EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 3, No. 1, 2010, 51-80
ISSN 1307-5543 – www.ejpam.com

Stochastic Complexity, Histograms and Hypothesis Testing of Homogeneity

Guoqi Qian

Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia.

Abstract. Information contained in a sample of quantitative data may be summarized or described by a nonparametric histogram density function. An interesting question is how to construct such a histogram density to express the data information with minimum stochastic complexity. The stochastic complexity is a pseudonym of Rissanen's minimum description length (MDL) which gives the length of a sequence of decipherable binary code resulted from optimally encoding the data information using a probability distribution based code-book. Here we have derived an optimal generalized histogram density estimator to provide both predictive and non-predictive coding description of a data sample. We have also obtained uniform and almost sure asymptotic approximations for the lengths of both descriptions. As an application of this result to statistical inference a new procedure for hypothesis testing of distribution homogeneity is proposed and is proved to have an asymptotic power of 1.

2000 Mathematics Subject Classifications: 62G07, 62G20, 60G10

Key Words and Phrases: Histogram density estimation, Minimum description length, Model selection, Quantization, Stochastic complexity, Test of homogeneity

1. Introduction

In digital data-transmission systems, input signals are first converted into digital form at the transmitter, then transmitted through a communication channel and finally reconstructed into output signals at the receiver. At the transmitter a quantization procedure is often executed in which the whole range of input amplitudes is divided into a finite number of amplitude sub-ranges and the input amplitudes in each sub-range are converted into the same digits. Such input digits are further encoded into a sequence of prefix binary digits for transmission. (Prefix codes are spontaneously and uniquely decipherable to where they are processed.) In order to achieve a cost-efficient transmission, an optimal encoding system is necessary by which the input binary digits sequence is as short as possible.

By Rissanen's stochastic complexity theory or principle of minimum description length (MDL) [18, 15, 13, 12, 11, 9], finding an optimal encoding system is equivalent to finding

Email address: g.qian@ms.unimelb.edu.au. (G. Qian)

the probability distribution underlying the input signals. [4, 3, 2] has developed an alternative theory of prequential analysis which implies the same conclusion. Another related is Bozdogan's information complexity criterion (ICOMP) (cf. [1]). However, the probability distribution for the input signals is mostly unknown and has to be estimated. In the context of digital data-transmission involving quantization aforementioned, it is sufficient to find a histogram density estimator of the probability distribution for the input data. A histogram density estimator is specified by a sequence of subintervals partitioning the range of the data (as corresponding to the input amplitude sub-ranges in quantization), and the probability values over all such subintervals.

Once a histogram density estimator is properly obtained, it determines the quantization of the input data and further enables the construction of an encoding system to encode the quantized input data. The length of the input binary codes obtained under this encoding system then measures the amount of information to be transmitted to the receiver, and we call it a summary *description* of the input data. The shorter this description is the more cost-efficient data transmission it would imply. Clearly, the optimal histogram density estimator is the one that would result in the shortest description of the input data. The discussions so far manifest the core of stochastic complexity theory — the principle of minimum description length — in the context of digital data-transmission.

The principles of MDL and maximum likelihood together provide a way for finding the best histogram density estimator. Suppose the input for digital transmission is a finite data-string $X^n = (X_1, \cdots, X_n)$ from a system involving chance, and we wish to estimate the probability distribution of this system by a histogram density for X^n. When number and locations of the subintervals to be used are specified for a histogram density estimator, the optimal probability over each subinterval can be determined by the maximum likelihood principle. When only the number of subintervals used in a histogram density estimator is specified, the optimal locations of the subintervals can also be determined by the maximum likelihood principle together with a recursive method. Thus for each specified number of the subintervals, a temporary histogram density estimator can be constructed which then provides a description of X^n with an appropriate prefix code-length. The optimal number of subintervals and accordingly the best histogram density estimator are therefore obtained from finding the shortest prefix code-length for data description.

Having seen the relationship between the best histogram density estimator and the shortest prefix code-length for data description, we now focus on investigating the shortest description length of X^n. Note that the prefix codewords of X^n can be obtained by either a non-predictive two-step or a predictive manner (see Chapter 3 of [13]). Even though the predictive coding requires longer codewords for encoding X^n, it enables the data-transmission system for self-adjustment and updating by using the data in a progressive way.

In Section 2 below we first discuss an optimal quantization scheme of the data for optimal description. The scheme provides a system of recursive equations for determining the optimal locations of the subintervals in the histogram estimator. Then lengths of both two-step and predictive codewords for the description of X^n are given using the MDL principle. Finally, uniform almost sure asymptotic expansion and the almost sure lower and upper bounds for both code lengths are derived and the results are list in Theorem 2 to Theorem 4.

In [7] and [20], the same type of stochastic complexity based histogram estimation is considered under the assumption of equal subinterval widths. Our results agree with theirs when this assumption applies.

As an application of stochastic complexity for optimal data description, in Section 3 we consider the problem of testing of homogeneity, i.e. the testing of the hypothesis that several independent samples are generated from the same population. A test procedure is proposed in which we use difference of the shortest predictive code lengths under the null and the alternative hypotheses respectively as a universal test statistic. The size of the test procedure is shown to be determined by the part of the code lengths which is used to describe the parameters in the histogram densities. The asymptotic power of the test procedure is shown to be 1.

2. Data Quantization for Optimal Information Description

Suppose $X^n = (X_1, \cdots, X_n)$ is a simple random sample from an unknown density function f on $[s, t]$, where s and t are finite real numbers. If f were known, the description of the sample could be accomplished by constructing a string of predictive binary codes for X^n under the information source determined by f where the description length is proportional to $-\log f(X^n)$. (See [13]; also the logarithm is in base 2 throughout this paper unless stated otherwise.) In other words, describing the sample is the same as finding a predictive probability density for the sample.

To estimate the unknown density f a frequently used method is based on data quantization: first quantize X^n by partitioning $[s, t]$ into a sequence of subintervals and then construct a histogram on the partition. The choice of the partition and the estimate of the probability for each subinterval may be determined by the maximum likelihood method if the number of subintervals is fixed.

Let $q^m = (q_{0,m}, q_{1,m}, \cdots, q_{m,m})$ be an increasing sequence of numbers, partitioning the interval $[s, t]$ into m subintervals $[q_{0,m}, q_{1,m}]$, $(q_{1,m}, q_{2,m}]$, $\cdots$, $(q_{m-1,m}, q_{m,m}]$, written as $Q_{1,m}, Q_{2,m}, \cdots, Q_{m,m}$, where $q_{0,m} = s$, $q_{m,m} = t$ and m is a fixed integer satisfying $m \leq n$. Denote $r_{i,m} = q_{i,m} - q_{i-1,m}$ as the length of $Q_{i,m}$ and $r = t - s$, the range of X^n. Consider the histogram densities defined by

$$f(x|p^m, q^m, s, t) = \sum_{i=1}^{m} \frac{p_{i,m}}{r_{i,m}} I_{Q_{i,m}}(x) \tag{1}$$

where $p^m = (p_{1,m}, p_{2,m}, \cdots, p_{m,m})$ denotes a sequence of nonnegative parameters with the sum equal 1, and $I_{Q_{i,m}}(\cdot)$ is the usual indicator function. The class of densities of the form (1) is denoted by H_m.

With the above notation, the log-likelihood function of the sample X^n under H_m is

$$L(X^n; H_m) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{m} \frac{p_{i,m}}{r_{i,m}} I_{Q_{i,m}}(X_j)\right) = \sum_{i=1}^{m} n_{i,m} \log \frac{p_{i,m}}{r_{i,m}} \tag{2}$$

where $n_{i,m} = \sum_{j=1}^{n} I_{Q_{i,m}}(X_j)$ is the number of data points falling into $Q_{i,m}$. If $n_{i,m}$ equals zero, the corresponding $p_{i,m}$ may not be uniquely optimized through maximization of $L(X^n; H_m)$. This difficulty may be overcome by introducing m numbers $y_1, \cdots, y_m$ (abbreviated as y^m, where y_i is regarded as an observation from the uniform distribution on $Q_{i,m}$, and mixing them thoroughly with the n observations X^n as if both y^m and X^n were generated from the same distribution. Then the log-likelihood function of X^n and y^m is

$$L_1(X^n; H_m) = \sum_{i=1}^{m} (n_{i,m} + 1) \log \frac{p_{i,m}}{r_{i,m}} \tag{3}$$

which does not depend on the particular values of y^m, and can, therefore, be regarded as the log-likelihood function of X^n.

Applying the maximum likelihood principle the optimal partition q^m and probabilities p^m for a fixed m are the ones which maximize $L_1(X^n, H_m)$ subject to the conditions that $\sum p_{i,m} = 1$ and $\sum r_{i,m} = r$. Denoting

$$F = \sum_{i=1}^{m} (n_{i,m} + 1) \log \frac{p_{i,m}}{r_{i,m}} + \lambda_1 (\sum_{i=1}^{m} p_{i,m} - 1) + \lambda_2 (\sum_{i=1}^{m} r_{i,m} - r), \tag{4}$$

differentiating F with respect to $p_{i,m}$'s and setting the derivatives equal to zero we have

$$\frac{\partial F}{\partial p_{i,m}} = \frac{n_{i,m} + 1}{p_{i,m}} \log e + \lambda_1 = 0, \quad i = 1, 2, \cdots, m \tag{5}$$

from which $p_{i,m} = (n_{i,m} + 1)/(n + m)$. Differentiating F with respect to $p_{i,m}$'s twice, the resulting second derivative matrix

$$\left(\frac{\partial^2 F}{\partial p_{i,m} \partial p_{j,m}} \right) = (\log e) \mathrm{diag} \left(-\frac{n_{1,m} + 1}{p_{1,m}^2}, \cdots, -\frac{n_{m,m} + 1}{p_{m,m}^2} \right) \leq 0. \tag{6}$$

Therefore a necessary condition for the maximization of (4) is that the probabilities $p_{i,m}$ are equal to the relative frequencies $(n_{i,m} + 1)/(n + m)$.

Both the allocation of $n_{i,m}$'s and the ranges $r_{i,m}$'s depend on the partition q^m. Thus the function F is not continuous with respect to $r_{i,m}$'s unless the allocation of $n_{i,m}$'s does not change. Under such allocation the local maximum/minimum value of $L_1(X^n; H_m)$ is achieved or converged to when $r_{i,m}$'s approach to their boundary values, i.e. where the resultant allocation of $n_{i,m}$'s would just about to change. Therefore, the global maximization of F with respect to $r_{i,m}$'s may not exist. In the light of the above discussions and in order to keep the code length needed for model description short, we may reasonably impose the restriction that the end points of every subinterval $Q_{i,m}$, except the two end points s and t, i.e. the sequence of the break points $q_{i,m}, \cdots, q_{m-1,m}$, should be at least d units away from the nearest observations, where $d > 0$ is half of the precision of X^n. In other words, if the locations of the sample X^n are expressed in an ascending order $z^N = z_1 < z_2 < \cdots < z_N$, where $N \leq n$ because of possible ties, then q^m is a subsequence of the $2N + 2$ long sequence

$$s, z_1 - d, z_1 + d, z_2 - d, z_2 + d, \cdots, z_N - d, z_N + d, t, \tag{7}$$

denoted as $s(X^n) = s_1, s_2, \cdots, s_{2N+2}$, with $q_{0,m} = s$ and $q_{m,m} = t$, such that the selected q^m achieves the largest likelihood $L_1(X^n; H_m)$ among all the selections.

