statistics **20** (1989) 3, 357–376 Akademie-Verlag Berlin

Nonparametric Density Estimates with Improved Performance on Given Sets of Densities

Luc Devroye

School of Computer Science, McGill University, Montreal

Summary. We consider the problem of choosing between two density estimates, a nonparametric estimate with the the standard properties of nonparametric estimates (universal consistency, robustness, but not extremely good rate of convergence) and a special estimate designed to perform well on a given set $\boldsymbol{T}$ of densities. The special estimate can often be thought of as a parametric estimate. The selection we propose is based upon the L_1 distance between both estimates. Among other things, we show how one should proceed to insure that the selected estimate matches the special estimate's rate on $\boldsymbol{T}$, and that it matches the nonparametric estimate's rate off $\boldsymbol{T}$.

AMS 1980 *subject classifications:* 62 G 05, 62 H 99, 62 G 20.

Key words: Density estimation, kernel estimate, minimax, theory, consistency, nonparametric estimation, normal density, asymptotic optimality, model selection.

1. Problem statement

There is a strong demand for density estimates that adapt to the situation at hand: they should be of a simple parametric nature if the data fit a given parametric model, and yet they should be flexible enough to handle any density if the parametric model or models fail. In the former case, they should be accurate. In the latter case, the estimates should behave like solid nonparamtric estimates, i.e. they should be consistent and robust (but possibly less accurate).

The data are used to decide between two or more types of estimates. We are faced with a particular model selection problem in which the models are extremely heterogeneous: in one case, a small "target" class of densities, $\boldsymbol{T}$, is envisaged (typically, this class can be described by virtue of a finite number of parameters such as the class of all gamma densities with unknown shape and scale parameters), and in the other case, a huge ocean of densities, typically the complement of $\boldsymbol{T}$, is considered. The small target class, or classes, can be regarded as small islands in this big ocean of densities.

In this note, we study a rather primitive method of selecting one of several density estimates. It is based upon a nonparametric estimate g_n which has the

Research of the author was supported by NSERC Grant A3456 and FCAR Grant EQ-1679.

desired consistency and robustness properties. Roughly speaking, when g_n is close (in the L_1 sense) to one of the estimates on one of the islands, we select that island and its corresponding estimate. If g_n is far away from all islands, the nonparametric estimate is used. Put another way, a halo or sphere of influence is put around all estimates, one for each target class. If g_n falls outside all halos, it is selected. Otherwise, one of the estimates on one of the islands is picked according to some rule. The sizes of the halos can differ from target class to target class.

The advantage of this scheme is that it is computationally simple (no numerical optimization is required), and that its properties are easy to derive in very global settings. We will see below that nearly all the results are valid without restrictions on densities, due in part to our choice of metric. Other universal metrics, such as the Hellinger metrics, could of course be used with equal ease.

The idea of picking a close better-looking estimate is certainly not new. COVER has advocated this as early as 1972, and BARRON (1985) has refined COVER's work. In the method of sieves (GRENANDER, 1981; GEMAN and HWANG, 1982), one picks a density from a growing class of densities so as to maximize the likelihood product. The difference here is that we already have certain estimates, and that we are merely asked to choose between them.

One is tempted to employ the maximum likelihood method for such a selection, possibly with cross-validation or based upon a sample splitting scheme. See e.g. SCHUSTER and YAKOWITZ (1985) or OLKIN and SPIEGELMAN (1987). Unfortunately, the likelihood products are very sensitive to areas of small or zero density, and the sensitivity is enhanced by the fact that we are working with products of density estimates, not just products of densities. As we will see below, the L_1 distance introduces just the right amount of robustness to the selection procedure.

Let us continue our short historical tour. For many parametric target classes, there exist excellent tests for deciding whether the density of the data is in the target class, see e.g. the recent book by D'AGOSTINO and STEPHENS (1986). Upon rejection of the parametric hypothesis, one would then use a nonparametric estimate. While such an approach could be useful for certain small classes, it is not so easy to apply with the kind of generality we are looking for. For example, how would one proceed if the target class consisted of all log-concave densities with mode at zero and modal value equal to one? Furthermore, the same estimate g_n is used both for decision making and estimating. The added homogeneity can only be helpful.

BERAN (1977, 1981) has studied the properties of estimates that are projections of nonparametric estimates onto target classes $\boldsymbol{T}$. (A projection of a density onto $\boldsymbol{T}$ is any density in $\boldsymbol{T}$ that is closest to the given density among all densities in $\boldsymbol{T}$.) These estimates inherit the robustness of the nonparametric estimate, when robustness is considered in the sense described in BICKEL (1976) and YATRACOS (1985). BERAN's approach differs from ours in two respects: he is not

interested in the performance of the density estimate outside $\boldsymbol{T}$ (except possibly in a small neighborhood of $\boldsymbol{T}$ when he studies robustness); and he is not concerned with the selection problem between two given density estimates. Note however that we could study the selection rule which decides between a nonparametric estimate g_n and its projection onto $\boldsymbol{T}$. This often introduces numerical problem for the practitioner. However, it might lead to a useful way of selecting an estimate. The general theorems below are also valid for this case but no worked example for these projection estimates is given here.

The basic technical tool needed to provide a simple analysis is related to the variation of $J_n = \int |g_n - f|$ around its mean, $\mathsf{E}(J_n) = \mathsf{E}\left(\int |g_n - f|\right)$ for all f. As shown in DEVROYE (1988), $\sqrt{n}\,|J_n - \mathsf{E}(J_n)|$ is stochastically bounded from above by a random variable which does not depend upon f, n or the smoothing factor, when g_n is the ordinary kernel estimate (ROSENBLATT, 1956; PARZEN, 1962; CACOULLOS, 1966). Thus the oscillation of g_n in the ocean of densities is controlled in a uniform manner; g_n is virtually anchored. It lives in an L_1 shell centered at f with radius bounded by $c\,\mathsf{E}(J_n)$, roughly speaking. The shell's thickness is about $1/\sqrt{n}$. The shell is so narrow that for intuitive purposes, we can think of $\int |g_n - f|$ as being equal to $\mathsf{E}\left(\int |g_n - f)|\right)$. With the aid of this tool, and a few other results, we will be able to show that in many cases, the expected L_1 distance between the selected estimate f_n and density f tends to zero at the rate of the target class estimate if f is indeed in $\boldsymbol{T}$, and at the rate of the nonparametric estimate otherwise.

The present method is not intended to be used for deciding between two or more nonparametric estimates without specification of a target class for one of them. It is also not suitable for choosing the smoothing factor in an automatic fashion. Nevertheless, we will be able to present a flavor of the usefulness, by illustrating the technique on a couple of simple examples.

Finally, we note that the methods presented here are certainly not limited to the L_1 space. To define balls and distances among densities, we could have used other metrics, such as L_p metrics, HELLINGER metrics, or a KULLBACK-LEIBLER based metric. DEVROYE (1987) explains what the advantages are of the L_1 approach. Perhaps the main reason in the present context is that for all sets (events) A, $\left|\int_A f_n - \int_A f\right| \leqq \frac{1}{2}\int |f - f_n|$, which is an absolute number between zero and one (SCHEFFÉ, 1947). This universal interpretation of the distance is extremely useful in the definition of the radii of the halos; e.g. a radius larger than 2 is nonsensical, and for "practical" sample sizes, radii smaller than 0.0005 may be equally unrealistic. Actually, in the absence of all a priori information, one could often set the radii about equal to the errors one is expected to accept; e.g. it is known that for kernel estimates with positive kernel, errors of 0.01 to 0.03 are close to best possible when $n \leqq 10{,}000$ (DEVROYE and GYÖRFI, 1985).

