AN ASYMPTOTICALLY OPTIMAL HISTOGRAM SELECTION RULE[1]

BY

CHARLES J. STONE

TECHNICAL REPORT NO. 34

JUNE 1984

[1]RESEARCH PARTIALLY SUPPORTED BY
NATIONAL SCIENCE FOUNDATION GRANT MCS83-01257

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

Par
ana
QA :
A (
Th
Ne
M,

AN ASYMPTOTICALLY OPTIMAL HISTOGRAM SELECTION RULE[1]

by Charles J. Stone

University of California, Berkeley

July, 1983; Revised June, 1984

Abstract. A random sample is available from a multivariate distribution having a bounded density, which is assumed to satisfy a mild additional condition. A finite collection of histogram estimates of the unknown density is constructed, whose cardinality increases algebraically fast with respect to the size of the random sample. A histogram selection rule is introduced, which is shown to be asymptotically optimal relative to integrated squared error loss.

[1]This research was supported in part by National Science Foundation Grant MCS83-01257.

AMS 1980 subject classifications. Primary 62G30; secondary 62G99.

Key words and phrases. Density estimation, histogram, selection rule, asymptotic efficiency.

<u>1. Statement of the main result</u>. Let $X_1, X_2, \ldots$ be independent $\mathbb{R}^d$-valued random variables having common absolutely continuous distribution P with bounded density p. Let P_n denote the empirical distribution of $X_1, \ldots, X_n$, defined by

$$P_n(A) = \frac{1}{n}\#\{i: 1 \leq i \leq n \text{ and } X_i \in A\} .$$

Let $\mathbb{R}_+^d$ denote the collection of d-tuples of positive numbers. Choose $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $b = (b_1, \ldots, b_d) \in \mathbb{R}_+^d$; set $h = (a,b)$. Consider the <u>histogram estimate</u> p_{nh} of p defined as follows: Let $\ell = (\ell_1, \ldots, \ell_d)$ denote an arbitrary d-tuple of integers. Set

$$I_{h\ell} = \mathop{\times}_1^d [a_j + (\ell_j - 1)b_j,\ a_j + \ell_j b_j) .$$

Each d-dimensional interval $I_{h\ell}$ has volume $v_h = \Pi_1^d b_j$; the collection of all such intervals forms a partition of $\mathbb{R}^d$. Finally, set

$$p_{nh} = \frac{P_n(I_{h\ell})}{v_h} \text{ on } I_{h\ell} .$$

(See page 21 of Kendall and Stuart, 1977, for a picture of a bivariate histogram based on a sample of size $n = 9{,}440$.) The <u>integrated squared error loss</u> of p_{nh} as an estimate of p is given by

$$L_{nh} = \int (p_{nh} - p)^2 = \frac{1}{v_h} \sum_\ell P_n^2(I_{h\ell}) - \frac{2}{v_h} \sum_\ell P_n(I_{h\ell}) P(I_{h\ell}) + \int p^2 .$$

Let H_n denote a finite subset of $\mathbb{R}^d \times \mathbb{R}_+^d$ whose cardinality increases algebraically fast with n; that is, $\lim_n n^{-c}\#(H_n) = 0$ for some $c > 0$. A <u>histogram selection rule</u> h_n is an H_n-valued function of $X_1, \ldots, X_n$. Clearly

$$\frac{L_{nh_n}}{\min_h L_{nh}} \geq 1 ;$$

here it is understood that $h \in H_n$. The selection rule h_n is said to be <u>asymptotically</u> <u>optimal</u> if

$$\lim_n \left[\frac{L_{nh_n}}{\min_h L_{nh}}\right] = 1 \quad \text{with probability one .}$$

Set

$$K_{nh} = \frac{1}{v_h}\left(\frac{2}{n} - \sum_\ell P_n^2(I_{h\ell})\right)$$

(see Section 2 for motivation). Let $\hat{h}_n$ be a value of h that minimizes K_{nh}. It will be shown below, under a mild condition on p, that the histogram selection rule $\hat{h}_n$ is asymptotically optimal.

CONDITION 1. There are positive constants α and β such that $\int (p_h - p)^2 \geq \alpha(v_h^\beta \wedge 1)$ for $n \geq 1$ and $h \in H_n$.

