The Annals of Statistics
1986, Vol. 14, No. 4, 1612–1618

ON IMPROVING DENSITY ESTIMATORS WHICH ARE NOT BONA FIDE FUNCTIONS

BY LESŁAW GAJEK

Technical University of Łódź

In order to improve the rate of decrease of the IMSE for nonparametric kernel density estimators with nonrandom bandwidth beyond $O(n^{-4/5})$ all current methods must relax the constraint that the density estimate be a bona fide function, that is, be nonnegative and integrate to one. In this paper we show how to achieve similar improvement without relaxing any of these constraints. The method can also be applied for orthogonal series, adaptive orthogonal series, spline, jackknife, and other density estimators, and assures an improvement of the IMSE for each sample size.

1. Introduction. Several techniques of nonparametric estimation of an unknown density have been proposed by a number of researchers during the last 25 years; see Tapia and Thompson (1978). Since the shape of the density is of most interest, the integrated mean square error (IMSE) is an appropriate criterion for comparison. In general, it is known that the best possible mean square error (MSE) convergence rate for a density estimate, which is uniform over the Sobolev space of functions with $m-1$ derivatives absolutely continuous and mth derivative in L_p, $p \geq 1$, is not better then $n^{-\varphi(m,\varepsilon)}$, where

$$\varphi(m,\varepsilon) = \left(2m - \frac{2}{p+\varepsilon}\right) \Big/ \left(2m + 1 - \frac{2}{p+\varepsilon}\right),$$

for arbitrary small $\varepsilon > 0$ [see Wahba (1975)]. Some estimates which offer the rate of convergence close to the optimal one are reviewed below.

(1) For the kernel type estimators with nonrandom bandwidth, Parzen (1962) proved that one can choose kernels so that the IMSE is like $O(n^{-2m/(2m+1)})$ for densities with $m-1$ derivatives absolutely continuous and mth derivative in L_∞ [see also Bartlett (1963), Rosenblatt (1971) and Nadaraya (1974)]. However, if $m > 2$ the kernel estimate is not nonnegative. Thus for bona fide (i.e., nonnegative and integrable to 1) kernels the method is limited by the rate $O(n^{-4/5})$. One should note here that Abramson (1982) has presented a kernel type bona fide estimator with bandwidth depending on the sample, the MSE of which is of order $o(n^{-4/5})$.

(2) Schucany and Sommers (1977) employed the generalized jackknife method to reduce the asymptotic and small sample MSE and bias of the kernel type estimators [see also Koronacki (1984)]. Since their estimator can be produced by the initial use of a single kernel that is not bona fide, this method does also relax the nonnegativity constraint.

Received November 1984; revised December 1985.
AMS 1980 *subject classification*. Primary 62G05.
Key words and phrases. Nonparametric density estimation, kernel estimation, orthogonal series estimation, rates of convergence.

(3) Wahba (1975) has shown that rates of convergence of the MSE like $O(n^{-\varphi(m)})$ are possible for the Kronmal–Tarter and spline estimators. However these estimators are not bona fide densities.

(4) An adaptive-data version of the Kronmal–Tarter estimator has been constructed by Anderson and Figueiredo (1980). This method offers reduced bias in comparison to the conventional Kronmal–Tarter estimator, but again, the nonnegativity constraint is relaxed.

(5) Walter (1977) has shown that for the Hermite series estimators of density functions with compact support the MSE is of order $O(n^{-1+1/m})$. Thus in the case for which the trigonometric system is natural, the Hermite series estimator does almost as well, while for the case when the support is not compact (which is natural for the Hermite series estimator), the trigonometric system cannot be even used. However, the Hermite orthogonal series estimator also is not nonnegative.

It is of interest to us that essentially all the above authors seem to ignore the nonnegativity constraint whereas, for techniques like likelihood ratio estimation, the presence of negative values poses theoretical as well as practical problems. A negative hazard rate implies the spontaneous reviving of the dead. Therefore Terrell and Scott (1980) proposed a particular class of estimators based on ordinary kernel estimators that achieve the goal of faster rates of convergence by relaxing the integral constraint rather than the nonnegativity one.

Another approach was proposed by Silverman (1982) who had applied the maximum penalized likelihood method for estimating the logarithm of an unknown density instead of the density itself. Silverman's approach gives bona fide estimators; however, it is restricted to the case of densities having bounded support.

In this note we propose a simple algorithm which improves any density estimator that is not bona fide with respect to weighted IMSE and which converges to a bona fide density estimator. The only condition on the weight function $h(t)$ in the IMSE is that $\int dt/h(t)$ is finite, that is, the tails have to be weighted strongly enough. All results of this note concern improving the weighted IMSE for arbitrary sample size n (contrary to the Terrell–Scott method which is of asymptotic type only) and may be useful in small sample estimation. Compared with Silverman's approach, our method is much more immediately practicable, and also is not restricted to a bounded interval.