2. Definitions

Let $X_1, \ldots, X_n$ be iid random vectors with common unknown density f in I Let g_n be a nonparametric estimate, consistent for all f (i.e. $\mathsf{E}(\int|g_n-f|)\to 0$ for f). For example, g_n could be a kernel estimate

$$g_n(x)=\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i),$$

where $h>0$ is a function of the data for which

$$h\to 0, \quad nh^d\to\infty \text{ in probability as } n\to\infty,$$

and K is a kernel, i.e. $\int K=1$ (Parzen, 1962; Rosenblatt, 1956).

Let $\boldsymbol{T}$ be a target class, for which we have a good density estimate t_n at han We will assume throughout that $t_n\in\boldsymbol{T}$. The goodness of this estimate is of cour conditional on f being a member of $\boldsymbol{T}$. The purpose is to choose between g_n an t_n. To do this, we require an explicitly known number q_n, and define the **hal based estimate** as follows:

$$f_n=\begin{cases} t_n & \text{if } \int|t_n-g_n|<q_n \\ g_n & \text{otherwise}\end{cases}.$$

For selection between more than two classes, one can choose between sever estimates in case of overlapping halos. For example, this could be done by selec ing the t_n with the smallest halo.

It should be noted that the evaluation of the L_1 distance between t_n and g requires a numerical integration routine. We assume throughout that this distanc can be evaluated with infinite precision. It should be noted here that for specifi forms of t_n and g_n (e.g. when both are kernel estimates with polynomial kernel of compact support) the integral can be rewritten as a finite easy-to-evaluat sum with $O(n)$ terms.

The choice of q_n is crucial to the analysis. We realize that it is not sufficient t give asymptotics for q_n. Useful techniques require explicit formulas, valid fo all n. As a first general guiding principle, one should take q_n slightly larger tha

$$R_n(\boldsymbol{T})\stackrel{\Delta}{=}\sup_{f\in\boldsymbol{T}}\mathsf{E}(\int|g_n-f|),$$

where the term "slightly" refers to the variability in $\int|g_n-f|$ and $\int|t_n-f|$ uni formly over $f\in\boldsymbol{T}$. Note that q_n should tend to zero for our method to be efficient Since q_n is greater than the minimax error for $\boldsymbol{T}$,

$$\inf_{g_n}\sup_{f\in\boldsymbol{T}}\mathsf{E}(\int|g_n-f|),$$

the method is only applicable when $\boldsymbol{T}$ is not too massive. This excludes many classes, such as the class of all densities with support on [0, 1], bounded by 2, or the class of all unimodal infinitely many times continuously differentiable densities, or the class of all normal scale mixtures, or the class of all densities

within L_1 (or HELLINGER) distance ε from a central density f_0 (see DEVROYE (1983, 1987), BIRGÉ (1985, 1986, 1987)). Luckily, nearly all parametric classes of interest to practising statisticians are included, as well as most L_1 totally bounded classes (YATRACOS, 1985). Examples of such classes include the class of all monotone densities on [0, 1] bounded by a given constant c, the class of all concave densities with mode at the origin, or the class of all densities on [0, 1] with $s-1$ absolutely continuous derivatives, for which $f^{(s)}$ satisfies the following Lipschitz condition:

$$|f^{(s)}(x)-f^{(s)}(y)| \leqq C\,|x-y|^{\alpha},$$

where $\alpha \in (0, 1]$ and $C>0$. For other examples, see BRETAGNOLLE and HUBER (1979).

3. The main theorems

Lemma 1. The basic inequalities. *Let f_n be the halo-based estimate with threshold $q_n > R_n(\boldsymbol{T})$.*

A. *Then*

$$\sup_{f\in\boldsymbol{T}} [\mathsf{E}\,(\int|f_n-f|)-\mathsf{E}\,(\int|t_n-f|)]$$
$$\leqq (1+\int|g_n|) \inf_{u,v:\,0\leqq u,\,0\leqq v,\,u+v\leqq q_n-R_n(\boldsymbol{T})} \{\sup_{f\in\boldsymbol{T}} \mathsf{P}\,(|\int|g_n-f|-\mathsf{E}\,(\int|g_n-f|)|>u)$$
$$+\sup_{f\in\boldsymbol{T}} \mathsf{P}\,(\int|t_n-f|>v)\}\,.$$

The same inequality is true without the suprema over $\boldsymbol{T}$.

B. *In addition, for $f\notin\boldsymbol{T}$, $\inf_{g\in\boldsymbol{T}} \int|g-f| \stackrel{\Delta}{=} L_1(f, \boldsymbol{T})$,*

$$\mathsf{E}\,(\int|f_n-f|) \leqq \mathsf{E}\,(\int|g_n-f|)+(1+\int|t_n|)\,\mathsf{P}\,(\int|g_n-f|>L_1(f, \boldsymbol{T})-q_n)\,.$$

C. *Finally, for all f, we have*

$$\mathsf{E}\,(\int|f_n-f|) \leqq \mathsf{E}\,(\int|g_n-f|)+q_n\,.$$

Proof. Part A is obtained by the triangle inequality: fix any $f\in\boldsymbol{T}$, and any nonnegative u, v with $u+v\leqq q_n-E_n(\boldsymbol{T})$. Then

$$\begin{aligned}\mathsf{E}\,(\int|f_n-f|) &\leqq \mathsf{E}\,(\int|t_n-f|)+(1+\int|g_n|)\,\mathsf{P}\,(\int|g_n-t_n|\geqq q_n)\\ &\leqq \mathsf{E}\,(\int|t_n-f|)+(1+\int|g_n|)\,\mathsf{P}\,(\int|g_n-f|\geqq R_n(\boldsymbol{T})+u)\\ &\quad+(1+\int|g_n|)\,\mathsf{P}\,(\int|t_n-f|\geqq v)\\ &\leqq \mathsf{E}\,(\int|t_n-f|)+(1+\int|g_n|)\,\mathsf{P}\,(\int|g_n-f|-\mathsf{E}\,(\int|g_n-f|\geqq u)\\ &\quad+(1+\int|g_n|)\,\mathsf{P}\,(\int|t_n-f|\geqq v)\,.\end{aligned}$$

Part B can be shown as follows:

$$\begin{aligned}\mathsf{E}\,(\int|f_n-f|) &\leqq \mathsf{E}\,(\int|g_n-f|\,I_{\int|g_n-t_n|\geqq q_n})+\mathsf{E}\,(\int|t_n-f|\,I_{\int|g_n-t_n|<q_n})\\ &\leqq \mathsf{E}\,(\int|g_n-f|)+(1+\int|t_n|)\,\mathsf{P}\,(\int|g_n-t_n|<q_n)\\ &\leqq \mathsf{E}\,(\int|g_n-f|)+(1+\int|t_n|)\,\mathsf{P}\,(\int|g_n-f|>L_1(f, T)-q_n)\,.\end{aligned}$$

Part C is trivially true, since t_n is only picked when it is within distance q_n of f_n; hence $\int|f_n-g_n| \leqq q_n$. ■

Observe that inequality B for $f \notin \boldsymbol{T}$ is not uniform, as nontrivial uniform bounds over the complement of a small class $\boldsymbol{T}$ do not exist. Inequality A, in contrast, is uniform over $\boldsymbol{T}$.

The proof of Lemma 1 also provides us with bounds on the probabilities of error, i.e. the probability of deciding $f \in \boldsymbol{T}$ when $f \notin \boldsymbol{T}$ and vice versa. While these probabilities are important in a hypothesis testing situation ("test whether $f \in \boldsymbol{T}$"), they are of secondary importance in density estimation problems ("try at all costs to make $\int|f_n-f|$ as small as possible, given the information at hand").

Each of the inequalities in Lemma 1 has its particular use for us. Inequality A can be used to show that the halo-based estimate inherits the minimax properties from g_n in many cases. Inequality B describes the behavior of the halo-based estimate when $f \notin \boldsymbol{T}$, and provides us with sharp bounds for the probability of (erroneously) picking t_n over g_n. Inequality C is rather naive but universally applicable. It can be used to derive the consistency of f_n. Each of these inequalities is now illustrated, starting with inequality C.

