On the consistency property of the data-based histogram density estimator

by

Jeesen Chen
Department of Mathematical Sciences
University of Cincinnati

and

Herman Rubin
Department of Statistics
Purdue University

Technical Report #84-11

Department of Statistics
Purdue University
West Lafayette, Indiana

March, 1984

On the consistency property of the data-based histogram density estimator

by

Jeesen Chen
Department of Mathematical Sciences
University of Cincinnati

and

Herman Rubin
Department of Statistics
Purdue University

ABSTRACT

We provide a sufficient condition for a data-based histogram density estimator to be consistent in the sense that the mean absolute deviation of the estimator and the density function converges to zero for any density function as the sample size increases to infinity.

AMS 1980 Subject classification: Primary 62G05 Secondary 62F12, 62H12,
Key Words: density estimator, data-based, consistency

The purpose of this paper is to provide a sufficient condition under which a data-based histogram density estimator is consistent with respect to mean absolute deviation loss.

Suppose we want to estimate the unknown density function f of a sequence of i.i.d. $\mathbb{R}^d$ valued observations $X_1, X_2, \ldots, X_n \ldots$. We could proceed in this way: First choose a sequence of partitions $\pi_n = \{\Delta_{n,k}: 1 \leq k \leq N_n\}$, then use $f_n(x)$ as an estimator for f(x) where:

$$f_n(x) \equiv \begin{cases} \dfrac{q_{n,k}}{n\lambda(\Delta_{n,k})} & \text{if } x \in \Delta_{n,k} \text{ and } 0 < \lambda(\Delta_{n,k}) < \infty , \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

$q_{n,k} = \#\{i : X_i \in \Delta_{n,k} \quad i = 1,2,\ldots,n\}$ is the number of sample points which fall in the n-cell $\Delta_{n,k}$ and λ is Lebesgue measure on $\Omega = \mathbb{R}^d$. The choices of $\{\pi_n\}$ might or might not depend on the observations $(X_1, X_2, \ldots, X_n)$. We call $\{f_n\}$ a histogram density estimator for f with respect to the partition $\{\pi_n\}$. A histogram density estimator is called data-based or fixed celled depending on whether the choice of $\{\pi_n\}$ is or is not based on the observations.

Some examples of data-based histogram estimators are Van Ryzin's unsymmetric histogram estimator (see Van Ryzin [8] p. 495-496), Scott's data-based histogram estimator (see Scott [7] p. 608) and the Abou-Jaoude random partition estimator (see Abou-Jaoude [2] p. 300-301).

The main reason that we want to study a data-based density estimator (data-based means the smoothing factors, which are the lengths or the volumes in the histogram estimator case, depend on the observations) comes from our belief that in order to accurately estimate a density function, the observations should primarily affect the estimation only locally. With a suitable data-based density estimator, one might get a fairly good estimator for f even if the sample size is just moderately large.

Abou-Jaoude [1], [2] studied the properties of histogram estimators (data-based or fixed celled), and especially the convergence properties of the risk when the absolute deviation is ued as the loss function and the sample size approaches infinity.[1] The mean absolute deviation, which we consider a reasonable "global measure" to indicate how good a density estimator is, enjoys at least the following good properties: first, it is invariant if we change the way we measure data, and second, if the estimator converges to the true density function pointwise almost everywhere and the risk is uniformly bounded, then the mean absolute deviation converges to zero as the sample size increases to infinity.[2]

In the following, we only consider the case that each cell $\Delta_{n,k}$ is a rectangle. We need the notations (2) - (5) to state our main result, Theorem 1.

$$\Delta_n(x) = \Delta_{n,k} \text{ if } \breve{x} \in \Delta_{n,k} \tag{2}$$

$$z_n(x) = \int_{\Delta_n(x)} f(t)\ \lambda(dt) \tag{3}$$

$$q_n(x) = z_n(x)/\lambda(\Delta_n(x)). \tag{4}$$

$$B_n(\varepsilon) = \{x: \text{ diameter of } \lambda(\Delta_n(x)) > \varepsilon\} \tag{5}$$

---

1 Abou-Jaoude[2] showed $\int|f_n-f| \to 0$ in probability. But this result is equivalent to $E\int|f_n-f| \to 0$.

2 In fact, it is easy to show that is $f_n \to f$ in $\lambda x P$ measure, and $E\int f_n \to 1$, then $E\int|f_n-f| \to 0$.

$\Delta_n(x)$ is the n-cell which contains the point x, $z_n(x)$ is the mass contained in the (random) n-cell $\Delta_n(x)$, $q_n(x)$ is the average mass at the point y over the cell $\Delta_n(x)$, and $B_n(\varepsilon)$ is the union of those n-cells having diameter greater or equals to ε. Of course, $B_n(\varepsilon)$ contains the union of infinite length (or volume) n-cells.