2. The main result. Let P be a probability measure on the real line absolutely continuous with respect to Lebesgue measure λ and denote $f = dP/d\lambda$. Let $\hat{f}(\cdot, x^n)$ be an estimator of f for the sample $x^n = (x_1, \ldots, x_n)$ and define the weighted IMSE for $\hat{f}$ by

$$R(\hat{f}, f) = \int\int [\hat{f}(t, x^n) - f(t)]^2 P^n(dx^n) h(t)\, dt, \tag{1}$$

where the weight function h is nonnegative and Borel measurable.

Denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the inner product and the norm of the unitary space $L_2^h(S) = \{g\colon S \to \mathbb{R} \mid \int g^2 h\, dx < \infty\}$, where S is a given Borel subset of $\mathbb{R}$.

Let $\mathscr{F}^+$ be the subset of all functions of $L_2^h(S)$ which are positive a.e. $[\lambda]$ on S; similarly let $\mathscr{F}^1$ denote the subset of all functions of $L_2^h(S)$ which integrate to one on S. Denote $\mathscr{F}^* = \mathscr{F}^+ \cap \mathscr{F}^1$ and observe that $\mathscr{F}^+$, $\mathscr{F}^1$, and $\mathscr{F}^*$ are all closed and convex subsets of $L_2^h(S)$. For the remainder of this note, attention will be restricted to the case where $\mathscr{F}^*$ is a class of densities under consideration.

Suppose that $\hat{f}$, an estimator of $f \in \mathscr{F}^*$, is not a bona fide density (i.e., it is not a function which is simultaneously nonnegative and does integrate to one). We shall investigate the following simple iterative procedure which corrects $\hat{f}$:

P-ALGORITHM.

1. *Set* $f_0(\cdot, x^n) = \hat{f}(\cdot, x^n)$ *and* $k = 0$.
2. *Set* $f_{k+1}(\cdot, x^n) = \max(0, f_k(\cdot, x^n))$ *and check* $C_{k+1} = \int_S f_{k+1}(t)\,dt$. *If* $C_{k+1} = 1$ *stop*.
3. *Set* $f_{k+2}(\cdot, x^n) = f_{k+1}(\cdot, x^n) - (C_{k+1} - 1)/[h(\cdot)\int_S 1/h(t)\,dt]$.
4. *Set* $k = k + 2$ *and go to* 2.

REMARK 1. For the sake of brevity we assume that Step 3 is always well defined. In Section 3 it will be shown that f_1 is the orthogonal projection of $\hat{f}$ onto $\mathscr{F}^+$, f_2 is the orthogonal projection of f_1 onto $\mathscr{F}^1$ (if it exists) and so on. The existence of f_2 is equivalent to the conditions $\int_S 1/h(t)\,dt < \infty$ and $|C_1| < \infty$. However if $\hat{f} \in L_2^h(S)$ and $L_2^h(S)$ is assumed to be a Hilbert space, then all projections exist and are unique, since $\mathscr{F}^+$ and $\mathscr{F}^1$ are closed and convex. For our purpose so strong an assumption is not necessary because we can directly verify the conditions $\int_S 1/h(t)\,dt < \infty$, $|C_1| < \infty$, which assure that Step 3 is well defined for each iteration.

The following theorem states the main result.

THEOREM 1. *Assume that* $\infty > \int_S \hat{f}(t, x^n)\,dt \geq 1$ *a.e.* $[\lambda^n]$,

$$\int_S 1/h(t)\,dt < \infty \tag{2}$$

and $R(\hat{f}, f) < \infty$ *for every* $f \in \mathscr{F}^*$. *Then*

(i) *for each iteration* f_k *of the P-Algorithm*

$$R(\hat{f}, f) \geq R(f_k, f) + \sum_{i=1}^{k} E\|f_i - f_{i-1}\|^2, \tag{3}$$

for all $f \in \mathscr{F}^*$.

(ii) *There exists a unique real number* α *such that the P-Algorithm converges pointwise and in norm to a function* $f^* \in \mathscr{F}^*$ *of the form*

$$f^*(\cdot, x^n) = \max(0, \hat{f}(\cdot, x^n) - \alpha/h(\cdot)). \tag{4}$$

(iii) f^* *is the best of all uniform over* $\mathscr{F}^*$ *corrections of* $\hat{f}$, *i.e., for any other function* $\bar{f}(\cdot, x^n) \in \mathscr{F}^*$, *depending on the sample* x^n *through* $\hat{f}$ *only, we have*

$$\inf_{f \in \mathscr{F}^*} [R(\hat{f}, f) - R(f^*, f)] \geq \inf_{f \in \mathscr{F}^*} [R(\hat{f}, f) - R(\bar{f}, f)].$$

REMARK 2. Observe that if the support S is compact and h is continuous, then hypothesis (2) will be satisfied. In the case where S is not compact the condition $\int dt/h(t) < \infty$ means that the weight h should increase for "large" arguments. A similar effect of increasing the importance of the tails could be caused by estimating the logarithm of the density f instead of f itself [see Silverman (1982)].

