The Annals of Statistics
1984, Vol. 12, No. 3, 880–886

ADAPTIVE DENSITY FLATTENING—A METRIC DISTORTION PRINCIPLE FOR COMBATING BIAS IN NEAREST NEIGHBOR METHODS[1]

BY IAN S. ABRAMSON

University of California at San Diego

With a wide variety of approaches to density estimation, it is profitable to perturb the data so as to make 2nd order derivatives of their density vanish. An adaptive transformation to local uniformity for instance will (for unchanged variance) lower bias to a vanishing fraction of what a Rosenblatt-Parzen or nearest neighbor estimator on the raw data yields; fractional pilot sampling, a common technical device of little practical appeal, can be shown by an embedding argument to be dispensable. An upshot is that MSE can be lowered by attacking the variance directly through extra smoothing, without the usual penalty from inflated bias.

1. Introduction. Consider a nonparametric density estimation problem. Many estimator forms and loss conventions indicate somehow balancing (through a smoothing parameter) two measures: one of bias and one of variability. This can impose a disappointing ceiling on the performance available; certainly error rates fall short of the regular parametric ones.

Mack and Rosenblatt (1979) point out that bias decay can be particularly slow in regions of low density or in high dimensional problems, and the practitioner's experience bears this out; high dimensional spaces are hard to sample representatively with samples of familiar sizes. Friedman (1981) spoke of a tendency rather for the near neighbors of a point to string themselves out along the line of steepest ascent of the density.

A transformation to uniformity is at the heart of our proposal; it springs from an observation that for many approaches to the problem, the bias is locally driven by the curvature of the density (or a combination of second partials when working in several dimensions).

The method is not claimed to bring about any strong uniform optimality (in the sense of Stone, 1980, say). Simply, given a user's naive procedure of a rather general kind, fully specified even down to choice of smoothing parameter, this data transformation will improve performance by reducing bias. Some degree of extra smoothing could then lower variability too, but we advance no rigorous guidelines for doing this optimally.

The nearest neighbor framework makes a particularly natural one for applying

Received October 1982; revised March 1984.

[1] Research supported by Regents Faculty Research Fellowship, U.C.S.D. and N.I.H. Grant PHSCA 26666-03.

AMS 1980 *subject classifications.* Primary 62G05, 62G20; secondary 62G99.

Key words and phrases. Bias reduction, density flattening and straightening, nearest neighbor and kernel estimates, metric distortion, probability integral transform, adaptation, fractional sampling, 2-pass method, tightness in *C*.

these ideas—it enjoys a certain closure property when the transformation and estimation phases are put in tandem—the notion of metric distortion is introduced for this purpose.

The theoretical aspects are complicated by admitting two-pass methods—reusing the data on which the adaptation is based. Path analysis of an error process resolves the difficulties, but our findings are incomplete for higher dimensions.

Work is underway on how analogous ideas can indicate optimal design transformations in regression studies.

2. The problem and the transformation. A Lebesgue density f on $\mathbb{R}^p$ gives rise to a sample $X_1, \cdots, X_n$ on which to base a point estimate of f at 0, say; we impose familiar local smoothness requirements on f, viz. continuous second order partial derivatives $D_{jk}f$ near 0.

Assume $|D_{jk}f(x)| \leq U_2$ and $f(0) \geq L_0$ where U_2 and L_0 are positive user-chosen constants. This implicitly defines upper bounds U_0 and U_1 on $f(x)$ and the $|D_jf(x)|$ respectively. Call the Sobolev-like class of permissible densities $\mathscr{S}_p''(U_2, L_0)$, or simply $\mathscr{S}_p''$.

