ESTIMATION OF THE ENTROPY FUNCTIONAL FROM DEPENDENT SAMPLES

Johan Lim

Department of Applied Statistics

Yonsei Univeristy

ShinChon Dong 134, SeoDaeMun Gu,

Seoul, 120-749, Korea

johanlim@yonsei.ac.kr

Key Words: Dependent samples; differential entropy; entropy estimation; histogram; $\phi-$ mixing.

ABSTRACT

The differential entropy is importantly used in many disciplines, where the estimation of entropy is often the main research objective or the first step toward it. To estimate entropy, plug–in estimators, such as histogram based entropy estimators or kernel based entropy estimators, are commonly used. Especially, though the histogram itself performs poorly in estimating density, the histogram based entropy estimator is often employed due to its computational benefit. Much efforts have been made to understand the properties of the histogram based entropy estimator theoretically, but most of such efforts are restricted to the case of independently and identically distributed (IID) samples. In this paper, we show that two histogram–based entropy estimators by Györfi and van der Meulen (1987) are almost surely consistent when samples are from a $\phi-$ mixing process. A limited simulation study is implemented to compare those two estimators and to investigate their performance for varying intensity of dependency. In addition, we discuss the extension of $\sqrt{n}-$consistency of the estimators in IID setting by Hall (1993) to the case of dependent samples.

MATHEMATICA SUBJECT CLASSIFICATION: Primary 62G05; Secondary 62G20.

1. INTRODUCTION

Let $f(x)$ be an unknown probability density function with $x \in \mathbb{R}^d$. The differential entropy functional $\mathbf{H}(f)$, defined by $-\int_{-\infty}^{\infty} f(x) \log f(x) dx$, is used in many disciplines. To list some representative examples in engineering, the normality of observed signal can be tested using the maximal entropy property of the normal distribution (among distributions that have the same variance) (Vasicek, 1976; Dudewicz and van der Meulen, 1981); the error exponents in binary decision problems are functions of the entropy of the sources (Cover and Thomas, 1991); and, entropy can be used to recover several independent sources from sole observations of the signal (Bercher and Vigan, 2000). In these examples, estimating entropy functional is often the main research objective or the first step toward it.

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed (IID) random variables taking values in $\mathbb{R}^d$ with the density function $f(x)$ and the distribution function $F(x)$. To estimate $\mathbf{H}(f)$, several methods have been proposed. When $d = 1$, based on the other representation of the entropy

$$\mathbf{H}(f) = \int_0^1 \log\Big\{\frac{d}{dp} F^{-1}(p)\Big\} dp,$$

Vasicek (1975) proposed the estimator

$$\mathbf{H}^{\mathrm{v}}_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\Big\{\frac{n}{2m}\big(x_{(i+m)} - x_{(i-m)}\big)\Big\},$$

where $x_{(1)} \leq x_{(2)} \leq \cdots, \leq x_{(n)}$ are the order statistics and m is the positive integer with $m/n \to 0$ and $m \to \infty$.

Plug–in estimators are also commonly used which are defined by

$$\mathbf{H}_{n1} = -\int_{A_n} \widehat{f}_n(x) \log \widehat{f}_n(x) dx \tag{1}$$

or by

$$\mathbf{H}_{2n} = -\sum_{i=1}^{n} \log \widehat{f}_n(Z_i) dx \tag{2}$$

with $A_n \to \mathbb{R}^d$ as $n \to \infty$ and $\{Z_i\}_{i=1}^n$ are IID samples from f. Here, several different choices of the nonparametric function estimator $\widehat{f}_n$ are suggested in the previous literature. The two

most popular are the kernel estimator (Prakasa Rao, 1983; Joe, 1989; Hall and Morton, 1993; Eggermont and LaRiccia, 1999) and the histogram estimator (Györfi and van der Meulen, 1987). Hall (1993) shows the $\sqrt{n}$ consistency of both the histogram based entropy estimator ($\mathbf{H}_{1n}$) and the kernel based entropy estimator under some regularity conditions to f. Eggermont and LaRiccia (1999) suggest using the double exponential kernel for $\widehat{f}_n$ in the kernel based entropy estimator and show the asymptotic normality of $\mathbf{H}_{1n}$. On the other hand, the histogram is also popularly employed as a nonparametric density estimator since it allows an explicit expression of the finial estimate unlike the kernel denisty estimators; thus, it does not require the numerical integration in (1). Györfi and van der Meulen (1987) propose the histogram based entropy estimators and show almost surely (a.s.) consistency under the finiteness of $\mathbf{H}(f)$ which is quite minimal.

Finally, Grassberger (1996) suggests an estimator for finite alphabets based on the deep connection between the entropy rate of a stationary ergodic process and the asymptotic behavior of the longest match lengths along a process realization (Kontoyiannis, Algoet, Suhov, and Wyner (1998) and references therein).

Unlike IID samples, the estimation of entropy functional from dependent samples are mostly limited to the case of finite alphabets and even the performance of the simple histogram based entropy estimators are not well understood. However, it is generally conjectured that a certain level of weakly dependency does not severely hurt the performance of estimators. In compliance to this conjecture, this paper shows that the histogram based entropy estimators in (1) and (2) are a.s. consistent when the samples are from an $\phi-$mixing process. A limited simulation study is implemented under the first order Gaussian autoregressive (AR) model with various AR coefficients. The simulation study shows that $\mathbf{H}_{1n}$ performs better than $\mathbf{H}_{2n}$ in terms of the mean squared error (MSE). Both estimators are negatively biased as pointed out in Hall and Morton (1993) (see Table 1 in their paper) and perform worse as the magnitude of the dependence increases.

