Report: Exploring Non-Parametric Entropy Estimators and Bootstrap Enhancements

Introduction

Understanding and measuring entropy, a fundamental concept in information theory, is crucial in various fields. This study delves into the evaluation of non-parametric of Shannon entropy estimators and their Bootstrap enhancements. Non-parametric estimators offer a flexible approach to parameter estimation, while Bootstrap technique provide a robust tool for refining these estimators and quantifying their uncertainty. Using various sample sizes and replications, different estimators are evaluated based on bias, variance, mean squared error (MSE), and confidence intervals. The results highlight the trade-off between bias and variance, shedding light on the efficiency and reliability of each estimator.

Non-Parametric Entropy Estimators

The problem of non-parametric estimating of $H(X)$ has been considered by many authors, who proposed estimators based on spacings.

Vasicek (1976), used $f(z) = p$ to express $H(X)$ as

$$H(X) = \int_0^1 \log \left(\frac{\mathrm{d}}{\mathrm{d}p} Q(p) \right) \mathrm{d}p,$$

where $Q(p) = F^{-1}(p) = \inf \{x : F(X) \leqslant p\}$ is the quantile function. The derivative of $F^{-1}(p)$ is estimated by a function of the order statistics.

Assuming that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the distribution $F(x)$, the estimator is defined as:

$$\hat{H}_V(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^n \log \left[\frac{n}{2m} \left(X_{(i+m)} - X_{(i-m)} \right) \right], \tag{1}$$

where $m < n/2$ is a positive integer, $X_{(i+m)} - X_{(i-m)}$ is the m-spacing and $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$ are the order statistics and $X_{(i)} = X_{(1)}$ if $i < 1$, $Z_{(i)} = X_{(n)}$ if $i > n$.

Multiple authors have presented adaptations to Vasicek's estimator, including Van Es (1992), who proposed a new estimator of entropy given by:

$$\hat{H}_{\mathrm{VE}}(\boldsymbol{X}) = \frac{1}{n-m} \sum_{i=1}^{n-m} \log \left[\frac{n+1}{m} \left(X_{(i+m)} - X_{(i)} \right) \right] + \sum_{k=m}^n \frac{1}{k} + \log \frac{m}{n+1}. \tag{2}$$

Van Es demonstrated that, under some conditions, this estimator exhibits consistency and asymptotic normality.

In 1995 Correa suggested a modification of Vasicek's estimator. In estimation the density f of F in the interval $\left(X_{(i-m)}, X_{(i+m)} \right)$ he used a local linear model based on $2m+1$ points: $F\left(X_{(j)} \right) = \alpha + \beta X_{(j)} + \varepsilon, j = m - i, \ldots, m + i$. This yields a following estimator

$$\hat{H}_C(\mathbf{X}) = -\frac{1}{n} \sum_{i=1}^n \ln (b_i), \tag{3}$$

where

$$b_i = \frac{\sum_{j=i-m}^{i+m} \left(X_{(j)} - \bar{X}_{(i)} \right) (j - i)}{n \cdot \sum_{j=i-m}^{i+m} \left(X_{(j)} - \bar{X}_{(i)} \right)^2},$$

$$\bar{X}_{(i)} = \frac{1}{2m+1} \sum_{j=i-m}^{i+m} X_{(j)},$$

m is a positive integer smaller than $\frac{n}{2}$, $X_{(i)} = X_{(1)}$ for $i < 1$ and $X_{(i)} = X_{(n)}$ for $i > n$.

Ebrahimi et al. (1994), adjusted the weights of Vasicek's estimator, in order to take into account the fact that the differences are truncated around the smallest and the largest data points. i.e. $X_{(i+m)} - X_{(i-m)}$ is replaced by $X_{(i+m)} - X_{(1)}$ when $i \leqslant m$ and $X_{(i+m)} - X_{(i-m)}$ is replaced by $X_{(n)} - X_{(i-m)}$ when $i \geqslant n-m+1$. Their estimator is given by

$$\widehat{H}_E(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{n}{c_i m} \left(X_{(i+m)} - X_{(i-m)} \right) \right], \tag{4}$$

where

$$c_i = \begin{cases} 1 + \frac{i-1}{m}, & 1 \leqslant i \leqslant m \\ 2, & m+1 \leqslant i \leqslant n-m, \\ 1 + \frac{n-i}{m}, & n-m+1 \leqslant i \leqslant n \end{cases}$$

Noughabi and Arghami suggested an entropy estimator given by

$$\widehat{H}_{NA}(\mathbf{X}) = \frac{1}{n}(\mathbf{X}) \sum_{i=1}^{n} \log \left[\frac{n}{a_i m} \left(X_{(i+m)} - X_{(i-m)} \right) \right], \tag{5}$$

where

$$a_i = \begin{cases} 1, & 1 \leqslant i \leqslant m, \\ 2, & m+1 \leqslant i \leqslant n-m, \\ 1, & n-m+1 \leqslant i \leqslant n, \end{cases}$$

and $X_{(i-m)} = X_{(1)}$ for $i \leqslant m$ and $X_{(i+m)} = X_{(n)}$ for $i \geqslant n-m$.

Al-Omari suggested the following estimators of entropy:

$$\widehat{H}_{AO_1}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{n}{\omega_i m} \left(X_{(i+m)} - X_{(i-m)} \right) \right] \tag{6}$$

where

$$\omega_i = \begin{cases} 3/2 & \text{if } 1 \leqslant i \leqslant m \\ 2 & \text{if } m+1 \leqslant i \leqslant n-m, \\ 3/2 & \text{if } n-m+1 \leqslant i \leqslant n, \end{cases}$$

in which $X_{(i-m)} = X_{(1)}$ for $i \leqslant m$, and $X_{(i+m)} = X_{(n)}$ for $i \geqslant n-m$. And

$$\widehat{H}_{AO_2}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{n}{v_i m} \left(X_{(i+m)} - X_{(i-m)} \right) \right] \tag{7}$$

where

$$v_i = \begin{cases} 1 + (i-1)/m & \text{if } 1 \leqslant i \leqslant m \\ 2 & \text{if } m+1 \leqslant i \leqslant n-m, \\ 1 + (n-i)/2m & \text{if } n-m+1 \leqslant i \leqslant n \end{cases}$$

in which $X_{(i-m)} = X_{(1)}$ for $i \leqslant m$, and $X_{(i+m)} = X_{(n)}$ for $i \geqslant n-m$.