We adopt (with little loss) squared error loss at 0. For a wide variety of popular estimation methods, each indexed by some "smoothing parameter" λ say, we can abstract a common form for the MSE:

$$E[f_n(0) - f(0)]^2 = A^2(f(0))\sigma^2(n, \lambda) + B^2(\{D_{jk}f(0)\})\beta^2(n, \lambda) + o(\sigma^2(n, \lambda)) + o(\beta^2(n, \lambda)) \quad \text{as} \quad n \to \infty, \tag{1}$$

a sum of a variance and a squared bias. Moreover, the functional B^2 is a quadratic form in $\{D_{jk}f(0)\}$, vanishing when they do, and the indexing by λ may be arranged to make σ^2 decrease and β^2 increase in λ, as if, for instance, λ were a kernel bandwidth.

Wahba (1975) has assembled results to this effect in one dimension at least, for spline based methods, orthogonal series estimators, and kernel estimators of the Rosenblatt-Parzen type. Nearest neighbor methods share the property too, as we see.

Asymptotic minimization of (1) in λ leads to a balancing of decay rates in $\sigma^2(n, \lambda)$ and $\beta^2(n, \lambda)$, determining a dependence of λ on n up to a proportion; the optimal multiplier depends on f, but two-stage adaptations can often be justified (Woodroofe, 1970; Krieger and Pickands, 1981; Abramson, 1982a).

While the variance term in (1) seems generally unassailable, there are several ways of eliminating the bias term. This allows balance to be struck at a smaller value of the variance, but not knowing the precise rate in the remainder $o(\beta^2(n, \lambda))$ may prevent an improvement uniform over $\mathscr{S}_p''$.

Two such methods, which are documented, are specific to kernel methods: the technique of bandwidth variation (Abramson, 1982a) is one; the other requires vanishing second moments of the kernel, and, untruncated, entails nonpositive curve estimates. Density flattening is a third approach, and the focus of this

paper. The rationale is as follows:

An estimator with MSE properties (1) would have bias $o(\beta(n, \lambda))$ if applied to data without local curvature in their density function. Restricting ourselves to one dimension, consider an approximate probability integral transformation of the $\{X_i\}$; i.e., if g is a pilot estimate of f near 0 (consistency requirements to be imposed as necessary), define

$$G(x) = \int_0^x g(\xi)\, d\xi$$

an estimate of the cumulative distribution $F(x)$, but shifted for convenience to fix 0.

Let $$Y_i = g(0)^{-1}G(X_i); \quad i = 1, \cdots, n.$$

$\{Y_i\}$ is a sample from an approximately uniform distribution.

If $\{X_i\}$ is independent of the pilot data used to construct g, then conditionally on the pilot sample, the $\{Y_i\}$ are distributed according to density

$$u(y) = g(0)g(G^{-1}(g(0)y))^{-1}f(G^{-1}(g(0)y)) = (\text{at } y = 0) \quad f(0) \quad \text{exactly.}$$

We now estimate this quantity by sending the $\{Y_i\}$ through the original routine, to which their near uniformity tailors them particularly well. This estimate, $\hat{f}(0)$ say, is our proposal for a refined estimate of $f(0)$. There is a version of the theorem below, which would assert its superiority, but in the interest of simplicity, we formulate it for a less transparently chosen transformation—one that achieves a vanishing second derivative without constraining the first, or a density straightener rather than a flattener.

We first introduce the following notion to keep calculations tidy.

DEFINITION. A sequence of estimators T_n based on samples of size n is said to be determined on b_n-neighborhoods of 0 iff there exists $k > 0$ such that $T_n[x_1, \cdots, x_n] = (1/n) \sum_{i=1}^n t_n(x_i)$ with $t_n(x)$ supported on $|x| \leq kb_n$.

In the sequel, we take $b_n t_n$ and $b_n^3 t_n''$ to be even and bounded. Regular delta-type estimates are the natural examples, but other nonparametric estimators will usually acquire the property on an innocuous modification.