The remainder of the paper is organized as follows. Section 2 provides the definitions of various mixing processes and some preliminary results which will be used in showing the

main results. Section 3 presents and proves the main results on the consistency of $\mathbf{H}_{1n}$ and $\mathbf{H}_{2n}$. A limited simulation study is implemented to investigate the finite sample performance of both estimators in Section 4. Finally, Section 5 discusses the extension of $\sqrt{n}-$ consistency of $\mathbf{H}_{1n}$ in IID samples by Hall (1990) to dependent samples.

2. PRELIMINARIES

Let $\{X_i\}_{i=1}^n$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{A}$ and $\mathcal{B}$ be sub-σ fields of $\mathcal{F}$, and let $\mathcal{L}_2(\mathcal{A})$ be a set of all $\mathcal{A}-$ measurable random variables with finite second moments. Define the measure of dependence

$$\phi(\mathcal{A}, \mathcal{B}) = \sup_{A\in\mathcal{A}, B\in\mathcal{B}, \mathbf{P}(A)>0} \big|\mathbf{P}\big(B\big|A\big) - \mathbf{P}\big(B\big)\big|$$

and

$$\rho(\mathcal{A}, \mathcal{B}) = \sup_{X\in\mathcal{L}_2(\mathcal{A}), Y\in\mathcal{L}_2(\mathcal{B})} \frac{\big|\mathrm{E}XY - \mathrm{E}X\cdot\mathrm{E}Y\big|}{\sqrt{\mathrm{Var}X\cdot\mathrm{Var}Y}}.$$

Based on the above dependence measures, $\phi-$mixing and $\rho-$mixing are defined as follows:

Definition 1. *A sequence $\{X_n, n\geq 1\}$ is a $\phi-$mixing process if $\phi(n) = \sup_{k\in N}\phi(\mathcal{F}_1^k, \mathcal{F}_{n+k}^\infty)$ decreases to 0, as $n\to\infty$, where $\mathcal{F}_a^b = \sigma(X_a, X_{a+1}, \ldots, X_b)$.*

Further, define a regular $\phi-$mixing process $\{X_i\}_{i=1}^\infty$ as a $\phi-$ mixing process with mixing rates satisfying $\phi\big(m(n)\big)\cdot n/m(n) \sim o(1)$ and $m(n)/n \sim o(1)$ for some sequence $\{m(n)\}_{n=1}^\infty$.

Definition 2. *A sequence $\{X_n, n\geq 1\}$ is a $\rho-$mixing process if $\rho(n) = \sup_{k\in N}\phi(\mathcal{F}_1^k, \mathcal{F}_{n+k}^\infty)$ decreases to 0, as $n\to\infty$, where $\mathcal{F}_a^b = \sigma(X_a, X_{a+1}, \ldots, X_b)$.*

The following lemma is from Lemma 1.2.8 in Lin and Lu (1996) with $p = q = 2$, and it shows that $\rho(n) \leq 2\phi^{1/2}(n)$. Thus, every $\phi-$mixing process is a $\rho-$mixing process.

Lemma 1. *Let $\{X_i\}_{i=1}^\infty$ be a $\phi-$mixing sequence, $X\in\mathcal{L}_2(\mathcal{F}_{-\infty}^k)$ and $Y\in\mathcal{L}_2(\mathcal{F}_{k+n}^\infty)$. Then,*

$$\big|EXY - EX\cdot EY\big| \leq 2\big(\phi(n)\big)^{1/2}\big(EX^2\big)^{1/2}\big(EY^2\big)^{1/2} \tag{3}$$

The following generalization of Bernstein inequality to $\phi-$mixing processes was proved by Columbo (1984); Carbon (1983) also proved and a similar result on $\alpha-$mixing. More details on various mixing conditions can be found in Lin and Lu (1996).

Lemma 2. (Bernstein inequality for a $\phi-$mixing sequence)
Let $X_i, i \geq 1$, be a $\phi-$mixing sequences satisfying $EX_i = 0, |X_i| \leq B, E|X_i| \leq L$, and $EX_i^2 \leq V$ for all $i \in N$. Set

$$\overline{\phi_m} = \sum_{i=1}^{m} \phi(k) \quad for\ all\ m,$$

where $\{\phi(k)\}_{k=1}^{\infty}$ are the mixing coefficients which are independent of n. Then for every $\epsilon > 0$ and every n

$$P\Big\{\Big|\sum_{i=1}^{n} X_i\Big| > \epsilon\Big\} \leq c \cdot \exp\{-\alpha\epsilon + \alpha^2 nC\},$$

where $C = 6\big(V + 4LB\overline{\phi(m)}\big)$ and $c = 2\exp\{3e^{1/2}n\phi_m/m\}$, where α and m are respectively any positive real and integer values less than n verifying $(\alpha \cdot m \cdot B) \leq \frac{1}{4}$. The number α, m, B, L, and V may also depend on n.

Consider a sequence of partitions of $\mathbb{R}^d$ $\mathcal{P}_n = \{A_{nj},\ j = 1, 2, \cdots, k_n\},\ n \geq 1$, where $\lambda\big(A_{nj}\big) \in (0, \infty)$ and A_{nj} is the Borel set in $\mathbb{R}^d$ for all n and j. Here, λ is the Lebesgue measure on $\mathbb{R}^d$. For $A \in \mathbb{R}^d$, let

$$\mu_n\big(A\big) = \sum_{i=1}^{n} \mathrm{I}\big(X_i \in A\big)/n$$

and $\mu\big(A\big)$ be its expectation

$$\int_A f(x)dx.$$

Then, the histogram density estimator by the partition $\mathcal{P}_n$ and its expectation becomes, respectively,

$$f_n(x) = \mu_n\big(A_{ni}\big)/\lambda(A_{ni)} \quad \text{and} \quad g_n(x) = Ef_n(x) = \int_{A_{ni}} f(x)dx\Big/\lambda(A_{ni)}, \quad \text{if} \quad x \in A_{ni}. \quad (4)$$

The following lemmas are extensions of the results in Györfi and van der Meulen (1987) to a stationary $\phi-$mixing sequence $\big\{X_i\big\}_{i=1}^{\infty}$.