Theorem 1:

If a (data-based) histogram density estimator $\{f_n\}$ satisfies conditions (A) and (B), then $\{f_n\}$ is consistent for f: i.e.

$$EL_1(f_n,f) \equiv E\int_{\Omega} |f_n(x) - f(x)|\lambda(dx) \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty . \qquad (6)$$

(A) For any $\varepsilon > 0$, there exists a fixed compact set $K = K_\varepsilon$ such that

$$\int_{\Omega \sim K_\varepsilon} f(x)\ \lambda(dx) \leq \varepsilon, \qquad (7)$$

the diameters of $\Delta_n(x)$ in all directions approach zero almost surely for all $x \varepsilon K$, (8)

and

The expected mass contained in the random set $B_n(\varepsilon)$ converges to zero as the sample size n approaches infinity for any fixed $\varepsilon > 0$, i.e.

$$E\int_{B_n(\varepsilon)\cap K_\varepsilon} f(x)\lambda(dx) \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty, \qquad (9)$$

(B) there are at most $s(n,K)$ n-cells intersecting K, where

$$s(n,K) = \begin{cases} o_p\ (n) & \text{for } d = 1, \\ O_p\ (n^{1/2}) & \text{for } d > 1. \end{cases} \qquad (10)$$

Heuristically speaking, the condition (A) guarantees that the expected total mass of the estimator, $E \int f_n$, converges to 1, the condition (A) (9) guarantees that the bias due to using $q_n(x)$ to approximate $f(x)$, $\int |q_n - f|$, approaches 0, and the conditions (A) (8) and (B) (10) imply the variation of $|f_n - q_n|$, $\int \text{var}(|f_n - q_n|)$ is small.

We need Lemmas 1 and 2 to prove Theorem 1.

Lemma 1:

Suppose $X_0, X_1, X_2, \ldots, X_m$ are i.i.d. random variables. If $k \le m$, then there exists a universal constant C such that

$$E \sup_{\pi} \sum_{i=1}^{k} |N_i - m\, p_i| \le C \sqrt{mk} \,, \tag{11}$$

where the supremum runs over all possible partitions with k components,

$N_i = N_i(\pi) = \#\{j: 1 \le j \le m,\ X_j \text{ belongs to the ith component of } \pi\}$, and

$p_i = p_i(\pi) = P(X_0 \text{ belongs to the ith components of } \pi)$.

This is a result due to Chen, Davis and Rubin, [3].

Lemma 2:

Suppose $X_1, X_2, \ldots, X_m$ are i.i.d. random vectors with continuous c.d.f. Then the empiric process, $\sqrt{m}(F_m - F)$, converges to a process with a.s. continuous path if $m \to \infty$.

This result should be well-known, but we have been unable to locate a good reference. We indicate a possible way to reach this conclusion: apply Rubin's Lemma [6], then follow the steps of Gastwirth and Rubin [5] using the moment of order 2d + 2.

Proof of Theorem 1:

First, observe that for $B_n = B_n(\infty)$ and any choice of K_n^*,

$$E \int |f_n - f|$$

$$\leq E \int_{B_n} |f_n - f| + E \int_{\Omega \sim K_n^* \sim B_n} |f_n| + E \int_{\Omega \sim K_n^* \sim B_n} |f|$$

$$+ E \int_{K_n^* \sim B_n} |f_n - q_n| + E \int_{K_n^* \sim B_n} |q_n - f|$$

$$\equiv Q_1(n) + Q_2(n) + Q_3(n) + Q_4(n) + Q_5(n), \text{ say,} \tag{12}$$

For any given positive number ε, $K = K_\varepsilon$ is the compact set mentioned in condition (B). If we let K_n^* be the union of those n-cells that intersect the compact set K, i.e. $K_n^* = K_n^*(\omega) = \{x : \Delta_n(x,\omega) \cap K \neq \phi\}$, then as shown in (I) - (V) below, each $Q_i(n)$ eventually is less then or equal to some constant multiple of ε. Therefore $E \int |f_n - f| \to 0$.

Define the integer sets $J_n^*(K)$ and $\tilde{J}_n$ as following:

$$J_n^*(K) = \{k: \ \Delta_{n,k} \cap K \neq \phi\}, \tag{13}$$

$$\tilde{J}_n = \{k: \ \lambda(\Delta_{n,k}) = \infty\}. \tag{14}$$

(I)

$Q_1(n) = E \int_{B_n} f \leq \varepsilon$ for sufficient large n, by condition (7).