There are $\binom{2N}{m-1}$ different selections for q^m within which the optimal sequence is to be found. In the following we provide a recursive method for finding the optimal q^m as well as the associated maximum likelihood value. A similar technique is used in [14]. Let

$$L_1^*(X^n; m) = \max_{q^m \subset s(X^n)} \max_{p^m : \sum p_{i,m} = 1} L_1(X^n; H_m) = \max_{q^m \subset s(X^n)} \sum_{i=1}^{m} (n_{i,m} + 1) \log \frac{n_{i,m} + 1}{(n+m) r_{i,m}}. \tag{8}$$

It is easy to see that

$$\begin{aligned} L_1^*(X^{n(\tau)}; m) = & \max_{s_{m-1} \le q_{m-1,m} \in s(X^{n(\tau)})} \left\{ \max_{\{q_{1,m}, \cdots, q_{m-2,m}\} \in s(X^{n(q_{m-1,m})})} L_1(X^{n(q_{m-1,m})}; H_{m-1}) \right. \\ & \left. + (n(\tau) - n(q_{m-1,m}) + 1) \log \frac{n(\tau) - n(q_{m-1,m}) + 1}{(n(\tau) + m) r_{m,m}} \right\} \\ = & \max_{s_{m-1} \le \nu \in s(X^{n(\tau)})} \left\{ L_1^*(X^{n(\nu)}; m-1) + (n(\tau) - n(\nu) + 1) \log \frac{n(\tau) - n(\nu) + 1}{(n(\tau) + m) r_{m,m}} \right\} \end{aligned} \tag{9}$$

where $X^{n(\nu)}$ denotes the sequence of the observations falling within $[s, \nu]$, and $n(\nu)$ denotes the number of the observations in $X^{n(\nu)}$. The recursive equation (9) are to be solved for $m \ge 1$ and $\nu \in s(X^{n(\tau)})$ with $\tau \le t$ until the desired range includes all the observations. That is, the following maximum log-likelihood functions need to be solved in sequence

$$\begin{array}{cccc} L_1^*(X^{n(s_2)}, 1), & L_1^*(X^{n(s_3)}, 1), & \cdots, & L_1^*(X^{n(t)}, 1), \\ L_1^*(X^{n(s_3)}, 2), & L_1^*(X^{n(s_4)}, 2), & \cdots, & L_1^*(X^{n(t)}, 2), \\ & \cdots & \cdots & \\ L_1^*(X^{n(s_{m+1})}, m), & L_1^*(X^{n(s_{m+2})}, m), & \cdots, & L_1^*(X^{n(t)}, m), \end{array} \tag{10}$$

for $m \le n$, where

$$L_1^*(X^{n(s_i)}, 1) = (n(s_i) + 1) \log \frac{1}{s_i - s}, \quad 2 \le i \le 2N + 2 \tag{11}$$

and

$$L_1^*(X^{n(s_k)}, k-1) = \sum_{i=1}^{k-1} (n(s_{i+1}) - n(s_i) + 1) \log \frac{n(s_{i+1}) - n(s_i) + 1}{(n(s_k) + k - 1)(s_{i+1} - s_i)}, \tag{12}$$

for $2 \le k \le n + 1$. For any fixed $m \le n$, the evaluation of (9) gives the maximum log-likelihood for X^n as well as the optimal partition $\{\tilde{Q}_{i,m}\}$ with about $m(4N + 3 - m)/2 \le 2m(n+2) - m^2/2$ operations. The corresponding optimal sequence of break points will be denoted by $\tilde{q}^m = (\tilde{q}_{1,m}, \cdots, \tilde{q}_{m,m})$, and the width of the subintervals by $\tilde{r}_{1,m}, \cdots, \tilde{r}_{m,m}$. In this paper data quantization will always be based on the optimal partition $\{\tilde{Q}_{i,m}\}$ (except in the case of equal width quantization). For sake of simple presentation the number of the data points falling into $\{\tilde{Q}_{i,m}\}$ will still be denoted as $n_{i,m} = \sum_{j=1}^{n} I_{\tilde{Q}_{i,m}}(X_j)$.

With an optimal procedure for the quantization of the data, we are now in a position to find the description of the data X^n. Following [13], the description length of the data X^n, for fixed m and corresponding $\tilde{q}^m$, is defined as a two-part code length

$$-L_1^*(X^n;m) + L_2(\tilde{q}^m, m, \delta) \tag{13}$$

where the first part $-L_1^*(X^n;m)$ can be interpreted as the code length needed to describe the data X^n under the given partition and histogram density estimator, and the second part L_2 is the code length needed to describe the functional form of the partition and histogram model employed. L_2 can be obtained by first truncating the parameter m and $\tilde{q}^m$ to a prescribed precision δ and then encoding the resulting integers with the technique introduced in [5] and [13]. Denote $\bar{a} = [a/\delta]$ as the nearest integer to a/δ, then

$$\begin{aligned} L_2(\tilde{q}^m, m, \delta) \quad = \quad & \log \binom{\sum_{i=1}^{m-1} \left|\overline{\tilde{r}_{i,m} - \frac{r}{m}}\right| + m - 2}{m-2} \\ & + \log 2.865 + \log^*(\bar{m} + |\bar{s}| + \bar{r} + 1) + \\ & \log \frac{(\bar{m} + |\bar{s}| + \bar{r} + 3)!}{(\bar{m} + |\bar{s}| + \bar{r})!2!} + \log \frac{4!}{a_+!(3 - a_+)!} + |\log \delta|. \end{aligned} \tag{14}$$

Here $\log^*(a) = \log a + \log\log a + \cdots$, with the sum including all the positive iterates, and a_+ is the number of nonnegative items in $\{\bar{m}, \bar{s}, \bar{r}\}$.

The length function (14) consists of three parts. Since the encoding of $\tilde{q}^m$ is equivalent to the encoding of $\tilde{r}_{1,m} - r/m, \cdots, \tilde{r}_{m-1,m} - r/m$, this will be achieved by a binary string beginning with $\overline{\tilde{r}_{1,m} - r/m}$ 0's and a 1, followed by $\overline{\tilde{r}_{2,m} - r/m}$ 0's and a 1, and so on until $\overline{\tilde{r}_{m-1,m} - r/m}$ 0's being added, but without attaching a 1 at the end, provided that m, s, t and d are given. Under this non-prefix encoding procedure the first term of (14) gives the code length of $\tilde{q}^m$. The second to the fifth terms of (14) are the code length needed for encoding $\bar{m}, \bar{s}$ and $\bar{t}$ (equivalent to $\bar{m}, \bar{s}$ and $\bar{r}$) in a prefix manner. In general we can encode a set of integers $\{\theta_1, \cdots, \theta_b\}$ in a prefix manner with about

$$L_3(\theta_1, \cdots, \theta_b) = \log 2.865 + \log^*(\theta + 1) + \log \frac{(\theta + b)!}{\theta!(b-1)!} + \log \frac{(b+1)!}{b_+!(b - b_+)!} \tag{15}$$

bits. Here $\theta = \sum_i |\theta_i|$, and b_+ is the number of nonnegative items in $\{\theta_1, \cdots, \theta_b\}$. The last term of (14) gives us the code length for encoding the truncation precision δ. Since a_+ equals either 2 or 3, the fifth term of (14) can be replaced by 1 reflecting the fact that one digit is needed to tell if $\bar{s}$ is negative or nonnegative.

With the description length defined by (13) the shortest code length for the data X^n by the above coding procedure is

$$\min_m \{-L_1^*(X^n;m) + L_2(\tilde{q}^m, m, \delta)\}$$

$$= -\sum_{i=1}^{m^*}(n_{i,m^*}+1)\log\frac{n_{i,m^*}+1}{(n+m^*)\tilde{r}_{i,m^*}} + L_2(\tilde{q}^{m^*}, m^*, \delta) \tag{16}$$

where the minimization is done by searching for an optimal integer $m^* \le n$ and δ is a prescribed precision.

If the sequence of break points are distributed uniformly in the interval $[s, t]$, then $\tilde{r}_{i,m} = r/m$ and the first term of (14) becomes zero. The code length (16) turns out to be

$$\min_m \left\{ -\sum_{i=1}^{m}(n_{i,m}+1)\log\frac{(n_{i,m}+1)m}{(n+m)r} + L_3(m, \bar{s}, \bar{r}) + |\log\delta| \right\}. \tag{17}$$

An alternative to (16) is to use the idea of the shortest predictive code length. This idea involves the ordering of the data X^n, either by location or by time of arrival, then finding the histogram density estimate based on the past and making appropriate modifications each time a new observation comes [13, 20]. In our situation the data X^n is ordered by location, as $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. For any fixed $m \le n$, an optimal sequence of break points $\tilde{q}^m$ is obtained by solving the recursive equation (9). Let $i(X_{(j)})$ be the unique integer i such that $X_{(j)} \in \tilde{Q}_{i,m}$, and $n_{i,m}(v) = \sum_{X_l \le v} I_{\tilde{Q}_{i,m}}(X_l)$ be the number of those X_l's satisfying $X_l \le v$ and falling into the i-th subinterval $\tilde{Q}_{i,m}$. The histogram density estimator based on the first j observations $X_{(1)}, \cdots, X_{(j)}$ can be written as

$$\tilde{f}(x|X_{(1)}, \cdots, X_{(j)}, m) = \sum_{i=1}^{m}\frac{n_{i,m}(X_{(j)})+1}{(j+m)\tilde{r}_{i,m}} I_{\tilde{Q}_{i,m}}(x) \tag{18}$$

and the likelihood of X^n can be constructed in a predictive manner as

$$\begin{aligned} \tilde{f}(X^n; m) &= \prod_{j=1}^{n}\tilde{f}(X_{(j)}|X_{(1)}, \cdots, X_{(j-1)}, m) \\ &= \prod_{j=1}^{n}\frac{n_{i(X_{(j)}),m}(X_{(j-1)})+1}{(j-1+m)\tilde{r}_{i(X_{(j)}),m}} \\ &= \frac{(m-1)!}{(n+m-1)!}\prod_{i=1}^{m}\frac{n_{i,m}!}{\tilde{r}_{i,m}^{n_{i,m}}}. \end{aligned} \tag{19}$$

In [14] $-\log\tilde{f}(X^n; m)$ is defined as the (predictive) stochastic complexity of X^n under the given partition. Now the shortest predictive code length for the data X^n is

$$\min_m \left\{ -\log\tilde{f}(X^n; m) + L_2(\tilde{q}^m, m, \delta) \right\}$$

$$= -\sum_{i=1}^{\hat{m}}\log n_{i,\hat{m}}! + \sum_{i=1}^{\hat{m}} n_{i,\hat{m}}\log\tilde{r}_{i,\hat{m}} - \log\frac{(\hat{m}-1)!}{(n+\hat{m}-1)!} + L_2(\tilde{q}^m, m, \delta) \tag{20}$$

where the minimization is achieved at $\hat{m} \leq n$ and δ is a prescribed precision. In particular, when the subintervals are of equal length, the expression (20) becomes

$$\min_{m} \left\{ n \log \frac{r}{m} + \log \binom{n}{n_{1,m}, \cdots, n_{m,m}} + \log \binom{n+m-1}{n} + L_3(m, \bar{s}, \bar{r}) + |\log \delta| \right\}. \qquad (21)$$

Having obtained the shortest code length (16) and shortest predictive code length (20), it is natural to ask how they differ in values. An asymptotic result is given below with its proof to be presented in Section 4.