Results for Nonparametric Estimators with $\Gamma_{\mathbf{SAR}}$

Our experimental setup involves an analysis of bias and MSE for each estimator using varying sample sizes, denoted as $n \in \{9, 25, 49, 81, 121\}$. For each specific sample size, we generate 100 random samples to ensure robust statistical insights. Additionally, we consider distinct parameter values, where μ varies within $\{1, 3, 10, 100\}$ and L within $\{1, 3, 5, 18\}$. This parameter variation aims to explore potential impacts on estimator performance. Figures 1, 2, 3, and 4 present the bias and MSE for each of the seven nonparametric entropy estimators.
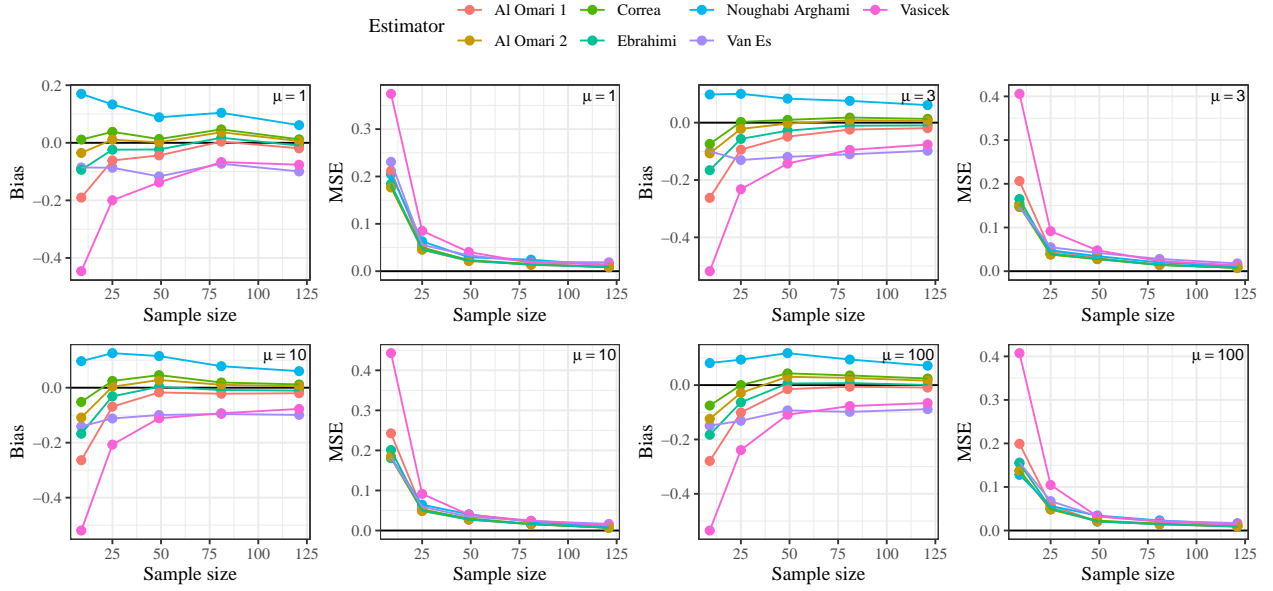
Figure 1: Bias and MSE, for $L = 1$.
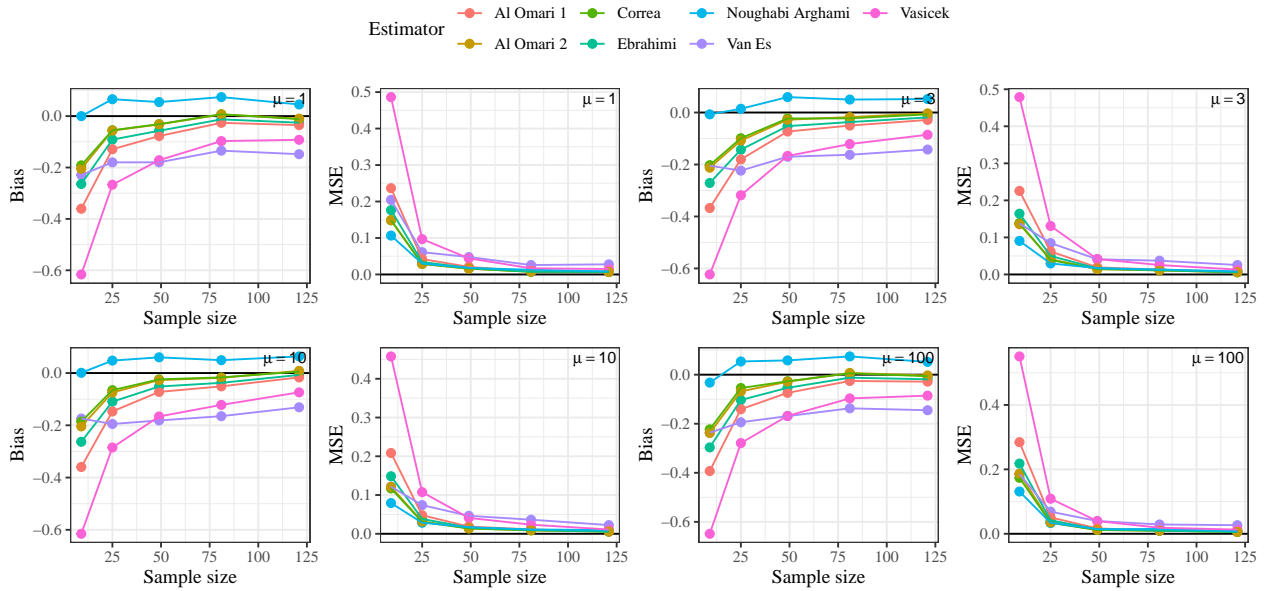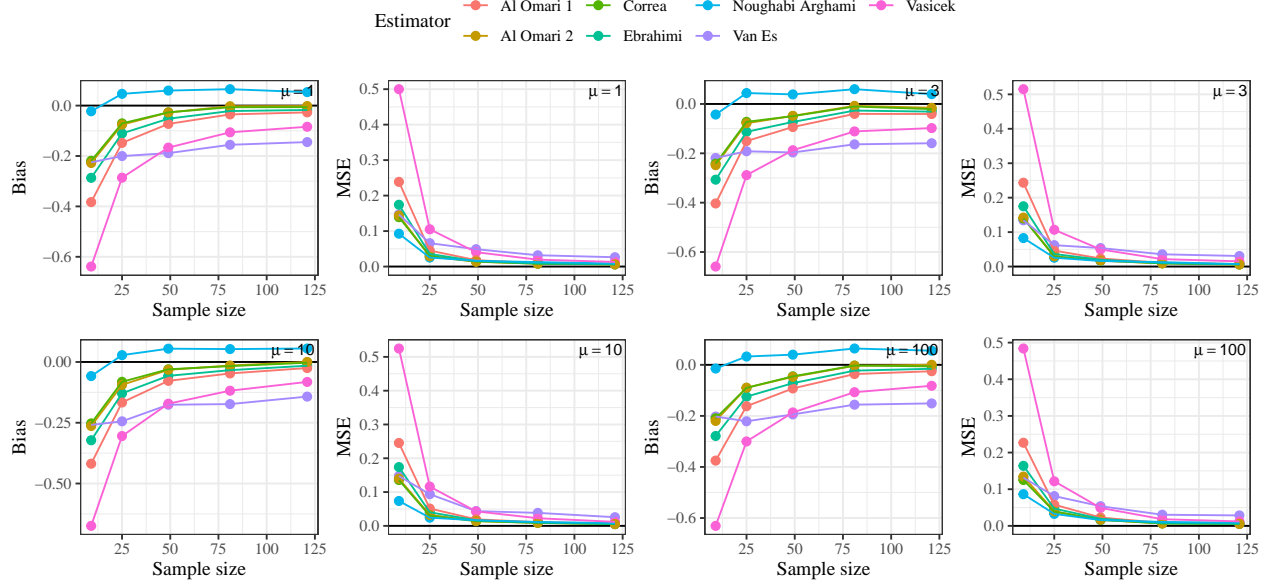

Figure 2: Bias and MSE, for $L = 3$.
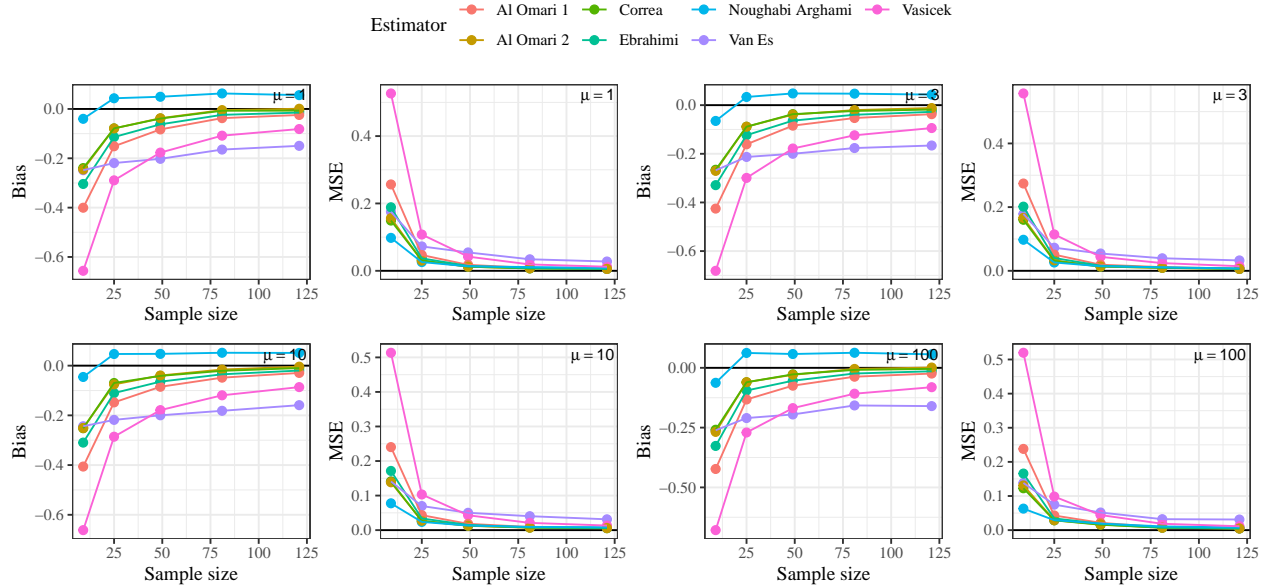
Figure 3: Bias and MSE, for $L = 5$.