THEOREM. *Let $T_n: \mathbb{R}^n \to [L_0, U_0]$ be a sequence of functions. Suppose there exist $A: [L_0, U_0] \to [0, \infty)$, $B \geq 0$, $\beta(n) \to 0$ with $n\sqrt{\beta(n)} \to \infty$, such that:* (i) *$T_n$ is determined on $\sqrt{\beta(n)}$-neighborhoods of* 0, (ii) *whenever $x_1, \cdots, x_n$ is a random sample from a density $h \in \mathscr{S}_1''$,*

$$\limsup_{n\to\infty} \beta(n)^{-1} |ET_n[x_1, \cdots, x_n] - h(0)| \leq B|h''(0)|$$

and

$$\limsup_{n\to\infty} \beta(n)^{-2} \operatorname{var} T_n[x_1, \cdots, x_n] \leq A(h(0))^2.$$

Let $f \in \mathscr{S}_1''$, and θ_n be a bounded consistent sequence of estimators of $\theta =$

$\frac{1}{6}f''(0)/f(0)$. *Independently, let* $X_1, X_2, \cdots \sim f$ *i.i.d. For each* n, *define* Y_{ni} *by* $X_i = Y_{ni} - \theta_n Y_{ni}^3$, $i \leq n$. *Then* $\limsup_{n\to\infty}\beta(n)^{-1} \mid ET_n[Y_{n1}, \cdots, Y_{nn}] - f(0) \mid = 0$, *and* $\limsup_{n\to\infty}\beta(n)^{-2}\text{var } T_n[Y_{n1}, \cdots, Y_{nn}] \leq A(f(0))^2$.

PROOF. Let $T_n[x_1, \cdots, x_n] = (1/n)\sum t_n(x_i)$ with t_n supported in $[-k\beta(n)^{1/2}, k\beta(n)^{1/2}]$. Define a transformation $D_t\colon \mathbb{R} \to \mathbb{R}$ implicitly, by $x = D_t(x) - tD_t(x)^3$; t real. D_t is invertible, and $Y_{ni} = D_{\theta_n}(x_i)$; $i = 1, \cdots, n$. Let $u_t(y)$ be the density of $D_t(x_t)$, so that verifiably, $u_t(y) = f(y - ty^3) \mid 1 - 3ty^2 \mid$, and $u_\theta''(0) = 0$. Letting $\mathcal{F}_n$ denote the σ-field of θ_n, the absolute conditional bias of $T_n[Y_{n1}, \cdots, Y_{nn}]$ may be written $\mid E_{\mathcal{F}_n} t_n(Y_{n1}) - f(0) \mid$, and bounded above by $\mid Et_n(Y_1) - f(0) \mid + \mid E_{\mathcal{F}_n}(t_n(Y_{n1}) - t_n(Y_1)) \mid$, where Y_1 denotes the unobservable $D_\theta(X_1)$. The second term is at most

$$\left| \int_{|y|\leq k\beta(n)^{1/2}} t_n(y)(u_{\theta_n}(y) - u_\theta(y))\, dy \right|,$$

and since u_{θ_n} and u_θ agree at the origin, and t_n is an even function, we have in turn a bound of

$$\frac{1}{2}\left| \int_{|y|\leq k\beta(n)^{1/2}} t_n(y)y^2(u_{\theta_n}''(\varepsilon_n y) - u_\theta''(\varepsilon y))\, dy \right|$$

where $\mid \varepsilon_n \mid, \mid \varepsilon \mid \leq 1$, but vary with y, and are random through θ_n. By explicitly computing the second derivatives appearing, we readily find a constant $M > 0$ so that

$$\mid E_{\mathcal{F}_n} T_n[Y_{n1}, \cdots, Y_{nn}] - f(0) \mid \leq \mid Et_n(Y_1) - f(0) \mid + M\beta(n) \mid \theta_n - \theta \mid.$$

Finally,

$$\begin{aligned}
&\limsup_{n\to\infty}\beta(n)^{-1} \mid ET_n[Y_{n1}, \cdots, Y_{nn}] - f(0) \mid \\
&\qquad \leq \limsup_{n\to\infty}\beta(n)^{-1}E \mid E_{\mathcal{F}_n} T_n[Y_{n1}, \cdots, Y_{nn}] - f(0) \mid \\
&\qquad \leq \limsup_{n\to\infty}\beta(n)^{-1}[\mid Et_n(Y_1) - f(0) \mid + EM\beta(n) \mid \theta_n - \theta \mid] \\
&\qquad \leq \limsup_{n\to\infty}\beta(n)^{-1} \mid Et_n(Y_1) - f(0) \mid + M \lim_{n\to\infty} E \mid \theta_n - \theta \mid \\
&\qquad = 0, \quad \text{as required.}
\end{aligned}$$