Lemma 3. *If* $\mathbf{H}(f)$ *is finite and* $\lim_{n\to\infty} h_n = 0$ *where* $h_n = \lambda\big(A_{ni}\big)$, *then*

$$\lim_{n\to\infty} -\sum_i \mu(A_{ni}) \log\Big(\mu(A_{ni})/\lambda(A_{ni})\Big) = \mathbf{H}\big(f\big)$$

where μ *is the probability measure of a single observation* X_1.

Lemma 4. *Let*

$$D\big(f_1||f_2\big) = \int_{-\infty}^{\infty} f_1(x) \log\Big(f_1(x)/f_2(x)\Big)dx$$

be the informational divergence of f_1 *with respect to* f_2, *which is well defined if* $f_1(x) = 0$ *whenever* $f_2(x) = 0$. *Then*

$$\int \Big|\log\Big(f_1(x)/f_2(x)\Big)\Big| f_1(x)dx \le D\big(f_1||f_2\big) + 2^{3/2}\log e\sqrt{D\big(f_1||f_2\big)}. \tag{5}$$

Lemma 5. *For each* $\epsilon > 0$ *and every set* A, *there exist a positive* γ *satisfying*

$$P\Big(\Big|\log\big(\mu(A)/\mu_n(A)\big)\Big| > \epsilon\Big) \le \exp\big(-\gamma\cdot n\big)$$

Proof. From Equation (2.3) in Györfi and van der Meulen (1987),

$$\Big\{\Big|\log\big(\mu(A)/\mu_n(A)\big)\Big| > \epsilon\Big\} \in \Big\{\big|\mu_n(A) - \mu(A)\big| > \mu(A)(1-2^{-\epsilon})\Big\}.$$

Hence,

$$\begin{aligned} P\Big(\Big|\log\frac{\mu(A)}{\mu_n(A)}\Big| > \epsilon\Big) &\le P\Big(\big|\mu_n(A)-\mu(A)\big| > \mu(A)(1-2^{-\epsilon})\Big) \\ &\le c\cdot\exp\big\{n\cdot\big(-\alpha\epsilon' + \alpha^2 nC\big)\big\}, \end{aligned} \tag{6}$$

where $\epsilon' = \mu(A)(1-2^{-\epsilon})$. c and C are defined as in Lemma 2 with $B = L = 1$. Therefore, by choosing a sufficiently small α, Equation (6) is smaller than $\exp\big(-\gamma n\big)$ for some positive γ. □

Finally, the following lemma is a version of Abou-Jaoude (1976) for $\phi-$mixing processes, which is proved using a generalized Bernstein inequality (Lemma 2).

Lemma 6. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a regular $\phi-$mixing process. Let $\mathcal{P}_n = \{A_{nj},\ j = 1, 2, \cdots, k_n\},\ n \geq 1$, be a partition of $\mathbb{R}^d$ and let $\sigma\big(\mathcal{P}_n\big)$ be the $\sigma-$algebra generated by $\mathcal{P}_n$. Let $\mathbf{A}$ be a measurable set with $0 < \lambda\big(\mathbf{A}\big) < \infty$. Assume that (i) there exists $A_n \in \sigma(\mathcal{P}_n\big)$ such that $\lambda\big(\mathbf{A}\Delta A_n\big) < \epsilon$ for all sufficiently large n and for every $\epsilon > 0$, and (ii)*

$$\sup_{S\in\mathbb{R}^d} \limsup_{n\to\infty} \lambda\Bigg(\bigcup_{\lambda\big(A_{nj}\cap S\big)\leq h_n} A_{nj} \cap S \Bigg) = 0.$$

for a sequence $h_n \approx O(1/n)$. Then,

$$\mathbf{P}\bigg(\int \Big| f_n(x) - f(x)\Big| dx > \epsilon \bigg) \leq \exp\big(-\gamma \cdot n \cdot h_n^2\big), \qquad \textit{for some} \quad \gamma > 0. \tag{7}$$

Proof. Let $\mathbf{S}$ be a fixed compact set in $\mathbb{R}^d$, $\mathcal{F}_n = \big\{ i \ \big|\ \lambda\big(A_{ni}\big) \leq (1+\epsilon)h_n \big\}$, and $g_n(x)$ be the expectation of $f_n(x)$ as defined in (4). Note that $\int \big| g_n - f\big|$ is a deterministic sequence converging to 0, which becomes arbitrary small by choosing a sufficiently large n. Thus, from

$$\int \Big| f_n(x) - f(x)\Big| dx \leq \int \Big| g_n(x) - f(x)\Big| dx + \int \Big| f_n(x) - g_n(x)\Big| dx,$$

it suffices to show $\int \big| f_n - g_n\big|$ converges to 0 exponentially fast in probability.