REMARK 3. Clearly f^* given by (4) has at least the same rate of convergence of the IMSE as its original $\hat{f}$. Furthermore, Lemma 4 below implies that for every $f \in \mathscr{F}^*$

$$\|\hat{f} - f\|^2 \geq \|f^* - f\|^2 \quad \text{a.e. } [\lambda^n]$$

and therefore f^* has also at least the same rate of consistency as $\hat{f}$.

REMARK 4. In practice f^* will be evaluated numerically, for example by the use of the P-Algorithm. Then we can stop the algorithm whenever $|C_k - 1| < \varepsilon$, for a given real $\varepsilon > 0$, and estimate the improvement of the weighted IMSE by (3).

REMARK 5. Terrell and Scott (1980) have proposed another method which assures good asymptotics and nonnegativity of the estimator by relaxing the integral constraint. Evidently the estimator f^* defined by (4) is much more simple than the Terrell–Scott one, which is the ratio of two nonnegative kernel estimators. Furthermore, their result is restricted to kernel type estimators and is only of asymptotic kind, contrary to the above theorem.

3. Proof of Theorem 1. The proof proceeds through a series of lemmas.

LEMMA 1. *Let $\mathscr{F}_0$ be a given convex subset of $L_2^h(S)$ and $\hat{f} \in L_2^h(S)$. Then $f_0 \in \mathscr{F}_0$ is the orthogonal projection of $\hat{f}$ onto $\mathscr{F}_0$ if and only if*

$$\|\hat{f} - f\|^2 \geq \|f - f_0\|^2 + \|f_0 - \hat{f}\|^2,$$

for all $f \in \mathscr{F}_0$.

PROOF. It is well known that f_0 is the orthogonal projection of $\hat{f}$ onto $\mathscr{F}_0$ if and only if for every $f \in \mathscr{F}_0$

$$\langle \hat{f} - f_0, f - f_0 \rangle \leq 0.$$

Since

$$\|\hat{f} - f\|^2 = \|f_0 - f\|^2 - 2\langle f - f_0, \hat{f} - f_0 \rangle + \|\hat{f} - f_0\|^2,$$

therefore the result follows. □

Suppose that $\hat{f}$ is an estimator of $f \in \mathscr{F}^*$ which does not satisfy the condition $\hat{f}(\cdot, x^n) \geq 0$. Then it is easy to verify directly that the corrected version of $\hat{f}$ defined by

$$f^+(\cdot, x^n) = \max(0, \hat{f}(\cdot, x^n)) \tag{5}$$

has smaller (or equal) IMSE than $\hat{f}$. Furthermore, the definition of the orthogonal projection prompts

LEMMA 2. *f^+ defined by* (5) *is the orthogonal projection of $\hat{f}$ onto $\mathscr{F}^+$.*

Lemmas 1 and 2 show optimality of Step 2 of the P-Algorithm.

LEMMA 3. *Assume that $\int_S 1/h(t)\,dt < \infty$. If $\hat{f}(\cdot, x^n)$ integrates to a finite constant a.e. $[\lambda^n]$, then*

$$f^1(\cdot, x^n) = \hat{f}(\cdot, x^n) - \left(\int_S \hat{f}(t, x^n)\,dt - 1\right) \Big/ \left[h(\cdot)\int_S 1/h(t)\,dt\right] \tag{6}$$

is the orthogonal projection of $\hat{f}$ onto $\mathscr{F}^1$.

PROOF. Let us notice that f^1 is well defined and for every $f \in \mathscr{F}^1$

$$\begin{aligned}\int[\hat{f}(t, x^n) - f(t)]^2 h(t)\,dt = \int[\hat{f}(t, x^n) - f^1(t, x^n)]^2 h(t)\,dt \\ +2\int[\hat{f}(t, x^n) - f^1(t, x^n)] \\ \times[f^1(t, x^n) - f(t)]h(t)\,dt \\ +\int[f^1(t, x^n) - f(t)]^2 h(t)\,dt.\end{aligned} \tag{7}$$

It follows from (6) that $[\hat{f}(\cdot, x^n) - f^1(\cdot, x^n)]h(\cdot) = \text{const.}$, and since $f^1(\cdot, x^n)$ integrates to one, the second integral on the right-hand side of (7) is equal to 0. Therefore $\|\hat{f} - f^1\|^2 = \inf_{\mathscr{F}^1}\|\hat{f} - f\|^2$ and this completes the proof. □

Let us notice that Lemmas 1–3 imply for every $f \in \mathscr{F}^*$

$$\|\hat{f} - f\|^2 \geq \|f_k - f\|^2 + \sum_{i=1}^{k} \|f_i - f_{i-1}\|^2. \tag{8}$$

Thus (i) follows from (8) and the Fubini theorem.