As for the variance,

$$\begin{aligned}
&\limsup_{n\to\infty}\beta(n)^{-2}\text{var } T_n[Y_{n1}, \cdots, Y_{nn}] \\
&\qquad \leq \limsup_{n\to\infty}\beta(n)^{-2}E \text{ var}_{\mathcal{F}_n} T_n[Y_{n1}, \cdots, Y_{nn}] \\
&\qquad\quad + \limsup_{n\to\infty}\beta(n)^{-2}\text{var } E_{\mathcal{F}_n} T_n[Y_{n1}, \cdots, Y_{nn}].
\end{aligned}$$

The second term is no larger than the limiting expected square of the scaled conditional bias, which is zero as before; conditional independence among $\{Y_{ni}\}$ allows the first term to be written as

$$\limsup_{n\to\infty} n^{-1}\beta(n)^{-2}E \text{ var}_{\mathcal{F}_n}[t_n(Y_1) + (t_n(Y_{n1}) - t_n(Y_1))],$$

and since Y_1 has density $f(0)$ at 0, it is thus sufficient to show that

$$\lim \sup_{n\to\infty} n^{-1}\beta(n)^{-2}E E_{\mathcal{F}_n}(t_n(Y_{n1}) - t_n(Y_1))^2 = 0.$$

The conditional expectation may be written in the form

$$(\theta_n - \theta)^2 \int_{-2k\beta(n)^{1/2}}^{2k\beta(n)^{1/2}} x^2 t_n''(D_{\tilde{\theta}_n}(\varepsilon_n x))^2 \frac{\partial}{\partial\theta} D_{\tilde{\theta}_n}(x)^2 \, dx$$

where $\tilde{\theta}_n = \tilde{\theta}_n(x)$ lies between θ and θ_n, and $|\varepsilon_n| \leq 1$ as before.

From the nature of the transformation D_t, and the bound laid down on $\beta(n)^{3/2}t_n''(\cdot)$, this is bounded in turn by a multiple of $(\theta_n - \theta)^2 n\beta(n)^{3/2}$, and mean square convergence of θ_n to θ completes the proof.

3. Nearest neighbors. When T_n has the kernel format, the transformation involved destroys one of the appealing features of the methods: the curve contributions of which $\hat{f}_n$ is a pointwise average are no longer identical, even up to scale. We can argue, however, that nearest neighbor methods present no such drawback. (Loftsgaarden and Quesenberry, 1965, gave a prototype without the option to let the influence of the more distant near neighbors fade away rather than be truncated at a sphere boundary.) Moore and Yackel (1977) proposed an estimator

$$f_n(0) = n^{-1}R_n^{-p} \sum_{j=1}^n w(R_n^{-1}X_j),$$

where R_n denotes the Euclidean distance to the kth nearest neighbor of 0, and w is an even kernel function which we take to be compactly supported. Mack and Rosenblatt (1979) investigated the asymptotic properties of $f_n(0)$.

We make a natural extension by observing that there is nothing sacrosanct about the Euclidean metric, and admitting a locally invertible componentwise distortion of the data:

$$X_j \mapsto d(X_j) = [d_1(X_{1j}) \cdots d_p(X_{pj})]^T$$

where X_{rj} denotes the rth component of X_j, and $d_r(0) = 0$; $d_r'(0) = 1$, for close local agreement with the original data.