As in Theorem 3.2. of Devroye and Györfi (1984) (p 20),

$$\int \Big| f_n(x) - g_n(x)\Big| dx \leq \sum_{j\in\mathcal{F}_n} \Big| \mu_n(A_{nj}) - \mu(A_{nj})\Big| + 2\cdot\mu(C_n) + \Big| \mu_n(C_n) - \mu(C_n)\Big|,$$

where $C_n = \bigcup_{j\in\mathcal{F}_n^c} A_{nj}$. Thus,

$$\begin{aligned} &\mathbf{P}\bigg(\int \Big| f_n(x) - g_n(x)\Big| dx > 3\epsilon \bigg) \leq \\ &\qquad \sum_{j\in\mathcal{F}_n} \mathbf{P}\bigg(\Big| \mu_n(A_{nj}) - \mu(A_{nj})\Big| > \frac{\epsilon}{|\mathcal{F}_n|} \bigg) + \mathbf{P}\bigg(\Big| \mu(C_n)\Big| > \frac{\epsilon}{2} \bigg) + \mathbf{P}\bigg(\Big| \mu_n(C_n) - \mu(C_n)\Big| > \epsilon \bigg). \end{aligned}$$

Using Lemma 2,

$$\begin{aligned} \mathbf{P}\bigg(\Big| \mu_n(A_{nj}) - \mu(A_{nj})\Big| > \frac{\epsilon}{|\mathcal{F}_n|} \bigg) &\leq c\cdot\exp\Big\{ -\alpha\frac{\epsilon n}{|\mathcal{F}_n|} + \alpha^2\cdot n\cdot C\Big\} \\ &= \exp\Big(-\gamma_0\cdot n \Big/ |\mathcal{F}_n|^2\Big), \end{aligned}$$

where $C = 6(2 + 16\overline{\phi_m})$, $c = 2 \cdot \exp\big(3 \cdot e^{1/2} \cdot n \cdot \phi_m/m\big)$, and $\alpha < \epsilon/\big(C \cdot \big|\mathcal{F}_n\big|\big)$. It is easy to show that $\mu\big(C_n \cap \mathbf{S}\big)$ is a deterministic function decreasing to 0 from (ii). The absolute continuity of μ with respect to λ can also be shown. Thus, $\mu(C_n) = \mu\big(C_n \cap S\big) + \mu(C_n \cap S^c)$ decreases to 0 as n $\to \infty$, and $\mathbf{P}\big(\big|\mu(C_n)\big| > \epsilon/2\big) = 0$, for all sufficiently large n. Finally, noting that $\big|\mathcal{F}_n\big| \le \big|\mathcal{P}_n\big|$,

$$\begin{aligned}
\mathbf{P}\Big(\int \Big|f_n(x) - g_n(x)\Big| > 3\epsilon\Big) &\le \Big(\big|\mathcal{F}_n\big| + 1\Big) \exp\Big(-\gamma_0 \cdot n \Big/ \big|\mathcal{F}_n\big|^2\Big) \\
&\le \exp\big(-\gamma_1 \cdot n/\big|\mathcal{P}_n\big|^2\big) \\
&\le \exp\big(-\gamma_1 \cdot n \cdot h_n^2\big). \qquad (8)
\end{aligned}$$

The inequality in (8) is from $\big|\mathcal{P}_n\big| \cdot h_n < 1$ and γ_1 is chosen to be sufficiently smaller than γ_0. □

The following lemma is from Lemma 2.2.6 in Lin and Lu (1996) and will be used in extending the $\sqrt{n}-$consistency of the histogram based entropy estimator by Hall and Morton (1993) to that in dependent samples (see Discussion).

Lemma 7. *Let $\{X_n\}_{n=1}^{\infty}$ be a stationary $\rho-$mixing sequence such that*

$$EX_n = 0, \quad E|X_n|^q < \infty \text{ with } q \ge 3, ES_n^2 \le nh_n \cdot EX_n^2,$$

where $S_n = X_1 + X_2 + \cdots + X_n$ and h_n satisfies

$$\max\Big(h\big([n/2]\big), h\big(n - [n/2]\big)\Big) \le \theta \cdot h\big(n\big) \quad \text{and} \quad h(n) \ge \frac{1}{C} \exp\Big\{-C \sum_{i=0}^{[\log n]} \rho^{2/q}\big(2^i\big)\Big\}$$

with $0 < \theta < 2^{1/3}$ and for some $C > 0$. Then, there exists a constant K, such that

$$E\big|S_n\big|^q \le K\Big\{\big(n \cdot h(n) \cdot EX_1^2\big)^{q/2} + n \exp\Big\{K \sum_{i=1}^{[\log n]} \rho\big(2^i\big)\Big\} \cdot E\big|X_1\big|^q\Big\}. \qquad (9)$$

3. HISTOGRAM BASED ENTROPY ESTIMATORS

This section proves that two entropy estimators in Györfi and van der Meulen (1987) are still consistent when $\{X_i\}_{i=1}^{n}$ is a regular $\phi-$mixing process. The overall structure of the proof

follows that of Gyrófi and van der Meulen (1987) with a generalized Bernstein inequality (Lemma 2).

Two histogram based entropy estimators were proposed by Gyrófi and van der Meulen (1987) for IID samples $\{X_i\}_{i=1}^n$. First, let $\widehat{f}_n(x)$ be the histogram based density estimator based on the partition $\mathcal{P}_n\{A_{n1}, A_{n2}, \cdots\}$ which is defined in (4). When plugging $\widehat{f}_n(x)$ into (1), $\mathbf{H}_{1n}$ becomes

$$\mathbf{H}_{1n} = -\sum_{i\in\mathcal{F}_n} \mu_n(A_{ni}) \cdot \log\left(\frac{\mu_n(A_{ni})}{\lambda(A_{ni})}\right), \tag{10}$$

where $\mathcal{F}_n = \{i : \mu_n(A_{ni}) \geq a_n h_n\}$ with a sequence $\{a_n\}_{n=1}^\infty$ deceasing to zero. Second, decompose the samples $X_1, \ldots, X_n$ into two subsamples $Y = \{Y_i\}_{i=1}^{[n/2]}$ and $Z = \{Z_i\}_{i=1}^{[n/2]}$, where $Y_i = X_i$ and $Z_i = X_{[n/2]+i}$ for $i = 1, 2, \ldots, [n/2]$. Subsequently, compute the histogram density estimate using $\{Z_i\}_{i=1}^{[n/2]}$ and approximate the numerical integration for evaluating the entropy with the samples $\{Y_i\}_{i=1}^{[n/2]}$. Finally, the proposed second estimator becomes

$$\mathbf{H}_{2n} = -\frac{1}{[n/2]} \sum_{i=1}^{[n/2]} \log \widehat{f}_n(Y_i)\ I\big(\widehat{f}_n(Y_i) > a_n\big). \tag{11}$$

In the remainder of the paper, although $\mu_n(\,\cdot\,)$ in $\mathbf{H}_{1n}$ and $\mathbf{H}_{2n}$ are different, we use the same notation for simplicity.