Here $s \wedge t = \min(s,t)$. Condition 1 is satisfied if, say, there is some nonempty open subset of $\mathbb{R}^d$ on which the derivative of p exists and is continuous and nonzero. For an alternative set of assumptions which guarantees that this condition is satisfied, at least when $d = 1$, see Freedman and Diaconis (1981).

THEOREM 1. *If Condition* 1 *holds, then* $\hat{h}_n$ *is asymptotically optimal.*

For other theoretical results on the selection of a histogram see Freedman and Diaconis (1981); Chow, Geman and Wu (1981, 1983); and Burman (1984). For an analogous result on kernel density estimates see Stone (1984). The latter two papers were written after the original version of this paper.

2. Motivation for K_n. Ideally, h should be chosen to minimize

$$L_{nh} - \int p^2 = \frac{1}{v_h} \sum_\ell P_n^2(I_{h\ell}) - \frac{2}{v_h} \sum_\ell P_n(I_{h\ell})P(I_{h\ell}) ,$$

but the quantity $P(I_{h\ell})$ is unknown. The estimate $P_n(I_{h\ell})$ of $P(I_{h\ell})$ leads to the biased estimate $P_n^2(I_{h\ell})$ of $P_n(I_{h\ell})P(I_{h\ell})$. It is easily checked that

$$\frac{n}{n-1} P_n^2(I_{h\ell}) - \frac{P_n(I_{h\ell})}{n-1}$$

is an unbiased estimate of $P_n(I_{h\ell})P(I_{h\ell})$; that is,

$$E\left[\frac{n}{n-1} P_h^2(I_{h\ell}) - \frac{P_n(I_{h\ell})}{n-1}\right] = E[P_n(I_{h\ell})P(I_{h\ell})] = P^2(I_{h\ell}) .$$

This leads to the following histogram selection rule: choose h to minimize

$$K'_{nh} = \frac{1}{v_h} \sum_\ell P_n^2(I_{h\ell}) - \frac{2}{v_h} \sum_\ell \left[\frac{n}{n-1} P_n^2(I_{h\ell}) - \frac{P_n(I_{h\ell})}{n-1}\right]$$

$$= \frac{1}{v_h} \left(\frac{2}{n-1} - \frac{n+1}{n-1} \sum_\ell P_n^2(I_{h\ell})\right) .$$

An inessential simplifying approximation leads to the formula for K_{nh} given in Section 1. For an alternative motivation in terms of cross-validation see Rudemo (1982).

3. Proof of Theorem 1. Recall that p is assumed to be bounded and that the cardinality of H_n increases algebraically fast with n. Define the density p_h on $\mathbb{R}^d$ by $p_h(x) = P(I_{h\ell})/v_h$ for $x \in I_{h\ell}$. Set

$$G_{nh} = \frac{1}{n} \sum_1^n p_h(X_i) - Ep_h(X) ,$$

$$G_n = \frac{1}{n} \sum_1^n p(X_i) - Ep(X) ,$$

$$J_{nh} = \int (p_h - p)^2 + \frac{1}{nv_h} ,$$

and

$$J_{nhr} = v_h^r \wedge 1 + \frac{1}{nv_h} \quad \text{for} \quad r > 0 .$$

LEMMA 1. *If Condition* 1 *holds, then* $\lim_n \max_h \frac{|G_{nh} - G_n|}{J_{nh}} = 0$ *with probability one.*

LEMMA 2. *For all* $r > 0$

$$\lim_n \max_h \frac{1}{J_{nhr}} \left| \int (p_{nh} - p_h)^2 - \frac{1}{nv_h} \right| = 0 \quad \textit{with probability one.}$$

The proofs of these two lemmas will be given at the end of the paper.

To prove that $\hat{h}_n$ is asymptotically optimal it suffices to show that

$$\lim_n \max_{h,h'} \frac{|L_{nh'} - L_{nh} - (K_{nh'} - K_{nh})|}{L_{nh} + L_{nh'}} = 0 \quad \text{with probability one.} \tag{1}$$

To verify (1) it suffices to show that

$$\inf_n \min_h \frac{L_{nh}}{J_{nh}} > 0 \quad \text{with probability one} \tag{2}$$

and

$$\lim_n \max_{h,h'} \frac{|L_{nh'} - L_{nh} - (K_{nh'} - K_{nh})|}{J_{nh} + J_{nh'}} = 0 \quad \text{with probability one .} \tag{3}$$

Observe that

$$L_{nh} = \int (p_{nh} - p)^2 = \int (p_{nh} - p_h)^2 + \int (p_h - p)^2 .$$

It now follows easily from Condition 1 and Lemma 2 that (2) holds.