The metric by which we gauge and order the nearest neighbor distances is then Euclidean on the transformed data, and setting $R_{d,n} = k$th such distance, we define

$$\hat{f}_n(0) = n^{-1}R_{d,n}^{-p} \sum_{j=1}^n w(R_{d,n}^{-1}d(X_j)).$$

Mack and Rosenblatt (1979) have found the MSE for the Moore-Yackel form, and our extension is narrow enough for their findings to carry over:

THEOREM. *In the present context, the bias and variance of* $\hat{f}_n(0)$ *are given by*:

$$\begin{aligned}\text{Bias}(\hat{f}_n(0)) &= (2\pi)^{-1}\Gamma(1 + p/2)^{2/p}f(0)^{-2/p}Q[f_d](0)(k/n)^{2/p} \\ &\quad + \pi^{p/2}\Gamma(1 + p/2)^{-1}f(0) \int_{\|u\|=1} w(u) \sum (du)k^{-1} \\ &\quad + o((k/n)^{2/p}) + o(k^{-1}),\end{aligned}$$

where $\sum$ *is uniform distribution on the surface of the unit p-sphere, and* f_d *is the density of the transformed data* $\{d(X_j)\}$;

$$\operatorname{Var}(\hat{f}_n(0)) = \pi^{p/2}\Gamma(1 + p/2)^{-1}f(0)^2 \int_{\mathbb{R}^p} w(u)^2\, du\, k^{-1} + o(k^{-1}).$$

For proof we note merely that the transformed data has density given by

$$f_d(y) = f(d^{-1}(y)) \left| \det\left(\left(\frac{\partial d(d^{-1}(y))}{\partial y}\right)\right) \right|$$

$$= f(d^{-1}(y))(1 + o(1)) \quad \text{as} \quad \| y \| \to 0,$$

agreeing as needed with $f(0)$ at 0.

The results are immediate from the corresponding expressions of Mack and Rosenblatt (1979).

The smoothing parameter (earlier λ) is of course k here. There is an unfamiliar term in the bias which stems from forcing the kth nearest neighbor to lie exactly on the skin of the near neighbor sphere, but the MSE expansion will still conform to our prototype (1), so long as the other term dominates, i.e. $k^{-1} = o((k/n)^{2/p})$, or $k^{-1} = o(n^{-2/(p+2)})$. Violations of this may be excluded without loss, since generally, $k^{-1} \propto n^{-4/(p+4)}$ gives the optimal rate.

In one dimension, $f_n(0)$ and $\hat{f}_n(0)$ will be related to each other as $T_n[X_1, \cdots, X_n]$ and $T_n[Y_{n1}, \cdots, Y_{nn}]$ in our theorem, provided d is a density straightening distortion in the sense that $y - \theta y^3 \mapsto y$ was one in that context.

4. On a two-pass method with matching asymptotic performance. The apparent need for an independent fractional sample to implement our proposal is a shortcoming, but to admit arbitrary consistent forms for the pilot estimate raises some daunting technical difficulties, and fully relaxing the independence assumption seems to necessarily entail a weakened conclusion. By modifying the notion of risk, the problem of justifying data reuse has been solved in one dimension by path analysis of an error process. Details are available from the author.

REFERENCES

ABRAMSON, I. S. (1982a). Arbitrariness of the pilot estimator in adaptive kernel methods. *J. Multivariate Anal.* **12** 562–567.

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

BICKEL, P. and WICHURA, M. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670.

CACOULLOS, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.* **18** 179–189.

FRIEDMAN, J. H. (1981). Lecture course, U.C. Berkeley.

KRIEGER, A. M. and PICKANDS, J., III. (1981). Weak convergence and efficient density estimation at a point. *Ann. Statist.* **9** 1066–1078.

LOFTSGAARDEN, D. O., and QUESENBERRY, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36** 320–326.

MACK, Y. P. and ROSENBLATT, M. (1979). Multivariate k-nearest neighbor density estimates. *J. Multivariate Anal.* **9** 1–15.

MOORE, D. S. and YACKEL, J. W. (1977). Consistency properties of nearest neighbor density estimates. *Ann. Statist.* **5** 143–154.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

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

WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.* **41** 1665–1671.

DEPARTMENT OF MATHEMATICS C-012
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093