First, Theorem 1 extends the result of Gyórfi and van der Meulen (1987) to $\phi-$mixing processes.

Theorem 1. *Under the assumptions* (i) $\lim_{n\to\infty} a_n = 0$, (ii) $\lim_{n\to\infty} h_n = 0$, *and* (iii) *the finiteness of* $\sum_{n=1}^\infty \exp\big(-n\cdot h_n^2\big)$ (*equivalently, the finiteness of* $\sum_{n=1}^\infty \exp\big(-n\cdot h_n^2\cdot a_n^2\big)$), *both* $\mathbf{H}_{1n}$ *and* $\mathbf{H}_{2n}$ *are a.s. consistent, when the observations* $\{X_i\}_{i=1}^n$ *are from a regular* $\phi-$*mixing process.*

Proof. We first show the convergence of $\mathbf{H}_{1n}$. As in Gyrófi and van der Meulen (1987,1991),

$\mathbf{H}_{1n}$ are represented as a sum of four components $\mathbf{U}_{1n}, \mathbf{V}_{1n}, \mathbf{W}_{1n}$, and $\mathbf{Z}_{1n}$, where

$$\begin{aligned}
\mathbf{U}_{1n} &= \sum_{i\in\mathcal{F}_n} \big\{ -\mu_n(A_{ni}) + \mu(A_{ni})\big\} \cdot \log\left(\frac{\mu_n(A_{ni})}{\lambda(A_{ni})}\right) \\
\mathbf{V}_{1n} &= \sum_{i\in\mathcal{F}_n} \mu(A_{ni}) \cdot \big\{ \log\mu(A_{ni}) - \log\mu_n(A_{ni})\big\} \\
\mathbf{W}_{1n} &= \sum_{i\in\mathcal{F}_n^c} \mu(A_{ni}) \cdot \big\{ \log\mu(A_{ni}) - \log\lambda(A_{ni})\big\} \\
\mathbf{Z}_{1n} &= \sum_{i} \mu(A_{ni}) \cdot \big\{ \log\lambda(A_{ni}) - \log\mu(A_{ni})\big\} - \mathbf{H}(f).
\end{aligned}$$

It suffices to show that each component a.s. converges to 0. Here, the last two components $\mathbf{W}_{1n}$ and $\mathbf{Z}_{1n}$ are irrelevant to observations and, accordingly, their a.s. convergence to 0 follows from the case of IID sample in Gyrófi and van der Meulen (1987). Thus, we only show the convergence of $\mathbf{U}_{1n}$ and $\mathbf{V}_{1n}$ in the following.

First, we show that $\mathbf{U}_{1n}$ converges to 0 a.s. Simple arithmetic shows

$$\Big|\mathbf{U}_{1n}\Big| \quad\leq\quad \max_{j\in\mathcal{F}_n}\left|\frac{\mu_n(A_{nj})}{\mu(A_{nj})} - 1\right| \times \sum_{i\in\mathcal{F}_n} \mu(A_{nj}) \cdot \left|\log\left(\frac{\mu_n(A_{ni})}{\lambda(A_{ni})}\right)\right| \equiv \alpha_n \times \beta_n, \tag{12}$$

where we denote the first component and the second component of right-hand side in (12) be α_n and β_n, respectively. Hence, it suffices to show that α_n converge to 0 a.s. and β_n converges to $\mathbf{H}\big(f\big)$. To show the convergence of α_n, applying Lemma 2 with a sufficiently small α, we have

$$\begin{aligned}
\mathbf{P}\Big(\alpha_n > \epsilon\Big) &\leq \sum_{j\in\mathcal{F}_n} \mathbf{P}\Big(\Big|\mu_n(A_{nj}) - \mu(A_{nj})\Big| > \epsilon\cdot\mu_n(A_{nj})\Big) \\
&\leq c\cdot\big|\mathcal{F}_n\big|\cdot\exp\big(-\gamma_2\cdot n\cdot\min_{j\in\mathcal{F}_n}\mu_n(A_{nj})^2\big), \\
&\leq c\cdot\exp\big(-\gamma_3\, n\, a_n^2\, h_n^2\big),
\end{aligned}$$

for some positive γ_2 and γ_3. Therefore, Borel-Cantelli lemma with the assumption (iii) proves that α_n converges to 0 a.s. On the other hand, $\beta_n = \beta_{1n} + \beta_{2n}$, where

$$\begin{aligned}
\beta_{1n} &= \left|\sum_{i\in\mathcal{F}_n} \mu(A_{ni}) \cdot \big\{ \log\mu(A_{ni}) - \log\lambda(A_{ni})\big\}\right| \\
\beta_{2n} &= \left|\sum_{i\in\mathcal{F}_n} \mu(A_{ni}) \cdot \big\{ \log\mu_n(A_{ni}) - \log\mu(A_{ni})\big\}\right|.
\end{aligned}$$

Then, β_{1n} converges to $\mathbf{H}\big(f\big) < \infty$ by Lemma 3, and β_{2n} is $\big|\mathbf{V}_{n1}\big|$ whose convergence to 0 will be shown in the next.

