi}}\frac{1}{(1+\frac12 h^2)^{1/2}}\Big[\alpha_0 + \sum_{j=1}^m \frac{\alpha_{2j}}{2^j j!}(-1)^j\Big\{\frac{1-\frac12 h^2(1-h_H^2)/h_H^2}{1+\frac12 h^2}\Big\}^j\Big].$$

There are simplifications for the Hermite expansion standard choice $h_H = 1$, but we shall typically want to use smaller values. Recall here that the α_{2j}s depend on h_H.

Observe further that normality corresponds to all α_{2j}s being equal to zero except for $\alpha_0 = 1$, in which case we get back the formula used in Section 3.1. One natural proposal is therefore to use bandwidth $\widehat{h}$, depending on both $\widehat{\sigma}$ and the $\widehat{\alpha}_{2j}$s that have been chosen for inclusion, defined as the minimiser of the easily programmed function

$$\begin{aligned}\widehat{\mathrm{DNA}}(h) &= (2\sqrt{\pi})^{-1}(nh)^{-1} + (1-n^{-1})\,\widehat{\sigma}^{-1}\widehat{S}_0(\widehat{\sigma}^{-1}h) - 2\,\widehat{\sigma}^{-1}\widehat{T}_0(\widehat{\sigma}^{-1}h)\\ &= \frac{1}{2\sqrt{\pi}}\Big[\frac{1}{nh} + \frac{1-n^{-1}}{(\widehat{\sigma}^2+h^2)^{1/2}}\sum_{j=0}^m \frac{\widehat{\alpha}_{2j}}{2^j j!}(-1)^j\Big\{\frac{\widehat{\sigma}^2 - h^2(1-h_H^2)/h_H^2}{\widehat{\sigma}^2+h^2}\Big\}^j\\ &\qquad - 2\frac{1}{(\widehat{\sigma}^2+\frac12 h^2)^{1/2}}\sum_{j=0}^m \frac{\widehat{\alpha}_{2j}}{2^j j!}(-1)^j\Big\{\frac{\widehat{\sigma}^2-\frac12 h^2(1-h_H^2)/h_H^2}{\widehat{\sigma}^2+\frac12 h^2}\Big\}^j\Big].\end{aligned} \tag{4.5}$$

Here $\widehat{S}_0(\cdot)$ and $\widehat{T}_0(\cdot)$ are as above but with estimates inserted for the α_{2j}s included.

Remark 1. The h selection scheme laid out here requires the prior selection of two pilot parameters, the expansion order m and the Hermite approximation bandwidth h_H. Some advice on this will be implicit in results of the next section, and a more careful study will take place in future work. The next section also discusses certain modifications that lead to better performance than that achievable using the direct method above.

Remark 2. A sometimes unfortunate side effect of additive expansion procedures like the one developed above is that the density estimator occasionally may become negative. Thus $\widehat{S}_0(\widehat{\sigma}^{-1}h)$ and $\widehat{T}_0(\widehat{\sigma}^{-1}h)$ above are formed employing integrands involving say $\widehat{g}_{2m}(y)$ that once in a while is negative. While this is annoying from a density estimation point of view, like the situation with higher order kernels, it does not disturb our method. If $\widehat{g}_{2m}$ is negative in places a remedy is to lift it up the entire curve the required amount. But this fortunately does not change the minimiser of (4.5) at all. Another view supporting the notion that occasional negativity of the density estimator is not a disturbing matter for the present method is as follows. As discussed in the following section, the precision of a method for finding the best h is essentially determined by the implicit estimate of the fourth g-derivative at zero. Thus the actual level of the estimated density is unimportant, and, specifically, it is immaterial whether the density has been lifted or not.

REMARK 3. There are alternative estimators for α_{2j} that are sometimes favourable, having to do with separate treatment of the diagonal terms. Some initial discussion is given in Section 6.

5. Analysing behaviour of the new h selectors

This section analyses properties of semiparametric bandwidth selection methods that employ the Hermite machinery of the previous section. The results will also be used in Section 6 to suggest various amendments and modifications to the direct method that uses minimisation of (4.5).

5.1. Background on competing methods. Estimating h_n of (1.2) must for large n in one way or another be related to estimating the quantity $R(f'')$, see e.g. (1.3) and (3.4) above. Some of the best estimators in the literature for this crucial parameter use kernel smoothing in various forms. The nature of our own results, to be described below, makes it relevant to describe some features of these other constructions. Hall and Marron (1987) gave estimators with bias $O(\beta^{2r})$ and

$$\text{variance} = \frac{4}{n}\Big\{\int (f^{(4)})^2 f\,\mathrm{d}x - R(f'')^2\Big\} + O(\beta^2 n^{-1} + (n^2\beta^9)^{-1}), \tag{5.1}$$

in terms of a bandwidth parameter β for a kernel L of order $2r$, that is, $k_j(L) = \int u^j L(u)\,\mathrm{d}u$ is zero for $j = 1, \ldots, 2r-1$ while $k_{2r}(L) > 0$. See also Bickel and Ritov (1988). It turns out that the bias term here is always negative, so instead of choosing β to balance squared bias with variance, Jones and Sheather (1991) 'added back bias' using suitable extra 'diagonal terms' in a modified estimator construction. Their estimator, say $\widehat{R}_D(\beta)$, using an ordinary second order kernel, has

$$\mathrm{E}\widehat{R}_D(\beta) = R(f'') - \tfrac{1}{2}\beta^2 k_2(L) R(f''') + L^{(4)}(0)(n\beta)^{-5} + o(h^2)$$

and variance of the same size as given above for the Hall and Marron estimators. This was successfully utilised in Sheather and Jones (1991) to construct an h selector which appears to be one of the very best in existence, cf. extensive simulations in Jones, Marron and Sheather (1996). It works as follows. The leading terms of the bias cancel out when β is a certain $c_0 n^{-1/7}$, say, which they re-express as $c h_{n,a}^{5/7}$ using the AMISE-optimal $h_{n,a}$ given in (3.4). This invites putting $\widehat{\beta}(h) = \widehat{c} h^{5/7}$, say, in the equation governing the final choice of h. Here $\widehat{c}$ involves first stage estimation of both $R(f'')$ and $R(f''')$ but with only lower-level precision required, see their paper for details. The final equation determining their recommended $\widehat{h}_{\mathrm{SJ}}$ is

$$h = \{R(K)/k_2^2\}^{1/5} \widehat{R}_D(\widehat{c} h^{5/7})^{-1/5} n^{-1/5}. \tag{5.2}$$

5.2. Bias for the Hermite approximation rules. As a benchmark we first consider the normal reference distribution, for which we find

$$g^{(4)}(0) = R(N(\mu, \sigma^2)'') = 3\sigma^{-5}/(8\sqrt{\pi}); \tag{5.3}$$

this is what leads to the ordinary form of the normal reference rule. We will see how the Hermite expansion model (4.2) is able to make its way from the normal-based approximation (5.3) to the completely general $R(f'')$, through a combination of choosing m larger and h_H smaller. In other words, the approximation that at each stage involves only a finite and typically small number of parameters is able to bridge the full gap from simplistic parametrics to full nonparametrics.