Theorem 1. *Let X^n be a simple random sample from an unknown density function f on $[s,t]$. Suppose the following conditions are satisfied:*

(i). $0 < c_1 \leq f \leq c_2 < \infty$, *where c_1 and c_2 are constants;*

(ii). The number of subintervals m in the quantization of X^n satisfies

$$n^{\gamma_1} \leq m \leq n^{\gamma_2}, \qquad (22)$$

where γ_1 and γ_2 are two constants satisfying $0 < \gamma_1 < \gamma_2 < 1$*;*

(iii). The width $\tilde{r}_{i,m}$ of each optimal subinterval $\tilde{Q}_{i,m}$ satisfies

$$b_1 m^{-\alpha_1} \leq \tilde{r}_{i,m} \leq b_2 m^{-\alpha_2} \qquad (23)$$

uniformly for integers m in $[n^{\gamma_1}, n^{\gamma_2}]$, where b_1, b_2, α_1, α_2 are constants satisfying $1 \leq \alpha_1 < \frac{1}{2} + \frac{1}{2\gamma_2}$, *and* $\max\{0, 2\alpha_1 - \frac{1}{\gamma_2}\} < \alpha_2 \leq 1$.

Then uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$, the difference between the shortest code length and the shortest predictive code length of X^n is

$$-\log \tilde{f}(X^n; m) + L_1^*(X^n; m) = \alpha' m \log m + \frac{1}{2} m \log n + O(m) \quad \text{a.s.} \qquad (24)$$

where $-\frac{1}{2}\alpha_1 \leq \alpha' \leq -\frac{3}{2} + \alpha_1$.

Note that if the support of the density f is finite, then $\alpha_2 \leq 1 \leq \alpha_1$ is necessary for condition (iii) to hold. Also conditions (ii) and (iii) imply that $\alpha_2 \leq 1 \leq \alpha_1 < \frac{1}{\gamma_2} < \frac{1}{\gamma_1}$. Further, the righthand side of (24) becomes $\frac{1}{2} m \log \frac{n}{m} + O(m)$ a.s. if $\alpha_1 = \alpha_2 = 1$.

From Rissanen (2007) we know that the Shannon complexity is $-\log f^n(X^n) = -\sum_{i=1}^n \log f(X_i)$ if X^n is a simple random sample. The expectation of Shannon complexity represents the shortest code length of X^n on average if the underlying density f is known. The following three theorems show how the shortest code length (16) and the shortest predictive code length (20) differ from the Shannon complexity. The proof of these theorems will be presented in Section 4.

Theorem 2. *In addition to the conditions (i), (ii) and (iii) in Theorem 1, suppose that* *(iv). f is absolutely continuous with derivative $\dot{f}$ a.e. such that $|\dot{f}(x)| \leq c_3$ with c_3 a constant.*

Then uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ we have

(1).

$$\begin{aligned} & -Am^{\alpha_1} + (\alpha_2 - \alpha_1)m\log m + o(nm^{-2\alpha_2} + m\log m) \\ \leq \; & -L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) \\ \leq \; & (\alpha_1 - 1)m\log m + C_f nm^{-2\alpha_2} + o(nm^{-2\alpha_2} + m\log n) \quad \text{a.s.} \end{aligned} \tag{25}$$

if either $\alpha_1 \neq 1$ or $\alpha_2 \neq 1$; and

$$-L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) = O(nm^{-2} + m\log n) \quad \text{a.s.} \tag{26}$$

if $\alpha_1 = \alpha_2 = 1$.

(2).

$$\begin{aligned} & -Am^{\alpha_1} + \tfrac{1}{2}m\log n + (\alpha_2 - \tfrac{3}{2}\alpha_1)m\log m + o(nm^{-2\alpha_2} + m\log m) \\ \leq \; & -\log\tilde{f}(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) \\ \leq \; & \tfrac{1}{2}m\log n + (2\alpha_1 - \tfrac{5}{2})m\log m + C_f nm^{-2\alpha_2} + o(nm^{-2\alpha_2} + m\log n) \;\; \text{a.s.} \end{aligned} \tag{27}$$

if either $\alpha_1 \neq 1$ or $\alpha_2 \neq 1$; and

$$\begin{aligned} & -\log\tilde{f}(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) = \\ & \frac{1}{2}m\log\frac{n}{m} + C_f' nm^{-2} + o(nm^{-2} + m\log n) \quad \text{a.s.} \end{aligned} \tag{28}$$

if $\alpha_1 = \alpha_2 = 1$.

Here $\log f^n(X^n) = \prod_{j=1}^n f(X_j)$, $C_f = 24^{-1} b_2 \int_s^t \dot{f}^2 f^{-1}$, $A > 0$ is a constant and C_f' is a constant between $C_f b_1 b_2^{-1}$ and C_f.

The upper bounds in (25) and (27) imply that, for a given number of subintervals m, the shortest predictive code length (20) is likely to involve more redundant code length in encoding the unknown f than the shortest code length (16).

Theorem 3. *Under the conditions of Theorem 1 and Theorem 2 and having either $\alpha_1 \neq 1$ or $\alpha_2 \neq 1$, we have*

$$\begin{aligned} & -M_1(n^{\alpha_1\gamma_2} + n^{\gamma_2}\log n) \\ \leq \; & \min_{m \in [n^{\gamma_1}, n^{\gamma_2}]}\{-L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta)\} + \log f^n(X^n) \\ \leq \; & M_2 n^{\frac{1}{1+2\alpha_2}} (\log n)^{\frac{2\alpha_2}{1+2\alpha_2}} \quad \text{a.s.} \end{aligned} \tag{29}$$

and

$$-M_3(n^{\alpha_1\gamma_2} + n^{\gamma_2}\log n)$$

$$\begin{aligned} &\leq \min\nolimits_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-\log\tilde{f}(X^n;m)+L_2(\tilde{q}^m,m,\delta)\}+\log f^n(X^n) \\ &\leq M_4 n^{\frac{1}{1+2\alpha_2}}(\log n)^{\frac{2\alpha_2}{1+2\alpha_2}} \quad \text{a.s.} \end{aligned} \tag{30}$$

where M_1, M_2, M_3, M_4 are positive constants depending on f.

Theorem 3 implies that, even though for a fixed m the predictive code length (27) may be longer than the code length (25) by an infinite number of digits as $n \to \infty$, both of them have the minimax bounds of the same order. Theorem 3 can be further refined when $\alpha_1 = \alpha_2 = 1$ is assumed, which is given below.

Theorem 4. *Under the conditions of Theorem 1 and Theorem 2 and that $\alpha_1 = \alpha_2 = 1$, the following statements hold.*

(a).

$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-L_1^*(X^n;m)+L_2(\tilde{q}^m,m,\delta)\}+\log f^n(X^n) = O(n^{\frac{1}{3}}(\log n)^{\frac{2}{3}}) \quad \text{a.s.} \tag{31}$$

(b).

$$\begin{aligned} &\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-\log\tilde{f}(X^n;m)+L_2(\tilde{q}^m,m,\delta)\}+\log f^n(X^n) \\ &\qquad = M_5 n^{\frac{1}{3}}(\log n)^{\frac{2}{3}}(1+o(1)) \quad \text{a.s.} \end{aligned} \tag{32}$$

(c).

$$m^* = O((n/\log n)^{\frac{1}{3}}) \quad \text{a.s.} \tag{33}$$

(d).

$$\hat{m} = M_6(n/\log n)^{\frac{1}{3}}(1+o(1)) \quad \text{a.s.} \tag{34}$$

where M_5 and M_6 are positive constants depending on f.

Note that $\alpha_1 = \alpha_2 = 1$ implies the width of each subinterval in the histogram, although still being variable, is of the same order as m^{-1}. The results (32) and (34) are the same as (ii) and (iv) of Theorem 2.4 of [20] where they use predictive histogram estimator of equal width subintervals. This shows that using a variable subinterval-width optimal histogram density estimator, if the widths are of the same order, achieves the same order of shortest code length for description of X^n as using an equal width optimal histogram density estimator. From Theorem 4 we also see that the results for the shortest predictive code length are more definite than for the shortest code length. Therefore, we will focus our study on the predictive code length in the next section.

3. Hypothesis testing for homogeneity

One of the basic problems in statistical inference is testing the equality of two distributions where two independent samples are observed from; and more generally, testing the homogeneity of k distributions with $k > 2$ where k independent samples are observed from. Using the data quantization method developed in Section 2 we propose a stochastic complexity based procedure for testing the homogeneity of k distributions. First, a shortest predictive code length, based on a class of histogram density estimators with variable width subintervals, is computed for each of the k independent samples. This, when minimized, gives the optimal number of subintervals and their locations, the associated density estimator and the proper measurement of the information contained in each sample. Second, the shortest predictive code length is computed for the pooled sample, which when minimized gives the histogram estimator of the associated mixture distribution. Finally, the shortest predictive code length of the pooled sample is compared with the sum of the shortest predictive code lengths of all the k samples. If the former one is smaller, then the hypothesis that the k distributions are the same is not rejected, but rejected otherwise.

Let $(X_{11},\cdots,X_{1n_1}),(X_{21},\cdots,X_{2n_2}),\cdots,(X_{k1},\cdots,X_{kn_k})$ (abbreviated as $X_1^{n_1}, X_2^{n_2}, \cdots, X_k^{n_k}$) be k independent random samples with sizes $n_1, n_2, \cdots, n_k$ and $\sum_{i=1}^k n_i = n$. The respective unknown population density functions are $f_1(x), f_2(x), \cdots, f_k(x)$, all with the same support $[s,t]$. The underlying problem is the testing of the hypothesis

$$\begin{aligned} H_0: &\quad f_1 = f_2 = \cdots = f_k \quad \text{versus} \\ H_a: &\quad \text{at least two of them are not equal.} \end{aligned} \tag{35}$$

Under the alternative hypothesis H_a, we should describe the information of the k samples $X_1^{n_1}, X_2^{n_2}, \cdots, X_k^{n_k}$ separately, i.e. we should find the shortest code length for each sample $X_i^{n_i}$ with density f_i. By (20) the total predictive code length for the k samples is

$$\min_{m_1,\cdots,m_k}\left\{-\sum_{i=1}^{k}\log\tilde{f}_i(X_i^{n_i};m_i)+\sum_{i=1}^{k}L_2(\tilde{q}_i^{m_i},m_i,\delta)\right\} \tag{36}$$

provided that the parameter truncation is based on the same precision δ. Here $\tilde{f}_i(X_i^{n_i};m_i)$ is the likelihood function of the i-th sample $X_i^{n_i}$ defined as (19), i.e.

$$\tilde{f}_i(X_i^{n_i};m_i)=\frac{(m_i-1)!}{(n_i+m_i-1)!}\prod_{j=1}^{m_i}\frac{n_{i,j,m_i}!}{\tilde{r}_{i,j,m_i}^{n_{i,j,m_i}}} \tag{37}$$

where $\tilde{r}_{i,j,m_i}$'s are the widths of the optimal partition $\{\tilde{Q}_{i,j,m_i}\}$ of the i-th sample. These are obtained by applying the maximum likelihood principle (4) to the i-th sample with fixed number of subintervals m_i. Then n_{i,j,m_i} is the number of data points falling into the j-th subinterval $\tilde{Q}_{i,j,m_i}$.