Theorem 1. Consistency. *If g_n is consistent at f (i.e. $\mathsf{E}\left(\int|g_n-f|\right) \to 0$ as $n \to \infty$), and $q_n \to 0$, then f_n is consistent at f.*

Theorem 1 implies that if the kernel estimate g_n is used, with smoothing parameter $h \to 0$, $nh^d \to \infty$ as $n \to \infty$, and $q_n \to 0$, f_n is consistent for all f (DEVROYE, 1983). However, carelessly putting $q_n = 1/\sqrt{n}$ (for example) can have a detrimental effect on the rate of convergence when $f \in \boldsymbol{T}$. Hence the need for a deeper study regarding the rate with which we should let q_n tend to zero. In first instance, this can be done via the concept of asymptotic optimality introduced below.

We say that f_n is **asymptotically optimal** for a class $\boldsymbol{G}$ of densities f containing $\boldsymbol{T}$ (i.e. $\boldsymbol{T} \subseteqq \boldsymbol{G}$), when for all $f \notin \boldsymbol{T}$, $f \in \boldsymbol{G}$,

$$\mathsf{E}\left(\int|f_n-f|\right) \sim \mathsf{E}\left(\int|g_n-f|\right),$$

and for all $f \in \boldsymbol{T}$,

$$\mathsf{E}\left(\int|f_n-f|\right) \sim \mathsf{E}\left(\int|t_n-f|\right).$$

In many cases, we will not only establish the asymptotic optimality of f_n, but also provide inequalities about how close the expressions in the definition are to each other.

Theorem 2. Asymptotic optimality. *f_n is asymptotically optimal for the class of **all** densities, when g_n is the kernel estimate with deterministic h, and each of these conditions is satisfied:*

(i) *the complement of $\boldsymbol{T}$ is an open set;*

(ii) $R_n(\boldsymbol{T}) \leqq q_n \to 0$;

(iii) $\inf_{u,v:0\leq u,0\leq v,u+v\leq q_n-R_n(\boldsymbol{T})} \{2\exp(-nu^2/(32\int^2|K|))+\mathsf{P}(\int|t_n-f|>v)\}$
$=o(\mathsf{E}(\int|t_n-f|))$ *for all* $f\in\boldsymbol{T}$.

(iv) $\int|t_n|$ *is uniformly bounded in* n.

The conditions of Theorem 2. We do not explicitly require that $h\to 0$ and $nh^d\to\infty$. However, for (ii) to hold, it is necessary that these conditions are met, at least for nearly all kernels K, the exceptions being kernels whose FOURIER transform is one in an open neighborhood of the origin (DEVROYE, 1987). Condition (ii) in effect tells us that $\boldsymbol{T}$ can't be large, especially since we require h to be deterministic. For data-based h, as we will see further on in some examples, condition (ii) is much less restrictive, since h can adapt itself more easily to the underlying densities. As it stands, (ii) states that $\boldsymbol{T}$ can only contain densities from a certain small neighborhood of a fixed density.

Conditions (i) and (iv) are usually trivially satisfied. For example, (iv) always holds when t_n is a bona fide density. The technical condition (iii) governs how much bigger we can take q_n than $R_n(\boldsymbol{T})$. The difference can be split up into $u+v$ at will. However, to better understand what is going on, we can assume that $u=v=(q_n-R_n(\boldsymbol{T}))/2$. It is known that for most small classes $\boldsymbol{T}$, $\mathsf{E}\int|t_n-f|\geq c/\sqrt{n}$ (but this is by no means a universal rule!). In such cases, the u-part of the condition holds when u tends to zero slower than $1/\sqrt{n}$. This condition in fact ensures that the kernel estimate is relatively stable. The v-part of (iii) is satisfied if $\mathsf{P}(\int|t_n-f|>v)=o(1/\sqrt{n})$, a condition that is often easy to check. This condition too insures that the "oscillations" of $\int|t_n-f|$ are negligible relative to the crucial difference $q_n-R_n(\boldsymbol{T})$. ■

Proof of Theorem 2. We recall three properties of the kernel estimate with deterministic $h=h_n$. First,

$$\inf_{f,h,K}\mathsf{E}(\int|g_n-f|)\geq\sqrt{\frac{1}{528n}}$$

(DEVROYE, 1986). Secondly, $\mathsf{P}(\int|g_n-f|>\varepsilon)\leq e^{-n\varepsilon^2/3}$ for all $\varepsilon>0$ and all n large enough (DEVROYE, 1983). Thirdly, again for all f and all $h>0$,

$$\mathsf{P}(|\int|g_n-f|-\mathsf{E}(\int|g_n-f|)|>u)\leq 2e^{-\frac{nu^2}{32\int|K|}}$$

(DEVROYE, 1988).

Consider first $f\notin\boldsymbol{T}$. Since the complement of $\boldsymbol{T}$ is open, we have $L_1(f,\boldsymbol{T})>0$. This, together with part B of Lemma 1, and (iv), shows that $\mathsf{E}(\int|f_n-f|)\sim$ $\sim\mathsf{E}(\int|g_n-f|)$.

Consider next $f\in\boldsymbol{T}$. By inequality A of Lemma 1, and the fact that $\int|g_n|=\int|K|$, it suffices to show that

$$\inf_{u,v:0\leq u,0\leq v,u+v\leq q_n-R_n(\boldsymbol{T})}\{\mathsf{P}(|\int|g_n-f|-\mathsf{E}(\int|g_n-f|)|>u)+\mathsf{P}(\int|t_n-f|>v)\}$$
$$=o(\mathsf{E}(\int|t_n-f|))$$

for all $f \in \boldsymbol{T}$. The first term on the left-hand-side, for fixed u, does not exceed $2 \exp(-nu^2/(32 \int^2 |K|))$ for all n and u. ■

Let us finally turn to the problem of preserving minimax optimality. Assume that an estimate t_n is **minimax-optimal** for a given target class $\boldsymbol{T}$, i.e. there exists a constant C such that

$$\sup_{f \in \boldsymbol{T}} \mathsf{E}(\int |t_n - f|) \leqq C \inf_{f_n} \sup_{f \in \boldsymbol{T}} \mathsf{E}(\int |f_n - f|) .$$

For $f \notin \boldsymbol{T}$, t_n is often inconsistent or poor. It usually is not as reliable as the universal estimate g_n. If we apply our halo-based selection rule, then we would like to inherit the minimax optimality, at the very least. With little work we are able to take a minimax-optimal estimate t_n (for $\boldsymbol{T}$), possibly non-consistent,, outside $\boldsymbol{T}$, and obtain another minimax-optimal estimate f_n, which is guaranteed to converge for all f. All that is needed is a simple nonparametric estimate with good uniformly bounded mean error over $\boldsymbol{T}$ (not necessarily minimax-optimal for $\boldsymbol{T}$!), and with uniformly bounded variation of the error about its mean.