The Hermite model density can be written $(\sqrt{2}\sigma)^{-1} g_{0,2m}((\sqrt{2}\sigma)^{-1} y)$, where $g_{0,2m}(y)$ is $\phi(y) \sum_{j=0}^{m} \{\alpha_{2j}/(2j)!\}\, H_{2j}(h_H^{-1} y)$. Hence its fourth derivative at zero becomes $(\sqrt{2}\sigma)^{-5}$ $g_{0,2m}^{(4)}(0)$. And the fourth derivative of $g_{0,2m}$ at zero can be written

$$\sum_{j=0}^{m} \frac{\alpha_{2j}}{(2j)!} \{\phi^{(4)}(0) H_{2j}(0) + 6\phi''(0) H_{2j}''(0)/h_H^2 + \phi(0) H_{2j}^{(4)}(0)/h_H^4\}.$$

This can be nicely simplified using various detectable facts about Hermite polynomials. The first fact is that $H_{2j}'' = 2j(2j-1) H_{2j-2}''$. Furthermore, the successive values of $H_{2j}(0)$ are $1, -1, 3, -3\cdot 5, 3\cdot 5\cdot 7, -3\cdot 5\cdot 7\cdot 9$ and so on. This leads after some further simplifications to an expression for the central quantity of the form

$$\begin{aligned} g_{2m}^{(4)}(0) &= \frac{3\sigma^{-5}}{8\sqrt{\pi}} \Big[\alpha_0 + \sum_{j=1}^{m} \frac{\alpha_{2j}}{2^j j!} (-1)^j \{1 + 4j/h_H^2 + (4/3) j(j-1)/h_H^4\}\Big] \\ &= \frac{3}{(\sqrt{2}\sigma)^5} \sum_{j=0}^{m} \frac{\alpha_{2j}}{(2\pi)^{1/2}} \frac{(-1)^j}{2^j j!} \{1 + 4j/h_H^2 + (4/3) j(j-1)/h_H^4\}. \end{aligned} \tag{5.4}$$

This also illustrates one way in which the normal reference rule can be corrected for non-zero coefficients. Note that this method cleverly estimates $R(f'')$ without having to use odd-numbered coefficients.

We are now in position to reach the following result. The proof is relegated to the Appendix.

Bias Proposition. *For the expansion model (4.2) of order $2m \geq 4$ and Hermite bandwidth h_H, let $R_{2m} = g_{2m}^{(4)}(0)$ be the implicit approximation of $R = R(f'')$. Assume that g has finite variance and $2m+2$ derivatives at zero. Then, as h_H gets small, the following holds:*

$$R_{2m} = R(f'') + (-1)^m \frac{1}{(\sqrt{2}\sigma)^5} \frac{G^{(2m+2)}(0)}{2^{m-1}(m-1)!}\, h_H^{2m-2} + o(h_H^{2m-2}),$$

where $G(x)$ is the function $\exp(\frac{1}{2}x^2) g(\sqrt{2}\sigma x)\sqrt{2}\sigma$.

The error term here is $O(h_H^{2m})$ if g has an additional two derivatives at zero. Note the similarity of this result to those available for the $R(f'')$ estimators mentioned in Section 5.1.

To find out more about the size of the modelling bias involved, let us write $g(y) = (\sqrt{2}\sigma)^{-1} g_0((\sqrt{2}\sigma)^{-1}y)$ in terms of g_0, the normalised Y density with variance one. We need successive derivatives of the $G(x)$ function at zero, and to this end note that the $(2i)$th derivative of zero of the $\exp(\frac{1}{2}x^2)$ function can be recognised from Hermite polynomial coefficients, and is $(2i)!/(2^i i!)$. The odd ones disappear. Hence

$$G^{(2j)}(0) = \sum_{i=0}^{j} \binom{2j}{2i} \frac{(2j-2i)!}{2^{j-i}(j-i)!} g^{(2i)}(0)(\sqrt{2}\sigma)^{2i+1} = \sum_{i=0}^{j} \frac{(2j)!}{(2i)!2^{j-i}(j-i)!} g_0^{(2i)}(0).$$

The coefficients here are precisely of the form $w(2j, j-i)$, where these are defined and used in the proof of the bias proposition, see the Appendix, and are those entering Hermite polynomials, but here without sign corrections. Thus the implicit modelling bias involved when using the (4.2) expansion is seen via

$$R_{2m} \doteq R(f'') + \frac{1}{(\sqrt{2}\sigma)^5} \Big\{ \frac{(-1)^m}{2^{m-1}(m-1)!} \sum_{i=0}^{m+1} w(2m+2, 2m+2-2i) g_0^{(2i)}(0) \Big\} h_H^{2m-2}, \quad (5.5)$$

the error term being $o(h_H^{2m-2})$ if g has $2m-2$ derivatives at zero and actually $O(h_H^{2m})$ if it possesses an additional two. For example, the approximate model biases when using respectively coefficients up to order four, up to order six and up to order eight are

$$\begin{aligned}
\mathrm{bias}_4 &\doteq \frac{1}{2} \frac{G^{(6)}(0)}{(\sqrt{2}\sigma)^5} h_H^2 = \frac{1}{2} \frac{1}{(\sqrt{2}\sigma)^5} (15 g_0 + 45 g_0^{(2)} + 15 g_0^{(4)} + g_0^{(6)})\, h_H^2, \\
\mathrm{bias}_6 &\doteq -\frac{1}{8} \frac{G^{(8)}(0)}{(\sqrt{2}\sigma)^5} h_H^4 \\
&= -\frac{1}{8} \frac{1}{(\sqrt{2}\sigma)^5} (105 g_0 + 420 g_0^{(2)} + 210 g_0^{(4)} + 28 g_0^{(6)} + g_0^{(8)})\, h_H^4, \\
\mathrm{bias}_8 &\doteq \frac{1}{48} \frac{G^{(10)}(0)}{(\sqrt{2}\sigma)^5} h_H^6 \\
&= \frac{1}{48} \frac{1}{(\sqrt{2}\sigma)^5} (945 g_0 + 4725 g_0^{(2)} + 3150 g_0^{(4)} + 630 g_0^{(6)} + 45 g_0^{(8)} + g_0^{(10)})\, h_H^6
\end{aligned}$$

(the functions being evaluated at zero), smoothness of g_0 around zero permitting. Thus estimating only four coefficients from the data already has the potential of the small bias of size h_H^6, estimation of five coefficients gives the potentially very small bias of size h_H^8, and so on. It is helpful to note that all these complicated-looking bias term expressions are actually zeroed under normality, also indicating that they will not be too big if g is within a reasonable vicinity of the normal. See also Section 6.