Because all of the k samples are encoded simultaneously, the second term of (36) could be further reduced by a more efficient encoding process as given by

$$L_4(\tilde{q}_1^{m_1},\cdots,\tilde{q}_k^{m_k},m_1,\cdots,m_k,\delta)=\sum_{i=1}^{k}\log\binom{\sum_{j=1}^{m_i-1}\left|\overline{\tilde{r}_{i,j,m_i}-\frac{r}{m_i}}\right|+m_i-2}{m_i-2}$$
$$+L_3(m_1,\cdots,m_k,\bar{s},\bar{r})+|\log\delta| \qquad (38)$$

where $\tilde{q}_i^{m_i}$ is the sequence of break points corresponding to the optimal partition $\{\tilde{Q}_{i,j,m_i}\}$. The efficiency lies in the fact that the set of integers $\{m_1,\cdots,m_k,\bar{s},\bar{r}\}$ is encoded in a prefix manner, which requires $L_3(m_1,\cdots,m_k,\bar{s},\bar{r})$ bits instead of $\sum_{i=1}^{k}L_3(m_i,\bar{s},\bar{r})$ bits. Therefore under the hypothesis H_a the total predictive code length (36) for the k samples can be replaced by a shorter code length

$$C(X_1^{n_1},\cdots,X_k^{n_k})=\min_{m_1,\cdots,m_k}\{-\sum_{i=1}^{k}\log\tilde{f}_i(X_i^{n_i};m_i)$$
$$+L_4(\tilde{q}_1^{m_1},\cdots,\tilde{q}_k^{m_k},m_1,\cdots,m_k,\delta)\} \qquad (39)$$

where the minimum is attained at $\hat{m}_1,\cdots,\hat{m}_k$.

If the null hypothesis H_0 is true, that is the k samples are drawn from the same unknown distribution, we can describe the information in the k samples using only the optimal codewords required to encode the pooled sample $X^n=(X_1^{n_1},\cdots,X_k^{n_k})$. By regarding the pooled sample X^n as being drawn from a mixed distribution with density $f_{\text{mix}}=\sum_{i=1}^{k}\frac{n_i}{n}f_i$, the shortest predictive code length for encoding X^n is the one defined by (20):

$$C(X^n)=\min_m\{-\log\tilde{f}_{\text{mix}}(X^n;m)+L_2(\tilde{q}_{\text{mix}}^m,m,\delta)\}$$
$$=-\sum_{j=1}^{\hat{m}}\log n_{j,\hat{m}}!+\sum_{j=1}^{\hat{m}}n_{j,\hat{m}}\log\tilde{r}_{j,\hat{m}}-\log\frac{(\hat{m}-1)!}{(n+\hat{m}-1)!}+L_2(\tilde{q}_{\text{mix}}^{\hat{m}},\hat{m},\delta) \qquad (40)$$

where the minimum is attained at $\hat{m}$.

According to the theory of stochastic complexity, under the right probabilistic model (here the density function), or the right constraints inside the probabilistic pattern of the observations, the corresponding encoding process is expected to produce a shorter code length than the one under a wrong model, or the one that ignores the right constraints in the underlying model. Therefore, in the problem of testing the hypotheses (35), when H_0 is true the shortest predictive code length (40) is not expected to be greater than (39), the shortest predictive code length obtained under the encoding process ignoring the constraint $f_1=f_2=\cdots=f_k$. And vise versa, when H_a is true the corresponding code length (39) is expected to be less than the code length (40), which is obtained by the encoding process ignoring the difference among the f_i's. We summarize this property in the following theorem.

Theorem 5. *Let $X_1^{n_1},\cdots,X_k^{n_k}$ be simple random samples, respectively, drawn from the unknown density functions $f_1,\cdots,f_k$ on $[s,t]$, and $X^n=(X_1^{n_1},\cdots,X_k^{n_k})$ the pooled sample. Suppose that the conditions (i) to (iv) listed in Theorem 1 and Theorem 2 are satisfied for each $X_i^{n_i}$ and the corresponding f_i. Then the following statements hold.*

(a). If at least two of $f_1, \cdots, f_k$ are not equal, there exists a constant $\eta < 0$ such that

$$\frac{1}{n}[C(X_1^{n_1}, \cdots, X_k^{n_k}) - C(X^n)] < \eta \quad \text{a.s.} \tag{41}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$ satisfying $\liminf_{n_1 \to \infty} \frac{n_1}{n} > 0$, $\cdots$, $\liminf_{n_k \to \infty} \frac{n_k}{n} > 0$.

(b). If $f_1 = f_2 = \cdots = f_k$ a.s., then

$$\frac{1}{n}[C(X_1^{n_1}, \cdots, X_k^{n_k}) - C(X^n)] \to 0 \quad \text{a.s.} \tag{42}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$.

The proof of Theorem 5 will be given in Section 4.

From part (a) of the above theorem we know that when using the test procedure stated in the beginning of this section, the asymptotic power is 1 in the limit as the sample sizes tend to infinity; namely, almost surely the shortest predictive code length under H_a is less than that under H_0 when H_a is true. On the other hand, when the null hypothesis H_0 is true, the difference of the two shortest predictive code lengths per observation converges to zero almost surely as the sample sizes go to infinity. When the sample sizes are finite, the size of the test is essentially determined by the part of the code lengths used for encoding the parameters. In the encoding process corresponding to (39) there are more parameters $(\tilde{q}_1^{m_1}, \cdots, \tilde{q}_k^{m_k}, m_1, \cdots, m_k, \delta)$ to be encoded than in the encoding process corresponding to (40) in which only $\tilde{q}_{\text{mix}}^m, m$ and δ are to be encoded. Thus the code length (39) has a high probability to be larger than (40) if the null hypothesis H_0 is true, implying the test has a firm control of type I error. Simulation study could be done to investigate how well the two types of errors are controlled in the proposed test. But we will not get into the details in this paper. A simulation study was done in [10] to assess the finite sample performance of the homogeneity test proposed in this section in the special case of using equal width histogram density estimators. The results there show that the method is competitive in comparison to the existent methods such as the two sample t test and Smirnov test when the data can be analyzed by all these methods. But the proposed method is more efficient and powerful in some situations such as that the data come from different families of distributions, whereas the other methods do not perform well.

4. Proofs of the theorems

In this section we provide proofs for all the theorems listed in this paper. For the sake of simplicity, the logarithms in the proofs are all natural logarithms.

From (8) and (19),

$$-\log \tilde{f}(X^n; m) + L_1^*(X^n; m) =$$

$$-\log\left\{\frac{(m-1)!}{(n+m-1)!}\prod_{i=1}^{m}\frac{n_{i,m}!}{\tilde{r}_{i,m}^{n_{i,m}}}\right\} + \sum_{i=1}^{m}(n_{i,m}+1)\log\frac{n_{i,m}+1}{(n+m)\tilde{r}_{i,m}}. \tag{43}$$

By Stirling's formula $n! = \sqrt{2\pi n} n^n e^{-n} e^{\theta_n}$ $(0 < \theta_n < (12n)^{-1})$, (43) can be rewritten as

$$
\begin{aligned}
-\log \tilde{f}(X^n; m) + L_1^*(X^n; m) = {} & -\sum_{i=1}^{m} \log \tilde{r}_{i,m} + \sum_{n_{i,m}>0} \log \left(1 + \frac{1}{n_{i,m}}\right)^{n_{i,m}+1} \\
& + \frac{1}{2} \sum_{n_{i,m}>0} \log n_{i,m} - m \log m + O(m). \qquad (44)
\end{aligned}
$$

We will show that the first term of (44) is $\alpha_{12} m \log m + O(m)$ where $1 \le \alpha_{12} \le \alpha_1$, the second term is $O(m)$ and the third term is $\frac{m}{2} \log \frac{n}{m^{\alpha_{22}}}$ where $1 \le \alpha_{22} \le \alpha_1$. The following lemmas will be needed.

Lemma 1. *Suppose that $n_{i,m}$'s have a multinomial distribution with probabilities $\pi_{i,m}$'s such that $\sum_{i=1}^{m} \pi_{i,m} = 1$, $\pi_{i,m} \ge b_1 c_1 m^{-\alpha_1}$ and $\sum_{i=1}^{m} n_{i,m} = n$. Then for each integer w, there exists a constant a_w such that*

$$
E\left\{\sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right\}^{2w} \le a_w n^{\frac{1}{2}-w} m^{2\alpha_1 w}. \qquad (45)
$$

Proof. Denote $T_1 = \sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}$. By the definition of multinomial distribution and from Stirling's formula we have

$$
\begin{aligned}
E(T_1^{2w}) &= \sum_{n_{1,m}+\cdots+n_{m,m}=n} T_1^{2w} \frac{n!}{\prod_{i=1}^{m} n_{i,m}!} \prod_{i=1}^{m} \pi_{i,m}^{n_{i,m}} \\
&= \sum_{n_{1,m}+\cdots+n_{m,m}=n} T_1^{2w} \prod_{i=1}^{m} \frac{(n\pi_{i,m})^{n_{i,m}}}{n_{i,m}!} e^{-n_{i,m}} \sqrt{2\pi n} e^{c_n} \\
&\le \sqrt{2\pi n} e \sum_{N=0}^{\infty} \sum_{n_{1,m}+\cdots+n_{m,m}=N} T_1^{2w} \prod_{i=1}^{m} \frac{(n\pi_{i,m})^{n_{i,m}}}{n_{i,m}!} e^{-n_{i,m}} \\
&= \sqrt{2\pi n} e \sum_{n_{1,m}=0}^{\infty} \cdots \sum_{n_{m,m}=0}^{\infty} T_1^{2w} \prod_{i=1}^{m} \frac{(n\pi_{i,m})^{n_{i,m}}}{n_{i,m}!} e^{-n_{i,m}} \\
&= \sqrt{2\pi n} e E'(T_1^{2w}) \qquad (46)
\end{aligned}
$$

where the final expectation $E'(T_1^{2w})$ is with respect to a series of independent Poisson random variables $\{n_{i,m}\}$ with parameters $\{n\pi_{i,m}\}$. This technique, used by [16] and [19], of approximating the multinomial distribution by Poisson distribution is called Poissonization. The constant $c_n = o((12n)^{-1})$. By Shiryayev's [17] Theorem 6 of section 2.12, the $2w$-th moment of T_1 can be written as a sum of its cumulants:

$$
E'(T_1^{2w}) = \sum_{j_1+\cdots+j_l=2w} \rho(j_1, \cdots, j_l) \prod_{k=1}^{l} \kappa_{jk}(T_1) \qquad (47)
$$

where $\rho(j_1,\cdots,j_l)=\frac{1}{l!}\frac{(2w)!}{j_1!\cdots j_l!}$ and $j_k\geq 1$, $l\leq 2w$. Because $n_{i,m}$'s are independent Poisson random variables, it follows from the section 1.5 of [8] that the j_k-th cumulant of T_1

$$\kappa_{jk}(T_1)=\sum_{i=1}^{m}\kappa_{jk}\left(\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right)=\sum_{i=1}^{m}\frac{n\pi_{i,m}}{(n\pi_{i,m})^{j_k}}\leq(b_1c_1)^{-j_k}n^{1-j_k}m^{\alpha_1 j_k} \tag{48}$$

if $j_k>1$ and $\kappa_{jk}(T_1)=0$ if $j_k=1$. Thus

$$E'(T_1^{2w})=\sum^{*}\rho(j_1,\cdots,j_l)\prod_{k=1}^{l}\kappa_{jk}(T_1)$$

$$\leq\sum^{*}\rho(j_1,\cdots,j_l)\prod_{k=1}^{l}(b_1c_1)^{-j_k}n^{1-j_k}m^{\alpha_1 j_k}\leq a_w n^{-w}m^{2\alpha_1 w} \tag{49}$$

where the summation $\sum^{*}$ is taken over all partitions of $2w$ such that $\sum_{k=1}^{l}j_k=2w$, $j_k\geq 2$ and $l\leq w$. Using the same notation for possibly different constants and substituting the last bound into (46) the lemma is proved. ◁

Lemma 2. *Suppose that N is a binomial random variable with mean np. Then for any integer $w>0$, there is a constant $a_w>0$ such that*

$$E(N-np)^{2w}\leq a_w n^{\frac{1}{2}+w}p^w. \tag{50}$$

Proof. By the same technique of Poissonization we have

$$E(N-np)^{2w}\leq a_w n^{\frac{1}{2}}E(N_1-np)^{2w} \tag{51}$$

where N_1 is a Poisson random variable with mean np. By the equation (47) and the fact that $\kappa_k(N_1-np)=np$ if $k>1$ and $\kappa_1(N_1-np)=0$ it follows that $E(N_1-np)^{2w}$ is a polynomial of order w, and therefore (50) must hold. ◁

Lemma 3. *Under the conditions that $\tilde{r}_{i,m}\geq b_1m^{-\alpha_1}$, $1\leq\alpha_1<1+(2\gamma_2)^{-1}$ and $f\geq c_1>0$,*

$$\sum_{i=1}^{m}\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}=o(m)\quad\text{a.s.} \tag{52}$$

uniformly in $m\in[1,n^{\gamma_2}]$ as $n\to\infty$, where $\pi_{i,m}=\int_{\tilde{Q}_{i,m}}f$.