Theorem 3. Preserving minimaxity. *Assume that* $q_n \geqq R_n(\boldsymbol{T})$ *is such that*

$$\inf_{u,v: 0 \leqq u, 0 \leqq v, u+v \leqq q_n - R_n(\boldsymbol{T})} \{\sup_{f \in \boldsymbol{T}} \mathsf{P}(|\int |g_n - f| - \mathsf{E}(\int |g_n - f|)| > u) + \sup_{f \in \boldsymbol{T}} \mathsf{P}(\int |t_n - f| > v)\} = o(\sup_{f \in \boldsymbol{T}} \mathsf{E}(\int |t_n - f|))$$

and that $\int |g_n|$ *is uniformly bounded in* n. *If* t_n *is minimax optimal for* $\boldsymbol{T}$, *then so is* f_n:

$$\sup_{f \in \boldsymbol{T}} \mathsf{E}(\int |f_n - f|) \leqq (1 + o(1)) \sup_{f \in \boldsymbol{T}} \mathsf{E}(\int |t_n - f|) .$$

Proof. The proof is immediate from part A of Lemma 1. ■

Note that the uniformity of the variation of the L_1 error for g_n is essential. If minimax optimality is our only concern, then we could take $q_n = \infty$ (which would imply that $f_n \equiv t_n$). Unfortunately, for reasons of consistency (Theorem 1) and asymptotic optimality (Theorem 2), it is necessary to take $q_n \to 0$. It is perhaps helpful to verify when the conditions of Theorem 3 are satisfied. This is the case if both g_n and t_n are kernel estimates with deterministic smoothing factors and absolutely integrable kernels. To see this, use an inequality used in the proof of Theorem 2 and the fact that

$$\inf_{f,h,K} \mathsf{E}(\int |g_n - f|) \geqq \sqrt{\frac{1}{528n}}$$

(Devroye, 1986). The halo-based estimate can thus be used to make a given estimate t_n more useful (i.e., robust, universally consistent); and we won't have to give up any of the nice properties of t_n on $\boldsymbol{T}$. One might for example consider a minimax-optimal monotone estimate t_n of a monotone density on $[0, 1]$ (which by definition, can't possibly be consistent for non-monotone f). Such estimates were obtained recently by Birgé (1987).

4. Example 1: Superperformance for a single density

It is well-known that in ordinary estimation problems for location or scale, one can modify existing estimates so that they become extremely good when the unknown parameter takes one particular value. The modification usually involves replacing the original estimate by the particular value if the difference between them is smaller than some threshold (which in turn tends to zero with n at some controlled rate). This can be done too in density estimation using the L_1 halo discussed in this paper. Consider as our target class the class $\boldsymbol{T}$ consisting of one density, f^*. Then, since formally $t_n \equiv f^*$, we have

$$f_n = \begin{cases} f^* & \text{if } \int |f^* - g_n| < q_n \\ g_n & \text{otherwise} \end{cases}.$$

Here g_n is for example the kernel estimate with K and h picked for satisfactory overall performance (assume that K is absolutely integrable and that $h \to 0$ and $nh^d \to \infty$). Note that asymptotic optimality cannot be hoped for here since t_n commits zero error on f^*, and a fixed positive error elsewhere. However, we have the following:

Theorem 4. *If $q_n \to 0$, then f_n is consistent for all f.*
If $f \neq f^*$ and $q_n \to 0$, then

$$\mathsf{E}\left(\int |f_n - f|\right) \leqq \mathsf{E}\left(\int |g_n - f|\right) + O(e^{-cn})$$

for some constant $c > 0$. Hence, $\mathsf{E}(\int |f_n - f|) \to \mathsf{E}\left(\int |g_n - f|\right)$ for ***all*** *$f \neq f^*$.*
Finally, if $f \equiv f^*$, and $q_n \geqq \mathsf{E}\left(\int |g_n - f^*|\right) = R_n(\boldsymbol{T})$,

$$\mathsf{E}\left(\int |f_n - f^*|\right) \leqq 2\,(1 + \int |K|)\, e^{-\frac{n(q_n - R_n(\boldsymbol{T}))^2}{32 \int^2 |K|}}.$$

Proof. The first statement is an immediate corollary of Theorem 1. The second statement follows from part B of Lemma 1, the exponential inequality of DEVROYE (1983) used in the proof of Theorem 2, and the fact that $\mathsf{E}\left(\int |g_n - f|\right) \geqq \geqq 1/\sqrt{528n}$ for all f, h, K and n (DEVROYE, 1986). The third statement is immediate from part A of Lemma 1, the fact that $t_n \equiv f^*$ and $\boldsymbol{T} = \{f^*\}$, and an inequality of DEVROYE (1988). ■

The compromises ahead of us are clear from Theorem 4. When $f \equiv f^*$, we would like to make q_n as large as possible, preferably infinite (see last part of the Theorem). On the other hand, when $f \neq f^*$, the second statement of the Theorem shows that a small q_n is called for. If $\mathsf{E}\left(\int |g_n - f^*|\right)$ is known (which seems plausible, since we know f^*), then taking q_n equal to this value plus $\sqrt{32 \int^2 |K| \log n / n}$ insures that when $f \equiv f^*$,

$$\mathsf{E}\left(\int |f_n - f^*|\right) \leqq \frac{2 + 2 \int |K|}{n}.$$

We can control the rate with which $\mathsf{E}\left(\int |f_n - f^*|\right)$ tends to zero by adjusting q_n.

Rates all the way down to (but not including) e^{-cn} are achievable, thus outperforming g_n on this particular density in a dramatic fashion.

On the other hand, the threshold suggested above is small enough so as not to upset the performance when $f \neq f^*$. Indeed, by part C of Lemma 1,

$$\mathsf{E}\left(\int |f_n - f|\right) \leqq \mathsf{E}\left(\int |g_n - f|\right) + \mathsf{E}\left(\int |g_n - f^*|\right) + \sqrt{\frac{32\, f^2 |K| \log n}{n}}\,.$$

This inequality is sometimes satisfactory for medium-sized n. Note the presence of the term involving f^* on the right-hand-side. In contrast, this term is missing in the exponential inequality of Theorem 4. The only disadvantage of the latter inequality is that the exponentially decreasing term is partially hidden from view due to the "big oh" format.

For small target classes, the only thing that changes is the rate of convergence of f_n when $f \in \boldsymbol{T}$. For target classes with infinitely many densities, it is rarely possible to achieve the exponential power rates of Theorem 4. However, it is usually true that the probability of not picking the target class estimate when $f \in \boldsymbol{T}$, or of not picking the nonparametric estimate when $f \notin \boldsymbol{T}$, tends to zero at an exponential power rate for some appropriate choice of q_n. This will be illustrated on a modest example in the next section.

5. A case study: The normal density

Assume in this section that $\boldsymbol{T}$ is the class of normal densities on the real line. Let us take a few paragraphs to discuss and analyze normal density estimates. Kolmogorov, and later Basu (1964), have shown that for the normal family with unknown mean μ and variance σ^2, the following density is an unbiased estimate at all x:

$$t_n(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi\,(n-1)}\,\hat{\sigma}} \left(1 - \frac{1}{(n-1)\,\hat{\sigma}^2}\,(x-\hat{\mu})^2\right)_+^{\frac{n-4}{2}}.$$

Here $\hat{\mu}$ and $\hat{\sigma}$ are the standard sample-based estimates of μ and σ. For other examples and more theoretical background on unbiased estimation, see Lumelskii and Sapozhnikov (1969), Wertz (1975), Guttmann and Wertz (1976) and Seheult and Quesenberry (1971), and the references found there. Another possible (but not unbiased) estimate is

$$t_n(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}}\, e^{-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}},$$

where $\hat{\sigma}$ and $\hat{\mu}$ are as above. It can be shown that for both estimates, $\sup_{f \in \boldsymbol{T}} \mathsf{E}\left(\int |t_n - f|\right) = O\left(1/\sqrt{n}\right)$. The second estimate itself is a member of $\boldsymbol{T}$, a feature that greatly facilitates the ensuing analysis. Furthermore, it has not been established

to date that the unbiased estimate dominates the second estimate in the expected L_1 sense. For this reason, we will consider as our parametric estimate the second normal estimate.