5.3. Relative variability of the Hermite rules. Implicit in our Hermite h selection scheme is an estimator of the fourth g-derivative at zero, which by (5.4) and (4.4) can be expressed as

$$\begin{aligned}\widehat{R}_{2m} &= \frac{3}{(\sqrt{2}\sigma)^5}\sum_{j=0}^{m}\frac{\widehat{\alpha}_{2j}}{(2\pi)^{1/2}}\frac{(-1)^j}{2^j j!}\{1+4j/h_H^2+(4/3)j(j-1)/h_H^4\} \\ &= \frac{3}{(\sqrt{2}\sigma)^5}\binom{n}{2}^{-1}\sum_{i<l}\Big\{\sum_{j=0}^{m}h_H^{-1}p_{2j}\Big(\frac{h_H^{-1}Y_{i,l}}{\sqrt{2}\sigma}\Big)c(j)\Big\},\end{aligned} \tag{5.6}$$

in which $c(j)=(-1)^j\{1+4j/h_H^2+(4/3)j(j-1)/h_H^4\}/(2^j j!)$ and $p_{2j}(u)$ is the function $H_{2j}(u)\phi(u)\exp(\frac{1}{2}h_H^2u^2)$. As noted after (4.4) it is not a necessity to use $\widehat{\sigma}$ estimated from data, the σ can be a pre-determined value. And it is easier to work with a fixed σ when attacking the variance question. The quite lengthy proof of the following result has been sent to the Appendix.

Variance Proposition. *The variance of $\widehat{R}_{2m}$ can be expressed as*

$$\frac{4}{n}\Big\{\int(f^{(4)})^2f\,\mathrm{d}x-R(f'')^2\Big\}+O(n^{-1}h_H^2+(n^2h_H^9)^{-1})$$

when h_H goes to zero slowly enough to let nh_H^9 grow to infinity.

Again this is quite and perhaps even surprisingly similar to results for competing estimators mentioned in Section 5.1. Importantly, the n^{-1} constant here is also the theoretically best possible, in a precise asymptotic sense; this may be shown following results of Fan and Marron (1992).

6. Combating the bias

Some simulation experience with the rules associated with minimisation of (4.5), working with finite normal mixtures of the Marron and Wand (1992) variety, suggested that the selected $\widehat{h}$ values tended to have low variability but that they too often erred on the positive side of the target h_n. Examination revealed that this was caused by underestimation of $R(f'')$ when using $\widehat{R}_{2m}$ in its direct form (5.6). In other words, in the notation of Section 5.2, the $G^{(6)}(0)$ quantity was typically negative, while $G^{(8)}(0)$ was positive, and so on.

One may consider several possible routes towards combating this negative bias. Let us write

$$\mathrm{E}\widehat{R}_{2m}=R(f'')+(-1)^m\{(\sqrt{2}\sigma)^5 2^{m-1}m!\}^{-1}b_{2m}h_H^{2m-2}+o(h_H^{2m}), \tag{6.1}$$

using the bias proposition, with $b_{2m}=G^{(2m+2)}(0)$. Results of Section 5 entail that b_{2m} may be expressed in terms of g's even derivatives at zero, in other words in terms of $R(f),R(f'),\dots,R(f^{(m+1)})$, and an estimator can be constructed for b_{2m} based on this.

It is simpler however to use the expansion methods already worked with above. Some analysis shows that

$$G^{(2m+2)}(0) = (2\pi)^{-1/2}(\alpha_{2m+2} - \tfrac{1}{2}\alpha_{2m+4} + \tfrac{1}{8}\alpha_{2m+6} - \cdots)/\widetilde{h}_H^{2m+2}, \tag{6.2}$$

in terms of a long enough expansion with a new and typically bigger Hermite bandwidth $\widetilde{h}_H$ and associated α_{2j} coefficients. Thus a simple estimate of b_{2m} is $(2\pi)^{-1/2}\widetilde{\alpha}_{2m+2}/\widetilde{h}_H^{2m+2}$, where $\widetilde{\alpha}_{2m+2}$ is as in (4.4) but with new bandwidth $\widetilde{h}_H$.

Another proposal, motivated by the success of the Jones and Sheather method, is to use a 'diagonals-in' sister version of estimator (5.6). That is,

$$\begin{aligned}\widehat{R}_{2m,D} &= \frac{3}{(\sqrt{2}\sigma)^5}\frac{1}{n^2}\sum_{i=1}^n\sum_{l=1}^n\Big\{\sum_{j=0}^m h_H^{-1}p_{2j}\Big(\frac{h_H^{-1}Y_{i,l}}{\sqrt{2}\sigma}\Big)c(j)\Big\}\\ &= (1-n^{-1})\widehat{R}_{2m}\\ &\qquad + \frac{1}{nh_H^5}\frac{3}{(\sqrt{2}\sigma)^5}\frac{1}{(2\pi)^{1/2}}\sum_{j=0}^m\frac{(2j)!}{2^{2j}(j!)^2}\{(4/3)j(j-1)+4jh_H^2+h_H^4\}.\end{aligned}$$

Thus the added quantity is positive, which is good in view of the non-negligible negative bias that bothers $\widehat{R}_{2m}$. The idea is to choose h_H to make the leading terms of the bias approximately cancel each other.

It is clear that there must be a great variety of co-existing natural specialisations of these ideas, and it is not at all clear a priori which will perform better than others in various situations. Here we are content to describe only some representatives that look promising on the basis of parallel experience for kernel based methods.

PROPOSAL 1: In view of the Taylor approximation to $\widehat{q}(h)$ given in Section 3.5, let $\widehat{h}_{\mathrm{I}}$ minimise

$$\begin{aligned}(nh)^{-1}R(K) + \tfrac{1}{4}k_2^2h^4\{\widehat{R}_4(h_H) - (2\widehat{\tau}^5)^{-1}\widehat{b}_4h_H^2\}(1-n^{-1})&\\ + \tfrac{1}{24}k_2k_4h^6\widehat{S}_6(\widetilde{h}_H)(1-n^{-1}),&\end{aligned} \tag{6.3}$$

where $\widehat{S}_6(\widetilde{h}_H)$ is an estimate of $g^{(6)}(0) = -R(f''')$ formed as for $R(f'')$, but using terms of order 0, 2, 4, 6, and using another and bigger pilot bandwidth $\widetilde{h}_H$. Here τ and $\widehat{\tau}$ are used for $\sqrt{2}\sigma$ and $\sqrt{2}\widehat{\sigma}$, for simplicity, and $\widehat{b}_4 = (2\pi)^{-1/2}\widetilde{\alpha}_6/\widetilde{h}_H^6$. A suitable pilot value for $\widetilde{h}_H$ needs to be decided on. One might also try to deduct bias for $\widehat{S}_6(\widetilde{h}_H)$ too, although this seems less crucial. One might drop the h^6 term here, aiming more directly at the approximate $h_{n,a}$ minimiser, but keeping it in is more in the spirit of our proclaimed non-asymptotic view using $q(h)$. One may also consider the option of putting $h_H = \widehat{c}h^{5/7}$ here, with $\widehat{c}$ as given in the course of Proposal 2, and then minimise the resulting expression over h. This would avoid having to select a pilot parameter value for h_H.