Proof. For any $\varepsilon>0$,

$$P\left(\max_{m\in[1,n^{\gamma_2}]}\left|\sum_{i=1}^{m}\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right|\geq\varepsilon m\right)\leq\sum_{m\in[1,n^{\gamma_2}]}P\left(\left|\sum_{i=1}^{m}\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right|\geq\varepsilon m\right)$$

$$\le \sum_{m\in[1,n^{\gamma_2}]} \varepsilon^{-2w} m^{-2w} E\left\{\sum_{i=1}^{m} \frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right\}^{2w} \tag{53}$$

where the last inequality comes by using Chebyshev's inequality. From Lemma 1,

$$\begin{aligned} P\left(\max_{m\in[1,n^{\gamma_2}]}\left|\sum_{i=1}^{m}\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right| \ge \varepsilon m\right) &\le \sum_{m\in[1,n^{\gamma_2}]} \varepsilon^{-2w} m^{-2w} a_w n^{\frac{1}{2}-w} m^{2\alpha_1 w} \\ &\le a_w \varepsilon^{-2w} n^{(2\alpha_1\gamma_2-2\gamma_2-1)w+\frac{1}{2}+\gamma_2}. \end{aligned} \tag{54}$$

By the condition that $\alpha_1 < 1+(2\gamma_2)^{-1}$, the sum of the series defined by the terms of form (54) converges as $n\to\infty$, if $w > \frac{3+2\gamma_2}{2+4\gamma_2-4\alpha_1\gamma_2}$. Hence (52) follows from applying the Borel-Cantelli lemma. ◁

Lemma 4. *Under the conditions that* $b_1 m^{-\alpha_1} \le \tilde{r}_{i,m} \le b_2 m^{-\alpha_2}$, *where* $2\alpha_1-\gamma_2^{-1} < \alpha_2 \le 1 \le \alpha_1 < 2^{-1}+(2\gamma_2)^{-1}$ *and* $0 < c_1 \le f \le c_2$,

$$\max_{1\le i\le m}\left|\frac{m^{\alpha_1}n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m}\right| = o(1) \quad \text{a.s.} \tag{55}$$

uniformly in $m \in [1,n^{\gamma_2}]$ *as* $n\to\infty$.

Proof. Denote

$$I_{m,n} = \max_{1\le i\le m}\left|\frac{m^{\alpha_1}n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m}\right|, \tag{56}$$

then for any $\varepsilon > 0$,

$$\begin{aligned} P\left(\max_{m\in[1,n^{\gamma_2}]} I_{m,n} > \varepsilon\right) &\le \sum_{m\in[1,n^{\gamma_2}]} P(I_{m,n} > \varepsilon) \\ &\le \sum_{m\in[1,n^{\gamma_2}]}\sum_{i=1}^{m} P\left(\left|\frac{m^{\alpha_1}n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m}\right|\right) \\ &\le \sum_{m\in[1,n^{\gamma_2}]}\sum_{i=1}^{m} \varepsilon^{-2w} E\left|\frac{m^{\alpha_1}n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m}\right|^{2w} \\ &= \sum_{m\in[1,n^{\gamma_2}]}\sum_{i=1}^{m} \varepsilon^{-2w} m^{2w\alpha_1} n^{-2w} E(n_{i,m}-n\pi_{i,m})^{2w}, \end{aligned} \tag{57}$$

where the last inequality is obtained by applying Chebyshev's inequality. From Lemma 2 and the property that $c_1 b_1 m^{-\alpha_1} \le \pi_{i,m} \le c_2 b_2 m^{-\alpha_2}$,

$$P\left(\max_{m\in[1,n^{\gamma_2}]} I_{m,n} > \varepsilon\right) \le \sum_{m\in[1,n^{\gamma_2}]}\sum_{i=1}^{m} \varepsilon^{-2w} m^{2w\alpha_1} n^{-2w} a_w n^{\frac{1}{2}+w} (c_2 b_2 m^{-\alpha_2})^{w}$$

$$\leq \quad a_w n^{2\gamma_2+\frac{1}{2}+(2\alpha_1\gamma_2-\alpha_2\gamma_2-1)w}. \tag{58}$$

From now on the same notation will be used for possibly different constants. By the condition that $\alpha_2 > 2\alpha_1 - \gamma_2^{-1}$, the sum of the series defined by the terms of form (58) converges as $n \to \infty$, if $w > \frac{3+4\gamma_2}{2+2\alpha_2\gamma_2-4\alpha_1\gamma_2}$. Hence (55) follows from applying the Borel-Cantelli lemma. ◃

Lemma 5. *Under the conditions that $\tilde{r}_{i,m} \leq b_2 m^{-\alpha_2}$ and $f \leq c_2$, we have*

$$\sum_{i=1}^{m}(n_{i,m} - n\pi_{i,m})^2 - n = o(n) \quad \text{a.s.} \tag{59}$$

uniformly in $m \in [n^{\gamma_1}, n]$ as $n \to \infty$.

Proof. Suppose $\{N_{i,m}\}$ are a sequence of independent Poisson random variables with means $\{n\pi_{i,m}\}$, and denote $T_2 = \sum_{i=1}^{m}(N_{i,m} - n\pi_{i,m})^2$. We first show that the j-th cumulants of T_2 satisfies

$$|\kappa_j(T_2)| \leq \alpha_j n^j m^{-\alpha_2(j-1)} \tag{60}$$

where a_j is a constant depending on j.

Because $\{N_{i,m}\}$ are independent, it follows that

$$\kappa_j(T_2) = \sum_{i=1}^{m}\kappa_j((N_{i,m} - n\pi_{i,m})^2). \tag{61}$$

By applying Theorem 6 of section 2.12 of Shiryayev's [17] again, the j-th cumulants of $(N_{i,m} - n\pi_{i,m})^2$ can be written as a sum of its moments:

$$\kappa_j((N_{i,m} - n\pi_{i,m})^2) = \sum_{j_1+\cdots+j_l=j} \zeta(j_1,\cdots,j_l)\prod_{k=1}^{l} E((N_{i,m} - n\pi_{i,m})^{2j_k}) \tag{62}$$

where $\zeta(j_1,\cdots,j_l) = \frac{(-1)^{l-1}}{l}\frac{j!}{j_1!\cdots j_l!}$ and $j_k \geq 1$, $l \leq j$. From Lemma 2 we know that $E((N_{i,m} - n\pi_{i,m})^{2j_k})$ is an order-j_k polynomial of $n\pi_{i,m}$, therefore

$$|\kappa_j(T_2)| \leq \sum_{i=1}^{m}\alpha_j(n\pi_{i,m})^j \leq \alpha_j n^j m^{-\alpha_2(j-1)} \tag{63}$$

for some constant a_j, hence (60) holds.

By (47) and the identities $\kappa_1(T_2 - n) = E(T_2 - n) = 0$ and $\kappa_j(T_2 - n) = \kappa_j(T_2)$ for $j \geq 2$, it can be seen that

$$E(T_2 - n)^{2w} = \sum^{*}\rho(l_1,\cdots,l_k)\prod_{j=1}^{k}\kappa_{l_j}(T_2) \tag{64}$$

where the summation $\sum^{*}$ is taken over all the partitions of $2w$ such that $\sum_{j=1}^{k} l_j = 2w$, $l_j \geq 2$ and $k \leq w$. By (60) it follows that

$$E(T_2 - n)^{2w} \leq \sum^{*} a_w n^{2w} m^{-\alpha_2(2w-k)} \leq a_w n^{2w} m^{-\alpha_2 w} \tag{65}$$

for some constant a_w depending on w.

Now for any $\varepsilon > 0$,

$$\begin{aligned}
&P\left(\max_{m\in[n^{\gamma_1},n]}\left|\sum_{i=1}^{m}(n_{i,m} - n\pi_{i,m})^2 - n\right| > \varepsilon n\right) \\
&\leq \sum_{m\in[n^{\gamma_1},n]} P\left(\left|\sum_{i=1}^{m}(n_{i,m} - n\pi_{i,m})^2 - n\right| > \varepsilon n\right) \\
&\leq \sum_{m\in[n^{\gamma_1},n]} \varepsilon^{-2w} n^{-2w} E\left|\sum_{i=1}^{m}(n_{i,m} - n\pi_{i,m})^2 - n\right|^{2w} \\
&\leq \sum_{m\in[n^{\gamma_1},n]} \varepsilon^{-2w} n^{-2w+\frac{1}{2}} E\left|\sum_{i=1}^{m}(N_{i,m} - n\pi_{i,m})^2 - n\right|^{2w}
\end{aligned} \tag{66}$$

by applying Chebyshev's inequality and the technique of Poissonization.

From (65) it follows that

$$\begin{aligned}
&P\left(\max_{m\in[n^{\gamma_1},n]}\left|\sum_{i=1}^{m}(n_{i,m} - n\pi_{i,m})^2 - n\right| > \varepsilon n\right) \\
&\leq \sum_{m\in[n^{\gamma_1},n]} \varepsilon^{-2w} n^{-2w+\frac{1}{2}} a_w n^{2w} m^{-\alpha_2 w} \leq a_w n^{\frac{3}{2}-\alpha_2\gamma_1 w}.
\end{aligned} \tag{67}$$

The sum of the series defined by the terms of form (67) converges as $n \to \infty$ if $w > \frac{5}{2\alpha_2\gamma_1}$, hence (59) follows by applying the Borel-Cantelli Lemma. ◁

Corollary 1. *Under the conditions that $b_1 m^{-\alpha_1} \leq \tilde{r}_{i,m} \leq b_2 m^{-\alpha_2}$ and $0 < c_1 \leq f \leq c_2$,*

$$\sum_{i=1}^{m} \frac{(n_{i,m} - n\pi_{i,m})^2}{(n\pi_{i,m})^2} = O(n^{-1}m^{2\alpha_1}) \quad \text{a.s.} \tag{68}$$

uniformly in $m \in [n^{\gamma_1}, n]$ as $n \to \infty$.