By elementary algebra

$$L_{nh} - K_{nh} - 2G_n - \int p^2 = 2(G_{nh} - G_n) + 2\int (p_{nh} - p_h)^2 - \frac{2}{nv_h} .$$

It now follows easily from Lemma 1 and Lemma 2 that (3) holds.

Thus the proof of Theorem 1 is complete once the two lemmas are verified. To prove Lemma 1 write

$$G_{nh} - G_n = \frac{1}{n}\sum_1^n Z_{ih} = \overline{Z}_{nh} ,$$

where

$$Z_{ih} = p_h(X_i) - p(X_i) - E(p_h(X_i) - p(X_i)) .$$

Then Z_{ih}, $i \geq 1$, are independent and identically distributed random variables having mean zero. Since p is bounded, there is a positive constant c independent of h such that $|Z_{ih}| \leq c$ and $\mathrm{Var}(Z_{ih}) \leq cu_h^2$, where $u_h^2 = \int(p_h-p)^2$. By Bernstein's inequality (see Hoeffding, 1963)

$\Pr(|\overline{Z}_{nh}| \geq t) \leq 2\exp[-\tau\lambda/2(1+\lambda/3)]$, where $0 \leq \lambda \leq t/u_h^2$ and $\tau = nt/c$. Choose $\varepsilon > 0$. Suppose that $u_h \geq n^{\varepsilon-\frac{1}{2}}$. Set $t = n^{\varepsilon-\frac{1}{2}}u_h$ and $\lambda = n^{\varepsilon-\frac{1}{2}}/u_h \leq 1$. Then $\lambda\tau = n^{2\varepsilon}/c$. Suppose instead that $u_h < n^{\varepsilon-\frac{1}{2}}$. Set $t = n^{2\varepsilon-1}$ and $\lambda = 1$. Again, $\lambda\tau = n^{2\varepsilon}/c$. Thus in either case it follows from Bernstein's inequality that

$$\Pr(|\overline{Z}_{nh}| \geq t) \leq 2\exp(-n^{2\varepsilon}/3c) .$$

Consequently

$$\lim_n \Pr(|\overline{Z}_{nh}| \geq n^{\varepsilon-\frac{1}{2}}u_h + n^{2\varepsilon-1} \text{ for some } h \in H_n) = 0 .$$

Thus to verify Lemma 1 it is enough to show that for some $\varepsilon > 0$

$$\lim_n \max_{u>0} \frac{n^{\varepsilon-\frac{1}{2}}u + n^{2\varepsilon-1}}{u^2 + 1/nu^{2/\beta}} = 0 ,$$

where β is from Condition 1. For $0 < \varepsilon < 1/2(1+\beta)$, this result is easily shown by considering separately: $0 < u \leq n^{\varepsilon-\frac{1}{2}}$, $n^{\varepsilon-\frac{1}{2}} < u < n^{-\beta/2(1+\beta)}$, and

$u > n^{-\beta/2(1+\beta)}$.

The simplest way to prove Lemma 2 is by means of the technique called "Poissonization." It was used by Rosenblatt (1975) in a related context.

LEMMA 3. *Let* N_ℓ *be independent Poisson random variables with mean* λ_ℓ *such that* $0 < \lambda = \sum_\ell \lambda_\ell < \infty$. *Set* $N = \sum_\ell N_\ell$, $P_\ell = \lambda_\ell/\lambda$ and $\overline{P} = \max_\ell P_\ell$. *For each positive integer* k *there is a finite positive universal constant* c_k *such that*

$$E[(\sum_\ell (N_\ell - NP_\ell)^2 - N)^{2k}] \leq c_k(\lambda + \lambda^k + \lambda^{2k}\overline{P}^k) .$$

This lemma follows in a straightforward manner from properties of cumulants summarized in Gnedenko and Kolmogorov (1954) or Kendall and Stuart (1977). (Observe that $E[(N-\lambda)^{2k}]$ is a polynomial in λ of degree k with zero constant term. The next step is to prove the desired conclusion with N replaced by λ.)