PROPOSAL 2: We now present an attempt to follow the chain of arguments and calculations used so successfully by Sheather and Jones, but exploiting our identity of

Section 2 rather than large-sample approximations, and using Hermite expansions rather than kernel estimators as auxiliary machinery. Consider in general terms

$$\widehat{g}_{2m,D}(y,h_H)=\phi\Big(\frac{y}{\tau}\Big)\frac{1}{\tau}\sum_{j=0}^{m}\frac{\widehat{\alpha}_{2j,D}}{(2j)!}H_{2j}\Big(\frac{y}{h_H\tau}\Big),$$

estimated from the (4.2) model, but now using 'diagonals in' versions of coefficient estimates. Then its fourth derivative at zero is precisely $\widehat{R}_{2m,D}$ considered above. A formula for $\widehat{q}_D(h,h_H)=\int A_K(v)\widehat{g}_{2m,D}(hv,h_H)\,dv$ is already available in Section 4, and indeed it makes sense to minimise the accompanying version of equation (4.5), for a suitable pilot value of h_H. It is more in spirit of Sheather and Jones' final recommendation, however, to use a special h_H that depends on h, as follows. Its motivation is that the leading terms of the bias of $\widehat{R}_{4,D}$ are

$$\tfrac{1}{2}\frac{1}{\tau^5}b_4h_H^2+\frac{1}{nh_H^5}\frac{3}{\tau^5}\frac{1}{(2\pi)^{1/2}}(1+5h_H^2+(15/8)h_H^4),$$

and these are made to cancel if $h_H=J_n(b_4)n^{-1/7}$, say, defined as the solution to

$$h_H=\{6/(2\pi)^{1/2}\}^{1/7}(-1/b_4)^{1/7}(1+5h_H^2+(15/8)h_H^4)^{1/7}n^{-1/7}.$$

Here one takes care to use a negative estimate of $b_4=G^{(6)}(0)$. A quick approximation is $h_H=\{6/(2\pi)^{1/2}\}^{1/7}(-1/b_4)^{1/7}n^{-1/7}$. This can be written in the form $h_H=c_nh_{n,a}^{5/7}$. Estimating b_4 as in Proposal 1 gives a suitable $\widehat{c}_n$. Now define $\widehat{h}_{\rm II}$ to be the minimiser of

$$\begin{aligned}\mathrm{DNA}_{\rm II}(h)&=(nh)^{-1}R(K)+\widehat{q}_D(h,\widehat{h}_H(h))\\&=(nh)^{-1}R(K)+\int A_K(v)\widehat{g}_{2m,D}(hv,\widehat{c}h^{5/7})\,dv.\end{aligned}$$

For a normal kernel this is exactly formula (4.5), but with $\widehat{c}h^{5/7}$ replacing h_H there, and with diagonals-in versions

$$\widehat{\alpha}_{2j,D}=(1-n^{-1})\widehat{\alpha}_{2j}+(nh_H)^{-1}(-1)^j(2j)!/(2^jj!).$$

An easy approximation, which might suffice in practice, is

$$\widehat{c}=\Big\{\frac{6/(2\pi)^{1/2})}{R(K)/k_2^2}\Big\}^{1/7}\frac{\widehat{R}(f'')^{1/7}}{(-\widehat{b}_4)^{1/7}}.$$

This proposal needs for its completion a sound and robust pilot estimate of b_4.

Proposal 3: The methods above used very short expansions containing α coefficients up to order $2m=4$. We now try using αs up to order $2m=6$. Then the parallel of Proposal 1 is to minimise

$$\begin{aligned}(nh)^{-1}R(K)&+\tfrac{1}{4}k_2^2h^4\{\widehat{R}_6(h_H)-(8\widehat{\tau}^5)^{-1}\widehat{b}_6h_H^4\}(1-n^{-1})\\&+\tfrac{1}{24}k_2k_4h^6\widehat{S}_8(\widetilde{h}_H)(1-n^{-1}),\end{aligned}$$

where $\widehat{b}_6=(2\pi)^{-1/2}\widetilde{\alpha}_8/\widetilde{h}_H^8$. There would be a parallel to Proposal 2, with some extra work.

Proposal 4: Yet another idea is to plug in the best Hermite estimates one can think of for $R(f'')$ and $R(f''')$, in the approximation formula for h_n given in Fan and Marron (1992).

Remark. It is important to study the large-sample behaviour of the explicit and implicit h selectors here, to understand better how well they perform, and to compare with other competing methods. This will be tended to in future work.

7. Concluding comments

7.1. Finite-sample modification of the Sheather–Jones rule? The Sheather and Jones (1991) rule in effect uses $\widehat{S}(\beta)=\widehat{g}_D^{(4)}(0,\beta)$ to estimate $R(f'')$, where $\widehat{g}_D(y,\beta)$ is a diagonals-in density kernel density estimate with bandwidth β. A 'finite-sample corrected' version of this, in the spirit of other methods developed here, would be as follows. Let $\widehat{\beta}(h)=\widehat{c}h^{5/7}$ where $\widehat{c}$ is constructed in suitable parallel to a similar object in Sheather and Jones (1991). Then let $\widehat{h}$ minimise

$$\mathrm{DNA_{SJ}}(h)=(nh)^{-1}R(K)+\int A_K(v)\widehat{g}(hv,\widehat{\beta}(h))\,\mathrm{d}v.$$

This might be worth pursuing.

7.2. Giving the g estimator a normal start. Whereas the UCV method uses the empirical distribution of $Y_{i,j}$s to estimate $q(h)=\mathrm{E}\,h^{-1}A_K(h^{-1}Y)$, the SCV method, viewed in the light of this paper, uses a smoother density estimate, giving $\widehat{q}(h)=\int h^{-1}A_K(h^{-1}y)\widehat{g}(y)\,\mathrm{d}y$. There are other versions of this that could easily perform better, through exploitation of the prior knowledge that $g(y)$ is often not far from being normal. One density estimator which can take such knowledge into account is the multiplicative method of Hjort and Glad (1995). Here it takes the form

$$\widehat{g}(y)=\phi\Big(\frac{y}{\sqrt{2}\widehat{\sigma}}\Big)\frac{1}{\sqrt{2}\widehat{\sigma}}\frac{1}{n(n-1)}\sum_{i\neq j}L_{\widetilde{h}}(Y_{i,j}-y)\Big/\phi\Big(\frac{Y_{i,j}}{\sqrt{2}\widehat{\sigma}}\Big)\frac{1}{\sqrt{2}\widehat{\sigma}}$$

in terms of a bandwidth parameter $\widetilde{h}$. The intended h selection mechanism is to let $\widehat{h}$ minimise

$$\widehat{\mathrm{DNA}}(h)=(nh)^{-1}R(K)+\widehat{q}(h)=(nh)^{-1}R(K)+(1-n^{-1})\widehat{S}(h)-2\widehat{T}(h),$$

where $\widehat{S}(h)=\int g_K(v)\widehat{g}(hv)\,\mathrm{d}v$ and $\widehat{T}(h)=\int K(v)\widehat{g}(hv)\,\mathrm{d}v$. With a normal kernel for L this can be seen to yield

$$\widehat{S}(h)=\binom{n}{2}^{-1}\sum_{i<j}\frac{1}{(2\pi)^{1/2}}\frac{\widetilde{\sigma}_2}{\sqrt{2}h}\exp\Big\{-\tfrac{1}{2}\Big(\frac{Y_{i,j}}{\widetilde{h}}\Big)^2\Big(1-\frac{h^2\widetilde{\sigma}_2^2}{\widetilde{h}^2}-\frac{\widetilde{h}^2}{2\sigma^2}\Big)\Big\},$$

$$\widehat{T}(h)=\binom{n}{2}^{-1}\sum_{i<j}\frac{1}{(2\pi)^{1/2}}\frac{\widetilde{\sigma}_1}{h}\exp\Big\{-\tfrac{1}{2}\Big(\frac{Y_{i,j}}{\widetilde{h}}\Big)^2\Big(1-\frac{h^2\widetilde{\sigma}_1^2}{\widetilde{h}^2}-\frac{\widetilde{h}^2}{2\sigma^2}\Big)\Big\},$$

in which

$$1/\widetilde{\sigma}_1^2 = 1 + h^2/(2\sigma^2) + h^2/\widetilde{h}^2 \quad \text{and} \quad 1/\widetilde{\sigma}_2^2 = 1/2 + h^2/(2\sigma^2) + h^2/\widetilde{h}^2.$$

