数理解析研究所講究録
第842巻 1993年 126-150

An optimal variable cell histogram based on sample spacings

金澤 雄一郎
Yuichiro Kanazawa

University of Tsukuba

April 7, 1993

1 Introduction

The purpose of the present paper is to explain how an idea of a new histogram density estimator comes about to the readers who are not familiar with density estimation. Specifically we try to explain why certain decisions were made in the course of our research and why we thought they were justified. To serve this purpose we restrict the technical aspect of the material to the minimum and we shall resort to heuristics if necessary. Before we get into the particular density estimator we propose, we first would like our readers to familiarlize themselves with what density estimation is trying to do and why it was developed.

Suppose that we have a set of observed data points $X_1, \ldots, X_n$ assumed to be a random sample from an unknown probability density function $f(x)$.

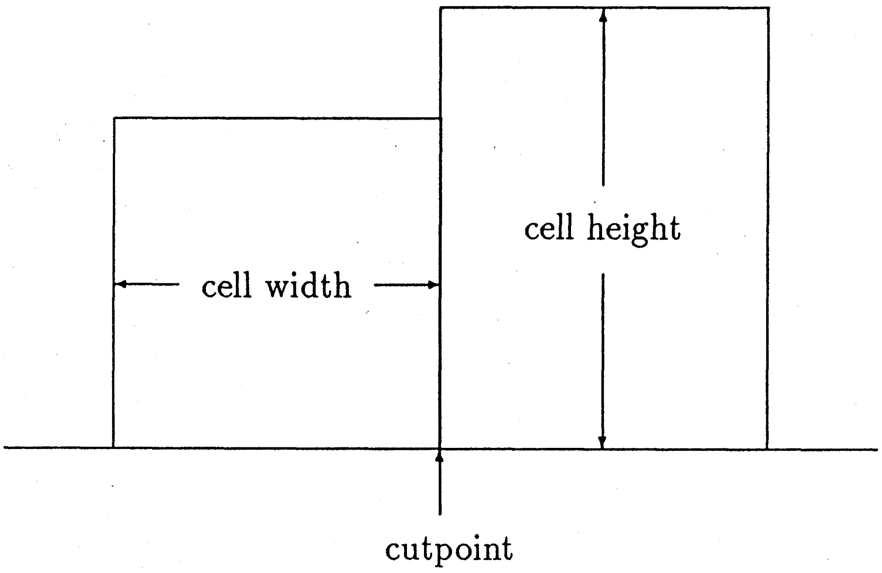
Density estimation concerns with the construction of an estimator of the density from the observed data, without assuming the data are drawn from one of a known *parametric* family of distributions.

Historically there had been two separate developments that were to merge in Rosenblatt(1956) to create density estimation as a new branch of statistics. One was the sequence of investigations by Grenander and Rosenblatt(1953,1956) to estimate locally the continuous part of the spectral density function for weakly stationary sequence of random variables. The other was the effort by Fix and Hodges(1951) to free discriminant analysis from rigid distributional assumptions for independent data. How the two developments merged and the kernel density estimator was born was recollected in Rosenblatt(1991).

Next we would like to summarize relative strengths and weaknesses for the two most important density estimators, the kernel estimator and histogram. The kernel estimator is considered to be mathematically more appealing since its expected discrepancy to the underlying density converges at the rate of $O(n^{-4/5})$, faster than that of the histogram, which converges at the rate of $O(n^{-2/3})$. The kernel estimator, however, often neglects boundary restrictions because it is based on smoothing. Accordingly it is rarely used in analyzing, let alone in presenting, survival data for which the data are restricted to be nonnegative and compactly supported. The histogram is by far the oldest and most widely used density estimator. Constructing the histogram, however, requires not only a cell width $|C|$, which is essential, but the leftmost cutpoint a. The choice of the latter can have a quite effect, a nuisance.

In this paper, we are concerned with the particular class of histogram density estimator. We would like to remind the reader that this is simply because the we were interested in applying the estimator to survival data for which boundary restrictions are facts of life. We do not think the histogram is always superior to the kernel estimator, but we do not believe the converse to be the case either.

As a preparation for presenting our histogram, we would like to define some basic terms and notations. We would like to use the term histogram or *empirical* histogram $f_n^{\circ}(x)$ for our density estimator that has more than or equal to one cell whose widths may vary. Each cell has width and height as seen in the picture below. Also cell heights may not have to be determined by the number of data that fall in the cell. In other words, we regard our histogram as a mixture of uniforms whose heights may be determined by some other algorithm.



2 Motivation

2.1 Closeness to the underlying density

First we would like our histogram to be close to the underlying density. In order to evaluate overall closeness to the underlying density, however, we need measures of discrepancy between $f_n^\circ(x)$ and $f(x)$.

Let $\hat{f}(x)$ be the histogram with constant cell width. Historically mean integrated square error(MISE)

$$MISE(\hat{f},f) = E\int\left\{\hat{f}(x) - f(x)\right\}^2 dx$$

was widely used in the previous work. Scott(1979) obtains theoretically optimal constant cell width $|C|$ that asymptotically minimizes $MISE(\hat{f},f)$. Rudemo (1982) proposed a data-based method of choosing the leftmost cutpoint and cell width $(a,|C|)$ called *least-squares cross-validation*(LSCV) by minimizing a criterion whose expected value is $MISE(\hat{f},f) - \int f(x)^2\,dx$. Stone(1984) showed Rudemo's estimators of $(a,|C|)$ is asymptotically close to $\min MISE(\hat{f},f)$. Kogure(1987) extended Rudemo's estimators of $(a,|C|)$ to a variable cell histogram.