Set $\tau = \sup p$ and $N_n(I_{h\ell}) = nP_n(I_{h\ell})$.

LEMMA 4. *For each positive integer* k *there is a universal constant* c_k' *such that* $E[(\sum_\ell (N_n(I_{h\ell}) - nP(I_{h\ell}))^2 - n)^{2k}] \leq c_k' n^k(1 + (n\tau v_h)^k)$.

PROOF. Let μ_n denote the $2k^{th}$ moment of

$$Z = \sum_\ell (N_n(I_{h\ell}) - nP(I_{h\ell}))^2 - n$$

and set $\mu_0 = 0$. Let $R(\lambda)$ denote the $2k^{th}$ moment of the random variable obtained through replacing n in the definition of Z by a Poisson number N having mean λ. Then

$$R(\lambda) = \sum_n \Pr(N=n)\mu_n = \sum_n \frac{\lambda^n}{n!} e^{-\lambda}\mu_n .$$

According to Lemma 3 and the well known connection between multinomial and independent Poisson random variables, $R(\lambda)$ is a polynomial of degree $2k$ in λ and

$$0 \leq \sum_{j=1}^{2k} \frac{R^{(j)}(0)}{j!} \lambda^j = R(\lambda) \leq c_k(\lambda+\lambda^k+\lambda^{2k}(\tau v_h)^k) \quad \text{for} \quad \lambda \geq 0 .$$

Thus there is a finite positive universal constant c_k'' such that

$$\sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} \lambda^j \leq c_k''(\lambda+\lambda^k+\lambda^{2k}(\tau v_h)^k) \quad \text{for} \quad \lambda \geq 0 .$$

(For suppose otherwise and note that for each fixed $c > 0$, if $\lambda > 0$ and

$$\frac{|R^{(j)}(0)|}{j!} \lambda^j \gg c_k(\lambda+\lambda^k+\lambda^{2k}(\tau v_h)^k) ,$$

then

$$\frac{|R^{(j)}(0)|}{j!} (c\lambda)^j \gg \sum_{j=1}^{2k} \frac{R^{(j)}(0)}{j!} (c\lambda)^j \geq 0 ;$$

by a compactness argument, there would then be a nonzero polynomial in c of degree $2k$ that equals zero at more than $2k$ distinct points.) Consequently,

$$\mu_n = \sum_{j=1}^{2k} \frac{n!R^{(j)}(0)}{(n-j)!j!} \leq \sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} n^j \leq c_k''(n+n^k+n^{2k}(\tau v_h)^k) \quad \text{for} \quad n \geq 0 ,$$

which yields the desired result.

To prove Lemma 2, observe that by Lemma 4 and Chebyshev's inequality,

$$\lim_n \max_h \frac{|\sum_\ell (P_n(I_{h\ell}) - P(I_{h\ell}))^2 - \frac{1}{n}|}{n^{\varepsilon-1}(v_h^{1/2} + n^{-1/2})} \overset{a.s.}{=} 0 \quad \text{for} \quad \varepsilon > 0 .$$

Let $r > 0$ be fixed. It is easily seen that for sufficiently small $\varepsilon > 0$,

$$\lim_n \max_v \frac{n^{\varepsilon-1}(v^{1/2} + n^{-1/2})}{v(v^r + \frac{1}{nv})} = 0 .$$

The desired conclusion follows from these two observations.

REFERENCES

BURMAN, P. (1984). A data dependent approach to density estimation, manuscript.

CHOW, Y.-S., GEMAN, S. and WU, L. D. (1981). Consistent cross-validated density estimation. *Reports in Pattern Analysis*, No. 110, Division of Applied Mathematics, Brown University.

CHOW, Y.-S., GEMAN, S. and WU, L. D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25-38.

FREEDMAN, D. and DIACONIS, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheorie und ver. Gebiete* **57** 453-476.

GNEDENKO, B. V. and KOLMOGOROV, A. N. (1954). *Limit Theorems for Sums of Independent Random Variables.* Cambridge, Mass.: Addison-Wesley.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13-30.

KENDALL, M. and STUART, A. (1977). *The Advanced Theory of Statistics*, Vol. 1, New York: Macmillan.

ROSENBLATT, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3** 1-14.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65-78.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** to appear.