Note that $\widetilde{h} \to 0$ gives UCV, that $\sigma \to \infty$ gives SCV, and that even UCV emerges if only $\widetilde{h} \to 0$ for fixed σ.

An alternative estimator of $R(f'') = g^{(4)}(0)$ emerges, as a separate bonus of this approach. It takes the form

$$\widehat{R} = \binom{n}{2}^{-1} \sum_{i<j} \exp\Big(\tfrac{1}{2}\frac{Y_{i,j}^2}{\tau^2}\Big)\Big\{\frac{3}{\tau^4}L\Big(\frac{Y_{i,j}}{\widetilde{h}}\Big)\frac{1}{\widetilde{h}} - \frac{6}{\tau^2}L''\Big(\frac{Y_{i,j}}{\widetilde{h}}\Big)\frac{1}{\widetilde{h}^3} + \frac{1}{\tau}L^{(4)}\Big(\frac{Y_{i,j}}{\widetilde{h}}\Big)\frac{1}{\widetilde{h}^5}\Big\},$$

where $\tau = \sqrt{2}\widehat{\sigma}$. Some additional work shows that also this estimator has variance of the optimal form (5.1), while its bias can be written

$$\mathrm{E}\widehat{R} - R(f'') = \tfrac{1}{2}k_2(L)\widetilde{h}^2(g_{\mathrm{in}}r^{(2)})^{(4)}(0) + \tfrac{1}{24}k_4(L)\widetilde{h}^4(g_{\mathrm{in}}r^{(4)})^{(4)}(0) + o(h^4),$$

where g_{in} is the initial best normal approximation, that is, the $N(0, 2\sigma^2)$ density, and $r = g/g_{\mathrm{in}}$. The bias here is potentially smaller than that of the more traditional kernel estimator that 'starts with nothing' rather than starting with a normal approximation. In particular the bias is close to zero when f is normal. Separate h selectors can be constructed using $\widehat{R}$.

This could be implemented and tried out. A good pilot value for $\tilde{h}$ could be worked out from calculations that to some extent would parallel those of Hjort and Glad (1995).

7.3. A local likelihood approach. The Hermite methods are essentially additive. Here we take a look at methods that are multiplicative in nature. As a local parametric model for $g(t)$, with t in the vicinity of a fixed y, employ a suitable $g(t, \theta)$. The local parameters θ are estimated by maximising

$$\frac{1}{n(n-1)} \sum_{i \neq j} L_b(y_{i,j} - y) \log g(y_{i,j}, \theta) - \int L_b(t - y) g(t, \theta)\,\mathrm{d}t,$$

where b is a bandwidth parameter for the kernel L. This is the local likelihood approach of Hjort and Jones (1996) and Loader (1996), here using a broad enough model for its fourth derivative at zero to be meaningfully expressed. The $Y_{i,j}$ data are partly dependent, but that does not disturb the basic motivation behind the method, which still aims at the best local parametric approximation in a suitable local Kullback–Leibler distance fashion. Here we aim directly at zero, around which g is symmetric, so a natural local model could be $\exp(a + bt^2 + ct^4)$. The procedure would then take the following form: maximise

$$\binom{n}{2}^{-1} \sum_{i<j} L_h(y_{i,j})(a + by_{i,j}^2 + cy_{i,j}^4) - \int L_h(t)\exp(a + bt^2 + ct^4)\,\mathrm{d}t$$

with respect to the three parameters. This gives the estimate

$$R^* = g^{(4)}(0,\widehat{a},\widehat{b},\widehat{c}) = \exp(\widehat{a})(24\widehat{c} + 12\widehat{b}^2).$$

The kernel L should have bounded support here.

7.4. Density estimates of g. The difference density g is instrumental in several of our ingredients. Estimating g also has separate interest, and the following might come in handy on a rainy day. Consider

$$\widehat{g}(y) = \frac{1}{n(n-1)} \sum_{i\neq j} L_h(Y_{i,j} - y) = \binom{n}{2}^{-1} \sum_{i<j} \bar{L}_h(Y_{i,j}, y),$$

in which $\bar{L}_h(y_{i,j}, y) = \frac{1}{2}\{L_h(y_{i,j} - y) + L_h(y_{i,j} + y)\}$. This is one of two canonical kernel density estimators for g. The other one comes from $g(y) = \int f(x)f(x+y)\,\mathrm{d}x$ and uses

$$\begin{aligned}
\widehat{g}_D(y) &= \int \widehat{f}(x)\widehat{f}(x+y)\,\mathrm{d}x \\
&= \frac{1}{n^2}\sum_{i,j}\int K_h(x - x_i)K_h(x + y - x_j)\,\mathrm{d}x \\
&= \frac{1}{n^2}\sum_{i,j}(K_h * K_h)(y_{i,j} - y),
\end{aligned}$$

leaving diagonal contributions in.

Let us analyse the first one. Its mean is

$$\begin{aligned}
e_h(y) &= \mathrm{E}\bar{L}_h(Y_{i,j}, y) \\
&= \tfrac{1}{2}\int \{L_h(z - y) + L_h(z + y)\}g(z)\,\mathrm{d}z \\
&= \tfrac{1}{2}\int L(u)\{g(y + hu) + g(y - hu)\}\,\mathrm{d}u,
\end{aligned}$$

which becomes of the familiar $g + \frac{1}{2}\lambda_2 h^2 g'' + \cdots$ kind. The variance is more difficult. From well-known results on U-statistics, see e.g. Serfling (1980, Ch. 5), we have

$$\mathrm{Var}\,\widehat{g}(y) = \frac{4}{n}\frac{n-2}{n-1}\mathrm{cov}_0(y) + \frac{2}{n(n-1)}\mathrm{Var}_0(y),$$

where $\mathrm{Var}_0(y)$ is the variance of $\bar{L}_h(Y_{1,2}, y)$ and $\mathrm{cov}_0(y)$ is the covariance between $\bar{L}_h(Y_{1,2}, y)$ and $\bar{L}_h(Y_{1,3}, y)$. The variance term can be written

$$\begin{aligned}
&\tfrac{1}{4}\int \{h^{-1}L(h^{-1}(z - y)) + h^{-1}L(h^{-1}(z + y))\}^2 g(z)\,\mathrm{d}z - e_h(y)^2 \\
&\quad= \tfrac{1}{4}h^{-1}\int L(u)^2\{g(y + hu) + g(y - hu)\}\,\mathrm{d}u \\
&\qquad\quad + \tfrac{1}{2}h^{-1}\int L(u)L(u + 2y/h)g(y + hu)\,\mathrm{d}u - e_h(y)^2.
\end{aligned}$$