We take different strategy. Instead of comparing $f_n^\circ(x)$ directly with $f(x)$, we introduce a concept of the k-cell *theoretical* histogram $g^\circ(x)$ that belongs to a class of k-cell histogram-type densities $g(x)$ which is defined on the same domain as $f(x)$, and that minimizes Hellinger distance(HD)

$$HD(g,f) = \int\left\{g(x)^{1/2} - f(x)^{1/2}\right\}^2 dx.$$

Using the concept of the theoretical histogram $g^\circ(x)$, we can define a

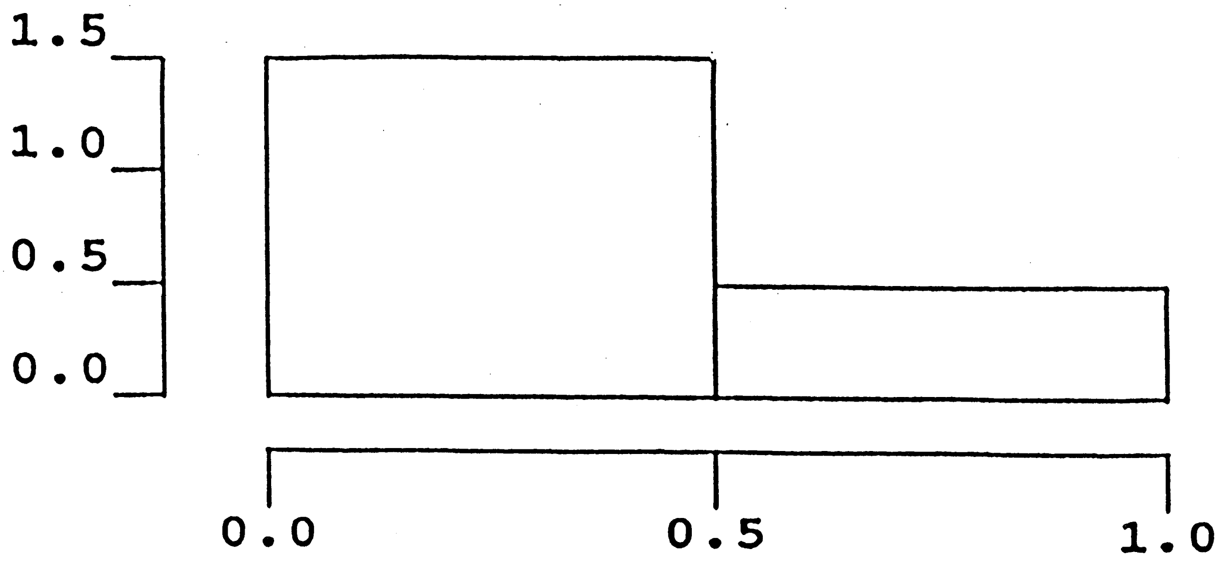


Figure 1: A finite mixture of uniforms. The theoretical histogram coincides with the underlying density.

histogram $f_n^\circ(x)$ to be close to its underlying density $f(x)$ when $f_n^\circ(x) \rightarrow g^\circ(x)$ as the number of sample increases. We give two examples of theoretical histogram.

EXAMPLE 1. An underlying density itself is a finite mixture of uniforms

$$f(x) = 3/2\{.0 \leq x \leq .5\} + 1/2\{.5 \leq x \leq 1.0\}.$$

It is shown in Figure 1. The theoretical histogram coincides with the underlying density.