Figure 4: Bias and MSE, for $L = 18$.

The simulation results for bias and MSE, where μ takes values in $\{1, 3, 10, 100\}$ and $L = 5$, are presented in Table 1.

Based on our experiments, the estimators with consistently lower bias and MSE across various scenarios are $\widehat{H}_C$, $\widehat{H}_{AO_2}$, and $\widehat{H}_E$. These estimators demonstrate robust performance and convergence under different parameter combinations.

For the Noughabi Arghami estimator, it demonstrates consistently low MSE across various scenarios. However, it is noteworthy that the bias remains constant and notably high for sample sizes larger than 25, with no indication of rapid convergence.

4

Table 1: Bias and MSE for different μ values, and $L = 5$.

		Bias							MSE						
μ	n	$\hat{H}_{VE}$	$\hat{H}_C$	$\hat{H}_E$	$\hat{H}_{NA}$	$\hat{H}_V$	$\hat{H}_{AO_1}$	$\hat{H}_{AO_2}$	$\hat{H}_{VE}$	$\hat{H}_C$	$\hat{H}_E$	$\hat{H}_{NA}$	$\hat{H}_V$	$\hat{H}_{AO_1}$	$\hat{H}_{AO_2}$
	9	-0.2251	-0.2190	-0.2866	-0.0225	-0.6387	-0.3829	-0.2278	0.1468	0.1391	0.1742	0.0926	0.5000	0.2387	0.1440
	25	-0.2003	-0.0705	-0.1105	0.0467	-0.2860	-0.1480	-0.0756	0.0659	0.0299	0.0354	0.0254	0.1050	0.0451	0.0289
1	49	-0.1887	-0.0267	-0.0521	0.0597	-0.1666	-0.0727	-0.0273	0.0489	0.0140	0.0154	0.0163	0.0405	0.0180	0.0135
	81	-0.1555	-0.0064	-0.0216	0.0652	-0.1060	-0.0350	-0.0024	0.0318	0.0079	0.0087	0.0124	0.0194	0.0094	0.0082
	121	-0.1447	-0.0061	-0.0173	0.0535	-0.0840	-0.0269	-0.0017	0.0263	0.0062	0.0062	0.0088	0.0130	0.0067	0.0059
	9	-0.2189	-0.2417	-0.3068	-0.0428	-0.6589	-0.4032	-0.2481	0.1343	0.1396	0.1750	0.0827	0.5150	0.2434	0.1424
	25	-0.1916	-0.0724	-0.1128	0.0443	-0.2884	-0.1503	-0.0780	0.0621	0.0289	0.0364	0.0256	0.1068	0.0462	0.0297
3	49	-0.1964	-0.0493	-0.0729	0.0388	-0.1875	-0.0936	-0.0482	0.0537	0.0174	0.0199	0.0161	0.0498	0.0234	0.0169
	81	-0.1636	-0.0104	-0.0268	0.0600	-0.1112	-0.0402	-0.0076	0.0356	0.0093	0.0098	0.0127	0.0215	0.0107	0.0092
	121	-0.1592	-0.0208	-0.0311	0.0398	-0.0977	-0.0406	-0.0154	0.0306	0.0059	0.0065	0.0072	0.0151	0.0072	0.0058
	9	-0.2589	-0.2527	-0.3221	-0.0581	-0.6742	-0.4185	-0.2634	0.1474	0.1356	0.1740	0.0736	0.5248	0.2454	0.1396
	25	-0.2442	-0.0816	-0.1291	0.0280	-0.3047	-0.1666	-0.0943	0.0937	0.0301	0.0402	0.0243	0.1163	0.0512	0.0324
10	49	-0.1759	-0.0301	-0.0571	0.0547	-0.1716	-0.0777	-0.0323	0.0436	0.0140	0.0166	0.0164	0.0428	0.0194	0.0144
	81	-0.1733	-0.0171	-0.0338	0.0530	-0.1182	-0.0471	-0.0146	0.0384	0.0088	0.0098	0.0115	0.0226	0.0109	0.0089
	121	-0.1425	-0.0032	-0.0154	0.0554	-0.0821	-0.0250	0.0002	0.0256	0.0051	0.0054	0.0082	0.0119	0.0058	0.0052
	9	-0.2025	-0.2094	-0.2786	-0.0146	-0.6307	-0.3750	-0.2198	0.1314	0.1251	0.1638	0.0864	0.4839	0.2268	0.1345
	25	-0.2211	-0.0901	-0.1246	0.0326	-0.3001	-0.1620	-0.0897	0.0812	0.0389	0.0471	0.0326	0.1216	0.0578	0.0396
100	49	-0.1942	-0.0450	-0.0720	0.0398	-0.1865	-0.0926	-0.0472	0.0537	0.0163	0.0194	0.0158	0.0490	0.0228	0.0165
	81	-0.1561	-0.0028	-0.0228	0.0640	-0.1072	-0.0362	-0.0036	0.0303	0.0066	0.0072	0.0107	0.0181	0.0079	0.0066
	121	-0.1511	-0.0060	-0.0154	0.0554	-0.0821	-0.0250	0.0002	0.0285	0.0053	0.0057	0.0085	0.0122	0.0060	0.0054

In contrast, the Van Es and Vasicek estimators have a larger bias, but the one with the highest MSE is undoubtedly the Vasicek estimator.

Enhanced Bootstrap Estimators

In this section, we extend our exploration of nonparametric entropy estimators by incorporating enhanced bootstrap methodologies. The integration of bootstrap techniques aims to refine the precision and reliability of entropy estimates, offering a robust framework for information-theoretic analyses.

Bootstrap methods, renowned for their versatility in handling complex data structures, have become invaluable tools for refining statistical inferences. In the realm of entropy estimation, the application of bootstrap techniques seeks to mitigate the inherent challenges posed by finite sample sizes. By repeatedly resampling from the observed data, bootstrap enhances the stability of estimates and provides a comprehensive understanding of the uncertainty associated with each estimator.

The bootstrap technique is a powerful resampling method widely employed in statistical analyses to comprehend the variability and uncertainty associated with sample data. It operates on the fundamental principle that a sample drawn from an underlying population is a representative reflection of the entire population. Bootstrap methodology involves generating numerous resamples by drawing observations with replacement from the original dataset.

Here's a step-by-step breakdown of the bootstrap process:

- Resampling with Replacement: A primary characteristic of bootstrap is the sampling with replacement, meaning that each observation selected during resampling is returned to the pool, allowing the same observation to be chosen multiple times or not at all.

- Creating Bootstrap Samples: Through the resampling process, multiple bootstrap samples are generated, each equivalent in size to the original dataset. This technique mimics the randomness inherent in the data collection process.

- Estimate Computation: The statistical quantity of interest (e.g., mean, variance, or entropy) is calculated for each bootstrap sample, yielding a distribution of estimates.

Theoretical Framework

We present an innovative approach to enhance the non-parametric entropy estimators denoted as $\hat{H}_i$. Acknowledging the inherent bias in these estimators, we propose bootstrap-enhanced versions denoted as $\tilde{H}_i$ to address the bias and refine the accuracy of the original estimations.