In our examples we need a suitable density for the global selection process. Prime candidates include the kernel estimate with data-based smoothing factor and nonnegative kernel K, or the kernel estimate with data-based smoothing factor and flattop kernel K (DEVROYE, 1987; a flattop kernel is a symmetric function, integrating to one, whose FOURIER transform is constant in an open neighborhood of the origin). There are many excellent schemes for choosing h as a function of the data, but to limit the analysis somewhat we will merely be concerned with an old-fashioned but common sense scale-invariant estimate, discussed e.g. in DEHEUVELS (1977): here K is a symmetric unimodal density on the real line, and h is defined by

$$h = cn^{-\frac{1}{5}} \hat{\sigma}$$

where $\hat{\sigma}^2$ is the sample-based variance,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 ,$$

and

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

The constant c is adapted to the density f we are estimating. It depends upon the shape of f only. A priori information about the shape of f should be used in the choice of c. Interestingly, its value is rather insensitive with respect to f. The value 1.2019409 ... for the normal density can be employed without too much loss for many bell-shaped curves. It should be stressed that by picking K nonnegative, we are limiting ourselves when f is very smooth. On the other hand, for oscillatory densities (such as densities with discontinuities), picking $h \approx n^{-1/5}$ is suboptimal, as a larger value is called for. Also, it is generally recommended to avoid averages when computing scale factors such as $\hat{\sigma}$. Instead, one should use robust quantile-based estimates. The technical reader will have no trouble adapting the results that follow to his particular situation.

Theorem 5. The normal class.

A. *Let f be a density for which $\hat{\sigma} n^{-1/5} \to 0$ and $\hat{\sigma} n^{4/5} \to \infty$ in probability as $n \to \infty$ (it suffices, for example, that f has finite second moment). If $q_n \to 0$, then f_n is consistent.*

B. *Let $\boldsymbol{G}$ be the class of all densities f for which $\mathsf{E}(|\hat{\sigma} - \sigma|) = o(n^{-2/5})$ (this class includes $\boldsymbol{T}$ and all densities for which $\int |x|^{(10+\varepsilon)/3} f(x)\, dx < \infty$ for some $\varepsilon > 0$). If $q_n - R_n(\boldsymbol{T}) \geqq c^* \sqrt{\log n/n}$ for some constant $c^* > 20$ (it suffices for example to require that $q_n \geqq Cn^{-2/5} + c^* \sqrt{\log n/n}$ where C, c and K are as in Lemma 5), and $q_n \to 0$, then f_n is asymptotically optimal on $\boldsymbol{G}$.*

C. *t_n is minimax-optimal for $\boldsymbol{T}$. So is f_n when $q_n - R_n(\boldsymbol{T}) \geqq c^* \sqrt{\log n/n}$ for some constant $c^* > 20$.*

D. *When $q_n \geqq R_n(\boldsymbol{T}) + (\sqrt{a} + \sqrt{b})\ \sqrt{\log n/n}$ for constants a, b, then, for $f \in \boldsymbol{T}$,*

$$\mathsf{E}\left(\int |f_n - f|\right) \leqq \mathsf{E}\left(\int |t_n - f|\right) + 3n^{-\frac{a}{32}} + 5n^{-\frac{b}{512}}$$

provided that $n \geqq 6$, and that $8\ (6\sqrt{2} + \sqrt{2/3})/\sqrt{n-5} \leqq \sqrt{b \log n/n} \leqq 4\ (\sqrt{8} - 2)$.

The proof of Theorem 5 is given in the appendix. Basically, Theorem 5 is obtained by a straightforward application of the previous theorems, but is complicated by the fact that g_n is not a kernel estimate with deterministic h (for which we have useful universal inequalities: see the proof of Theorem 2), but a kernel estimate with data-dependent h. However, since the data-dependence is more realistic than determinism, the study in its present form carries more weight.

A typical choice for q_n would be $Cn^{-2/5}$ (where C is defined in Lemma 5 below) plus the $\sqrt{\log n/n}$ term defined in parts B or D of Theorem 5. One could also take $(C+\varepsilon)\ n^{-2/5}$ for some constant $\varepsilon > 0$. Note that the inequality of part D can be used for moderate and even small values of n. It seems unwise to take q_n larger than these suggestions (from B and D), since that would decrease the performance when $f \notin \boldsymbol{T}$ (recall that q_n can be considered as a halo, and equivalently as the size of a discretization grid in the space of all densities).

6. Appendix: The proof of Theorem 5

6.1. Behavior of t_n

First we need a simple upper bound for the L_1 error committed with t_n. Let $f_{a,b}$ denote a normal density with mean a and standard deviation b. (Thus, $t_n \equiv f_{\hat{\mu},\hat{\sigma}}$.) Since the L_1 error is invariant under linear transformations of the axis, we can and do assume, without loss of generality, that $\mu = 0$, $\sigma = 1$.

Lemma 2.

$$\begin{aligned}\int |f_{a,b} - f_{0,1}| &\leqq 2 \log\left(\max\left(b, 1/b\right)\right) + \left(1 + \sqrt{\frac{2}{\pi}}\right)|a| \\ &\leqq 2\left(\max\left(b, 1/b\right) - 1\right) + \left(1 + \sqrt{\frac{2}{\pi}}\right)|a| \,.\end{aligned}$$

Proof.

$$\begin{aligned}\int |f_{a,b} - f_{0,1}| &\leqq \int |f_{a,b} - f_{a,1}| + \int |f_{a,1} - f_{0,1}| \\ &= \int |f_{0,b} - f_{0,1}| + \int |f_{a,1} - f_{0,1}| \,.\end{aligned}$$

For $b>1$,

$$f_{0\,b}=f_{0,1}+\int_1^b \frac{1}{u}\, f_{0,u}\left(\frac{x^2}{u^2}-1\right) du$$

by taking the derivative of f with respect to b. Thus,

$$\int |f_{0,b}-f_{0,1}| \leqq \int\int_1^b \frac{1}{u}\, f_{0,u}\left|\frac{x^2}{u^2}-1\right| du dx$$

$$\leqq \int_1^b \left(\int f_{0,u}\left|\frac{x^2}{u^2}-1\right| dx\right)\frac{1}{u}\, du \leqq 2\int_1^b u^{-1} du = 2\log b \leqq 2\,(b-1)\,.$$

A similar argument is valid for $b<1$, so that we obtain the first bound of the Lemma. By a similar line of reasoning, for $a>0$,

$$\int |f_{a,1}-f_{0,1}| \leqq \int \left|\int_0^a (x-a)\, f_{u,1} du\right| dx \leqq \int\int_0^a |x-a|\, f_{u,1} dx du$$

$$\leqq \min\left(2, \sqrt{\frac{2}{\pi}}\, a+\frac{a^2}{2}\right) \leqq \left(1+\sqrt{\frac{2}{\pi}}\right) a\,.$$

Combining all this proves Lemma 2. ■

It is known that $\hat{\mu}$ is normal with mean 0 and variance $1/n$, and that $\hat{\sigma}^2$ is $1/n$ times a chi-square random variable with $n-1$ degrees of freedom. This can be used to bound the L_1 error. In Lemma 3 below, we also establish that the estimate t_n is minimax-optimal for $\boldsymbol{T}$.

Lemma 3. *Let $\boldsymbol{T}$ be the class of all normal densities on the real line. The normal estimate t_n satisfies the following inequality, for all $0<u\leqq\sqrt{8}-2$:*

$$\sup_{f\in\boldsymbol{T}} \mathsf{P}\left(\int |t_n-f|>u\right) \leqq 3\, e^{-nu^2/32}\,.$$

Additionally, there exists a positive constant $\alpha>0$ such that

$$\frac{\alpha}{\sqrt{n}} \leqq \inf_{f\in\boldsymbol{T}} \mathsf{E}\left(\int |t_n-f|\right) = \sup_{f\in\boldsymbol{T}} \mathsf{E}\left(\int |t_n-f|\right) \leqq \frac{B}{\sqrt{n-5}}\,,$$

where $B=6\sqrt{2}+\sqrt{2/3}+\sqrt{2/\pi}+2/\pi$, and for the upper bound, it is assumed that $n\geqq 6$. Finally, for some positive constant $\beta>0$,

$$\inf_{f_n}\sup_{f\in\boldsymbol{T}} \mathsf{E}\left(\int |f_n-f|\right) \geqq \frac{\beta}{\sqrt{n}}\,.$$