The main step to show the convergence of $\mathbf{V}_{1n}$ is the log–sum inequality, which shows

$$\mathbf{V}_{1n} \quad \leq \quad \log\left(\frac{1}{\sum_{i\in\mathcal{F}_n}\mu(A_{ni})}\right). \tag{13}$$

Hence, it suffices to show $\sum_{i\in\mathcal{F}_n}\mu(A_{ni}) \to 1$ a.s.; equivalently, $\sum_{i\in\mathcal{F}_n^c}\mu(A_{ni}) \to 0$ a.s.

$$\begin{aligned}
\sum_{i\in\mathcal{F}_n^c}\mu(A_{ni}) &= \mu\Big(\bigcup_{i\in\mathcal{F}_n^c}A_{ni}\Big) \\
&= \mu\Big(\{x:\ \widehat{f}_n(x)\leq a_n\}\Big) = \int_{\widehat{f_n}(x)\leq a_n} f(x)dx \\
&\leq \int_{\frac{1}{2}f(x)\leq a_n} f(x)dx + \int_{\widehat{f_n}(x)\leq a_n\leq\frac{1}{2}f(x)} f(x)dx \\
&\leq \int_{\frac{1}{2}f(x)\leq a_n} f(x)dx + 2\int\big|f(x)-\widehat{f_n(x)}\big|dx,
\end{aligned}$$

which converges to 0 a.s. from Lemma 6 and assumption (i).

To show the convergence of $\mathbf{H}_{2n}$, again note $\mathbf{H}_{2n} = \mathbf{U}_{2n} + \mathbf{V}_{2n} + \mathbf{W}_{2n} + \mathbf{Z}_{2n}$, where

$$\begin{aligned}
\mathbf{U}_{2n} &= \mathbf{H}_{2n} - E\big(\mathbf{H}_{2n}\big|Z\big) \\
\mathbf{V}_{2n} &= \int_{\widehat{f_n}(x)\geq a_n}\Big(\log E\big(\widehat{f}_n(x)\big) - \log\widehat{f}_n(x)\Big)f(x)dx \\
\mathbf{W}_{2n} &= \int_{\widehat{f_n}(x)\leq a_n}\log f(x)\cdot f(x)dx \\
\mathbf{Z}_{2n} &= \int_{\widehat{f_n}(x)\leq a_n}\Big(\log f(x) - \log E\big(\widehat{f}_n(x)\big)\Big)\cdot f(x)dx
\end{aligned}$$

Here, $\mathbf{W}_{2n}$, and $\mathbf{Z}_{2n}$ do not depend on observations and $\mathbf{V}_{2n}$ euqals $\mathbf{V}_{1n}$. Thus, it suffices to show that $\mathbf{U}_{2n}$ converges to 0 a.s.

Define

$$\psi_i = -\log\widehat{f}_n(Y_i)\cdot\mathrm{I}\big(\widehat{f}_n(Y_i)\geq a_n\big) - \int_{\widehat{f_n}(x)\geq a_n}\big(-\log\widehat{f}_n(x)\big)f(x)dx,$$

for $i = 1, 2, \cdots, \big[n/2\big]$. Then, since $\big\{Y_i\big\}_{i=1}^{[n/2]}$ is a regular $\phi-$mixing process, $\big\{\psi_i\big\}_{i=1}^{[n/2]}$ is also regular $\phi-$mixing. According to Lemma 2 with a sufficiently small α

$$\mathbf{P}\Big(\big|\mathbf{H}_{2n} - E\big(\mathbf{H}_{2n}\big|Z\big)\big| > \epsilon\ \Big|Z\Big) = \mathbf{P}\Big(\Big|\sum_{i=1}^{[n/2]}\psi_i\Big| \geq n\epsilon\Big) \leq \exp\big(-\gamma_1\cdot n\big) \tag{14}$$

for some positive γ_1. Finally, (14) has a finite sum and the Borel-Cantelli lemma shows that $\mathbf{U}_{2n}$ converges to 0 a.s. □

4. SIMULATION STUDIES

In this section, we implement a limited simulation study to investigate the finite sample performance of two histogram based entropy estimators, $\mathbf{H}_{1n}$ and $\mathbf{H}_{2n}$.

First, the sequence $\{X_n\}_{n=1}^{\infty}$ is generated from a Gaussian autoregressive process (AR) $X_n = \rho X_{n-1} + \epsilon_n$, where ϵ_n is IID $N\big(0,1\big)$, where two hundreds data sets are generated for each for each n =200,400,600, 800, and 1000. In each n and ρ, H_{1n} and H_{2n} are evaluated from 200 data sets. Finally, using those 200 estimators from the 200 generated data sets, the mean squared error (MSE) and the bias of the estimators are evaluated (approximated). The results are presented in Table 1. Second, the sequence is generated from a Gaussian AR process with a coefficient of ρ where ϵ_n from IID $N\big(0,1-\rho^2\big)$. Here, the stationary marginal distribution is $N(0,1)$ for every $\rho \in (-1,1)$. As earlier, the MSE and the bias are evaluated for every n and ρ. The results are presented in Table 2.

From Table 1 and 2, we find that $\mathbf{H}_{1n}$ performs better than $\mathbf{H}_{2n}$ in every case. We can also read that both estimators underestimate the true entropy and it is consistent with the result in Hall and Morton (1993) which states that the bias of $\mathbf{H}_{1n}$ based on IID samples is negatively biased when the tails of the distribution are exponentially decreasing. Also, the Gaussian distribution has the maximum entropy among distributions with the same variance. Further note that the estimated density estimator is not Gaussian and may have an entropy smaller than the true entropy. Finally, MSE and bias increase when ρ decreases as expected. Table 1 shows that both MSE and bias seem to depend on true parameter values, the conditional variance σ^2, and the AR coefficient ρ. Especially, to see the single effect of the AR coefficient, Table 2 fixes the variance of the marginal distribution as 1 and shows that both the MSE and the bias still increase as ρ increases.