EXAMPLE 2. A quadratic underlying density

$$f(x) = 3x^2/7\{1.0 \leq x \leq 2.0\}.$$

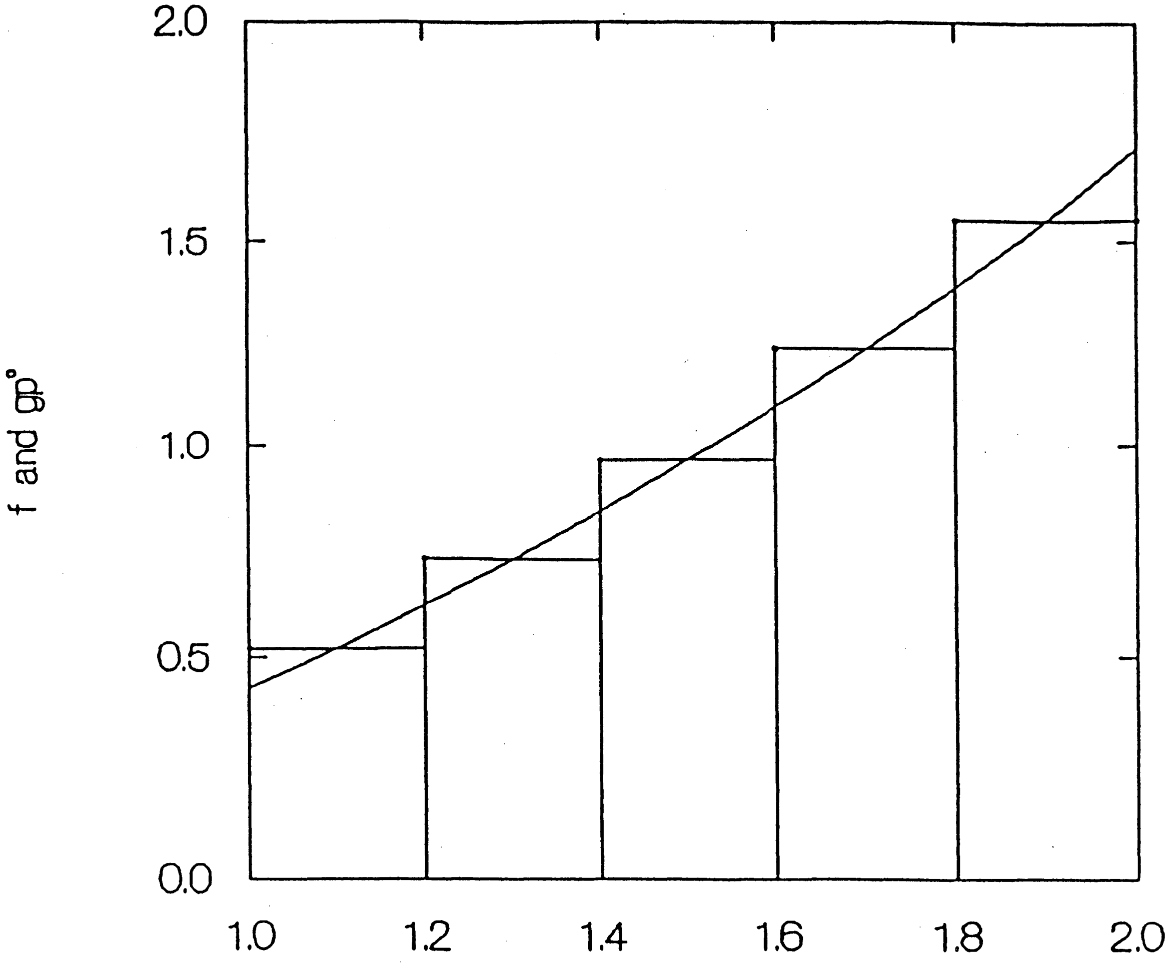


Figure 2: A quadratic underlying density and its theoretical histogram.

For g° with two cells let the mid-cutpoint be x where $1 < x < 2$. The resulting Hellinger distance $\propto$ $[(x+1)^2(x-1)+(2+x)^2(2-x)]$ has its maximum at $x = 3/2$. Thus the two cells have equal width. In general for g° with k cells, any two neighboring cells have the same width, and so all cells are of the same width. Therefore the theoretical histogram with five cells has the constant cell width as shown in Figure 2.

If the problem of determining how much smoothing should be done is important as it has been in density estimation, then the problem of choosing a proper measure of discrepancy must be important as well. The integarated square error(ISE) has generally been regarded as the *standard* measure of the discrepancy because it was advocated first, is mathematically tractable, and because the mean square error has been familiar to us. Since we decided to choose *non-standard* Hellinger distance over the more conventional ISE, the readers are entitled to have an explanation why we should do so.

Let us note first that the HD is asymptotically close to the weighted integrated square error(WISE)

$$\int \frac{\{f_n^\circ(x) - f(x)\}^2}{f(x)} dx,$$

in the sense that

$$\int \left\{f_n^\circ(x)^{1/2} - f(x)^{1/2}\right\}^2 dx \approx \frac{1}{4} \int \frac{\{f_n^\circ(x) - f(x)\}^2}{f(x)} dx.$$

Assuming $f_n^\circ(x)$ is sufficiently close to $f(x)$, we can see this from

$$\begin{aligned} \left\{f_n^\circ(x)^{1/2} - f(x)^{1/2}\right\}^2 &= f_n^\circ(x) + f(x) - 2\left\{1 + \frac{f_n^\circ(x) - f(x)}{f(x)}\right\}^{1/2} f(x) \\ \approx \quad & f_n^\circ(x) + f(x) - 2\left[1 + \frac{1}{2}\left\{\frac{f_n^\circ(x) - f(x)}{f(x)}\right\} - \frac{1}{8}\left\{\frac{f_n^\circ(x) - f(x)}{f(x)}\right\}^2\right] f(x) \\ = \quad & \frac{1}{4}\frac{\{f_n^\circ(x) - f(x)\}^2}{f(x)}. \end{aligned}$$

We shall give four reasons, in decreasing order of significance in our judgment, why we think the WISE is preferable to the ISE. The first and most important reason is derived from the following analogy. If the variances in error terms vary with observations in least-squares estimation, we know the residual sum of squares to be minimized must be adjusted by multiplying the inverse of its heteroscedastic, and thereby non-identity, variance-covariance matrix to obtain the best linear unbiased estimator. We think the same principle should be applied to in density estimation. The argument suggests that we use as our measure of the discrepancy

$$\int \frac{\{f_n^\circ(x) - f(x)\}^2}{V(f_n^\circ(x))} dx. \tag{1}$$

We know, for the kernel estimator $\widehat{f_{kernel}}(x)$ with the constant window width, its variance is

$$\mathrm{V}(\widehat{f_{kernel}}(x)) \approx \frac{f(x)}{nh} \int K(t)^2 dt,$$

where $K(t)$ is the kernel function, h the size of the window width, n the sample size. Similarly, for the histogram $\widehat{f_{histo}}(x)$ with the constant cell width, its variance is

$$\mathrm{V}(\widehat{f_{histo}}(x)) \approx \frac{f(x)}{nh},$$

where h is the size of the cell width. If we assume the size of the smoothing parameter such as the window or cell width does not vary much with x, then the variance $\mathrm{V}(f_n^\circ(x))$ in (1), whether its $f_n^\circ(x)$ is the histogram or kernel estimator, is proportional to $f(x)$.

In other words, the WISE properly discounts the contributions from high density regions where the variabilities are also high by giving there the weight of $1/f(x)$. The other distance-based measure of discrepancy such as the ISE, neglecting this essential operation, gives an equal weight to the regions regardless of the variabilities there.