(II)

$$Q_2(n) = E \int_{\Omega \sim K_n^*} f_n$$

$$= E \sum_{j \varepsilon J_n^*(K)} \int_{\Delta_{n,j}} f_n$$

$$\leq E \sum_{j \varepsilon J_n^*(K)} \frac{1}{n} \frac{q_{n,j}}{\lambda(\Delta_{n,j})} \lambda(\Delta_{n,j})$$

$$\leq \frac{1}{n} E\#\{i: \quad X_i \varepsilon \Omega \sim K, 1 \leq i \leq n\}$$

$$\to \int_{\Omega \sim K} f(x) \lambda(dx), \text{ by the Law of Large Numbers.}$$

Hence,

$Q_2(n) \leq \varepsilon$ for sufficiently large n.

(III)

Since $\Omega \sim K_n^* \sim B_n \subset \Omega \sim K$, we have:

$Q_3(n) \leq \int_{\Omega \sim K} f \leq \varepsilon$.

(IV)

Let $z_{n,j}$ denote the mass contained in the n-cell $\Delta_{n,j}$, i.e.

$$z_{n,j} = \int_{\Delta_{n,j}} f(x)\lambda(dx).$$

For d = 1, we apply Lemma 2 and condition (10),

$$Q_n(n) = E \sum_{j \varepsilon J_n^*(K)} \int_{\Delta_{n,j}} \left|\frac{1}{n} \frac{q_{n,j}}{\lambda(\Delta_{n,j})} - \frac{z_{n,j}}{\lambda(\Delta_{n,j})}\right|$$

$$= E \frac{1}{n} \sum_{j \varepsilon J_n^*(K)} |q_{n,j} - n\, z_{nj}|$$

$$\leq \frac{1}{n} C \sqrt{s(n,k)\cdot n}$$

$$\longrightarrow 0 \quad \text{as } n \longrightarrow \infty .$$

Hence

$Q_4(n) \leq \varepsilon$ for sufficient large n.

For $d > 1$, according to Lemma 3, $\sqrt{n}\,(F_n(x) - F(x))$ converges to an a.s. continuous stochastic process $Z(x)$. Let

$$G_n(x) = F_n(x) - F(x). \tag{16}$$

Since a continuous function on a compact set K is uniformly continuous for any ε, $\zeta > 0$, there exists a $\delta > 0$ and an integer n_0 such that

$$P(\sqrt{n}\,|G_n(x) - G_n(y)| \leq \varepsilon\, 2^{1-d} \text{ for all } x,y \text{ such that } |x-y|<\delta) < 1 - \zeta$$

$$\text{if } n \geq n_0 \tag{17}$$

Now, suppose the compact set K is covered by M non-overlaping rectangles one of whose sides is less than δ, $E_1,\ldots,E_m$.

Let

$$E_i = \{x\colon a^i_{k1} < x_k \leq a^i_{k2},\ \ k = 1, 2, \ldots, d\} . \tag{18}$$

$$F_n(E_i) = \sum_{k_1=1}^{2}\cdots\sum_{k_d=1}^{2}(-1)^{k_1+\,\cdots\,+k_d} F_n((a^i_{1k_1}, \ldots, a^i_{dk_d})), \tag{19}$$

and

$$F(E_i) = \sum_{j_1=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_1+\ldots+j_d} F((a^i_{1j_1},\ldots, a^i_{dj_d})). \quad (20)$$

Then

$$F_n(E_i) - F(E_i) = \sum_{j_1=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_1+\ldots+j_d} G_n((a^i_{1j_1},\ldots,a^i_{dj_d})) . \quad (21)$$

Since there are 2^{d-1} pairs of points, each having one positive and one negative coefficient, differering in a specific coordinate, we may use (17) to show that with probability $> 1 - \zeta$, all $|F_n(E_i) - F(E_i)| \leq n^{-1/2} \varepsilon$. Hence if $M = O(n^{1/2})$ and the rectangles get small (notice condition (7)), the total L_1 distance due to the rectangles which intersect K goes to 0 with probability greater than $1 - \zeta$. Since ζ is arbitrary, we have the conclusion: $Q_4(n) \leq \varepsilon$ for sufficient large n.