Proof. For the first inequality, it is clear that we can assume that $u\geqq 4/\sqrt{n}$. Furthermore, the distribution of $\int |t_n-f|$ is scale (and thus σ-) invariant, so we can and do assume that $\sigma=1$. Let N be a normal (0, 1) random variable, and let $\boldsymbol{G}$ be gamma $(n-1)/2$. Then

$$\mathsf{P}\left(\int |t_n-f|>u\right) \leqq \mathsf{P}\left(2\max\left(\hat{\sigma}, 1/\hat{\sigma}\right)>\frac{u}{2}\right)+\mathsf{P}\left((1+\sqrt{2/\pi})|\hat{\mu}|>\frac{u}{2}\right) \triangleq I+II\,.$$

Now, defining $\theta=1/(2\,(1+\sqrt{2/\pi}))$, and using $\theta\geqq 1/4$ and $u\geqq 4/\sqrt{n}$, we have

$$II = 2\,\mathsf{P}\,(N \geqq \theta\sqrt{n}u) \leqq \frac{2}{\vartheta\sqrt{2\pi n}\,u}\,\mathrm{e}^{-n\,u^2\vartheta^2/2}$$

$$\leqq \sqrt{\frac{2}{\pi}}\,\mathrm{e}^{-nu^2/32}.$$

Also,

$$I \leqq \mathsf{P}\left(\hat{\sigma} < \frac{1}{1+u/4}\right) + \mathsf{P}\,(\hat{\sigma} > 1+u/4) \leqq \mathsf{P}\left(\hat{\sigma}^2 < \frac{1}{1+u/2}\right) + \mathsf{P}\,(\hat{\sigma}^2 > 1+u/2)$$

$$= \mathsf{P}\left(G < \frac{n}{2+u}\right) + \mathsf{P}\left(G > \frac{n}{2} + \frac{nu}{4}\right)$$

$$= \mathsf{P}\left(G < \frac{n-1}{2} - \frac{nu-2-u}{4+2u}\right) + \mathsf{P}\left(G > \frac{n-1}{2} + \frac{nu+2}{4}\right)$$

$$\leqq \mathrm{e}^{-\frac{1}{2}\frac{n-1}{2}\left[\frac{nu-2-u}{(2+u)(n-1)}\right]^2} + \mathrm{e}^{-\frac{1}{2}\frac{n-1}{2}\left[\frac{nu+2}{2n-2}\right]^2\left(1+\frac{nu+2}{2n-2}\right)^{-1}}$$

$$\leqq \mathrm{e}^{-\frac{(n-1)u^2}{4(2+u)^2} + \frac{u}{(2+u)^2}} + \mathrm{e}^{-\frac{1}{16}(nu+2)^2\frac{2n-2}{(n-1)(2n+nu)}}$$

$$\leqq \mathrm{e}^{-\frac{nu^2}{4(2+u)^2} + \frac{u}{4(2+u)^2} + \frac{1}{8}} + \mathrm{e}^{-\frac{1}{16}(nu)^2\frac{2}{n\sqrt{8}}}$$

$$\leqq \mathrm{e}^{-nu^2/32 + \frac{1}{36} + \frac{1}{8}} + \mathrm{e}^{-nu^2/32}$$

by some tail inequalities for the gamma distribution, and the fact that $u \leqq \sqrt{8} - 2 \leqq 1$. Simplification of these bounds yields the bound

$$\left(\sqrt{\frac{2}{\pi}} + 1 + \frac{\mathrm{e}^5}{36}\right)\mathrm{e}^{-nu^2/32} \leqq 3\mathrm{e}^{-nu^2/32}\,.$$

Also, assuming again that f is normal (0, 1), we have

$$\mathsf{E}\,(\int |t_n - f|) \leqq 2\,\mathsf{E}\,(\max\,(\hat{\sigma},\, 1/\hat{\sigma}) - 1) \overset{\Delta}{=} I + II\,.$$

Now, $II = \frac{2}{\pi\sqrt{n}} + \sqrt{\frac{2}{\pi n}}$. Furthermore,

$$I = 2\,\mathsf{E}\left(\max\,\left(\sqrt{\frac{2G}{n}},\, \sqrt{\frac{n}{2G}}\right) - 1\right) \leqq 2\,\mathsf{E}\left(\sqrt{\frac{2G}{n}} - 1\right)_+ + 2\,\mathsf{E}\left(\sqrt{\frac{n}{2G}} - 1\right)_+$$

$$\leqq 2\,\mathsf{E}\left(\frac{2G}{n} - 1\right)_+ + 2\,\mathsf{E}\left(\frac{n}{2G} - 1\right)_+ \leqq 2\,\mathsf{E}\left(\frac{2\,(G - \mathsf{E}G)}{n} - \frac{1}{n}\right)_+$$

$$+ 2\,\mathsf{E}\left(\frac{n}{2G} - \mathsf{E}\left(\frac{n}{2G}\right) + \frac{3}{n-3}\right)_+$$

$$\leqq 2\,\mathsf{E}\left(\frac{2\,(G - \mathsf{E}G)}{n}\right)_+ + \frac{2}{n} + 2\,\mathsf{E}\left(\frac{n}{2G} - \mathsf{E}\left(\frac{n}{2G}\right)\right)_+$$

$$\leqq \frac{4}{n}\sqrt{\mathsf{Var}\,(G)} + \frac{2}{n} + n\sqrt{\mathsf{Var}\left(\frac{1}{G}\right)}\,,$$

where we used the CAUCHY-SCHWARZ inequality. We know that for a gamma random variable G with parameter α, $\mathsf{Var}\,(G) = \alpha$ and $\mathsf{Var}\,(1/G) = (\alpha-1)^{-2}\,(\alpha-2)^{-1}$

Thus, we have for $n \geqq 6$,

$$I \leqq \frac{4}{\sqrt{2n}} + \frac{2}{n} + \frac{2\sqrt{2}}{\sqrt{n-5}} + \frac{6\sqrt{2}}{(n-3)\sqrt{n-5}} \leqq \frac{A}{\sqrt{n-5}},$$

where $A = 6\sqrt{2} + \sqrt{2/3}$.

The minimax lower bound can be obtained by standard information-theoretic methods (see e.g. DEVROYE, 1987). Also, the lower bound on $\mathsf{E}(\int |t_n - f|)$ is a straight-forward exercise. ■

6.2. The scale-invariant kernel estimate

Let g_n be the kernel estimate with the data-based h given above, and let g_{n0} be the kernel estimate based upon the same data and h, but with h replaced by

$$h_0 = cn^{-\frac{1}{5}}\sigma,$$

where $\sigma = \sigma(f)$ is a scale factor for f which is equal to the standard deviation if it exists. It is assumed that $\hat{\sigma}$ is close to σ in some probabilistic sense. The closeness of g_n to g_{n0} is dealt with in the following lemma.

Lemma 4.

A. *We have*

$$\int |g_n - g_{n0}| \leqq 2\left(1 - \min\left(\frac{\hat{\sigma}}{\sigma}, \frac{\sigma}{\hat{\sigma}}\right)\right) \triangleq \Delta(\sigma, \hat{\sigma}).$$

B. *Let $\boldsymbol{T}$ be the class of all normal densities on the real line. For all $u \leqq \sqrt{8} - 2$:*

$$\sup_{f \in \boldsymbol{T}} \mathsf{P}(\Delta(\sigma, \hat{\sigma}) > u) \leqq 2e^{-nu^2/32}.$$

For all $n \geqq 6$,

$$\sup_{f \in \boldsymbol{T}} \mathsf{E}(\Delta(\sigma, \hat{\sigma})) \leqq \frac{6\sqrt{2} + \sqrt{2/3}}{\sqrt{n-5}}.$$

C. $\mathsf{E}(\Delta(\sigma, \hat{\sigma})) = o(n^{-2/5})$ when $\int |x|^{(10+\varepsilon)/3} f(x)dx < \infty$ *for some* $\varepsilon > 0$.