The couclusion above is also supported by the following. Let $|I|$ be the width of support of the density. Scott(1979) showed the optimal cell size for the histogram that asymptotically minimizes the MISE to be

$$\left[\frac{6}{\int f^{(1)}(x)^2 dx}\right]^{1/3} n^{-1/3}, \tag{2}$$

while Kanazawa(1993a) showed the optimal cell size that asymptotically minimizes mean Hellinger distance(MHD) to be

$$\left[\frac{12 \cdot |I|}{\int \{f^{(1)}(x)^2 / f(x)\} dx}\right]^{1/3} n^{-1/3}. \tag{3}$$

Rosenblatt(1956) showed the optimal window width for the kernel estimator with kernel $K(t)$ that asymptotically minimizes the MISE to be

$$\left[\frac{\int K(t)^2 dt}{\{\int t^2 K(t) dt\}^2}\right]^{1/5} \left\{\int f^{(2)}(x)^2 dx\right\}^{-1/5} n^{-1/5}, \tag{4}$$

while Kanazawa (1993b) showed the optimal window width that asymptotically minimizes the MHD to be

$$(4 \cdot |I|)^{1/5} \left[\frac{\int K(t)^2 dt}{\{\int t^2 K(t) dt\}^2}\right]^{1/5} \left\{\int \frac{f^{(2)}(x)^2}{f(x)} dx\right\}^{-1/5} n^{-1/5}. \tag{5}$$

Those based on the MISE use raw derivatives, $f^{(1)}$ and $f^{(2)}$, while those based on the MHD use derivatives adjusted by the variance of the density estimators, $f^{(1)}/f^{1/2}$ and $f^{(2)}/f^{1/2}$. This illustrates that the MHD tries to stabilize variance at a single point before it integrates out over the support, while the MISE does not.

The second reason is that Kullback-Leibler loss also measures the WISE asymptotically in the sense that

$$\int f(x) \log \frac{f(x)}{f_n^\circ(x)} dx \approx \frac{1}{2} \int \frac{\{f_n^\circ(x) - f(x)\}^2}{f(x)} dx.$$

We can see this by using Taylor's expansion to Kullback-Leibler loss,

$$\log\frac{f_n^\circ(x)}{f(x)} = \log\left\{1+\frac{f_n^\circ(x)-f(x)}{f(x)}\right\} \approx \frac{f_n^\circ(x)-f(x)}{f(x)} - \frac{1}{2}\left\{\frac{f_n^\circ(x)-f(x)}{f(x)}\right\}^2 .$$

Hence the HD is an essential link between distance-based and information-theoretic measures of discrepancy.

The third reason is that the HD, through its asymptotic equivalence with the KLL, has an well established and completely data-based method for choosing the size of histogram cells in Akaike's information criterion. The fact that the AIC is essentially a model selection rule based on KLL is evidenced by the Akaike's own writing in Akaike(1973):

> ...this [AIC] is equivalent to maximizing an information theoretic quantity which is given by the definition
>
> $$E\log\frac{f(x|\hat{\theta})}{f(x|\theta)} = E\int f(x|\theta)\log\frac{f(x|\hat{\theta})}{f(x|\theta)}dx.$$
>
> The integral in the right-hand side of the above equation gives the Kullback-Leibler's mean information for discrimination between $f(x|\hat{\theta})$ and $f(x|\theta)$....

The AIC must be equivalent to the KLL asymptotically under appropriate conditions on the density. The fact that the AIC is completely data-based and does not depend on the unknown density being estimated can be seen in the following specification of it by Taylor(1987) for the histogram;

$$\text{AIC} = J - \log\left[\Pi_{i=1}^{\#\text{ of cells}}\{\nu_i/(nh)\}^{\nu_i}\right]$$

where ν_i equal the number of data points in the i-th cell, h the (constant) bin width, and J the total number of histogram cells.

Fourth we would like to have tail-sensitive histogram because we are interested in applying the density estimator to survival data for which boundary restrictions cannot be avoided as we already stated. When one compares the MISE-optimal cell width in (2) with the MHD-optimal (3) for the histogram or the MISE-optimal window width in (4) with the MHD-optimal (5) for the kernel density estimator, one cannot fail to notice that those based on the MISE neglect the $|I|$, while those based on the MHD do not. This implies the MISE concentrates its focus on the center of distribution so much that it does not even care how wide its tail can be.

2.2 Handling of observations on cutpoints

Secondly we would like our histogram to be free of awkward problem often associated with constructing the histogram, the problem concerns with its handling of observations on cutpoints. When observations fall on the cutpoints, one needs to either move cutpoints or somehow classify observations on the cutpoints. We usually take the latter path and the histogram cells are chosen closed on the left and open on the right. For the kind of data we are interested in applying our histogram to, any density estimators are likely to be unacceptable if they give any weight to the region outside of its support. These two considerations led us to choose the leftmost cutpoint at $X_{(1)}$, rightmost at $X_{(n)}$ and to choose the other cutpoints from order statistics $X_{(2)}, \ldots, X_{(n-1)}$. This requirement makes our variable-cell histogram to be an uncensored Kaplan-Meier estimator whose number of jumps is reduced.

3 Heuristic derivation

To derive a variable cell histogram with these properties, we shall proceed as follows: (1) Compute a k-cell "theoretical histogram g^{o}"that minimizes the Hellinger distance $HD(g,f) = \int_I [g(x)^{1/2} - f(x)^{1/2}]^2 \, dx$ to a density f over a class of k-cell histogram-type density g that stays constant within a cell; (2) Derive a histogram f_n° based on the sample that estimates g°. The theoretical histogram g° depends on f but not on the sample.

We shall describe the two steps heuristically but with more details. First we note the height h_j

$$h_j = \left[\frac{\int_{I_j} f(x)^{1/2}\, dx}{|I_j|}\right]^2 \Big/ \sum_{j=1}^{k} \frac{\left[\int_{I_j} f(x)^{1/2}\, dx\right]^2}{|I_j|}, \qquad x \in I_j$$

of the j-th cell I_j of a class of k-cell histogram-type density g

$$\mathcal{L}(k) = \left\{ g : g(x) = \sum_{j=1}^{k} h_j \{x \in I_j\}, \sum_{j=1}^{k} h_j \, |I_j| = 1, h_j \geq 0 \right\},$$

where $|I_j|$ is the width of the j-th cell, minimizes the Hellinger distance to the density

$$HD(g,f) = \sum_{j=1}^{k} \int_{I_j} \left[h_j^{1/2} - f(x)^{1/2}\right]^2 dx.$$

This may be done by the method of Lagrange multipliers. The resulting Hellinger distance $HD(g,f)$ is

$$HD(g,f) = 2 - 2 \left[\sum_{j=1}^{k} \frac{\left[\int_{I_j} f(x)^{1/2}\, dx\right]^2}{|I_j|}\right]^{1/2}$$