Let $\hat{\theta}(X) = \hat{H}_i$ represent the non-parametric entropy estimators based on a sample X of size n. Let's assume that these estimators are inherently biased, that is, the bias is defined as:

$$\text{Bias}\left(\hat{\theta}(X)\right) = \mathbb{E}\left[\hat{\theta}(X)\right] - \theta. \tag{8}$$

Our objective is to devise unbiased estimators with reduced variance. To achieve this, we introduce an "ideal estimator" $\check{\theta}(X)$ using the bias information:

$$\check{\theta}(X) = \hat{\theta}(X) - \text{Bias}\left(\hat{\theta}(X)\right). \tag{9}$$

However, $\check{\theta}(X)$ is not an estimator, because it does not depend on the true parameter θ, prompting the formulation of a new unbiased "proper estimator" $\tilde{\theta}$, from (8) and (9) we have:

$$\tilde{\theta}(X) = \hat{\theta}(X) + \hat{\theta}(X) - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b(X^{(b)}),$$

$$= 2\hat{\theta}(X) - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b(X^{(b)}), \tag{10}$$

where B represents the number of replications in the bootstrap technique.

The proposed $\tilde{\theta}(X)$ leverages the bootstrap technique to refine the original estimators. For each iteration b (from 1 to B) in the bootstrap process, a sample $X^{(b)}$ is generated, and the corresponding bootstrap estimator $\hat{\theta}_b(X^{(b)})$ is calculated.

The final bootstrap-enhanced estimator $\tilde{H}_i$ is computed by incorporating these bootstrap estimators into the formulation:

$$\tilde{H}_i(X) = \tilde{\theta}(X)$$

This methodology significantly reduces bias, resulting in more accurate and robust entropy estimations. The notation $\tilde{H}_i$ denotes the improved versions of the original estimators, providing a comprehensive enhancement to their non-parametric counterparts.

MSE Results for Bootstrap Estimators

Similar to the analysis conducted on the original estimators, we consider two parameter values for $\mu \in \{1, 10\}$ and $L = 5$. To observe the behavior of the estimators under different scenarios, we perform experiments with

varying sample sizes $n \in \{9, 25, 49, 81, 121\}$. For each specific sample size, we generate 100 random samples and perform 100 replications for bootstrap.

In-depth comparisons of bias and MSE between original non-parametric entropy estimators and their respective bootstrap-enhanced versions are presented in Figures 5-10. The application of the bootstrap technique demonstrates notable improvements across the majority of methods.
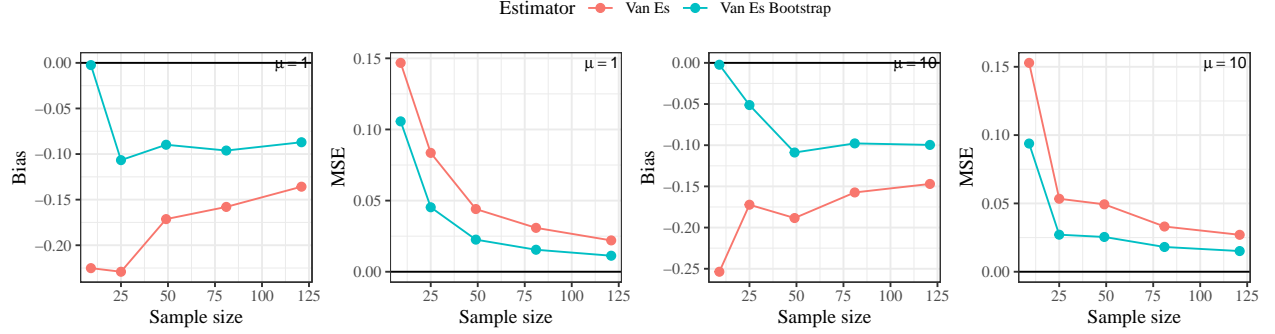