5. DISCUSSION

We conclude the paper with a discussion on $\sqrt{n}-$ consistency of $\mathbf{H}_{1n}$ (when $d=1$). Hall and Morton (1993) proved it for IID samples under the regularity conditions to the underlying density f. Here, we conjecture that under the same regularity conditions, Theorem 2.1 in Hall and Morton (1993) may be extended to $\rho-$mixing sequences, which is

$$\begin{aligned}
&f>0 \text{ on } \big(-\infty,\infty\big),\ f' \text{ exists and is continuous on } \big(-\infty,\infty\big),\\
&\text{and for constants } c_1,c_2>0 \text{ and } \alpha_1,\alpha_2>1,\ f'(x)\sim -c_1\alpha_1 x^{-\alpha_1-1}\\
&\text{and } f'\big(-x\big)\sim -c_2\cdot\alpha_2\cdot x^{-\alpha_2-1} \text{ as } x\to+\infty.
\end{aligned}$$

The following notations will be used. Let $\mathcal{P}=\big\{A_i\big\}_{i=\infty}^{\infty}$ be the partition of $\mathbb{R}$, where $A_i=\big((i-1)h,ih\big]$. Let $N_k=\sum_{i=1}^{n}\mathrm{I}\big(X_i\in A_k\big)$ and h, the bin size in the histogram, is in $\mathcal{H}_n\equiv\big\{m^{-1}\ :\ m \text{ is an integer and } n^{\delta}\le m\le n^{1-\delta}\big\}$ for $\delta\in(0,1/2)$. Define $U_{ij}=\mathrm{I}\big(X_j\in A_i\big)-p_i$ with $p_i=\mathbf{P}\big(X\in A_i\big)$. Then, under suitable mixing rates to the observed $\rho-$mixing process, we conjecture that

$$\mathbf{H}_{1n}\equiv\sum_{i=1}^{n}\log\widehat{f}_n(X_i)=\frac{1}{n}\sum_{i=1}^{n}\log f(X_i)+O\Big(\big(nh\big)^{-1+(1/\min(\alpha_1,\alpha_2))}\Big). \tag{15}$$

For brevity, we indicate and discuss only those where the proof differs significantly from that of Theorem 4.1 in Hall (1990) which is the source of Theorem 2.1 in Hall and Morton (1993).

As in Theorem 4.1 in Hall (1990),

$$\mathbf{H}_{1n}=\frac{1}{n}\sum_{j=1}^{n}\log\left(\frac{\widehat{f}_n(X_j)}{f(X_j)}\right)=\mathbf{S}_1+\mathbf{S}_2,$$

where $S_1\equiv\big(1/n\big)\sum_j\log\big\{\widehat{f}_n(X_j)/f(X_j)\big\}$ and $S_2\equiv\big(1/n\big)\sum_j\log\big\{\mu(X_j)/f(X_j)\big\}$. Since $\mathbf{S}_2$ does not depend on samples $\big\{X_i\big\}_{i=1}^{n}$, $\mathbf{S}_2\approx O\big(h^2\big)+o_p\big\{(nh)^{-1+1/\alpha_1}+(nh)^{-1+1/\alpha_2}+h^2\big\}$ as in Hall (1990). Thus, only $\mathbf{S}_1$ will be considered below.

In proving the rate of $\mathbf{S}_1$, it suffice to get the same rates of $\S_{kl}$ and T_{kl} for every (k,l) with observed $\rho-$mixing processes. Subsequently, to get the same rates of S_{kl}, T_{kl}, and their expectations with those in Hall (1990), it again may suffice to show that:

(1) for every k,

$$E\big|N_i-np_i\big|^k\approx E_{\mathrm{I}}\big|N_i-np_i\big|^k,$$

where $\mathrm{E_I}$ represents the expectation under the assumption that $\{X_i\}_{i=1}^n$ are independent of each other.

(2) For every k,

$$E\left|Z_1\right|^{2k} \approx E_{\mathrm{I}}\left|Z_1\right|^{2k}, \quad \text{and} \quad E\left|Z_2\right|^{2k} \approx E_{\mathrm{I}}\left|Z_2\right|^{2k}$$

where

$$\begin{aligned} Z_1 &= \sum_{j=1}^{n}\sum_{m=1}^{n}\sum_{i}^{(k,l)}\{U_{ij}U_{im} - \mathrm{E}(U_{ij}U_{im})\} \\ Z_2 &= \sum_{i}^{(k,l)}\{(N_i - np_i)^3 - \mathrm{E}(N_i - np_i)^3\} \end{aligned}$$

Suppose the $\rho-$mixing process has a geometric mixing rate $\rho(n) = q^n$ for some $0 < q < 1$. Then,

$$\begin{aligned} ES_n^2 &= n \cdot EX_1^2 + 2 \cdot \sum_{i<j}\rho(|i-j|) \\ &= n \cdot EX_1^2 + O(1). \end{aligned}$$

Thus, $h(n)$ in Lemma 5 is $O(1)$ and (1) is satisfied. However, it is still unknown what mixing rate can result in (2).

ACKNOWLEDGEMENT

We are grateful to the editor and referees for many helpful suggestions. Johan Lim was supported by Basic Science Research Fund from College of Economics at Yonsei University.

REFERENCES

Abou-Jaoude, S. (1976). Conditions necessaires et suffisantes de convergence L_1 en probabilite de I'histogramme pour une densite. *Annales de l'Institut Henri Poincare* 12:213-231.