The covariance term becomes

$$\begin{aligned}
&\tfrac14 h^{-2}\mathrm{cov}\{L(h^{-1}(Y_{1,2}-y))+L(h^{-1}(Y_{1,2}+y)),\\
&\qquad L(h^{-1}(Y_{1,3}-y))+L(h^{-1}(Y_{1,3}+y))\}\\
&=\tfrac14 h^{-2}\int\!\!\int\{L(h^{-1}(z_1-y))+L(h^{-1}(z_2-y))+L(h^{-1}(z_1-y))L(h^{-1}(z_2+y))\\
&+L(h^{-1}(z_1+y))L(h^{-1}(z_2-y))+L(h^{-1}(z_1+y))L(h^{-1}(z_2+y))\}\\
&\qquad \bar g(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2-e_h(y)^2\\
&=\tfrac14\int\!\!\int L(u_1)L(u_2)\{\bar g(y+hu_1,y+hu_2)+\bar g(y+hu_1,-y+hu_2)\\
&\qquad +\bar g(-y+hu_1,y+hu_2)+\bar g(-y+hu_1,-y+hu_2)\}\,\mathrm{d}u_1\,\mathrm{d}u_2-e_h(y)^2,
\end{aligned}$$

in which $\bar g(z_1,z_2)$ is the simultaneous density for two related differences (X_2-X_1,X_3-X_1). It follows that

$$\begin{aligned}
\mathrm{Var}\,\widehat g(y)&=\frac4n\frac{n-2}{n-1}g^*(y)+\frac{R(L)}{n(n-1)h}g(y)\\
&\quad-\frac{4n-6}{n(n-1)}e_h(y)^2+O(h^2/n)+O(n^{-2}h^{-1}g_L(2y/h)),
\end{aligned}$$

where $g^*(y)$ is the symmetrised $(1/4)\{\bar g(y,y)+\bar g(y,-y)+\bar g(-y,y)+\bar g(-y,-y)\}$.

There are a couple of points worth discussing briefly here. The first is that while the traditional U-statistics result says that the covariance term is most important and will dominate the variance term for large n, this does not happen here, due to the presence of the h which approaches zero with growing n. The covariance terms contribution is $O(n^{-1})$, as is that of the subtracted e_h^2 term, and the variance terms contribute markedly, namely $O((n^2h)^{-1})$, but this is still dominated by the $O(n^{-1})$ term if only $nh\to\infty$. Accordingly, g can be estimated with $1/n$ precision, unlike f, which can only be estimated with $1/n^{4/5}$ precision.

7.5. Correcting for longer tails. Suppose Y has somewhat longer tails than predicted by the normal, thus making the normal reference rule less than perfect. One way of correcting for such behaviour of data is via Hermite expansions again, with longtailedness showing up suitably in the coefficients, Another possibility is to model the density as a t density:

$$\frac{Y}{\sqrt2\sigma}\sim\Bigl(\frac{\nu-2}{\nu}\Bigr)^{1/2}t_\nu\sim(\nu-2)^{1/2}N(0,1)/(\chi^2_\nu)^{1/2}.$$

We write it in this form since the variance of $Y/(\sqrt2\sigma)$ must be one. The procedure is to estimate ν from data, thus leading to a suitable $\widehat q(h)$ by insertion in formula (2.4). The bandwidth to use in the end is the one minimising the consequent (3.1). Numerical integration seems necessary, but is not a serious obstacle in practice.

There are a couple of ways of estimating ν from the $Y_{i,j} = X_j - X_i$ data. One is to set the average $\widehat{\lambda}_4$ of all $Y_{i,j}^4/(4\widehat{\sigma}^4)$ equal to the value predicted by the t model, that is, to $3\{1 + 2/(\nu - 4)\}$. If the observed value of $\widehat{\lambda}_4$ is smaller than 3, then go back to normal reference after all. A second possibility is based on first finding z_0, the median of all $|Y_{i,j}|/(\sqrt{2}\widehat{\sigma})$. Thus 50% of the standardised differences are found in $(-z_0, z_0)$. Then determine ν from $G_\nu((\nu/(\nu-2))^{1/2} z_0) = \frac{3}{4}$, where G_ν is the cumulative distribution function for a t_ν.

7.6. Omitting non-significant coefficients. Another idea to explore is to test each hypothesis $H_0\colon \alpha_{2j} = 0$, at suitable significance levels, and only include those with a clearly visible presence.

7.7. Rules using Hermite expansion for f. Our Hermite-based rules are based on expansions for the difference density g, exploiting the fact that other aspects of the density f simply do not matter for the MISE, as seen from the proposition of Section 2. One could also work out expansions for f instead of g, still along the lines of Section 4. Some such was in fact carried out in Hjort and Jones (1995); see also Exercises 9, 20, 21 in the collection of Hjort (1993). The approach of the present paper is more immediately appealing and elegant, however, in that it actively exploits and benefits from the symmetry of g.

7.8. A theoretical question to ponder. We saw that g is more precisely estimated than f. Presumably also the density p_2 of $(X_1 + X_2)/\sqrt{2}$ can be estimated with $1/n$ precision, using a kernel estimate for all observed sums of pairs. And, presumably, the density p_k of $(X_1 + \cdots + X_k)/\sqrt{k}$ becomes more and more smooth and can be estimated better and better, as $k = 3, 4, 5, \ldots$ grows. It would have theoretical interest to pinpoint better this bridge into smoothness and the domain of the central limit theorem.

7.9. A World Cup contest. It would also be interesting to pit the many different estimators of $R_2 = \int (f'')^2 \,\mathrm{d}x$ against each other, for a variety of estimands f. The R_2 quantity is crucial for the smoothness problem, as we have seen, and also has independent interpretation as 'density roughness'.

Appendix: Bias and variance for the Hermite method

The development of Section 5 relied on two important propositions, concerned with the approximate bias and variance of the method's implicit $\widehat{R}_{2m}$ estimator of the fourth derivative of g at zero. The proofs of the propositions are given here.