Figure 5: Bias and MSE of $\widehat{H}_{VE}$ and $\widetilde{H}_{VE}$, for $L = 5$.
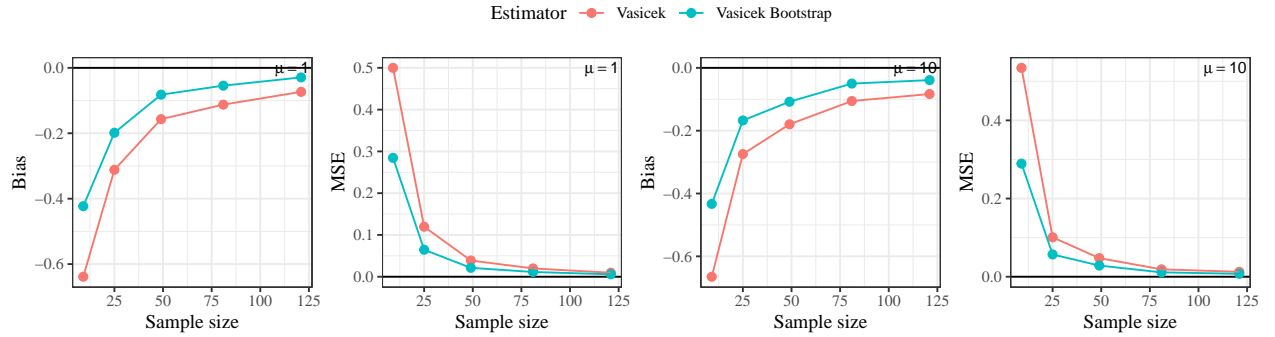
Figure 6: Bias and MSE, for $L = 5$.
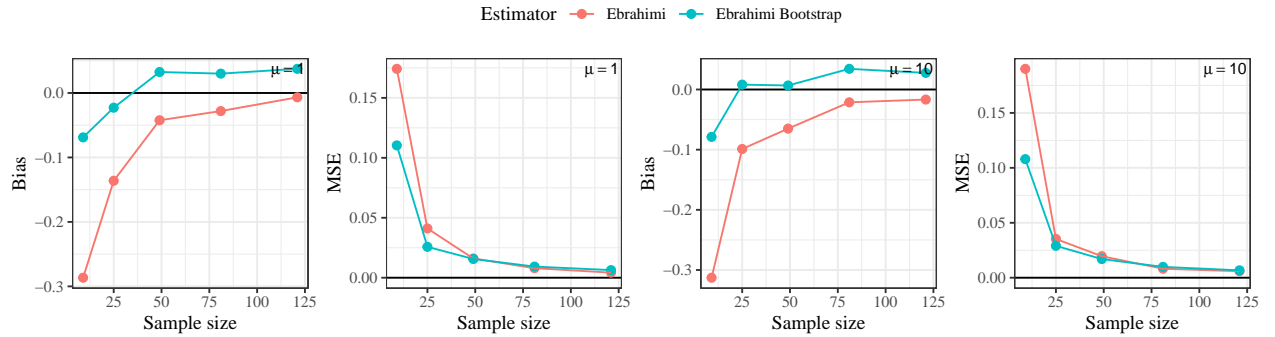
Figure 7: Bias and MSE, for $L = 5$.
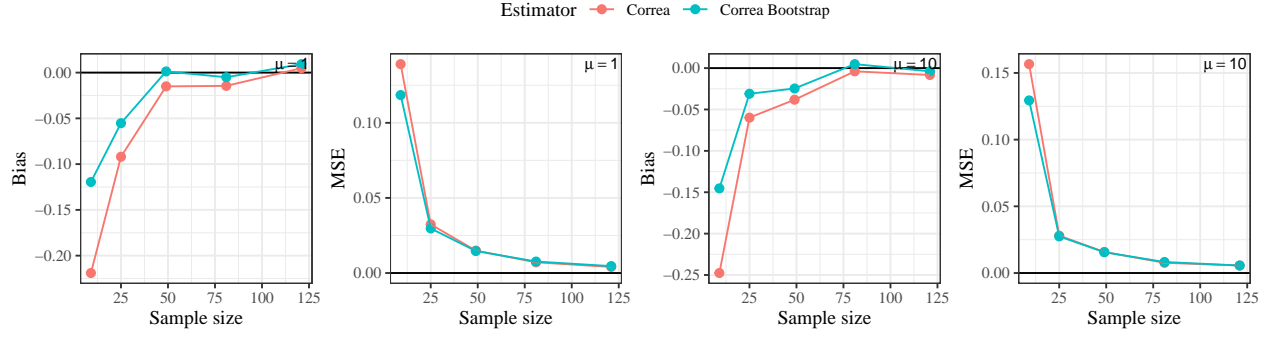
Figure 8: Bias and MSE, for $L = 5$.
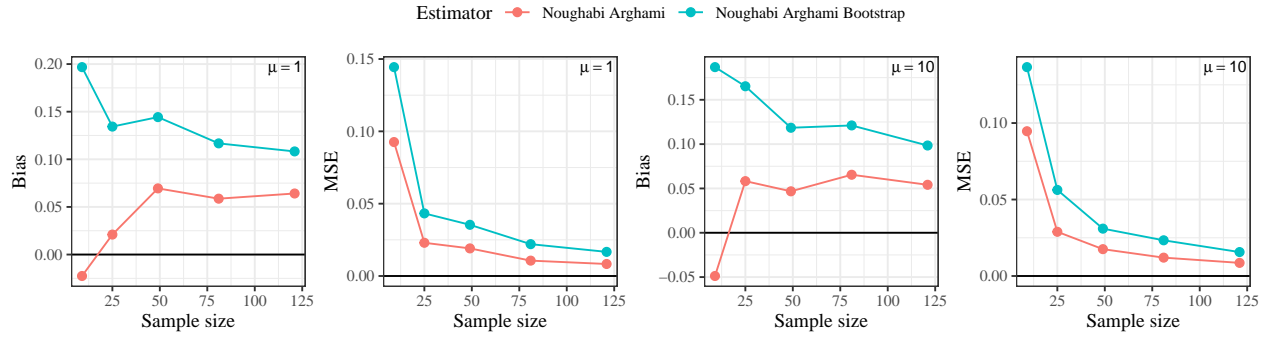
Figure 9: Bias and MSE, for $L = 5$.
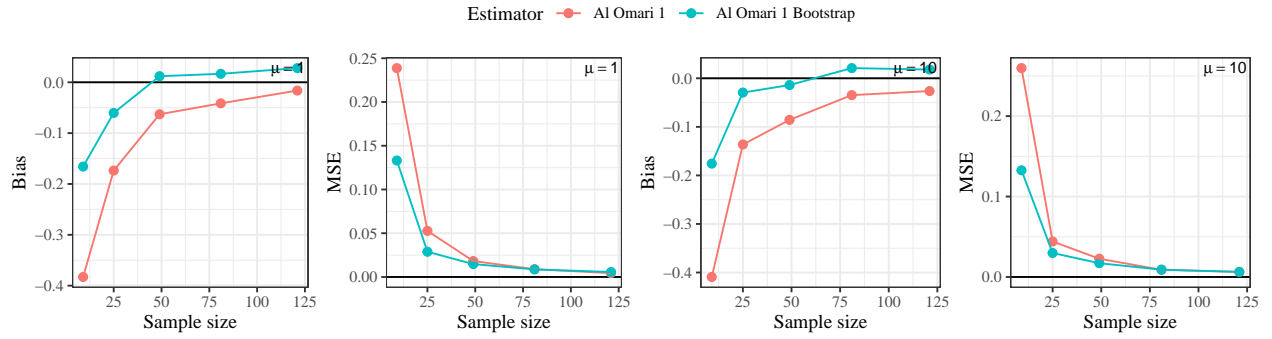
Figure 10: Bias and MSE, for $L = 5$.
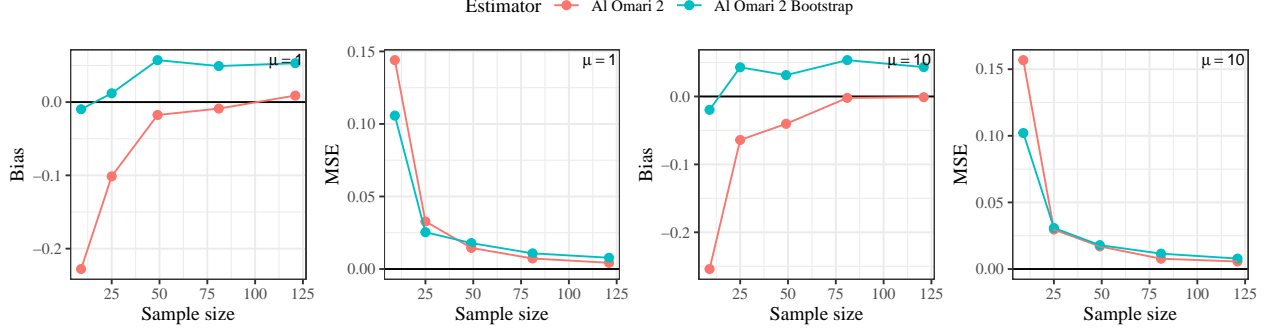
Figure 11: Bias and MSE, for $L = 5$.

The simulation results for Correa Bootstrap, Ebrahimi Bootstrap, and Al Omari 1 Bootstrap estimators are presented in Table 2, where μ takes values in $\{1, 10\}$ and $L = 5$.

Table 2: Bias and MSE for different μ values, and $L = 5$.

μ	n	Bias $\tilde{H}_C$	$\tilde{H}_E$	$\tilde{H}_{AO_1}$	MSE $\tilde{H}_C$	$\tilde{H}_E$	$\tilde{H}_{AO_1}$
	9	-0.1195	-0.0628	-0.1625	0.1185	0.1101	0.1305
	25	-0.0476	-0.0112	-0.0452	0.0312	0.0321	0.0334
1	49	-0.0042	0.0273	0.0067	0.0135	0.0144	0.0138
	81	-0.0047	0.0264	0.0116	0.0085	0.0095	0.0088
	121	0.0021	0.0343	0.0236	0.0054	0.0069	0.0065
	9	-0.1052	-0.0507	-0.1512	0.1191	0.1038	0.1301
	25	-0.0577	-0.0246	-0.0611	0.0378	0.0363	0.0399
10	49	-0.0018	0.0333	0.0126	0.0149	0.0151	0.0141
	81	-0.0098	0.0207	0.0091	0.0084	0.0090	0.0084
	121	-0.0008	0.0257	0.0170	0.0046	0.0054	0.0050

The precision of estimators, as evidenced by bias and MSE comparisons, benefits significantly from the bootstrap technique. Particularly, are the enhancements achieved by the Correa Bootstrap, Ebrahimi Bootstrap, and Al Omari 1 Bootstrap estimators. These methods exhibit remarkable precision, showcasing the efficacy of the bootstrap approach in refining their accuracy.

Conversely, the application of bootstrap did not yield favorable results for the Noughabi-Arghami Bootstrap and Al Omari 2 Bootstrap estimators. The observed lack of improvement suggests that, in certain cases, the bootstrap technique may not be universally effective for all entropy estimation methods.

Motivated by the superior performance of the Correa Bootstrap, Ebrahimi Bootstrap, and Al Omari 1 Bootstrap estimators, we proceed to a more detailed analysis of their variance and confidence intervals. This selection is driven by their consistent precision improvement through bootstrap enhancement, establishing them as promising candidates for further investigation.

Analysis of Variance and Confidence Intervals

The next phase of our investigation involves a detailed examination of the variance exhibited by three selected bootstrap-enhanced estimators: Correa Bootstrap, Ebrahimi Bootstrap, and Al Omari 1 Bootstrap. The primary objective is to assess the stability and consistency of these estimators under varying conditions.