Let H and $H^{(1)} = 1/f$ be the inverse of the distribution function F of the unknown density f and its first derivative respectively. If we denote the endpoints of $F(I_j)$ by $[p_j, p_{j+1}]$, elementary calculation shows that

$$P(H, \mathrm{p}) = \frac{\pi}{4} \sum_{j=1}^{k} \frac{\left[\int_{p_j}^{p_{j+1}} H^{(1)}(u)^{1/2}\, du\right]^2}{\int_{p_j}^{p_{j+1}} H^{(1)}(u)\, du} = \frac{\pi}{4} \sum_{j=1}^{k} \frac{\left[\int_{I_j} f(x)^{1/2}\, dx\right]^2}{|I_j|}.$$

Let $K = k+1$ be the number of cutpoints of the histogram-type density g. If the set of K-cutpoints $\mathrm{p}^\circ = (p_1^\circ, \ldots, p_K^\circ)$ maximizes $P(H, \mathrm{p})$, then it minimizes $HD(g, f)$. For this set of cutpoints p°, we obtain a set of k-heights $\mathrm{h}^\circ = (h_1^\circ, \ldots, h_k^\circ)$ by substituting p° in a formula of h_j above. Thus a pair $(\mathrm{p}^\circ, \mathrm{h}^\circ)$ determines the theoretical histogram g°.

Secondly, to find the histogram f_n° that best estimates g°, we shall proceed as follows: (1) Construct sample-based analogs $C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right)$ and h_{nj} of $P(H, \mathrm{p})$ and h_j respectively; (2) Find the set of cutpoints that maximizes $C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right)$; (3) Compute h_{nj} for the set of cutpoints. We expect the histogram f_n° constructed this way to converge to the theoretical histogram g°. Now we need a sample-based analog $C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right)$ of $P(H, \mathrm{p})$. Define the i-th spacing as $T_i = X_{(i+1)} - X_{(i)}$, $i = 1, \ldots, n-2$, where $X_{(i)}$ is an i-th order statistic of an independent and identically distributed sample $X_1, \ldots, X_{n-1}$ of size $n-1$. The inverse of the probability integral transformation gives $X_{(i)} = H(U_{(i)})$ where $U_{(i)}$ is the i-th order statistic from an independent and identically distributed sample of size $n-1$ from the uniform $[0, 1]$. Let $e_1, \ldots, e_n$ be independent exponentials with mean 1 and set $s_n = \sum_{i=1}^{n} e_i$. From the well-known relation between the uniform spacings and standardized exponentials e_i/s_n, we have

$$T_i \approx (U_{(i+1)} - U_{(i)}) H^{(1)}(i/n) = \frac{e_i}{s_n} H^{(1)}(i/n).$$

Then for a criterion $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ below, we obtain the following:

$$\begin{aligned} C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right) &= n^{-1}\sum_{j=1}^{k}\frac{\left[\sum_{i=n_j}^{-1+n_{j+1}}T_i^{1/2}\right]^2}{\sum_{i=n_j}^{-1+n_{j+1}}T_i} \\ &\approx \sum_{j=1}^{k}\frac{\left[n^{-1}\sum_{i=n_j}^{-1+n_{j+1}}e_i^{1/2}H^{(1)}(i/n)^{1/2}\right]^2}{n^{-1}\sum_{i=n_j}^{-1+n_{j+1}}e_iH^{(1)}(i/n)}\approx P\left(H,\mathrm{p}\right). \end{aligned}$$

We can construct a sample-based analog h_{nj} of h_j similarly.

We summarize the procedure to construct a variable k-cell histogram f_n°:

Step 1. Find a set of K-cutpoints $(X_{(n_1^\circ)},\ldots,X_{(n_K^\circ)})$ that maximizes

$$C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right) = n^{-1}\sum_{j=1}^{k}\frac{\left[\sum_{i=n_j}^{-1+n_{j+1}}T_i^{1/2}\right]^2}{\sum_{i=n_j}^{-1+n_{j+1}}T_i}. \tag{6}$$

Step 2. Compute the optimal height h_{nj}° from the cutpoints in **Step 1** by

$$h_{nj} = \left[\frac{\sum_{i=n_j}^{-1+n_{j+1}}T_i^{1/2}}{\sum_{i=n_j}^{-1+n_{j+1}}T_i}\right]^2 \Bigg/ \sum_{j=1}^{k}\frac{\left[\sum_{i=n_j}^{-1+n_{j+1}}T_i^{1/2}\right]^2}{\sum_{i=n_j}^{-1+n_{j+1}}T_i}.$$

We note the criterion requires a class of k-cell empirical histograms f_n with the j-th cell I_{nj}

$$\mathcal{L}_n(k) = \left\{f_n : f_n(x) = \sum_{j=1}^{k}h_{nj}\{x\in I_{nj}\}, \sum_{j=1}^{k}h_{nj}\left|I_{nj}\right| = 1, h_{nj}\geq 0\right\},$$

from which our histogram f_n° is chosen to have:

- variable cell widths for the cutpoints are chosen from $X_{(1)},\ldots,X_{(n-1)}$;
- a domain $[X_{(1)},X_{(n-1)}]$.