Lemma 6. *Under the conditions of Lemma 4, the following statement is true:*

$$\sum_{n_{i,m}>0} \log \frac{n_{i,m}}{n\pi_{i,m}} = O(m) \quad \text{a.s.} \tag{69}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$.

Proof. First note that

$$\sum_{n_{i,m}>0} \log \frac{n_{i,m}}{n\pi_{i,m}} = \sum_{n_{i,m}>0} \log\left(1 + \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right). \tag{70}$$

By Taylor expansion,

$$\sum_{n_{i,m}>0} \log \frac{n_{i,m}}{n\pi_{i,m}} = \sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} - \frac{1}{2} \sum_{n_{i,m}>0} (1+\xi_{i,m})^{-2} \frac{(n_{i,m} - n\pi_{i,m})^2}{(n\pi_{i,m})^2} + D \tag{71}$$

where $D = -\sum_{n_{i,m}=0} \frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}} < m$ and $|\xi_{i,m}| \le \left|\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right|$. Thus

$$\max_{1\le i\le m} |\xi_{i,m}| \le \max_{1\le i\le m} \left|\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right| \le (c_1 b_1)^{-1} \max_{1\le i\le m} \left|\frac{m^{\alpha_1} n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m}\right|, \tag{72}$$

and by Lemma 4

$$\max_{1\le i\le m} |\xi_{i,m}| = o(1) \quad \text{a.s.} \tag{73}$$

uniformly in $m \in [1, n^{\gamma_2}]$ as $n \to \infty$. By (73) and Corollary 1 it follows that the second term of the right hand side of (71) is bounded uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ by $O(n^{-1}m^{2\alpha_1})$ a.s.. The latter is $o(m)$ because $n > m^{\frac{1}{\gamma_2}}$ and $\alpha_1 < \frac{1}{2} + \frac{1}{2\gamma_2}$. Therefore by Lemma 3

$$\sum_{n_{i,m}>0} \log \frac{n_{i,m}}{n\pi_{i,m}} = O(m) \quad \text{a.s.} \tag{74}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$. ◁

Proof. [Proof of Theorem 1] By condition (i) and (iii) we can obtain an interval estimate, respectively, for $-\sum_{i=1}^{m} \tilde{r}_{i,m}$ and $\sum_{i=1}^{m} \log n\pi_{i,m}$ as follows

$$m\log m + O(m) \le -\sum_{i=1}^{m} \log \tilde{r}_{i,m} \le \alpha_1 m \log m + O(m) \tag{75}$$

$$m\log n - \alpha_1 m \log m + O(m) \le \sum_{i=1}^{m} \log n\pi_{i,m} \le m\log n - m\log m + O(m). \tag{76}$$

Hence there exists an α' satisfying $-\frac{1}{2}\alpha_1 \le \alpha' \le -\frac{3}{2} + \alpha_1$ such that

$$-\sum_{i=1}^{m} \log \tilde{r}_{i,m} + \frac{1}{2}\sum_{i=1}^{m} \log n\pi_{i,m} - m\log m = \alpha' m \log m + \frac{1}{2} m \log n + O(m). \tag{77}$$

Now we turn to the second term of (44). By Taylor expansion

$$\sum_{n_{i,m}>0} \log\left(1 + \frac{1}{n_{i,m}}\right)^{n_{i,m}+1}$$

$$= \sum_{n_{i,m}>0} (n_{i,m}+1)\left(\frac{1}{n_{i,m}} - \frac{1}{2}(1+\eta_{i,m})^{-2}\frac{1}{n_{i,m}^2}\right) = O(m), \tag{78}$$

where $0 \le \eta_{i,m} \le n_{i,m}^{-1}$.

From Lemma 6, (77), (78) and (44), it is easy to see that

$$-\log \tilde{f}(X^n;m) + L_1^*(X^n;m) = \alpha' m \log m + \frac{1}{2} m \log n + O(m) \quad \text{a.s.} \tag{79}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$. ◁

To prove Theorem 2 we need the following lemmas.

Lemma 7. *Under the condition (iii) of Theorem 1,*

$$L_2(\tilde{q}^m, m, \delta) = o(m). \tag{80}$$

Proof. From $b_1 m^{-\alpha_1} \le \tilde{r}_{i,m} \le b_2 m^{-\alpha_2}$ it follows that

$$\overline{\left|\tilde{r}_{i,m} - \frac{r}{m}\right|} \le \max\left\{\frac{b_2}{m^{\alpha_2}} - \frac{r}{m}, \frac{r}{m} - \frac{b_1}{m^{\alpha_1}}\right\} \le \frac{b_2 + r}{m^{\alpha_2}}. \tag{81}$$

From this (80) follows. ◁

Let $f(x|\tilde{q}^m)$ denote a density in H_m which assigns the same probability as f to each subinterval $\tilde{Q}_{i,m}$, i.e. for $x \in [s,t]$ let

$$f(x|\tilde{q}^m) = \sum_{i=1}^m \frac{\pi_{i,m}}{\tilde{r}_{i,m}} I_{\tilde{Q}_{i,m}}(x). \tag{82}$$

By Lemma 7 we have

$$\begin{aligned} &-L_1^*(X^n;m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) \\ &= -L_1^*(X^n;m) + \sum_{j=1}^n \log f(X_j|\tilde{q}^m) + \sum_{j=1}^n \frac{\log f(X_j)}{\log f(X_j|\tilde{q}^m)} + o(m). \end{aligned} \tag{83}$$

Lemma 8. *Under the condition of Theorem 1, there exist two positive constants A and B such that*

$$B m^{\alpha_2} \le \sum_{n_{i,m}>0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}} \le A m^{\alpha_1} \quad \text{a.s.} \tag{84}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$.

Proof. By Taylor expansion,

$$\sum_{n_{i,m}>0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}} = \sum_{n_{i,m}>0} n_{i,m} \log\left(1 + \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} n_{i,m} \left[\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} - \frac{1}{2}(1+\theta_{i,k})^{-2} \left(\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right)^2 \right] \\
&= \sum_{i=1}^{m} \frac{(n_{i,m} - n\pi_{i,m})^2}{n\pi_{i,m}} + \sum_{i=1}^{m} (n_{i,m} - n\pi_{i,m}) \\
&\quad - \sum_{i=1}^{m} \frac{1}{2}(1+\theta_{i,k})^{-2} \left(\frac{(n_{i,m} - n\pi_{i,m})^3}{(n\pi_{i,m})^2} + \frac{(n_{i,m} - n\pi_{i,m})^2}{n\pi_{i,m}} \right) \qquad (85)
\end{aligned}
$$

where $|\theta_{i,k}| \le \left| \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right|$, so that $\max_{1 \le i \le m} |\theta_{i,k}| = o(1)$ a.s. uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$. The argument is similar to the one used to establish (73). By Lemma 4, Lemma 5, the property $\pi_{i,m} \ge b_1 c_1 m^{-\alpha_1}$ and the following inequality obtained from (85)

$$
\begin{aligned}
&\left| \sum_{n_{i,m} > 0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}} \right| \\
&\le \sum_{i=1}^{m} \frac{(n_{i,m} - n\pi_{i,m})^2}{n\pi_{i,m}} \left[1 + \frac{1}{2}(1+\theta_{i,k})^{-2} \left(1 + \max_{1 \le i \le m} \left| \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right| \right) \right], \qquad (86)
\end{aligned}
$$

the lemma can easily be established. ◁

The following lemma can similarly be proved using Taylor expansion, Lemma 3, Lemma 4 and Corollary 1.

Lemma 9. *Under the conditions of Theorem 1,*

$$\sum_{n_{i,m} > 0} \frac{1}{n_{i,m}} = o(m) \quad \text{a.s.} \quad \text{and} \qquad (87)$$

$$\sum_{i=1}^{m} \log \frac{n_{i,m} + 1}{n\pi_{i,m} + 1} = o(m) \quad \text{a.s.} \qquad (88)$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ *as* $n \to \infty$.

Lemma 10. *Under the conditions of Theorem 1, there exists a positive constant A such that*

$$
\begin{aligned}
-Am^{\alpha_1} + (\alpha_2 - \alpha_1) m \log m + O(m) &\le -L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j | \tilde{q}^m) \\
&\le (\alpha_1 - 1) m \log m + O(m) \quad \text{a.s.} \qquad (89)
\end{aligned}
$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ *as* $n \to \infty$.

Proof. First note that

$$-L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j | \tilde{q}^m) = \sum_{i=1}^{m} n_{i,m} \log \frac{\pi_{i,m}}{\tilde{r}_{i,m}} - \sum_{i=1}^{m} (n_{i,m} + 1) \frac{n_{i,m} + 1}{(n+m)\tilde{r}_{i,m}}$$

$$= \sum_{i=1}^{m} \log \tilde{r}_{i,m} + \sum_{i=1}^{m} n_{i,m} \log \frac{(n+m)\pi_{i,m}}{n_{i,m}+1} + \sum_{i=1}^{m} \log \frac{n\pi_{i,m}+1}{n_{i,m}+1}$$

$$+m\log(n+m) - \sum_{i=1}^{m} \log(n\pi_{i,m}+1). \quad (90)$$

The second term of the righthand side of (90)

$$\sum_{i=1}^{m} n_{i,m} \log \frac{(n+m)\pi_{i,m}}{n_{i,m}+1} = \sum_{i=1}^{m} n_{i,m} \log \left[\frac{(n+m)\pi_{i,m}}{n_{i,m}+1} \cdot \frac{n_{i,m}}{n\pi_{i,m}} \cdot \frac{n\pi_{i,m}}{n_{i,m}} \right]$$

$$= n\log\left(1+\frac{m}{n}\right) - \sum_{n_{i,m}>0} n_{i,m} \log\left(1+\frac{1}{n_{i,m}}\right) - \sum_{n_{i,m}>0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}} \quad (91)$$

and

$$\sum_{n_{i,m}>0} n_{i,m} \log\left(1+\frac{1}{n_{i,m}}\right) = \sum_{n_{i,m}>0} n_{i,m} \left(\frac{1}{n_{i,m}} + \frac{1}{2}(1+\eta_{i,m})^{-2} \frac{1}{n_{i,m}^2} \right) \quad (92)$$

where $0 \le \eta_{i,m} \le 1$. By Lemma 8 and (87) of Lemma 9 we have

$$-Am^{\alpha_1} + O(m) \le \sum_{i=1}^{m} n_{i,m} \log \frac{(n+m)\pi_{i,m}}{n_{i,m}+1} \le O(m) \quad \text{a.s.} \quad (93)$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$.