D.

$$\int |g_n - f| \leqq \int |g_{n0} - f| + 2\Delta(\sigma, \hat{\sigma}),$$

$$\mathsf{E}(\int |g_n - f|) \leqq \mathsf{E}(\int |g_{n0} - f|) + 2\,\mathsf{E}(\Delta(\sigma, \hat{\sigma})),$$

and

$$\int |g_n - f| - \mathsf{E}(\int |g_n - f|) \leqq \int |g_{n0} - f| - \mathsf{E}(\int |g_{n0} - f|) + 2\Delta(\sigma, \hat{\sigma}) + 2\,\mathsf{E}(\Delta(\sigma, \hat{\sigma})).$$

Proof. Statement A follows from the unimodality of K, jointly with

$$\int |g_n - g_{n0}| \leqq \int |K_{\hat{\sigma}} - K_{\sigma}| \leqq 2\left(1 - \min\left(\frac{\hat{\sigma}}{\sigma}, \frac{\sigma}{\hat{\sigma}}\right)\right)$$

(DEVROYE and GYÖRFI, 1985, pp. 186–187).

Statement B is immediate from the proof of Lemma 3 and the fact that $1-\min(x, 1/x) \leqq \max(x, 1/x)-1$ for all $x>0$. Statement C can be proved by employing an inequality due to VON BAHR and ESSEEN (1965) (see also NAGAEV and PINELIS (1977) and MANSTAVICIUS (1982)) for sums of iid zero mean random variables $Z_1, \ldots, Z_n$:

$$\mathsf{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right|\right) \leqq \mathsf{E}^{\frac{1}{p}}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right|^p\right) \leqq 2^{\frac{1}{p}} n^{\frac{1}{p}-1}\, \mathsf{E}(|Z_1|^p) \quad (p \in [1, 2]) .$$

With $Z_i = X_i^2 - \sigma^2$, the upper bound is $o(n^{-2/5})$ for $p > 5/3$. This yields the condition that $\mathsf{E}(|X_i|^{2p}) < \infty$. A small additional argument is needed to bound $|\hat{\sigma} - \sigma|$ in terms of the Z_i's and asymptotically negligible terms. Statement D follows by the triangle inequality. ■

6.3. The centered kernel estimate g_{n0}

Having studied the closeness of g_n to g_{n0}, it is necessary to see how g_{n0} behaves in general. Since it is a kernel estimate with deterministic h, all the inequalities mentioned e.g. in the proof of Theorem 2 remain valid. Also, since $h \to 0$ and $nh \to \infty$, it is consistent for all f for which the estimate is well-defined, i.e. for those densities with finite variance.

Lemma 5.
Let f be the normal $(0, 1)$ density, and let g_{n0} be a kernel estimate with nonnegative symmetric kernel K with support in $[-1, 1]$, and with deterministic smoothing factor h. Then

$$\mathsf{E}\left(\int |g_{n0} - f|\right) \leqq \sqrt{\frac{2}{\pi \mathrm{e}}}\, h^2 \int x^2 K + \frac{\sqrt{\int K^2}\,[(8\pi)^{1/4} + 2h(2\pi)^{-1/4}]}{\sqrt{nh}}$$

With BARTLETT'S *kernel* $K(x) = \frac{3}{4}(1-x^2)_+$, *the estimate becomes*

$$\mathsf{E}\left(\int |g_{n0} - f|\right) \leqq \sqrt{\frac{2}{25\pi \mathrm{e}}}\, h^2 + \frac{\sqrt{\frac{3}{5}}\,[(8\pi)^{1/4} + 2h(2\pi)^{-1/4}]}{\sqrt{nh}} .$$

When

$$h = n^{-\frac{1}{5}}\left[\frac{225\pi^3 \mathrm{e}^2}{128}\right]^{\frac{1}{10}} \overset{\Delta}{=} cn^{-1/5} = 1.2019409 \ldots\, n^{-\frac{1}{5}},$$

the upper bound becomes $Cn^{-\frac{2}{5}} + Dn^{-\frac{3}{5}}$, where

$$C \overset{\Delta}{=} \left[\frac{2}{25\pi \mathrm{e}}\right]^{\frac{1}{2}}\left[\frac{225\pi^3 \mathrm{e}^2}{128}\right]^{\frac{1}{5}} + \left[\frac{3}{5}\right]^{\frac{1}{2}} (8\pi)^{\frac{1}{4}}\left[\frac{128}{225\pi^3 \mathrm{e}^2}\right]^{\frac{1}{20}},$$

$$D \overset{\Delta}{=} \left[\frac{3}{5}\right]^{\frac{1}{2}} 2(2\pi)^{-\frac{1}{4}}\left[\frac{225\pi^3 \mathrm{e}^2}{128}\right]^{\frac{1}{20}}.$$

Proof. By a uniform estimate for the bias given on p. 122 of DEVROYE and GYÖRFI (1985),

$$\int |f - f^* K_h| \leqq \frac{h^2 \int x^2 K \int |f''|}{2} .$$

Also (see last line of p. 124 of same reference):

$$\mathsf{E}\left(\int |g_{n0} - f^* K_h|\right) \leqq \frac{\int \sqrt{f^*(K^2)_h}}{\sqrt{nh}} .$$

The convolution integral in the numerator can further be bounded as follows, if we write K^2 instead of $(K^2)_h$:

$$\int f(y)\, K^2(x-y)\, dy \leqq \sup_{z:|z-x|\leqq h} f(z) \int K^2(x-y)\, dy \leqq \begin{cases} f(0) \int K^2 & |x| \leqq h \\ f(|x|-h) \int K^2 & |x| > h \end{cases} .$$

Thus,

$$\int \sqrt{f^*(K^2)_h} \leqq \sqrt{\int K^2}\, \left(\int \sqrt{f - 2h}\, \sqrt{f(0)}\right) .$$

Next, note that $f(0) = \frac{1}{\sqrt{2\pi}}$, $\int \sqrt{f} = (8\pi)^{\frac{1}{4}}$, and $\int |f''| = 4 \sup |f'| = 4 \sup |x|\, |f(x)| = \sqrt{\frac{8}{\pi e}}$. Combining all this gives us our first estimate. The second estimate is obtained after replacing $\int x^2 K$ by 1/5, and $\int K^2$ by 3/5. The optimization of the two main terms with respect to h is trivial. ■

6.4. The proof of Theorem 5

Theorem 1 implies the consistency of f_n whenever g_n is consistent. By a general theorem of DEVROYE and GYÖRFI for data based h, g_n is consistent when $h \to 0$ and $nh^d \to \infty$ in probability as $n \to \infty$ (DEVROYE and GYÖRFI, 1985, p. 148). This proves statement A.

For statement B, we have to extend Theorem 2 (which only applies when h does not depend upon the data). Consider first $f \notin \boldsymbol{T}$, and note that $L_1(f, \boldsymbol{T}) \stackrel{\Delta}{=} 2\delta > 0$. Now apply part B of Lemma 1, where it is clear that $\int |t_n| = 1$ for all n. Let n be so large that $q_n < \delta$. Then

$$\mathsf{E}\left(\int |f_n - f|\right) \leqq \mathsf{E}\left(\int |g_n - f|\right) + 2\,\mathsf{P}\left(\int |g_{n0} - f| > \delta\right) + 2\,\mathsf{P}\left(\int |g_n - g_{n0}| > \delta\right) .$$

Of the terms on the right-hand-side, the first one is at least $\mathsf{E}\int |g_{n0} - f|)$ minus $\mathsf{E}(\int |g_n - g_{n0}|)$, which is at least $(0.86 + o(1))\, n^{-2/5} - o(n^{-2/5})$ by an asymptotic inequality of DEVROYE and PENROD (1984) valid for symmetric $K \geqq 0$ and deterministic h, and part C of Lemma 4. The second term is $O(e^{-cn})$ for some $c > 0$ (see proof of Theorem 2). The third term does not exceed $\mathsf{E}\left(\int |g_n - g_{n0}|/\delta\right) = o(n^{-2/5})$ (part C of Lemma 4). Thus, $\mathsf{E}\left(\int |f_n - f|\right) \sim \mathsf{E}\left(\int |g_n - f|\right)$.