The intuition behind the choice of criterion (6) is that the information on the density is reflected in the spacings in such a way that regions with narrow spacings tends to have high density, while ones with wide spacings tends to have low density. Maximizing $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ is computationally simple through the *dynamic programming algorithm.*

4 Algorithm

We explain the dynamic programming algorithm to find the K-cutpoints that maximize the $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ through the example. Suppose that we want to choose four cutpoints from $X_{(1)},\ldots,X_{(6)}$. The leftmost and rightmost cutpoints are $X_{(n_1^\circ)} = X_{(1)}$ and $X_{(n_K^\circ)} = X_{(6)}$. Then:

```
Stage 3. Selection of the best 4
4   X(1)     X(2)  X(3)  X(4)  X(5)  X(6)
P    [--- Best 3 ---][                ] 1
T    [------ Best 3 ------][          ] 2
S    [--------- Best 3 ---------][    ] 3
Stage 2. Selection of the best 3
3   X(1)     X(2)  X(3)  X(4)  X(5)
P    [-Best 2-][    ]                   1
T    [-Best 2-][          ]             2
S    [--- Best 2 ---][    ]             2
     [-Best 2-][                ]       3
     [--- Best 2 ---][          ]       3
     [------ Best 2 ------][    ]       3
Stage 1. Selection of the best 2
2   X(1)     X(2)  X(3)  X(4)
P    [-Best 2-]                     1,2,3
T    [--- Best 2 ---]                 2,3
S    [------ Best 2 ------]             3
```

Since the best 2 and 3 are already computed at Stage 1 and 2, we only have to compute the increment in $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ brought in by adding one cutpoint at Stage 2 and 3.

Direct enumeration roughly requires $Kn!/K!(n-K)!$ calculations, while the dynamic programming roughly needs $Kn^2/2$. Hence the ratio of the former to the latter is roughly $2(n-2)!/K!(n-K)!$. For $K=5$ and 10, these are roughly $2n^3/5!$ and $2n^8/10!$, quite a saving.

5 Consistency with the number of cells fixed

The density f has to be "smooth" to substantiate the heuristic argument that $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ and $P(H,\mathrm{p})$ are close. Also the theoretical histogram g° has to be unique to establish the histogram f_n° converges to g°. In Kanazawa (1988a,b) we showed f_n° with the known number k of cells converges in probability to g° under the smoothness conditions **A1** through **A3** on f and the uniqueness condition **B1** on g°:

THEOREM 1.

Let the following conditions and **C1** be satisfied:

A1 F is twice continuously differentiable except for finite points;

A2 f is bounded away from 0 and ∞;

A3 f' is bounded away from ∞;

B1 A unique choice of cells $I_j, j=1,\ldots,k$ that maximize

$$P(H,\mathrm{p})=\sum_{j=1}^{k}\frac{\left[\int_{I_j} f(x)^{1/2}\,dx\right]^2}{|I_j|}.$$

Define

$$C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)=n^{-1}\sum_{j=1}^{k}\frac{\left[\sum_{i=n_j}^{-1+n_{j+1}}T_i^{1/2}\right]^2}{\sum_{i=n_j}^{-1+n_{j+1}}T_i}.$$

Then:

(1) For any set of indices $\mathrm{n}=(n_1,\ldots,n_K)$ of K-cutpoints $(X_{(n_1)},\ldots,X_{(n_K)})$

$$\max_{1=n_1<n_2<\ldots<n_K=n-1}\left|C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)-P\left(H,\mathrm{n}/n\right)\right|=O_p(n^{-1/2})$$

(2) The set of indices $\mathrm{n}^\circ=(n_1^\circ,\ldots,n_K^\circ)$ of the K-cutpoints $(X_{(n_1^\circ)},\ldots,X_{(n_K^\circ)})$ that maximizes $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ converges to the K-cutpoints $\mathrm{p}^\circ=(p_1^\circ,\ldots,p_K^\circ)$ that maximizes $P\left(H,\mathrm{p}\right)$ in the sense that $\mathrm{n}^\circ/n\to\mathrm{p}^\circ$ in probability as $n\to\infty$.

Proof

The idea of the proof is already covered in section 3.

Densities $f(x)$ in Example 1 and 2 on page 5 satisfy these conditions. For a uniform density on the interval $I$ of compact support, a theoretical histogram with $k>1$ cells does not exist because any choice of cells that covers the interval produce an identical $P\left(H,\mathrm{p}\right)$. This violates **B1**.

6 The optimal number of cells

For a density whose theoretical histogram g° has a "correct"and finite number of cells as in Example 1, consistency of f_n° with the known number of cells to g° in Kanazawa (1988a,b) in principle warrants the validity of the procedure, though the problem remains regarding how we actually identify the true number of cells for the density. For smoother densities such as the one in EXAMPLE 2, however, there is no "correct"number of cells and we are forced to determine the number of cells. As a global measure of error between a class $\mathcal{L}_n(k)$ of k-cell histograms f_n whose cutpoints are chosen from the order statistics $X_{(1)}, \ldots, X_{(n-1)}$ and the unknown density f, we use the mean Hellinger distance between f_n and f

$$MHD(f_n, f) = E\left[\int_I \left(f_n(x)^{1/2} - f(x)^{1/2}\right)^2 dx\right],$$

where E denotes the expectation with respect to the sample $X_1, \ldots, X_{n-1}$. We note that the mean Hellinger distance to f is defined for f_n, and not for f_n°, the k-cell histogram obtained by maximizing $C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right)$. In Kanazawa (1992a) we showed f_n° the mean Hellinger distance between f_n and f is minimized if we take the number of cells to be $k = O(n^{1/3})$ under an additional smoothness condition **A4** on the density and a constraint **C1** that prevents a small number of cells from dominating the other cells on the histogram as follows:

A4 f'' exists and is bounded away from ∞;

C1 For all $\Delta_j N = (n_{j+1} - n_j)/n$ where $j = 1, \ldots, k$ and some constants $C_\circ$ and C°

$$0 < \frac{C_\circ}{k} \leq \Delta_j N \leq \frac{C^\circ}{k} < 1.$$

We note that the cutpoints of f_n in **C1** do not involve maximizing the criterion (6) and are denoted by $(X_{(n_1)}, \ldots, X_{(n_K)})$, while those of f_n° obtained by maximizing (6) are denoted by $(X_{(n_1^\circ)}, \ldots, X_{(n_K^\circ)})$. We present the theorem in terms of H's.