It can also be seen that

$$\alpha_2 m \log m + O(m) \le -\sum_{i=1}^{m} \log\left(\pi_{i,m} + \frac{1}{n}\right) \le \alpha_1 m \log m + O(m). \quad (94)$$

From (75), (93), (88) of Lemma 9, and (94) it follows that

$$-Am^{\alpha_1} + (\alpha_2 - \alpha_1) m \log m + O(m) \le -L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j | \tilde{q}^m)$$

$$\le (\alpha_1 - 1) m \log m + O(m) \quad \text{a.s.} \quad (95)$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$. ◁

Lemma 11. *Under the conditions (i) to (iv) of Theorem 2 and $f \neq 1$, we have as $m \to \infty$*

$$E_f \log \frac{f}{f(\cdot|\tilde{q}^m)} = \sum_{i=1}^{m} \frac{1}{24} \tilde{r}_{i,m}^2 \int_{\tilde{Q}_{i,m}} \frac{\dot{f}^2}{f} + o(m^{-2\alpha_2}). \quad (96)$$

Proof. By the definition of $f(x|\tilde{q}^m)$

$$\lim_{m\to\infty}(f(x)-f(x|\tilde{q}^m))=\lim_{m\to\infty}\frac{1}{\tilde{r}_{i,m}(x)}\int_{\tilde{Q}_{i,m}(x)}(f(x)-f(y))dy=0 \tag{97}$$

uniformly in $x\in[s,t]$, where $\tilde{Q}_{i,m}(x)$ is the subinterval containing x, and $\tilde{r}_{i,m}(x)$ is the corresponding width. Now by Taylor expansion

$$\begin{aligned} E_f\log\frac{f}{f(\cdot|\tilde{q}^m)}&=\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}f\log\left(1+\frac{f-f(\cdot|\tilde{q}^m)}{f(\cdot|\tilde{q}^m)}\right)\\ =\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}f\frac{f-f(\cdot|\tilde{q}^m)}{f(\cdot|\tilde{q}^m)}&-\frac{1}{2}\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}f(1+\eta_i)^{-2}\left(\frac{f-f(\cdot|\tilde{q}^m)}{f(\cdot|\tilde{q}^m)}\right)^2 \end{aligned} \tag{98}$$

where $|\eta_i(x)|\le\left|\frac{f-f(\cdot|\tilde{q}^m)}{f(\cdot|\tilde{q}^m)}\right|$ and by (97) $\sup_x|\eta_i(x)|=o(1)$. Hence

$$\begin{aligned} E_f\log\frac{f}{f(\cdot|\tilde{q}^m)}&=\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}\frac{(f-f(\cdot|\tilde{q}^m))^2}{f(\cdot|\tilde{q}^m)}\\ -\frac{1}{2}(1+o(1))\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}\frac{(f-f(\cdot|\tilde{q}^m))^3}{f(\cdot|\tilde{q}^m)^2}&-\frac{1}{2}(1+o(1))\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}\frac{(f-f(\cdot|\tilde{q}^m))^2}{f(\cdot|\tilde{q}^m)}\\ &=\frac{1}{2}(1+o(1))\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}\frac{(f-f(\cdot|\tilde{q}^m))^2}{f(\cdot|\tilde{q}^m)} \end{aligned} \tag{99}$$

Now we apply the technique used in Proposition 2.7 of [6] to prove that

$$\sum_{i=1}^m\int_{\tilde{Q}_{i,m}}\frac{(f-f(\cdot|\tilde{q}^m))^2}{f(\cdot|\tilde{q}^m)}=\frac{1}{12}\sum_{i=1}^m\tilde{r}_{i,m}^2\int_{\tilde{Q}_{i,m}}\frac{\dot{f}^2}{f}+o(m^{-2\alpha_2}). \tag{100}$$

The lemma would follow from (100) and (99).

By denoting $z=x-\tilde{f}_{i-1,m}$ we have

$$\begin{aligned} \int_{\tilde{Q}_{i,m}}\frac{(f(x)-f(x|\tilde{q}^m))^2}{f(x|\tilde{q}^m)}dx&=\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}[f(z+\tilde{q}_{i-1})-f(z+\tilde{q}_{i-1}|\tilde{q}^m)]^2dz\\ =\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}&\left[\int_0^z\dot{f}(y+\tilde{q}_{i-1,m})dy-\frac{1}{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-y)\dot{f}(y+\tilde{q}_{i-1,m})dy\right]^2dz\\ =\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}&\left[\int_0^z\dot{f}(y+\tilde{q}_{i-1,m})dy\right]^2dz-\frac{1}{\pi_{i,m}}\left[\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-y)\dot{f}(y+\tilde{q}_{i-1,m})dy\right]^2\\ =\frac{\tilde{r}_{i,m}}{\pi_{i,m}}&\int_0^{\tilde{r}_{i,m}}\int_0^z\int_0^z\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudvdz \end{aligned}$$

$$
\begin{aligned}
&\quad -\frac{1}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-u)(\tilde{r}_{i,m}-v)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv\\
&=\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-u\vee v)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv\\
&\quad -\frac{1}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-u)(\tilde{r}_{i,m}-v)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv\\
&=\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v-\frac{1}{\tilde{r}_{i,m}}uv)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv \qquad (101)
\end{aligned}
$$

where $u\vee v=\max(u,v)$ and $u\wedge v=\min(u,v)$. Direct computation shows that

$$
\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v-\frac{1}{\tilde{r}_{i,m}}uv)dudv=\frac{1}{12}\tilde{r}_{i,m}^3. \qquad (102)
$$

Define $\bar{\dot{f}}_{i,m}=\frac{1}{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}\dot{f}(u+\tilde{q}_{i-1,m})du$. By (101)

$$
\begin{aligned}
&\sum_{i=1}^{m}\int_{\tilde{Q}_{i,m}}\frac{(f-f(\cdot|\tilde{q}^m))^2}{f(\cdot|\tilde{q}^m)}\\
&=\sum_{i=1}^{m}\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v-\frac{1}{\tilde{r}_{i,m}}uv)[\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}^2]dudv\\
&\quad+\sum_{i=1}^{m}\frac{\tilde{r}_{i,m}^3}{12\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}[\bar{\dot{f}}_{i,m}^2-\dot{f}^2(u+\tilde{q}_{i-1,m})]du\\
&\quad+\sum_{i=1}^{m}\frac{\tilde{r}_{i,m}^2}{12}\int_0^{\tilde{r}_{i,m}}\left(\frac{\tilde{r}_{i,m}\dot{f}^2(u+\tilde{q}_{i-1,m})}{\pi_{i,m}}-\frac{\dot{f}^2(u+\tilde{q}_{i-1,m})}{f(u+\tilde{q}_{i-1,m})}\right)du\\
&\quad+\frac{1}{12}\sum_{i=1}^{m}\tilde{r}_{i,m}^2\int_{\tilde{Q}_{i,m}}\frac{\dot{f}^2(x)}{f(x)}dx. \qquad (103)
\end{aligned}
$$

Note that $|u\wedge v-\frac{1}{\tilde{r}_{i,m}}uv|\le\tilde{r}_{i,m}$ and

$$
\begin{aligned}
&|\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}^2|\le\\
&|\dot{f}(u+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}||\dot{f}(v+\tilde{q}_{i-1,m})|+|\dot{f}(v+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}||\bar{\dot{f}}_{i,m}|,
\end{aligned}
$$

hence

$$
\begin{aligned}
&\left|\sum_{i=1}^{m}\frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v-\frac{1}{\tilde{r}_{i,m}}uv)[\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}^2]dudv\right|\\
&\qquad\le c_1^{-1}\sum_{i=1}^{m}\tilde{r}_{i,m}\int_0^{\tilde{r}_{i,m}}|\dot{f}(u+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}|\int_0^{\tilde{r}_{i,m}}|\dot{f}(v+\tilde{q}_{i-1,m})|
\end{aligned}
$$

$$
\begin{aligned}
&+c_1^{-1}\sum_{i=1}^{m}\tilde{r}_{i,m}\int_0^{\tilde{r}_{i,m}} |\dot{f}(v+\tilde{q}_{i-1,m})-\bar{\dot{f}}_{i,m}|\int_0^{\tilde{r}_{i,m}} |\bar{\dot{f}}_{i,m}| \\
\le\ & 2c_1^{-1}\sum_{i=1}^{m}\tilde{r}_{i,m}\int_{\tilde{Q}_{i,m}} |\dot{f}-\bar{\dot{f}}_{i,m}|\int_{\tilde{Q}_{i,m}} |\bar{\dot{f}}|. \qquad (104)
\end{aligned}
$$

Using the Cauchy-Schwartz inequality

$$
\begin{aligned}
&\sum_{i=1}^{m}\tilde{r}_{i,m}\int_{\tilde{Q}_{i,m}} |\dot{f}-\bar{\dot{f}}_{i,m}|\int_{\tilde{Q}_{i,m}} |\bar{\dot{f}}| \\
\le\ & \left[\sum_{i=1}^{m}\tilde{r}_{i,m}\left(\int_{\tilde{Q}_{i,m}} |\dot{f}-\bar{\dot{f}}_{i,m}|\right)^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{m}\tilde{r}_{i,m}\left(\int_{\tilde{Q}_{i,m}} |\bar{\dot{f}}|\right)^2\right]^{\frac{1}{2}} \\
\le\ & \left[\sum_{i=1}^{m}\tilde{r}_{i,m}^2\int_{\tilde{Q}_{i,m}} |\dot{f}-\bar{\dot{f}}_{i,m}|^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{m}\tilde{r}_{i,m}^2\int_{\tilde{Q}_{i,m}} |\bar{\dot{f}}|2\right]^{\frac{1}{2}} \le cm^{-2\alpha_2}\left(\int_{[s,t]}(\dot{f}-\bar{\dot{f}}_{i,m})^2\right)^{\frac{1}{2}}
\end{aligned}
$$

where $c=(t-s)b_2^2c_3$ is a constant. By (2.5) of [6]

$$
\int_{[s,t]}(\dot{f}-\bar{\dot{f}}_{i,m})^2 \to 0 \quad \text{as } m\to\infty. \qquad (105)
$$

Hence the first term of the righthand side of (103) is bounded by $o(m^{-2\alpha_2})$. Using the Cauchy-Schwartz inequality, (105) and the following result similar to (105)

$$
\int_{[s,t]}\left(\dot{f}-\frac{\pi_{i,m}}{\tilde{r}_{i,m}}\right)^2 \to 0 \quad \text{as } m\to\infty, \qquad (106)
$$

one can similarly show that the second and third terms of the righthand side of (103) are both bounded by $o(m^{-2\alpha_2})$. Hence (100) and accordingly the lemma is true. ◁

Lemma 12. *Under the conditions (i) to (iv) of Theorem 2 we have*

$$
\sum_{j=1}^{n}\log\frac{f(X_j)}{f(X_j|\tilde{q}^m)} = nE_f\log\frac{f}{f(\cdot|\tilde{q}^m)} + o(nm^{-2\alpha}+m\log n) \quad \text{a.s.} \qquad (107)
$$

as $n\to\infty$ uniformly for $m\in[n^{\gamma_1},n^{\gamma_2}]$, where α is a constant satisfying $\alpha_2\le\alpha<\alpha_2+\frac{1}{2}$.