Consider next $f \in \boldsymbol{T}$. Here we apply part A of Lemma 1, after observing that $\int |g_n| = 1$ for all n, and that $R_n(\boldsymbol{T}) \leqq Cn^{-2/5} + Dn^{-3/5} + B/\sqrt{n-5}$ for $n \geqq 6$ (apply

Lemma 5 and part B of Lemma 4). It should be stressed that C and D are the constants defined in Lemma 5 if h if $\hat{\sigma}n^{-1/5}$ with $c=1.2019409\ldots$ as suggested in that Lemma. Otherwise, the values of C and D are slightly different. Note further that $\sqrt{n}\ \mathsf{E}(\int|t_n-f|)$ is asymptotically sandwiched between two positive constants $\alpha<\beta$ (see Lemma 3). It suffices to establish that we can find sequences $u=u(n)>0$ and $v=v(n)>0$ such that $u+v\leqq c^*\sqrt{\log n/n}$,

$$\mathsf{P}\left(\left|\int|g_n-f|-\mathsf{E}\left(\int|g_n-f|\right)\right|>u\right)=o\left(\frac{1}{\sqrt{n}}\right)$$

and

$$\mathsf{P}\left(\int|t_n-f|>v\right)=o\left(\frac{1}{\sqrt{n}}\right).$$

The latter probability does not exceed $2\exp(-nv^2/32)$ (Lemma 3). This tends to zero at the required rate if we take $v=\sqrt{(16+\varepsilon)\log n/n}$ for some $\varepsilon>0$. The former probability is dealt with by a three-way decomposition as in the last part of part D of Lemma 4. The probability does not exceed

$$\begin{aligned}&\mathsf{P}\left(\left|\int|g_{n0}-f|-\mathsf{E}\left(\int|g_{n0}-f|\right)\right|>\frac{u}{4}\right)\\&\quad+\mathsf{P}\left(\Delta(\sigma,\hat{\sigma})>\frac{u}{4}\right)+\mathsf{P}\left(\mathsf{E}\,\Delta(\sigma,\hat{\sigma})>\frac{u}{8}\right)\\&\leqq 2e^{-\frac{nu^2}{16\times 32}}+3e^{-\frac{nu^2}{16\times 32}}+0\end{aligned}$$

provided that $n\geqq 6$, $u\leqq 4(\sqrt{8}-2)$, and $8(6\sqrt{2}+\sqrt{2/3})/\sqrt{n-5}\leqq u$ (apply Devroye (1988) (see also Theorem 2) and part B of Lemma 4). All of this is $o(1/\sqrt{n})$ when $u=\sqrt{(256+\varepsilon)\log n/n}$ for some $\varepsilon>0$. Thus, asymptotic optimality follows for $f\in\boldsymbol{T}$ if q_n is at least equal to $R_n(\boldsymbol{T})$ plus $c^*\sqrt{\log n/n}$ where $c^*>\sqrt{16}+\sqrt{256}=20$. It suffices, for example, that q_n is at least equal to $C_n^{-2/5}+c^*\sqrt{\log n/n}$. This concludes the proof of part B. Part C was also essentially proved when we obtained part B. Finally, the inequality in part D is obtained without work from the proof of part B. ■

Acknowledgement

This paper was written during a summer visit at the Department of Statistics of Stanford University. I would like to thank Ingram Olkin for showing me a preprint of a paper of him on mixing parametric and nonparametric estimates. I am also indebted to Art Owen for exciting discussions and constructive feedback. Finally, I am grateful to the referees for showing me how to better present the material.

References

VON BAHR, B., and ESSEEN, G. G. (1965). Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.* **36**, 299–303.

BARRON, A. R. (1985). Locally smooth density estimation. Technical Report TR 56, Department of Statistics, Stanford University.

BASU, A. P. (1964). Estimates of reliability for some distribution useful in life testing. *Technometrics* **6**, 215–219.

BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5**, 445–463.

BERAN, R. (1981). Efficient robust estimates in parametric models. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **55**, 91–108.

BICKEL, P. J. (1976). Another look at robustness: a review of reviews and some new developments. *Scand. J. Statist.* **3**, 145–168.

BIRGÉ, L. (1985). Non-asymptotic minimax risk for Hellinger balls. *Prob. Math. Statist.* **5**, 21–29.

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields* **71**, 271–291.

BIRGÉ, L. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15**, 1013–1022.

BIRGÉ, L. (1987). Estimating a density under order restrictions: nonasymptotic minimax risk. *Ann. Statist.* **15**, 995–1012.

BRETAGNOLLE, J., and HUBER, C. (1978). Lois empiriques et distance de Prokhorov. In *Seminaire de Probabilités XII*, vol. 649, 332–341, Springer-Verlag, New York.

CACOULLOS, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18**, 178–189.

COVER, T. M. (1972). A hierarchy of probability density function estimates. In *Frontiers in Pattern Recognition*, 83–98, Academic Press, New York.

D'AGOSTINO, R. B., and STEPHENS, M. A. (1986). *Goodness-of-fit Techniques*, Marcel Dekker, New York.

DEHEUVELS, P. (1977). Estimation non paramétrique de la densité par histogrammes generalisés. *Revue de Statistique Appliquée* **25**, 5–42.

DEVROYE, L. (1983). The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. *Ann. Statist.* **11**, 896–904.

DEVROYE, L. (1983). On arbitrarily slow rates of global convergence in density estimation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **62**, 475–483.

DEVROYE, L. (1986). A universal lower bound for the kernel estimate. Technical Report, School of Computer Science, McGill University.

DEVROYE, L. (1987). *A Course in Density Estimation*. Birkhäuser Verlag, Boston.

DEVROYE, L. (1988). The kernel estimate is relatively stable. *Probability Theory and Related Fields* **77**, 521–536.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L1 View*. John Wiley, New York.

DEVROYE, L., and PENROD, C. S. (1984). Distribution-free lower bounds in density estimation. *Ann. Statist.* **12**, 1250–1262.

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10**, 401–414.

GRENANDER, U. (1981). *Abstract Inference*. John Wiley, New York.

GUTTMANN, H., and WERTZ, W. (1976). Note on estimating normal densities. *Sankhya, Ser. B*, **38**, 231–236.

LUMELSKII. YA., and SAPOZHNIKOV, P. N. (1969). Unbiased estimates of density functions. *Theory of Probability and its Applications* **14**, 357–364.

MANSTAVICIUS, E. (1982). Inequalities for the p-th moment, $0<p<2$, of a sum of independent random variables. *Lithuanian Math.* **22**, 64–67.

NAGAEV, S. V., and PINELIS, N. F. (1977). Some inequalities for sums of independent random variables. *Theory of Probability and its Applications* **22**, 248–256.

OLKIN, I., and SPIEGELMAN, C. H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **82**, 858–865.

PARZEN, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33**, 1065–1076.

PETROV, V. V. (1975). *Sums of Independent Random Variables.* Springer-Verlag, Berlin.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–837.

SCHEFFÉ, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18**, 434–458.

SCHUSTER, E. F., and YAKOWITZ, S. (1985). Parametric/nonparametric mixture density estimation with application to flood frequency analysis. *Water Resources Bull.* **21**, 797–804.

SEHEULT, A. H., and QUESENBERRY, C. P. (1971). On unbiased estimation of density functions. *Ann. Math. Statist.* **42**, 1434–1438.

WERTZ, W. (1975). On unbiased density estimation. *An. Acad. Brasil. Cienc.* **47**, 65–72.

YATRACOS, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* **13**, 768–774.

Received September 1986; revised October 1987 and July 1988.

LUC DEVROYE
Schod of Computer Science
McGill University
805 Sherbrooke Street West
Montreal
Canada H3A 2K6