The Annals of Statistics
1996, Vol. 24, No. 2, 682–686

ON BICKEL AND RITOV'S CONJECTURE ABOUT ADAPTIVE ESTIMATION OF THE INTEGRAL OF THE SQUARE OF DENSITY DERIVATIVE

BY SAM EFROMOVICH[1] AND MARK LOW[2]

University of New Mexico and University of Pennsylvania

Bickel and Ritov suggested an optimal estimator for the integral of the square of the kth derivative of a density when the unknown density belongs to a Lipschitz class of a given order β. In this context optimality means that the estimate is asymptotically efficient, that is, it has the best constant and rate of risk convergence, whenever $\beta > 2k + 1/4$, and it is rate optimal otherwise. The suggested optimal estimator crucially depends on the value of β which is obviously unknown. Bickel and Ritov conjectured that the method of cross validation leads to a corresponding adaptive estimator which has the same optimal statistical properties as the optimal estimator based on prior knowledge of β.

We show for probability densities supported over a finite interval that when $\beta > 2k + 1/4$ adaptation is not necessary for the construction of an asymptotically efficient estimator. On the other hand, it is not possible to construct an adaptive estimator which has the same rate of convergence as the optimal nonadaptive estimator as soon as $k < \beta \le 2k + 1/4$.

1. Introduction. Suppose that $X_1, \dots, X_n$ are iid random variables with probability density $f(x)$ which vanishes beyond the unit interval $[0, 1]$. We assume that f belongs to the Lipschitz class of densities $\mathscr{L}(\beta)$ of order β, that is, $f(x) \ge 0$, $\int_{-\infty}^{\infty} f(x)\,dx = 1$, the mth derivative $f^{(m)}(x)$ satisfies the Hölder condition $|f^{(m)}(x + \varepsilon) - f^{(m)}(x)| \le \pm\varepsilon^{\beta - m}$ where m is the maximal integer which is not greater than β, $f^{(l)}(0) = f^{(l)}(1)$ for $l = 0, \dots, m$ and therefore our class of densities is the particular example of so-called periodic densities; see Devroye and Györfi (1985).

For the problem of estimating the quadratic functional $\theta_k(f) = \int_{-\infty}^{\infty} [f^{(k)}(x)]^2\,dx$ using the observations $X^n = (X_1, \dots, X_n)$, Bickel and Ritov (1988) suggested estimators $\hat{\theta}_{kn}(X^n, \beta)$, which depend on β, and which are asymptotically optimal in the following sense: (i) if $\beta > 2k + 1/4$ and f has $2k$ continuous and periodic derivatives, then the estimator is asymptotically efficient, that is,

$$\lim_{n\to\infty} \sup_{f\in\mathscr{L}(\beta)} \left[nE_f\left\{\left(\hat{\theta}_{kn}(X^n, \beta) - \theta_k(f)\right)^2\right\} - 4\,\mathrm{Var}_f\left(f^{(2k)}(X_1)\right)\right] = 0 \tag{1.1}$$

Received June 1993; revised July 1995.
[1]Supported in part by NSF Grant DMS-91-23956 and Sandia National Laboratories Grant AE-1679.
[2]Supported in part by an NSF Postdoctoral Research Fellowship.
AMS 1991 *subject classifications*. Primary 62C05; secondary 62E20, 62J02, 62G05, 62M99.
Key words and phrases. Functional estimation, adaptation, probability density.

and there are not any estimators with better first-order asymptotics; (ii) if $k < \beta \le 2k + 1/4$ the estimator is asymptotically rate optimal, that is,

$$\sup_{f \in \mathscr{L}(\beta)} n^{8(\beta-k)/(1+4\beta)} E_f\left\{\left(\hat{\theta}_{kn}(X^n, \beta) - \theta_k(f)\right)^2\right\} < C \tag{1.2}$$

and there are not any estimators with a better rate of convergence. Hereafter C's denote positive finite constants.