Figures 1, 2, and 3 present a comparative analysis of the variance for each estimator across different parameter values (μ and L) and sample sizes (n). Notably, the Correa Bootstrap, Ebrahimi Bootstrap, and Al Omari 1 Bootstrap estimators consistently demonstrate lower variance compared to their non-bootstrap counterparts.
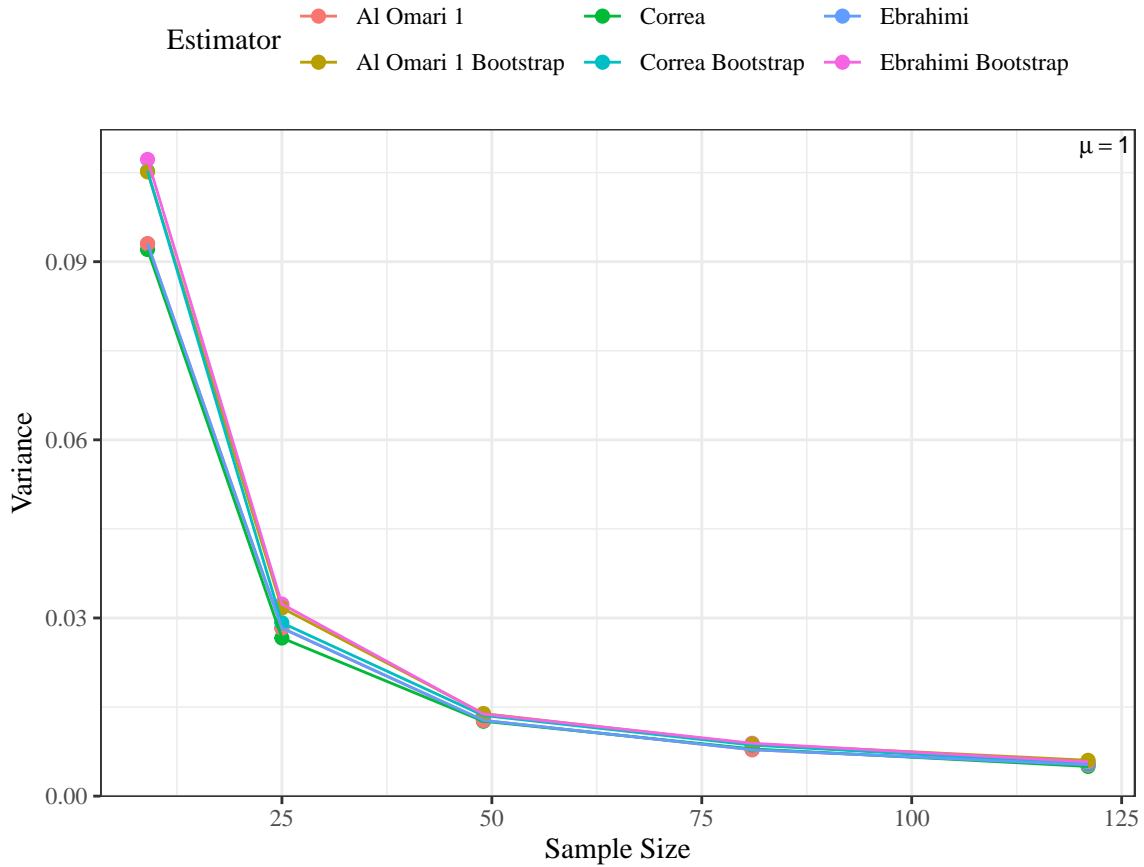
Figure 12: Variance for $L = 5$.
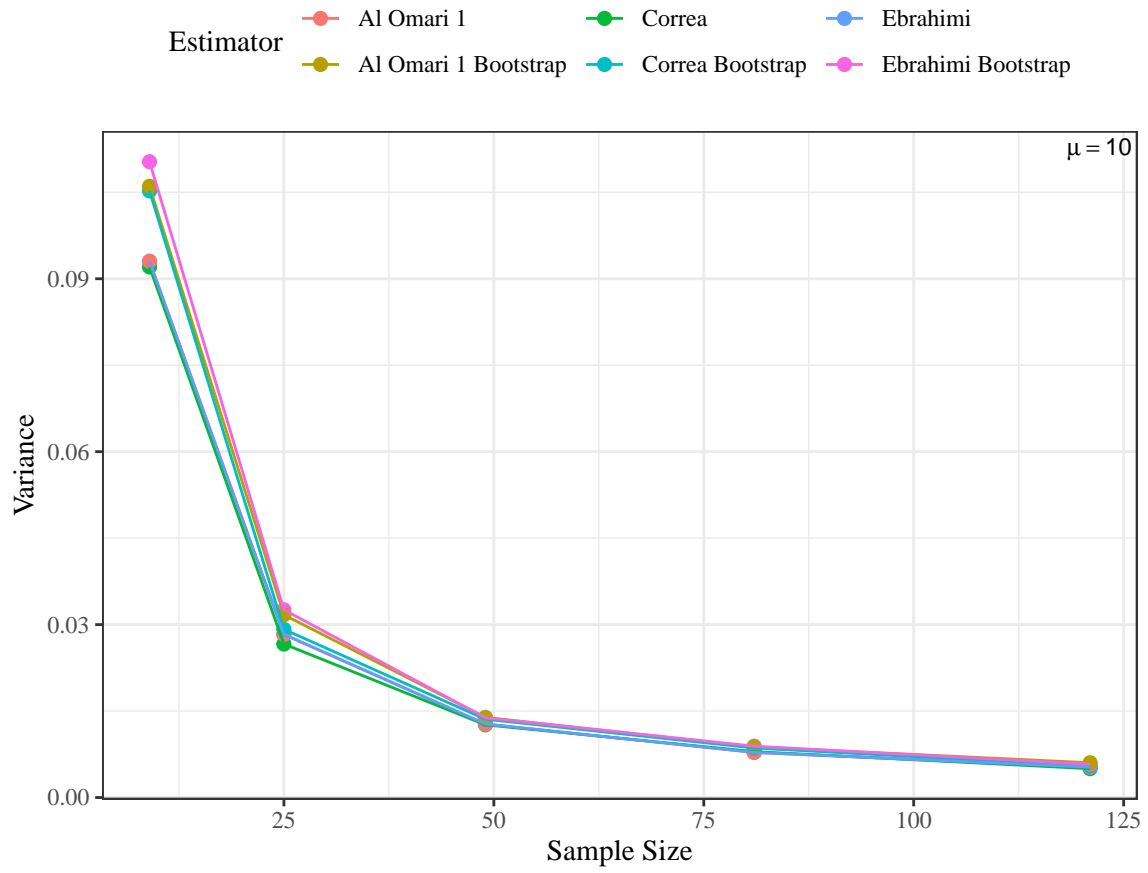
Figure 13: Variance for $L = 5$.
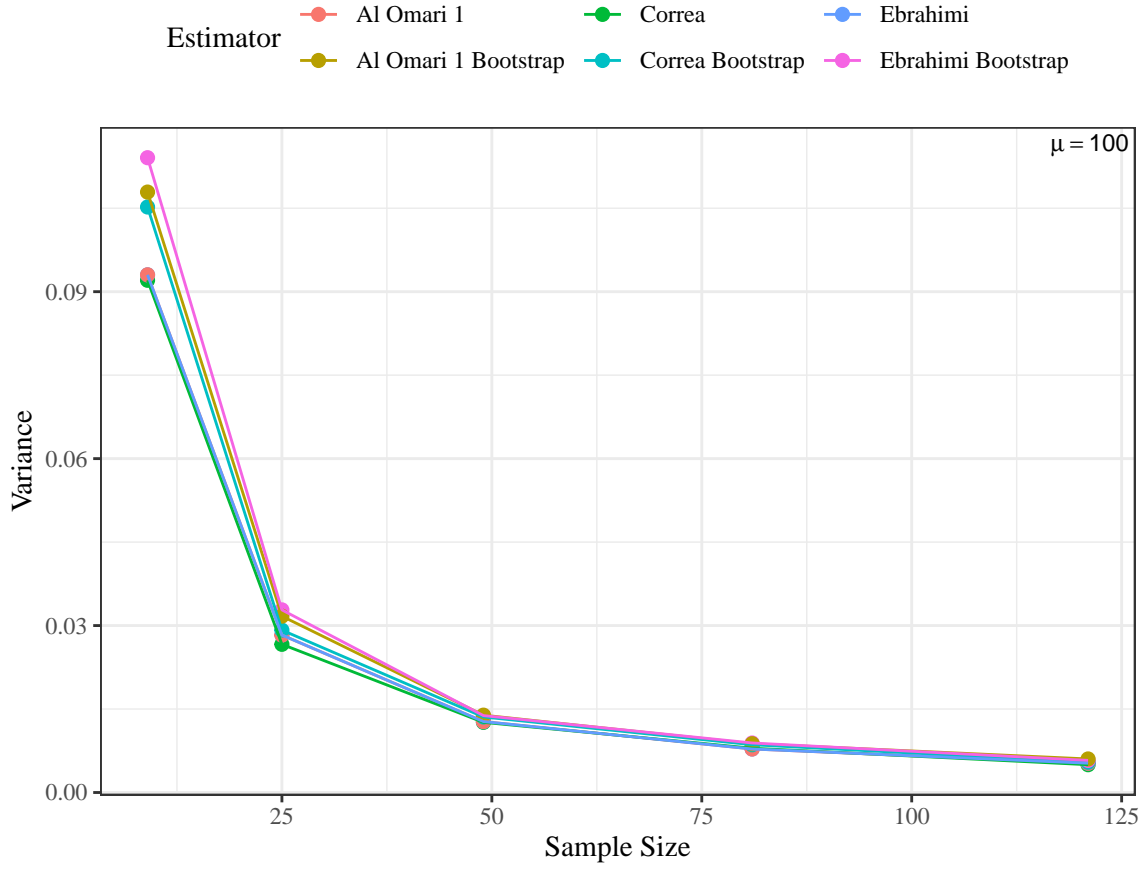
Figure 14: Variance for $L = 5$.