(V)

For $\mu > 0$, let $K(\mu)$ be the closed μ neighborhood of K. It is obvious that $\lambda(K(\mu)) < \infty$ for all $\mu > 0$. If we use $\tilde{d}(\xi)$ to denote the diameter of ξ , then by condition (9),

$$H_n \equiv \int_K P(\tilde{d}(\Delta_n(x)) > \varepsilon)\ \lambda(dx) < \varepsilon^2 \text{ for n sufficient large.} \quad (22)$$

But this means

$$P\{x: x \in K^*_n \sim K, \Delta_n(x) \cap K(\varepsilon) = \phi\} < \varepsilon^2 \text{ for n sufficient large.} \quad (23)$$

Hence

$$Q_5(n) \le E\int_{K_n^* \sim B_n \sim K} |q_n - f| + E\int_K |q_n - f| \tag{24}$$

$$\le E\int_{K_n^* \sim B_n \sim K} |q_n| + E\int_{K_n^* \sim B_n \sim K} f + E\int_K |q_n - f| .$$

The first term of (24)

$$E\int_{K_n^* \sim B_n \sim K} |q_n| \le \int_{\sim K} f + \frac{1}{\varepsilon}\int_{\sim K} f \quad x \quad P\{x:\ x\varepsilon K_n^* \sim K,\ \Delta_n(x) \cap K(\varepsilon) \ne \phi\}$$

$$\le \varepsilon + \frac{1-\varepsilon}{\varepsilon} \cdot \varepsilon^2$$

$$\le 2\varepsilon \quad \text{for n large enough.} \tag{25}$$

The second term of (24),

$$E\int_{K_n^* \sim B_n \sim K} f \le \int_{\sim K} f \le \varepsilon \tag{26}$$

The third term of (24) converges to zero by condition (9) and $\lambda(K) < \infty$.

$$E\int_K |q_n - f| \le \varepsilon \text{ for sufficient large n.} \tag{27}$$

Combining (24) - (27), we have

$$Q_5(n) \le \varepsilon + 2\varepsilon + \varepsilon \quad \text{for n sufficient large,} \qquad \text{q.e.d.}$$

Scott [6] proposed the data-based density estimator for 1-dimensional density function f: choose partitions $\pi_n = \{\Delta_{n,k} : 1 \le k < \infty\}$ satisfying

$$\lambda(\Delta_{n,k}) = h_n = 3.49\, s_n\ n^{-1/3} \quad \text{for all k , n.} \tag{28}$$

where s_n denotes the sample standard deviation of $\{X_1, \ldots, X_n\}$. It is a simple corollary of the Theorem 1 and a result of Chen and Rubin [3] that the Scott's data-based histogram estimator is consistent if $E|X_1|^{6/5} < \infty$. More generally, we have the following result:

Corollary 1: If f_n is a Scott-type data-based histogram estimator, there exists an α, $0 < \alpha < 1$ and a constant C such that

$$\lambda(\Delta_{n,k}) = h_n = C s_n n^{-\alpha} \qquad \text{for all } k, n, \tag{29}$$

then the density estimator f_n is a consistent estimator if $E|X_1|^{\frac{2}{1+2\alpha}} < \infty$.

Proof:

Applying the result of Chen and Rubin [4], Corollary 9, the condition $E|X_1|^{\frac{2}{1+2\alpha}} < \infty$. is equivalent to $h_n \to 0$ a.s. as $n \to \infty$. It is very easy to verify conditions (7) - (10). q.e.d.

Finally, we want to point out that:

Suppose we divide $(-\infty,\infty)$ in intervals of increasing length in both directions. For example, (0,1], (1,2], (2,4], (4,8],..., (-1,0], (-2,-1], (-4,-2],..., and use Scott's method in each interval which has at least $n^{1/3}$ observationists, and 0 in the other intervals.

Corollary 2: The above histogram estimator is consisten for any density.

This example shows some of the advantages of data-based estimators.

Corollary 3: Van Ryzin's unsymmetric histogram estimator ([8], p. 495-496) is consistent if $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$.

Proof: Verify conditions (A) and (B) of Theorem 1.

REFERENCES

[1] Abou-Jaoude, S. (1976) Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. Ann. Inst. H. Poincaré. Sect. B 12, 213-231.

[2] Abou-Jaude, S. (1976) Sur la convergence L_1 et L_∞ of l'estimateur de la partition aléatoire pour une densité. Ann. Inst. H. Poincaré, Sect. B 12, 299-317.

[3] Chen, J., Davis, B., and Rubin, H. (1984) How non-uniform can a uniform sample be? Purdue mimeograph.

[4] Chen, J. and Rubin, H. (1984) A note on the behavior of sample statistics when the population mean is infinite. Ann. Probability 12, 256-261.

[5] Gastwirth, J. L. and Rubin, H. (1975) Asymptotic distribution theory of the empiric c.d.f. for mixing stochastic processes. Ann. Statist. 3, 809-824.

[6] Rubin, H. (1968) A useful device for continuity theorems, (Abstract 17). Ann. Math. Statist. 39, 1090.

[7] Scott, D. W. (1979) On optimal and data-based histograms. Biometrika, 66, 605-610.

[8] Van Ryzin, J. (1973) A histogram method of density estimator. Common Statist. 2, 493-506