Bercher, J. and Vignat, C. (2000). Estimating the entropy of a signal with applications. *IEEE Trans. on Signal Processing* 48:1687-1694.

Carbon, M. (1983). InAgalitA de Bernstein pour les processus fortement mAlangeants non nAcessairement stationaire. *C.R. Acad. Scienc., Paris, I* 297:303-306.

Collomb, G. (1984). Propietes de convergence presque complete du predicteur a noyau. *Z. Wahrscheinlichkeitstheorie und verw. Gebiete* 66:441-460.

Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory.* New York: John Wiley & Sons. Inc.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* New York: John Willy & Sons.

Dudewicz, E.J. and van der Meulen, E.C. (1981). Entropy-based tests for uniformity. *Journal of the American Statistical Association* 76:967-974.

Eggermont, P.B. and LaRiccia, V.N. (1999). Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. on Information Theory* 45:1321-1326.

Grassberger, P. (1989). Estimating the information content of symbol sequences and efficient codes. *IEEE Trans. on Information Theory* 35:669-675.

Gyórfi, L. and van der Meulen, E. (1987). Density-free convergence properties of various estimators of entropy. *Computational Statistics and Data Analysis* 5: 425-436.

Gyórfi, L. and van der Meulen, E. (1991). On the nonparametric estimation of the entropy functional. *Nonparametric functional estimation and related topics*, 81-95. Boston, NY: Kluwer Academic Publishers.

Hall, P. (1990). Akaike's information criterion and Kullback–Leibler loss for histogram density estimation. *Probability Theory and Related Fields* 85:449-467.

Hall, P. and Morton, S. (1993). On the estimation of entropy. *Annals of Institute of Statistical Mathematics* 45:69-88

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Annals*

of Institute of Statistical Mathematics 41:683-697.

Kontoyiannis, I., Algoet, P.H., Suhov, Y.M., and Wyner, A.J. (1998). Nonparametric entropy estimation for stationary processes and random fields with application to English text. *IEEE Trans. on Information Theory* 44:1319-1327.

Lin, Z. and Lu, C. (1996). *Limit theory for mixing dependent random variables.* Boston. NY: Kluwer Academic Publishers.

Prakasa Rao, B.L.S. (1983). *Nonparametric functional estimation.* Olando. FL: Academic Press,

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B* 38:54-59.

		$\mathbf{H}_{1n}$			$\mathbf{H}_{2n}$		
	$(\rho,\ \sigma)$	$(0.1,\ 1.0)$	$(0.5,\ 1.0)$	$(0.9,\ 1.0)$	$(0.1,\ 1.0)$	$(0.5,\ 1.0)$	$(0.9,\ 1.0)$
$n = 200$	Bias	-0.1554	-0.1888	-0.3891	-0.2631	-0.3705	-1.0818
	$\sqrt{MSE}$	0.1632	0.1980	0.4063	0.2863	0.3895	1.0983
$n = 400$	Bias	-0.0847	-0.0996	-0.2133	-0.1037	-0.1345	-0.5276
	$\sqrt{MSE}$	0.1265	0.1598	0.5558	0.0930	0.1090	0.2322
$n = 600$	Bias	-0.0578	-0.0682	-0.1469	-0.0577	-0.0769	-0.2787
	$\sqrt{MSE}$	0.0830	0.1004	0.3163	0.0646	0.0778	0.1657
$n = 800$	Bias	-0.0450	-0.0505	-0.1125	-0.0359	-0.0511	-0.1888
	$\sqrt{MSE}$	0.0515	0.0595	0.1341	0.0615	0.0739	0.2202
$n = 1000$	Bias	-0.0340	-0.04219	-0.1000	-0.02705	-0.0469	-0.1387
	$\sqrt{MSE}$	0.0406	0.0497	0.1195	0.0481	0.0657	0.1711

Table 1: Performance of H_{1n} and H_{2n}. Both bias and MSE in each cell are evaluated from 200 data sets. The conditional variance σ^2 is fixed as 1 in every cell.

		$\mathbf{H}_{1n}$			$\mathbf{H}_{2n}$		
	$(\rho,\ \sigma)$	$(0.1,\ 1.01)$	$(0.5,\ 1.33)$	$(0.9,\ 5.26)$	$(0.1,\ 1.01)$	$(0.5,\ 1.33)$	$(0.9,\ 5.26)$
$n = 200$	Bias	-0.1590	-0.1623	-0.2200	-0.2806	-0.2867	-0.3800
	$\sqrt{MSE}$	0.1673	0.1729	0.2505	0.30503	0.3107	0.4271
$n = 400$	Bias	-0.0829	-0.0830	-0.1369	-0.0901	-0.1041	-0.1462
	$\sqrt{MSE}$	0.0901	0.0934	0.1727	0.1198	0.1321	0.2098
$n = 600$	Bias	-0.0565	-0.0630	-0.0787	-0.0488	-0.0626	-0.0660
	$\sqrt{MSE}$	0.0633	0.0728	0.1134	0.0743	0.0940	0.1360
$n = 800$	Bias	-0.0469	-0.0470	-0.0658	-0.0452	-0.0426	-0.0493
	$\sqrt{MSE}$	0.0529	0.0578	0.0994	0.0629	0.0665	0.1095
$n = 1000$	Bias	-0.0352	-0.0382	-0.0557	-0.0282	-0.0335	-0.0377
	$\sqrt{MSE}$	0.0416	0.0474	0.0819	0.0458	0.0545	0.0904

Table 2: Performance of H_{1n} and H_{2n}. Both bias and MSE in each cell are evaluated from 200 data sets. The conditional variance σ^2 is set to $1 - \rho^2$ in each cell.