The results indicate that, for small sample sizes, the original estimators exhibit lower variance compared to their bootstrap counterparts. However, as the sample size increases, the variance decreases for both the original and bootstrap estimators.

Confidence Interval Analysis

To complement our variance examination, we delve into the evaluation of confidence intervals associated with the selected bootstrap-enhanced estimators. Confidence intervals provide valuable insights into the precision and reliability of the estimators' predictions.

The narrower and more stable the confidence intervals, the higher the confidence in the accuracy of the estimators. Our analysis reveals that the Correa Bootstrap, Ebrahimi Bootstrap, and Al Omari 1 Bootstrap estimators consistently generate tighter and more reliable confidence intervals.
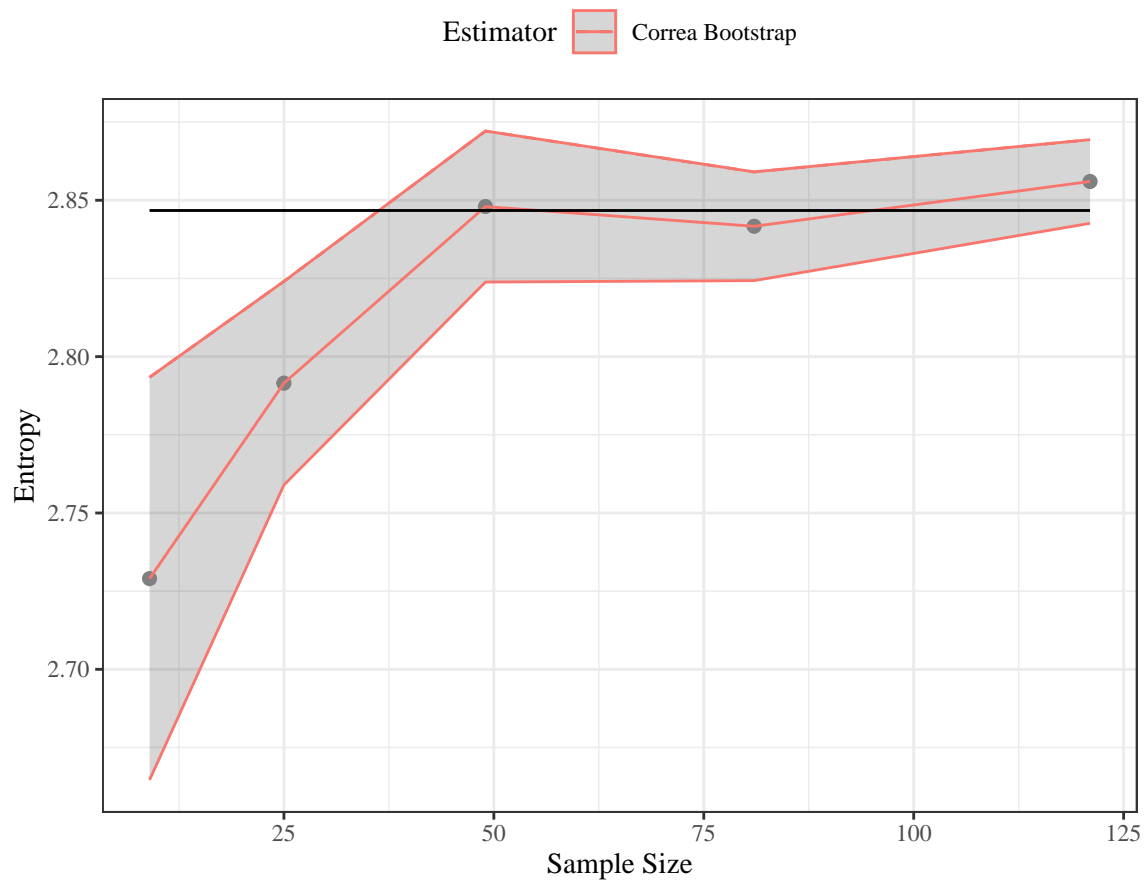
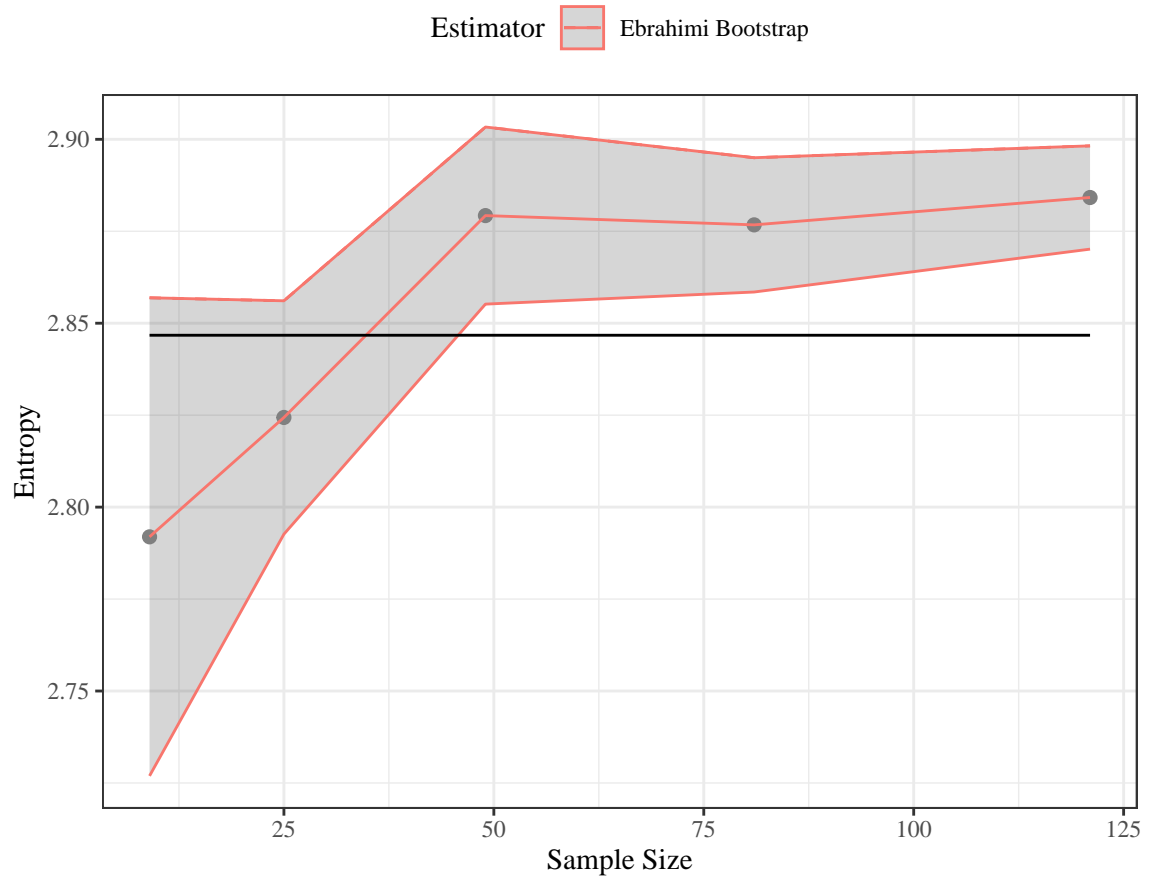Figure 15: Confidence Intervals for $L = 5$ and $\mu = 10$.