THEOREM 2

Let the following conditions and **C1** be satisfied:

A1' $H(u)$ is three times continuously differentiable except for finite points.

A2' $0 < m_1 \leq H^{(1)}(u) \leq M_1 < \infty, \qquad 0 \leq u \leq 1.$

A3' $|H^{(2)}(u)| \leq M_2 < \infty, \qquad 0 \leq u \leq 1.$

A4' $|H^{(3)}(u)| \leq M_3 < \infty, \qquad 0 \leq u \leq 1.$

As $n \to \infty$ the number of cells $\hat{k}$ that minimize the mean Hellinger distance between f_n and f, $MHD(f_n, f)$, satisfies:

$$\frac{\hat{k}}{n^{1/3}} \to \left[\frac{\pi}{24(4-\pi)}\right]^{1/3} \int_0^1 \left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du.$$

As $n \to \infty$ the minimal $MHD(f_n, f)$ satisfies:

$$\min MHD(f_n, f)\; n^{2/3} \to \left[\frac{3(4-\pi)}{8\pi}\right]^{2/3} \int_0^1 \left[\frac{H^{(2)}(u)}{H^{(1)}(u)}\right]^{2/3} du.$$

PROOF

After a long and tedious computations, it can be shown that

$$MHD(f_n, f) = \text{constant} - O(k^{-2}) + O(k/n)$$

Balance the second term against the third term. □

For the quadratic density in EXAMPLE 2, $\hat{k}/n^{1/3}$ and $\min MHD(f_n, f)n^{2/3}$ converges to 0.64 and 0.26 respectively.

7 Consistency with the number of cells increasing

If we are to construct a consistent data-based method by extending **step 1** and **2**, g° must remain as the target of f_n° even if we increase the number of variable cells at $O(n^{1/3})$. From the second statement of THEOREM 2 we know the minimal MHD decreases at $O_p\left(n^{-2/3}\right)$ and thus the $HD(g^\circ, f)$ decreases as fast. If we show the maximal difference between $C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right)$ and $P\left(H, \mathrm{n}/n\right)$ with all the specifications of K-cutpoints considered, are $o_p\left(n^{-2/3}\right)$, then g° remains as the target of f_n°. The maximal difference is still $O_p\left(n^{-2/3}\right)$ after appropriate random variables and constants independent of cutpoint-specification are subtracted. A constraint that requires the neighboring cell widths of f_n° to vary slowly on **step 1**, along with that on the cell widths of g°, make it $o_p\left(n^{-2/3}\right)$. Then in Kanazawa we shall show the f_n° converges to g° in the sense that the maximal absolute difference between any corresponding cutpoints of f_n° and of g° converges to 0 in probability.

THEOREM 3

Let **A1** through **A4**, **B1'**, and **C1'** be satisfied.

B1' There is a unique choice of $\mathbf{p}^\circ = (p_1^\circ, \ldots, p_K^\circ)$ that maximize

$$P\left(H, \mathrm{p}\right) = \frac{\pi}{4} \sum_{j=1}^{k} \frac{\left[\int_{p_j}^{p_{j+1}} H^{(1)}(u)^{1/2}\, du\right]^2}{\int_{p_j}^{p_{j+1}} H^{(1)}(u)\, du}$$

subject to the constraint

$$0 < C_\circ n^{-1/3} \leq \Delta p_j{}^\circ = p_{j+1}{}^\circ - p_j{}^\circ \leq C^\circ n^{-1/3} < 1.$$

C1' A unique choice of K-cutpoints $(X_{(n_1^\circ)}, \ldots, X_{(n_K^\circ)})$ that maximize

$$C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right) = n^{-1} \sum_{j=1}^{k} \frac{\left[\sum_{i=n_j}^{-1+n_{j+1}} T_i^{1/2}\right]^2}{\sum_{i=n_j}^{-1+n_{j+1}} T_i},$$

with probability 1 subject to the constraints that

$$0 < C_{\circ} n^{-1/3} \leq \Delta_j N^\circ = N_{j+1}{}^\circ - N_j{}^\circ = \frac{n_{j+1}{}^\circ - n_j{}^\circ}{n} \leq C^\circ n^{-1/3} < 1, \qquad (7)$$

$$n_j{}^\circ - n_{j-1}{}^\circ - a_n \leq n_{j+1}{}^\circ - n_j{}^\circ \leq n_j{}^\circ - n_{j-1}{}^\circ + a_n, \qquad (8)$$

where $a_n = n^{1/3} \log n$, $j = 1, \ldots, k$, some constants $C_\circ$ and C°.

The number k of cells is set at $k = \lambda n^{1/3}$ where λ is a constant. For any indices $\mathrm{n} = (n_1, \ldots, n_K)$ of the K-cutpoints $(X_{(n_1)}, \ldots, X_{(n_K)})$ that satisfy (7) and (8) in **C1'** and for a quantity Q_n independent of the cutpoints,

$$\sup_{1=n_1<n_2<\ldots<n_K=n-1} \left| C\left(X_{(1)}, \ldots, X_{(n-1)}, \mathrm{n}\right) - P\left(H, \mathrm{n}/n\right) - Q_n \right| = o_p\left(n^{-2/3}\right).$$

The set of indices $\mathrm{n}^\circ = (n_1^\circ, \ldots, n_K^\circ)$ that satisfies **C1'** converges to the set $\mathrm{p}^\circ = (p_1^\circ, \ldots, p_K^\circ)$ that maximize $P(H, \mathrm{p})$ in the sense that, for $N_j{}^\circ = n_j{}^\circ/n$,

$$\sup_{1 \leq j \leq K} |N_j{}^\circ - p_j{}^\circ| \to 0 \text{ in probability.}$$

For $\Delta_j N^\circ = N_{j+1}{}^\circ - N_j{}^\circ$ and $\Delta p_j{}^\circ = p_{j+1}{}^\circ - p_j{}^\circ$, the stronger result holds:

$$\sum_{j=1}^{k} \left[\frac{\Delta_j N^\circ}{\Delta p_j{}^\circ} - 1\right]^2 \Delta p_j{}^\circ = o_p(1).$$