The estimators constructed by Bickel and Ritov (1988) crucially depend on β. However, Bickel and Ritov (1988) also made the conjecture that using cross validation it might be possible to construct an adaptive estimator which is based only on the observations X^n and which is equivalent to $\hat{\theta}_{kn}(X^n, \beta)$ in the sense of (1.1) and (1.2) when β is known.

In the following section we show that for the case $\beta > 2k + 1/4$ it is not necessary to use an adaptive estimate at all and that a relatively simple projection estimate is always asymptotically efficient. On the other hand, when $k < \beta \le 2k + 1/4$, there do not exist sequences of estimators which satisfy (1.1) and (1.2) simultaneously.

We would like to note that Bickel and Ritov (1988) consider densities on the real line rather than compactly supported densities and thus our setting is a subcase of theirs. Also, the interested reader can find a discussion of optimal adaptive estimation for the irregular case $k < \beta \le 2k + 1/4$ in Efromovich and Low (1996).

2. Main result. Suppose that all the assumptions of the Introduction are valid and that β is unknown but fixed. We first construct an estimator which satisfies (1.1) for all $\beta > 2k + 1/4$ simultaneously. Define the projection estimate

$$\tilde{\theta}_{kn}(X^n) = \delta(k) + 2\left[n(n-1)\right]^{-1} \sum_{1 \le l < s \le n} \sum (2\pi j)^{2k} \varphi_j(X_l)\varphi_j(X_s),$$

where $\delta(k) = 1$ if $k = 0$ and it is equal to 0 otherwise, the second summation is over $1 \le j \le n^{1/(4k+1)}/\ln(n)$ and $\{\varphi_j\}$ is the classical trigonometric Fourier basis over the unit interval $[0, 1]$, that is, $\varphi_0(t) = 1$, $\varphi_{2j-1}(t) = \sqrt{2}\sin(2\pi jt)$, $\varphi_{2j}(t) = \sqrt{2}\cos(2\pi jt)$. The underlying idea of the estimate is based on term-by-term differentiation of a Fourier series expansion of the density function, Parceval's identity and the generalized Lorentz inequality $\sum_{j>t} j^{2k} E_f(\varphi_j(X_1)\varphi_j(X_2)) < Ct^{-2(\beta-k)}$. See, for example, Section 12.5 of Devroye and Györfi (1985). The last inequality also explains our choice of the cutoff sequence $n^{1/(4k+1)}/\ln(n)$ for the projection estimate $\tilde{\theta}_{kn}(X^n)$ which guarantees that the squared bias of the estimate decreases as $o(1)n^{-1}$ as $n \to \infty$ uniformly over $f \in \mathscr{L}(\beta)$ whenever $\beta > 2k + 1/4$.

The reader might notice that a projection pseudoestimator with a cutoff depending on β is well known and has been studied by many authors; see, for example, Donoho and Nussbaum (1990) and Fan (1991).

To estimate the risk of the projection estimate, we use the well-known equality between the mean squared error and the sum of variance and squared bias, that is,

$$E_f\Big\{\big(\tilde\theta_{kn}(X^n) - \theta_k(f)\big)^2\Big\} = \mathrm{Var}_f\big\{\tilde\theta_{kn}(X^n)\big\} + \Big[E_f\big\{\tilde\theta_{kn}(X^n) - \theta_k(f)\big\}\Big]^2. \tag{2.1}$$

Note that the projection estimate $\tilde\theta_{kn}(X^n)$ is the U-statistic with a kernel $\phi_n(X_1, X_2) = \Sigma(2\pi j)^{2k}\varphi_j(X_1)\varphi_j(X_2)$ of degree 2 where the sum is over $1 \le j \le n^{1/(4k+1)}/\ln(n)$. Applying a standard technique for the calculation of the variance of the U-statistic [see Serfling (1980)] and making calculations similar to the proof of Lemma 3 in Efromovich (1985), it is easy to check that $\mathrm{Var}_f\{\tilde\theta_{kn}(X^n)\} = 4\,\mathrm{Var}_f\{f^{(2k)}(X_1)\}n^{-1}(1 + o(1))$ uniformly over $f \in \mathscr{S}(\beta)$. As we mentioned above, the second term on the right-hand side of (2.1) is equal to $o(1)n^{-1}$ uniformly over $f \in \mathscr{S}(\beta)$. Hence, the suggested projection estimate $\tilde\theta_{kn}(X^n)$ is asymptotically efficient, that is, (1.1) is valid with this estimate in place of $\theta_{kn}(X^n, \beta)$.