Proof. Denote $Z_{j,m}=\log\frac{f(X_j)}{f(X_j|\tilde{q}^m)}$ for each X_j, then $Z_{j,m}$'s are i.i.d. and

$$
|Z_{j,m}| \le \max_x\frac{|f-f(\cdot|\tilde{q}^m)|}{f(\cdot|\tilde{q}^m)} \le \frac{c_3}{c_1}\max_{1\le i\le m}\tilde{r}_{i,m}. \qquad (108)
$$

Thus

$$|Z_{j,m} - EZ_{j,m}| \le \frac{2c_3}{c_1} \max_{1 \le i \le m} \tilde{r}_{i,m} \overset{\text{def}}{=} B, \tag{109}$$

and

$$\sum_{j=1}^{n} \mathrm{Var}(Z_{j,m}) \le 4n \frac{c_3^2}{c_1^2} \max_{1 \le i \le m} \tilde{r}_{i,m}^2 \overset{\text{def}}{=} V. \tag{110}$$

By Bernstein's inequality, for arbitrary $\varepsilon > 0$

$$P\left(\left|\sum_{j=1}^{n} (Z_{j,m} - EZ_{j,m})\right| > \eta\right) \le 2\exp\left\{-\frac{\eta^2}{2(V + \frac{1}{3}B\eta)}\right\}, \tag{111}$$

where $\eta = n(m^{-2\alpha} + mn^{-1}\log n)\varepsilon$ and $\alpha_2 \le \alpha < \alpha_2 + \frac{1}{2}$. By the definition of B and V,

$$\begin{aligned} V + \frac{1}{3}B\eta &= 4n\frac{c_3^2}{c_1^2} \max_{1 \le i \le m} \tilde{r}_{i,m}^2 + \frac{2c_3}{3c_1} \max_{1 \le i \le m} \tilde{r}_{i,m} n(m^{-2\alpha} + mn^{-1}\log n)\varepsilon \\ &\le c' nm^{-2\alpha_2} + c'' m^{1-\alpha_2} \log n \end{aligned}$$

where c' and c'' are constants not depending on n and m. Therefore,

$$\begin{aligned} \frac{\eta^2}{V + \frac{1}{3}B\eta} &\ge \frac{1}{2} \frac{n^2(m^{-2\alpha} + mn^{-1}\log n)^2 \varepsilon^2}{\max\{c' nm^{-2\alpha_2}, c'' m^{1-\alpha_2} \log n\}} \\ = \min\{&c' n(m^{-2\alpha+\alpha_2} + m^{1+\alpha_2} n^{-1} \log n)^2, \\ &c'' n^2 (\log n)^{-1} (m^{-2\alpha - \frac{1}{2} + \frac{1}{2}\alpha_2} + m^{\frac{1}{2} + \frac{1}{2}\alpha_2} n^{-1} \log n)^2\} \end{aligned}$$

for any $m \in [n^{\gamma_1}, n^{\gamma_2}]$ and hence

$$\frac{\eta^2}{V + \frac{1}{3}B\eta} \ge O\left(n^{\frac{-2\alpha+2\alpha_2+1}{2\alpha+1}} (\log n)^{\frac{4\alpha-2\alpha_2}{2\alpha+1}}\right). \tag{112}$$

By (112) and (111), it follows that

$$\begin{aligned} &\sum_{n=1}^{\infty} \sum_{m \in [n^{\gamma_1}, n^{\gamma_2}]} P\left(\left|\sum_{j=1}^{n} (Z_{j,m} - EZ_{j,m})\right| > \eta\right) \\ \le 2 &\sum_{n=1}^{\infty} \sum_{m \in [n^{\gamma_1}, n^{\gamma_2}]} \exp\left\{-O\left(n^{\frac{-2\alpha+2\alpha_2+1}{2\alpha+1}} (\log n)^{\frac{4\alpha-2\alpha_2}{2\alpha+1}}\right)\right\} < \infty. \end{aligned}$$

From the Borel-Cantelli Lemma, (107) follows. ◁

Proof. [Proof of Theorem 2] The first part of the theorem, i.e. the equation (25) can be obtained from (83), Lemma 10, Lemma 11 and Lemma 12. Then the second part is straightforward from Theorem 1. ◁

Proof. [Proof of Theorem 3] Noting that

$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{(\alpha_1-1)m\log m+C_f nm^{-2\alpha_2}\} = M_2 n^{\frac{1}{1+2\alpha_2}}(\log n)^{\frac{2\alpha_2}{1+2\alpha_2}}, \tag{113}$$

$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-Am^{\alpha_1}+(\alpha_2-\alpha_1)m\log m\} = -M_1(n^{\alpha_1\gamma_2}+n^{\gamma_2}\log n), \tag{114}$$

the first part of Theorem 3 is obvious from Theorem 2. The second part can be proved similarly. ◁

Proof. [Proof of Theorem 4] Regarding m as a real value and taking the derivative of $\frac{1}{2}n\log\frac{n}{m}+C_f' nm^{-2}$ with respect to m, we get

$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\left\{\frac{1}{2}n\log\frac{n}{m}+C_f' nm^{-2}\right\}=M_5 n^{\frac{1}{3}}(\log n)^{\frac{2}{3}} \tag{115}$$

and the minimization is achieved at $m=M_6(n/\log n)^{\frac{1}{3}}$. By this result and Theorem 2, (a), (b), (c) and (d) are readily obtained. ◁

Proof. [Proof of Theorem 5] As in Lemma 7, it can be shown that

$$L_4(\tilde{q}_1^{m_1},\cdots,\tilde{q}_k^{m_k},m_1,\cdots,m_k,\delta)=o\left(\sum_{i=1}^k m_i\right). \tag{116}$$

If either $\alpha_1\neq 1$ or $\alpha_2\neq 1$, then by Theorem 3

$$\begin{aligned}-M_3\sum_{i=1}^k(n_i^{\alpha_1\gamma_2}+n_i^{\gamma_2}\log n_i)\leq C(X_1^{n_1},\cdots,X_k^{n_k})+\sum_{i=1}^k\log f_i^{n_i}(X_i^{n_i})\\ \leq M_4\sum_{i=1}^k n_i^{\frac{1}{1+2\alpha_2}}(\log n_i)^{\frac{2\alpha_2}{1+2\alpha_2}}\quad\text{a.s.}\end{aligned} \tag{117}$$

and

$$-M_3(n^{\alpha_1\gamma_2}+n^{\gamma_2}\log n)\leq C(X^n)+\log f_{\text{mix}}^n(X^n)\leq M_4 n^{\frac{1}{1+2\alpha_2}}(\log n)^{\frac{2\alpha_2}{1+2\alpha_2}}\quad\text{a.s.} \tag{118}$$

for some positive constants M_3 and M_4 depending on $f_1,\cdots,f_k$.

If $\alpha_1=\alpha_2=1$, then by Theorem 4 (b)

$$C(X_1^{n_1},\cdots,X_k^{n_k})+\sum_{i=1}^k\log f_i^{n_i}(X_i^{n_i})=O\left(\sum_{i=1}^k n_i^{\frac{1}{3}}(\log n)^{\frac{2}{3}}\right)\quad\text{a.s.} \tag{119}$$

and

$$C(X^n)+\log f_{\text{mix}}^n(X^n)=O\left(n^{\frac{1}{3}}(\log n)^{\frac{2}{3}}\right)\quad\text{a.s..} \tag{120}$$

It remains to prove that there exists a constant $\eta<0$ such that

$$\frac{1}{n}\left(\log f_{\text{mix}}^n(X^n)-\sum_{i=1}^k\log f_i^{n_i}(X_i^{n_i})\right)<\eta\quad\text{a.s.} \tag{121}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$ satisfying $\frac{n_1}{n} > \varepsilon_1 > 0, \cdots, \frac{n_k}{n} > \varepsilon_k > 0$ for any prescribed constants $\varepsilon_1, \cdots, \varepsilon_k$, if at least two of $f_1, \cdots, f_k$ are not equal almost surely, and

$$\frac{1}{n}\left(\log f_{\text{mix}}^n(X^n) - \sum_{i=1}^k \log f_i^{n_i}(X_i^{n_i})\right) \to 0 \quad \text{a.s.} \tag{122}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$ if $f_1 = f_2 = \cdots = f_k$ a.s..

Because

$$\sum_{i=1}^k \log f_i^{n_i}(X_i^{n_i}) = \sum_{i=1}^k \sum_{j=1}^{n_i} \log f_i(X_{ij}),$$
$$\log f_{\text{mix}}^n(X^n) = \sum_{i=1}^k \sum_{j=1}^{n_i} \log\left(\sum_{l=1}^k \frac{n_l}{n} f_l(X_{ij})\right)$$

and f_i's are bounded density functions, by the strong law of large numbers for i.i.d. random variables it follows that

$$\frac{1}{n}\sum_{i=1}^k \log f_i^{n_i}(X_i^{n_i}) - \sum_{i=1}^k \frac{n_i}{n}\int f_i \log f_i \to 0 \quad \text{a.s.} \tag{123}$$

and

$$\frac{1}{n}\log f_{\text{mix}}^n(X^n) - \int f_{\text{mix}} \log f_{\text{mix}} \to 0 \quad \text{a.s.} \tag{124}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$. By the convexity of $x \log x$,

$$\int f_{\text{mix}} \log f_{\text{mix}} \le \sum_{i=1}^k \frac{n_i}{n}\int f_i \log f_i \tag{125}$$

for any group of samples of sizes $n_1, \cdots, n_k$ satisfying $\sum_{i=1}^k n_i = n$, where the equality holds if and only if all the densities $f_1, \cdots, f_k$ are equal (except a set with measure zero). Therefore (122) is established by using (123) and (124). Also for any $\varepsilon_1 > 0, \cdots, \varepsilon_k > 0$, if $\frac{n_1}{n} > \varepsilon_1, \cdots, \frac{n_k}{n} > \varepsilon_k$ and if at least two of $f_1, \cdots, f_k$ are not equal almost surely, there exists a constant $\eta < 0$ depending on $\varepsilon_1, \cdots, \varepsilon_k$ such that

$$\int f_{\text{mix}} \log f_{\text{mix}} - \sum_{i=1}^k \frac{n_i}{n}\int f_i \log f_i < \eta \tag{126}$$

for any set of integers $\{n_i\}$ satisfying $\sum_{i=1}^k n_i = n$. Hence (121) follows from (123) and (124). Notice that $\alpha_1\gamma_2 < 1$, $\gamma_2 < 1$ and $\frac{1}{1+2\alpha_2} < 1$, (41) and (42) hold by (116) to (122). ◁

References

[1] Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, 44, 62-91.

[2] Dawid, A.P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith eds), Oxford University Press, 109-125 (with discussions).

[3] Dawid, A.P. (1991). Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. B* 53, 79-109 (with discussions).

[4] Dawid, A.P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *J. Roy. Statist. Soc. A*, 47, 278-292 (with discussions).

[5] Elias, P. (1975). Universal codeword sets and representations of the integers. *IEEE Trans. Information Theory* 21, 194-203.

[6] Freedman, D.A. and Diaconis, P. (1981). On the histogram as a density estimator: L^2 theory. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* 57, 453-475.

[7] Hall, P. and Hannan, E.J. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika*, 75, 705-714.

[8] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York.

[9] Qian, G. and Künsch, H.R. (1998). Some notes on Rissanen's stochastic complexity. *IEEE Trans Information Theory* 44, 782-786.

[10] Qian, G. Gabor, G. and Gupta, R.P. (1996). Test for homogeniety of several populations by stochastic complexity. *Journal of Statistical Planning and Inference*. 53. 133-151

[11] Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. Springer, New York.

[12] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, 42, 40-47.

[13] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, Teaneck, NJ.

[14] Rissanen, J., Speed, T.P. and Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Information Theory* 38, 315-323.

[15] Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.*, 14, 1080-1100.

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

[17] Shiryayev, A.N. (1995).*Probability (2nd Edition)*. Springer-Verlag, New York.

[18] Solomonoff, R.J. (1978). Complexity-based induction system: comparison and convergence theorems. *IEEE Trans. Information Theory* 24, 422-432.

[19] Stone, C.J. (1985). An asymptotic optimal histogram selection rule. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer (ed. by Le Cam, L.M. and Ohshen, R.A.), Volume II*, 513-520. Wadsworth, Belmont, CA.

[20] Yu, B. and Speed, T.P. (1992). Data compression and histograms. *Probability Theory and Related Fields* 92, 195-229.