A.1. Proof of the bias proposition. We must work further with expression (5.4). The intention is to let h_H be at least moderately small, and look for expansions in terms h_H^2, h_H^4 and so on. The starting point is α_{2j} of (4.3), which we rewrite as

$$\alpha_{2j} = \int H_{2j}(u) \exp(-\tfrac{1}{2}u^2) G(h_H u) \,\mathrm{d}u,$$

with G being the smooth and symmetric function given in the proposition; note that normality of g is the same as saying that G is identical to the constant $(2\pi)^{-1/2}$. This may be expanded via Taylor expansion for $G(h_H u)$, for small h_H, provided only that g has the required number of derivatives. We find

$$\frac{\alpha_{2j}}{(2\pi)^{1/2}} = \sum_{i\geq 0} \frac{1}{(2i)!} \lambda_{2j,2i} G^{(2i)}(0)\, h_H^{2i} = \sum_{i\geq j} \frac{1}{(2i)!} \lambda_{2j,2i} G^{(2i)}(0)\, h_H^{2i},$$

writing $\lambda_{2j,2i}$ for $\int H_{2j}(u)\phi(u)u^{2i}\,du$. Note that these vanish for $j > i$ by orthogonality of the Hermite polynomials. Hence the sum appearing in the second expression of (5.4) can be divided into three parts, like Gallia;

$$\sum_{j=0}^{m} \sum_{i\geq j} \frac{\lambda_{2j,2i}}{(2i)!} G^{(2i)}(0) \frac{(-1)^j}{2^j j!} h_H^{2i} + \sum_{j=1}^{m} \sum_{i\geq j} \frac{\lambda_{2j,2i}}{(2i)!} G^{(2i)}(0) \frac{(-1)^j 4j}{2^j j!} h_H^{2i-2}$$
$$+ \sum_{j=2}^{m} \sum_{i\geq j} \frac{\lambda_{2j,2i}}{(2i)!} G^{(2i)}(0) \frac{(-1)^j (4/3) j(j-1)}{2^j j!} h_H^{2i-4}.$$

Rearranging terms suitably gives

$$\begin{aligned}
&\sum_{i\geq 0} \frac{1}{(2i)!} \Big\{ \sum_{0\leq j\leq \min(i,m)} (-1)^j \frac{\lambda_{2j,2i}}{2^j j!} \Big\} G^{(2i)}(0)\, h_H^{2i} \\
&\qquad + \sum_{i\geq 0} \frac{1}{(2i+2)!} \Big\{ \sum_{0\leq j\leq \min(i,m-1)} (-1)^{j-1} 2 \frac{\lambda_{2j+2,2i+2}}{2^j j!} \Big\} G^{(2i+2)}(0)\, h_H^{2i} \\
&\qquad + \sum_{i\geq 0} \frac{1}{(2i+4)!} \Big\{ \sum_{0\leq j\leq \min(i,m-2)} (-1)^j \frac{1}{3} \frac{\lambda_{2j+4,2i+4}}{2^j j!} \Big\} G^{(2i+4)}(0)\, h_H^{2i} \\
&\qquad = A_0 + A_2 h_H^2 + A_4 h_H^4 + A_6 h_H^6 + \cdots,
\end{aligned}$$

say.

All this simplifies to a spectacular degree. The leading constant term is found to be $G(0) - 2G''(0) + (1/3)G^{(4)}(0)$ (assuming that at least $\alpha_0, \alpha_2, \alpha_4$ are included in the expansion), and by some calculations this is seen to be the same as $(1/3)g^{(4)}(0)(\sqrt{2}\sigma)^5$. The pleasant and surprising simplification is that all other terms simply vanish, if only the size of m permits full sums $0 \leq j \leq i$ to be taken. That is,

$$\sum_{0\leq j\leq i} (-1)^j \frac{\lambda_{2j,2i}}{2^j j!} = 0 \quad \text{for } i = 1, 2, \ldots, \tag{A.1}$$

and the same thing happens for the two other sums involved. Proving this involves rather long calculations, where a formula for $\lambda_{2j,2i}$ is necessary. One such exploits the fact that $H_{2j}(x) = \sum_{l=0}^{j} (-1)^l w(2j,l) x^{2j-2l}$, where $w(2j,l) = (2j)!/\{l!(2j-2l)!2^l\}$, see

Fenstad and Hjort (1997); an alternative route starts with re-expressing $\alpha_{2j}/(2\pi)^{1/2}$ as $\int \phi(u)G^{(2j)}(h_H u)\,\mathrm{d}u\, h_H^{2j}$, after $2j$ partial integrations. These facts can be combined with $\mathrm{E}N^{2k} = (2k)!/(k!2^k)$ for standard normal even moments to give

$$\lambda_{2j,2i} = \sum_{l=0}^{j} (-1)^l \frac{(2j)!}{l!(2j-2l)!2^l} \frac{(2j-2l+2i)!}{(j-l+i)!2^{j-l+i}} \quad \text{for } j \le i.$$

This can be seen to lead to (A.1) by a suitable laborious induction proof.

To demonstrate how this leads to the desired conclusion, let us illustrate the induction step for $m = 4$, so that the expansion includes coefficients of order 0, 2, 4, 6, 8, and assume the previous steps $m = 1, 2, 3$ have been taken care of. Then R_{2m} is equal to $R(f'')$ plus certain bias terms of the small order h_H^6 (the h_H^2 and h_H^4 terms have already disappeared by induction hypothesis). These terms are found by examining the general calculations above;

$$\frac{1}{6!} \sum_{0 \le j \le 3} (-1)^j \frac{\lambda_{2j,6}}{2^j j!} G^{(6)}(0)\, h_H^6 + \frac{1}{8!} \sum_{0 \le j \le 3} (-1)^{j-1} 2 \frac{\lambda_{2j+2,8}}{2^j j!} G^{(8)}(0)\, h_H^6$$
$$+ \frac{1}{10!} \sum_{0 \le j \le 2} (-1)^j \frac{1}{3} \frac{\lambda_{2j+4,10}}{2^j j!} G^{(10)}(0)\, h_H^6.$$

The two first sums are zero and the last would also have been had the sum been all the way to $j = 3$. This leads to the conclusion, via $\lambda_{10,10} = 10!$, and ends our demonstration.

A.2. Proof of the variance proposition. By the initial rewriting of $\widehat{R}_{2m}$ in Section 5.3, the problem of finding a useful expression for its variance may be attacked using tools of U-statistics. One has

$$\mathrm{Var}\, \widehat{R}_{2m} = \frac{4}{n} \frac{n-2}{n-1} \mathrm{cov}_{\mathrm{be}} + \frac{2}{n(n-1)} \mathrm{cov}_{\mathrm{wi}}, \tag{A.2}$$

featuring between- and within-covariance terms. We tend to each in turn, and start with the between-term. Here

$$\begin{aligned} \mathrm{cov}_{\mathrm{be}} &= \frac{9}{(\sqrt{2}\sigma)^{10}} \mathrm{cov}\Big\{ \sum_{j=0}^{m} h_H^{-1} c(j) p_{2j}\Big(\frac{h_H^{-1} Y_{1,2}}{\sqrt{2}\sigma}\Big), \sum_{l=0}^{m} h_H^{-1} c(l) p_{2l}\Big(\frac{h_H^{-1} Y_{1,3}}{\sqrt{2}\sigma}\Big) \Big\} \\ &= \frac{9}{\tau^{10}} \sum_{j=0}^{m} \sum_{l=0}^{m} \Big[c(j)c(l) \int\!\!\int H_{2j}(u_1)\phi(u_1) H_{2l}(u_2)\phi(u_2) \\ &\qquad\qquad \bar{G}(h_H u_1, h_H u_2)\, \mathrm{d}u_1\, \mathrm{d}u_2 \Big] - R_{2m}^2, \end{aligned}$$