The situation drastically changes as soon as β can take on values less than or equal to $2k + 1/4$.

PROPOSITION. *Suppose that $\hat\theta_{kn}(X^n)$ is an adaptive estimate such that, for some $\beta > 2k + 1/4$,*

$$\lim_{n\to\infty}\sup_{f\in\mathscr{S}(\beta)}\Big[nE_f\Big\{\big(\hat\theta_{kn}(X^n) - \theta_k(f)\big)^2\Big\} - 4\,\mathrm{Var}_f\big(f^{2k}(X_1)\big)\Big] = 0, \tag{2.2}$$

then for $\beta_k = 2k + 1/4$,

$$\lim_{n\to\infty}\sup_{f\in\mathscr{S}(\beta_k)} nE_f\Big\{\big(\hat\theta_{kn}(X^n) - \theta_k(f)\big)^2\Big\} = \infty. \tag{2.3}$$

Likewise if $\hat\theta_{kn}(X^n)$ is an adaptive estimate such that, for some $\beta > 2k + 1/4$,

$$\lim_{n\to\infty}\sup_{f\in\mathscr{S}(\beta)} nE_f\Big\{\big(\hat\theta_{kn}(X^n) - \theta_k(f)\big)^2\Big\} < \infty, \tag{2.4}$$

then for every $k < \alpha < 2k + 1/4$,

$$\lim_{n\to\infty}\sup_{f\in\mathscr{S}(\alpha)} \big(n^2/\ln(n)\big)^{4(\alpha-k)/(1+4\alpha)} E_f\Big\{\big(\hat\theta_{kn}(X^n) - \theta_k(f)\big)^2\Big\} > 0. \tag{2.5}$$

In other words, there does not exist an adaptive estimator which is equivalent to the nonadaptive optimal estimate $\hat\theta_{kn}(X^n, \beta)$ uniformly over $\beta > k$.

To prove this assertion, we set f_0 to be the density corresponding to the uniform distribution on $[0, 1]$, that is, $f_0(x) = 1$ when $x \in [0, 1]$ and $f_0(x) = 0$ otherwise. Note that for every α we have $f_0 \in \mathscr{S}(\alpha)$. Now, following Ingster (1986), we construct some perturbations of f_0. Let h be a function supported on $[0, 1]$ such that $\int_0^1 h(x)\,dx = 0$, $h \in \mathscr{S}(i)$, $i = k, k + 1, \ldots, 2k + 1$,

$\int_0^1 (h^{(k)}(x))^2\,dx = c > 0$ and $1 + h(x) \geq 0$ for all x. Let v_n be an increasing sequence of integers to be specified later, and write $\bar{a}$ for vectors $\bar{a} = (a_0, \ldots, a_{v_n - 1})$, where each a_i is equal to 1 or -1. Denote by A_n the set of all such distinct vectors. There are 2^{v_n} of them.

Corresponding to each such vector, define

$$f_{\bar{a}}(x) = f_0(x) + \sum_{i=0}^{v_n - 1} a_i v_n^{-\alpha} h(v_n x - i). \tag{2.6}$$

Note that $f_{\bar{a}}$ depends on α and if $k < \alpha \leq 2k + 1/4$, then $f_{\bar{a}} \in \mathscr{L}(\alpha)$. The latter is easy to check by taking the greatest integer less than α derivatives of the function $f_{\bar{a}}$ and using the assumption that $h \in \mathscr{L}(i)$, $i = k, k + 1, \ldots, 2k + 1$.