PROOF

It can be shown that

$$\begin{aligned} C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right) &= \frac{\pi}{4}+O_p(k^{1/2}/n^{1/2})-O_p(k^{-2})+O_p(k/n) \\ P\left(H,\mathrm{n}/n\right) &= \frac{\pi}{4}-O_p(k^{-2})+o_p(n^{-2/3}) \end{aligned}$$

where $O_p(k^{-2})$ terms in $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ and $P\left(H,\mathrm{n}/n\right)$ are identical. Subtract appropriate terms that do not depend on the choice of the K-cutpoints. □

B1' is different from **B1** in that **B1'** imposes a constraint on the cell widths of g°. **C1'** differs from **C1** in two aspects: the cutpoints $(X_{(n_1^{\circ})},\ldots,X_{(n_K^{\circ})})$ of f_n° in **C1'** involves maximization of $C\left(X_{(1)},\ldots,X_{(n-1)},\mathrm{n}\right)$ while those $(X_{(n_1)},\ldots,X_{(n_K)})$ of f_n in **C1** do not; **C1'** imposes an additional constraint (8) that requires the neighboring cell widths of f_n° to change slowly. Applying (8) repeatedly gives

$$1-\frac{(k-1)a_n}{n_2^{\circ}-n_1^{\circ}}\leq\frac{n_K^{\circ}-n_k^{\circ}}{n_2^{\circ}-n_1^{\circ}}\leq 1+\frac{(k-1)a_n}{n_2^{\circ}-n_1^{\circ}}.$$

The ratio of the indices of the last cell to those of the first, $(n_K^{\circ}-n_k^{\circ})/(n_2^{\circ}-n_1^{\circ})$, is $O_p(\log n)$ because $k=O\left(n^{1/3}\right)$ and $n_2^{\circ}-n_1^{\circ}=O_p\left(n^{2/3}\right)$. Hence the bound $a_n=n^{1/3}\log n$ is large enough to guarantee the maximization in **C1'**.

We need a maximal inequality because the cutpoints themselves depend on the sample and are random variables. It is difficult to control terms under $n^{-2/3}$ because of the huge number $(n/k)^k$ of possible choices of cutpoints. Asymptotics for the known number of cells do not work. We have to reduce the number of possible choices of cutpoints by insisting neighboring cell sizes do not vary too rapidly.

n	Theoretical Mean	Sample Mean	MSE
10	0.750	0.579	0.0705
50	0.750	0.677	0.0400
100	0.750	0.729	0.0104
500	0.750	0.749	0.0005

Table 1: Error between sample and theoretical cutpoint in 100 repetitions

8 Simulation

For the underlying density in EXAMPLE 1 where the theoretical histogram is identical to the density, we assume that we know there are two cells in advance. Then we compute n_2°/n where $N$ is the index of the center cutpoint for sample sizes $n$ = 10, 50, 100, and 500 for one hundred times. Then we compute the mean and MSE to the theoretical center cutpoint $p = 0.750$. The result is in table 8. Convergence of n_2°/n in probability to $p = 0.750$ is observed from the result.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petrov B.N. and Cźaki P., editors, *2nd International Symposium on Information Theory*, pages 267–281, Akademiai Kiadó, Budapest, Hungary, 1973.

[2] E. Fix and J.L. Hodges Jr. *Discriminatory analysis, nonparametric discrimination: Consistency properties.* Technical Report No.4, USAF School of Aviation Medicine, Project No.21–49–004, Randolph Field, Texas, 1951.

[3] U. Grenander and M. Rosenblatt. *Statistical analysis of stationary time series.* Almqvist and Wiksell, Stockholm, 1956.

[4] U. Grenander and M. Rosenblatt. Statistical spectral analysis of time series arising from stationary stochastic processes. *The Annals of Mathematical Statistics*, 24:537–558, 1953.

[5] Y. Kanazawa. *Hellinger cross-validation for the variable cell histogram that achieves the theoretically optimal rate $n^{1/3}$ of number of cells.* Discussion Paper No.489, University of Tsukuba, Tsukuba, Ibaraki, 1992.

[6] Y. Kanazawa. Hellinger distance and Akaike's information criterion for the histogram. *Statistics & Probability Letters*, 17(5):*To appear*, 1993.

[7] Y. Kanazawa. Hellinger distance and Kullback-Leibler loss for the kernel density estimator. *Statistics & Probability Letters*, 18(5):*To appear*, 1993.

[8] Y. Kanazawa. An optimal variable cell histogram. *Communications in Statistics, Part A: Theory and Methods*, 17:1401–1422, 1988a.

[9] Y. Kanazawa. An optimal variable cell histogram based on the sample spacings. *The Annals of Statistics*, 20:291–304, 1992.

[10] Y. Kanazawa. *An optimal variable cell histogram based on the sample spacings.* PhD thesis, Yale University, 1988b.

[11] A. Kogure. Asymptotically optimal cells for a histogram. *The Annals of Statistics*, 15:1023–1030, 1987.

[12] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837, 1956.

[13] M. Rosenblatt. Stochastic curve estimation. In *NSF-CBMS Regional Conference Series in Probability and Statistics Volume 3*, sponsored by the Conference Board of the Mathematical Sciences and supported by the National Science Foundation, published by the Institute of Mathematical Statistics and the American Statistical Association, Hayward, California and Alexandria, Virginia, 1991.

[14] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.

[15] D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

[16] C.J. Stone. An asymptotically optimal histogram selection rule. In Le Cam L.M. and Olshen R.A., editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol.II*, pages 513–520, Wadsworth, Monterey, CA, 1984.

[17] C.C. Taylor. Akaike's information criterion and the histogram. *Biometrika*, 74:636–639, 1987.