in which τ is short hand for $\sqrt{2}\sigma$ and

$$\bar{G}(x_1, x_2) = \exp(\tfrac{1}{2}x_1^2 + \tfrac{1}{2}x_2^2)\bar{g}(\tau x_1, \tau x_2)\tau^2,$$

defined in terms of the density

$$\bar{g}(y_1, y_2) = \int f(x + y_1) f(x + y_2) f(x)\,\mathrm{d}x$$

for a related pair of differences, say $(Y_{1,2}, Y_{1,3})$, or $(X_2 - X_1, X_3 - X_1)$. To progress further, employ $2j$ partial integrations in the u_1 direction and $2l$ similar operations in the u_2 direction. The double integral can then be expressed as

$$\mathrm{E}_N \bar{G}_{2j,2l}(h_H U_1, h_H U_2)\, h_H^{2j+2l} = \bar{G}_{2j,2l}(0,0)\, h_H^{2j+2l}\{1 + O(h_H^2)\},$$

in which $\bar{G}_{2j,2l}(x_1, x_2)$ is found by taking $2j$ derivatives in x_1 and $2l$ derivatives in x_2, and where the expectation is with respect to U_1 and U_2 being independent and standard normal.

This can be employed in the double sum above. Somewhat careful analysis is called for since the $c(j)$s themselves have $1/h_H^2$ and $1/h_H^4$ terms, and it becomes an accountant's challenge to keep track of all contributions. One has $c(0) = 1$, $c(1)h_H^2 = -2 - \frac{1}{2}h_H^2$, $c(2)h_H^4 = \frac{1}{3} + h_H^2 + \frac{1}{8}h_H^4$, and we may put this to use in

$$\sum_{k \geq 0} \sum_{i=0}^{k} c(k-i)c(i)\, \mathrm{E}_N \bar{G}_{2k-2i,2i}(h_H U_1, h_H U_2)\, h_H^{2k}.$$

Terms up to order $k = 4$ here, or order h_H^8, must be monitored in order to chase the leading constant. After some analysis the result is of the form $T_0 + \cdots + T_4$ plus $O(h_H^2)$ or smaller terms, where the five T_k terms involve various partial derivatives of the $\bar{G}$ function at $(0,0)$. These in turn must involve the derivatives of the difference pair density $\bar{g}$ at zero, that is, the quantities

$$R_{a,b} = \bar{g}_{a,b}(0,0) = \int f^{(a)} f^{(b)} f\,\mathrm{d}x.$$

One finds

$$\begin{aligned}
T_0 &= \bar{G}_{0,0} = R_{0,0}\tau^2, \\
T_1 &= -2(\bar{G}_{2,0} + \bar{G}_{0,2}) = -4(R_{0,0}\tau^2 + R_{0,2}\tau^4), \\
T_2 &= \frac{1}{3}(\bar{G}_{4,0} + \bar{G}_{0,4}) + 4\bar{G}_{2,2} = \frac{2}{3}(3R_{0,0}\tau^2 + 6R_{2,0}\tau^4 + R_{4,0}\tau^6) \\
&\qquad + 4(R_{0,0}\tau^2 + 2R_{0,2}\tau^4 + R_{2,2}\tau^6), \\
T_3 &= -2(\bar{G}_{4,2} + \bar{G}_{2,4}) = -\frac{4}{3}(3R_{0,0}\tau^2 + 9R_{0,2}\tau^4 + 6R_{2,2}\tau^6 + R_{0,4}\tau^6 + R_{2,4}\tau^8), \\
T_4 &= \frac{1}{9}\bar{G}_{4,4} = \frac{1}{9}(9R_{0,0} + 36R_{0,2}\tau^4 + 6R_{0,4}\tau^6 + 36R_{2,2}\tau^6 + 12R_{2,4}\tau^8 + R_{4,4}\tau^{10}).
\end{aligned}$$

And adding these most terms make their polite excuses and one is left with simply $(1/9)R_{4,4}\tau^{10}$. Thus the remaining leading term of the first part of the (A.2) variance is simply $(4/n)(R_{4,4} - R_{2m}^2)$ plus various h_H^2/n terms.

The second part of the (A.2) variance is easier to deal with. One finds that $\mathrm{cov}_{\mathrm{wi}}$ can be written

$$\frac{9}{\tau^{10}}\frac{1}{h_H}\int\Big\{\sum_{j=0}^{m} c(j)p_{2j}(u)\Big\}^2 g(h_H\tau u)\tau\,\mathrm{d}u - R_{2m}^2,$$

and this is of size $O(1/h_H^9)$ since $c(j)$s are of size $1/h_H^4$ for $j \geq 2$.

Acknowledgements. This work started out during a pleasant visit to the Institute of Statistics at the Université Catholique de Louvain-la-Neuve. My work related to this report has benefitted from and will continue to be influenced by stimulating conversations, in vivo and in electro, with Irène Gijbels and Chris Jones.

References

Bickel, P.J. and Ritov, Y. (1988). Estimating integrated squared density derivatives. *Sankhyā Series A* **50**, 381–393.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–360.

Byholt, M. and Hjort, N.L. (1999). Sometimes nonparametrics beat parametrics, even when the model is right. *American Statistician*, to appear.

Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica* **6**, 129–145.

Fan, J. and Marron, S.J. (1992). Best possible constant for bandwidth selection. *Annals of Statistics* **20**, 2057–2070.

Fenstad, G.U. and Hjort, N.L. (1997). Two Hermite expansion estimators, and a comparison with the kernel estimator. Unpublished manuscript.

Hall, P. and Marron, J.S. (1987). Estimation of integrated squared density derivatives. *Statistics and Probability Letters* **6**, 109–115.

Hall, P., Marron, J.S. and Park, B.U. (1992). Smoothed cross validation. *Probability Theory and Related Fields* **92**, 1–20.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263–269.

Hjort, N.L. (1993). *Density Estimation and Smoothing.* Lecture Notes compendium, Department of Mathematics, University of Oslo.

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Annals of Statistics.*

Hjort, N.L. and Jones, M.C. (1995). Better rules of thumb for choosing bandwidth in density estimation. Unpublished manuscript.

Hjort, N.L. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics* **24**, 1619–1647.

Jones, M.C. (1991). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters* **12**, 51–56.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* **11**, 337–381.

Jones, M.C. and Sheather, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability Letters* **11**, 511–514.

Kim, W.C., Park, B.U. and Marron, J.S. (1994). Asymptotically best bandwidth selectors in kernel density estimation. *Statistics & Probability Letters* **19**, 119–127.

Loader, C. (1996). Local likelihood density estimation. *Annals of Statistics* **24**, 1602–1618.

Marron, S. and Wand, M.P. (1992). Exact mean integrated squared error. *Annals of Statistics* **20**, 712–736.

Müller, H.-G. (1985). Empirical bandwidth choice for nonparametric kernel regressions by means of pilot estimators. *Statistics and Decisions*, supplement **2**, 193–206.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65–78.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society* **B 53**, 683–690.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag, New York.

Staniswalis, J. (1989). Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association* **84**, 284–288.

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimation. *Annals of Statistics* **12**, 1285–1297.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman & Hall, London.