Then it is easy to check that

$$\theta_k(f_{\bar{a}}) = \theta_k(f_0) + c v_n^{-2(\alpha - k)} \tag{2.7}$$

and that if we define the mixture density

$$\bar{f}_n(X_1, \ldots, X_n) = 2^{-v_n} \sum_{\bar{a} \in A_n} f_{\bar{a}}(X_1) \cdot \cdots \cdot f_{\bar{a}}(X_n),$$

then from the inequality

$$E_{fa}\left\{\left(\hat{\theta}_{kn}(X^n) - \theta_k(f_{\bar{a}})\right)^2\right\} < B, \tag{2.8}$$

which holds for all $\bar{a} \in A_n$, follows

$$E_{\bar{f}n}\left\{\left(\hat{\theta}_{kn}(X^n) - \theta_k(f_{\bar{a}})\right)^2\right\} < B. \tag{2.9}$$

Set $I_n = E_{f_0}\{\bar{f}_n^2(X_1, \ldots, X_n)\} = \mathrm{Var}_{f_0}(\bar{f}_n(X_1, \ldots, X_n) + 1$; I_n is a chi-square type measure which is needed to apply a constrained risk inequality given in Theorem 1 of Brown and Low (1996). This inequality gives a lower bound to the mean squared error at one point given an upper bound at another point. In our context the inequality states that if

$$nE_{f_0}\left\{\left(\hat{\theta}_{kn}(X^n) - \theta_k(f_0)\right)^2\right\} < \varepsilon_n, \tag{2.10}$$

then at the point $\bar{f}_n$,

$$v_n^{4(\alpha - k)} E_{\bar{f}_n}\left\{\left(\hat{\theta}_{kn}(X^n) - \theta_k(f_{\bar{a}})\right)^2\right\}$$
$$\geq v_n^{4(\alpha - k)}\left(\theta_k(f_{\bar{a}}) - \theta_k(f_0)\right)^2\left(1 - 2v_n^{2(\alpha - k)}\left[n^{-1}\varepsilon_n I_n\right]^{1/2}\right).$$

Now, from (2.7) we see that $v_n^{4(\alpha - k)}(\theta_k(f_{\bar{a}}) - \theta_k(f_0))^2 = c^2$ and hence

$$v_n^{4(\alpha - k)} E_{\bar{f}_n}\left\{\left(\hat{\theta}_{kn}(X^n) - \theta_k(f_{\bar{a}})\right)^2\right\} \geq c^2\left(1 - 2v_n^{2(\alpha - k)}\left[n^{-1}\varepsilon_n I_n\right]^{1/2}\right). \tag{2.11}$$

To apply this result for proving the proposition, we have to bound I_n for some specific v_n. This is a relatively complicated procedure but, fortunately, for our setting all necessary computations are made in Ingster (1986). For

$\alpha = \beta_k = 2k + 1/4$ we set $v_n = c_n n^{2/(4\alpha+1)}$, where $c_n \to 0$ as $n \to \infty$, and we get

$$\lim_{n\to\infty} nE_{\bar{f}_n}\left\{\left(\hat{\theta}_{kn}(X^n) - \theta_k(f_{\bar{a}})\right)^2\right\} = \infty, \tag{2.12}$$

which together with (2.8) and (2.9) yields (2.3).

Now suppose that (2.4) is true and that $k < \alpha < 2k + 1/4$. Set v_n to be the smallest integer such that $v_n^{-(4\alpha+1)} \le d \ln(n) n^{-2}$ for some sufficiently small d. Then, again applying Ingster (1986) in an absolutely similar way, the validity of (2.5) follows. The proposition is proved.

Acknowledgment. A conversation with L. Brown is gratefully appreciated.

REFERENCES

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BROWN, L. D. and LOW, M. G. (1993). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* To appear.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation*. Wiley, New York.

DONOHO, D. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

EFROMOVICH, S. (1985). Nonparametric estimation of a density with unknown smoothness. *Theory Probab. Appl.* **30** 557–568.

EFROMOVICH, S. and LOW, M. (1996). On optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **24** (3).

FAN, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294.

INGSTER, YU. V. (1986). On minimax testing of statistical hypotheses for probability distributions in L_p. *Theory Probab. Appl.* **31** 384–389.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 17131
E-mail: efrom@math.unm.edu

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
E-mail: mlow@stat.wharton.upenn.edu