Now we shall prove (ii). Let us define $S_i(x^n) = \{t \in S | f_i(t, x^n) > 0\}$ and observe that the assumption $\int\hat{f}(t, x^n)\,dt \geq 1$ implies $S_1 \supseteq S_2 \supseteq \cdots$ and therefore $f_i(\cdot, x^n) > 0$ for every $i = 1, 2, \ldots$ on the set $S^*(x^n) = \bigcap_{i=1}^{\infty} S_i(x^n)$. Furthermore, the sequence $f_i(\cdot, x^n)$ is nonincreasing on S^* and therefore it converges pointwise to the following function,

$$f^*(t, x^n) = \begin{cases} \hat{f}(t, x^n) - \alpha/h(t) & \text{for } t \in S^*, \\ 0 & \text{elsewhere,} \end{cases}$$

where α is a constant such that $\int_S f^*(t, x^n)\,dt = 1$. Convergence in norm follows

from the Lebesgue theorem because

$$\int_S [f_i(t, x^n) - f^*(t, x^n)]^2 h(t)\, dt$$

$$\leq \int_{S^*} [f_i(t, x^n) - f^*(t, x^n)]^2 h(t)\, dt$$

$$+ (C_{i-1} - 1)^2 \int_{S - S_i} dt/h(t) \left[\int_S dt/h(t)\right]^{-2}$$

$$+ \int_{S_i - S^*} [f_i(t, x^n)]^2 h(t)\, dt \quad \text{for } i > 1.$$

In order to prove (iii) we show

LEMMA 4. *Under the assumptions of Theorem* 1, f^* *defined by* (4) *is the orthogonal projection of* $\hat{f}$ *onto* $\mathcal{F}^*$.

PROOF. It is easy to verify directly that for every $k = 0, 1, \ldots$

$$\langle f_k - f_{k+1}, f_{k+1} - f^* \rangle = 0,$$

where $f_0 \equiv \hat{f}$. Therefore we can obtain

$$\|\hat{f} - f^*\|^2 = \sum_{i=1}^{k} \|f_i - f_{i-1}\|^2 + \|f_i - f^*\|^2.$$

Hence

$$\|\hat{f} - f^*\|^2 = \sum_{i=1}^{\infty} \|f_i - f_{i-1}\|^2, \tag{9}$$

because $\|f_i - f^*\| \to 0$. On the other hand, (8) implies

$$\|\hat{f} - f\|^2 \geq \|f^* - f\|^2 + \sum_{i=1}^{\infty} \|f_i - f_{i-1}\|^2. \tag{10}$$

The result follows from (9), (10), and Lemma 1. □

To prove (iii) it is sufficient to apply Lemma 4 and the Fubini theorem.

Acknowledgments. The author would like to thank Andrzej Kozek, Jacek Koronacki, and Wolfgang Wertz for helpful remarks. Gratitude is expressed also to the Associate Editor and the referees for detailed comments which considerably improved the presentation of the paper.

REFERENCES

ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10** 1217–1223.

ANDERSON, G. L. and FIGUEIREDO, R. J. P. (1980). An adaptive orthogonal-series estimator for probability density functions. *Ann. Statist.* **8** 347–376.

BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A* **25** 245–254.

KORONACKI, J. (1984). Kernel estimation of smooth densities using Fabian's approach. Preprint 302, Inst. Math. Polish Academy of Sciences.

KRONMAL, R. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63** 925–952.

MÜLLER, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12** 766–774.

NADARAYA, E. A. (1974). On the integral mean square error of some nonparametric estimates of a probability density. *Theory Probab. Appl.* **19** 131–139.

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

ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.

SCHUCANY, W. R. and SOMMERS, J. P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc.* **72** 420–423.

SILVERMAN, B. W. (1982). On the estimator of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation*. Johns Hopkins Univ. Press, Baltimore.

TERRELL, G. R. and SCOTT, D. W. (1980). On improving convergence rates for nonnegative kernel density estimators. *Ann. Statist.* **8** 1160–1163.

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29.

WALTER, G. S. (1977). Properties of Hermite series estimation of probability density. *Ann. Statist.* **5** 1258–1264.

INSTITUTE OF MATHEMATICS
TECHNICAL UNIVERSITY OF ŁÓDŹ
AL. POLITECHNIKI 11
90-924 ŁÓDŹ, POLAND