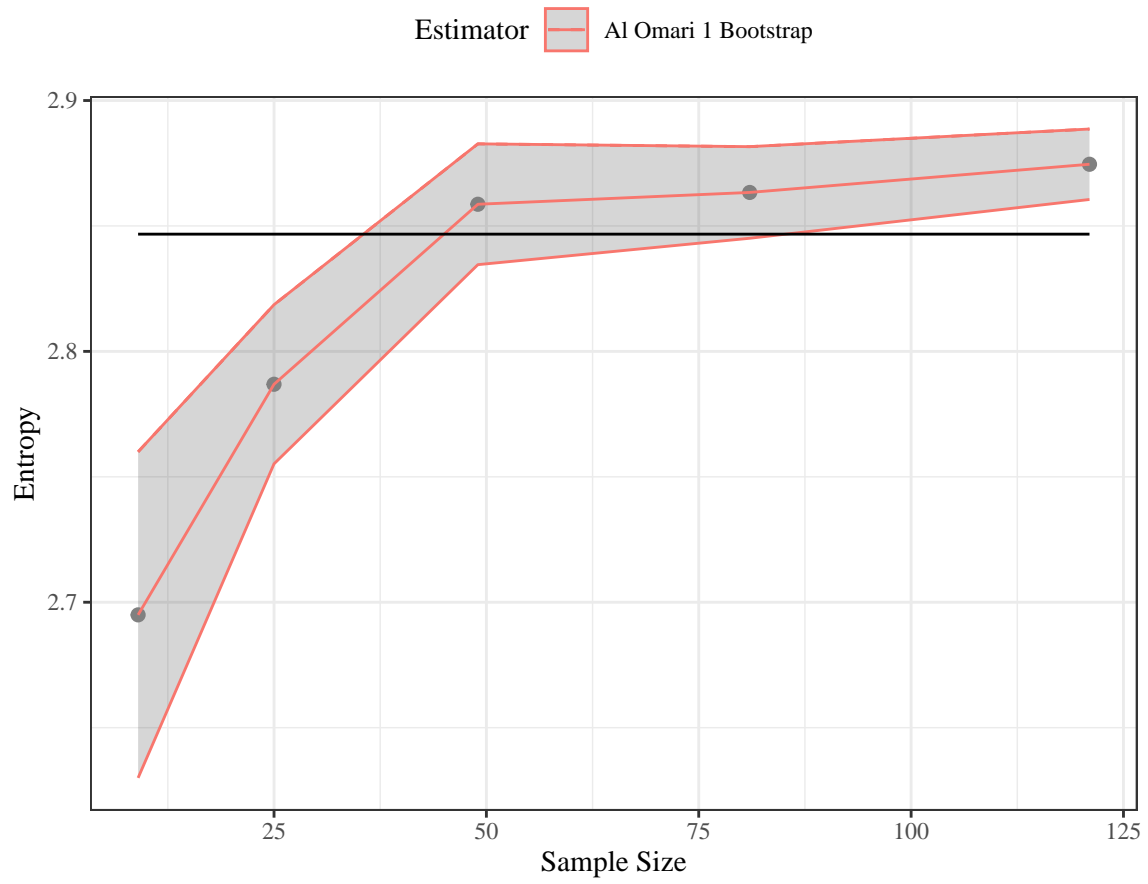Figure 16: Confidence Intervals for $L = 5$ and $\mu = 10$.

Figure 17: Confidence Intervals for $L = 5$ and $\mu = 10$.
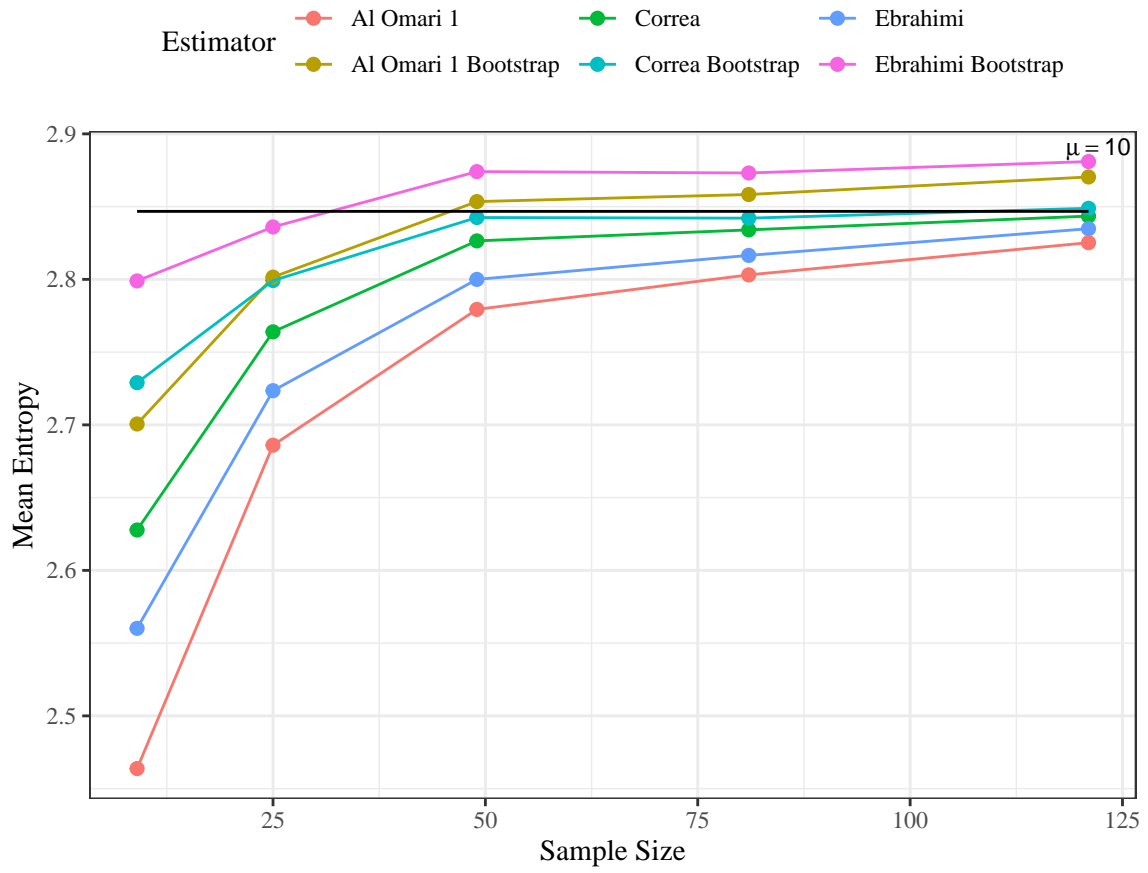
Mean Entropy for Estimators

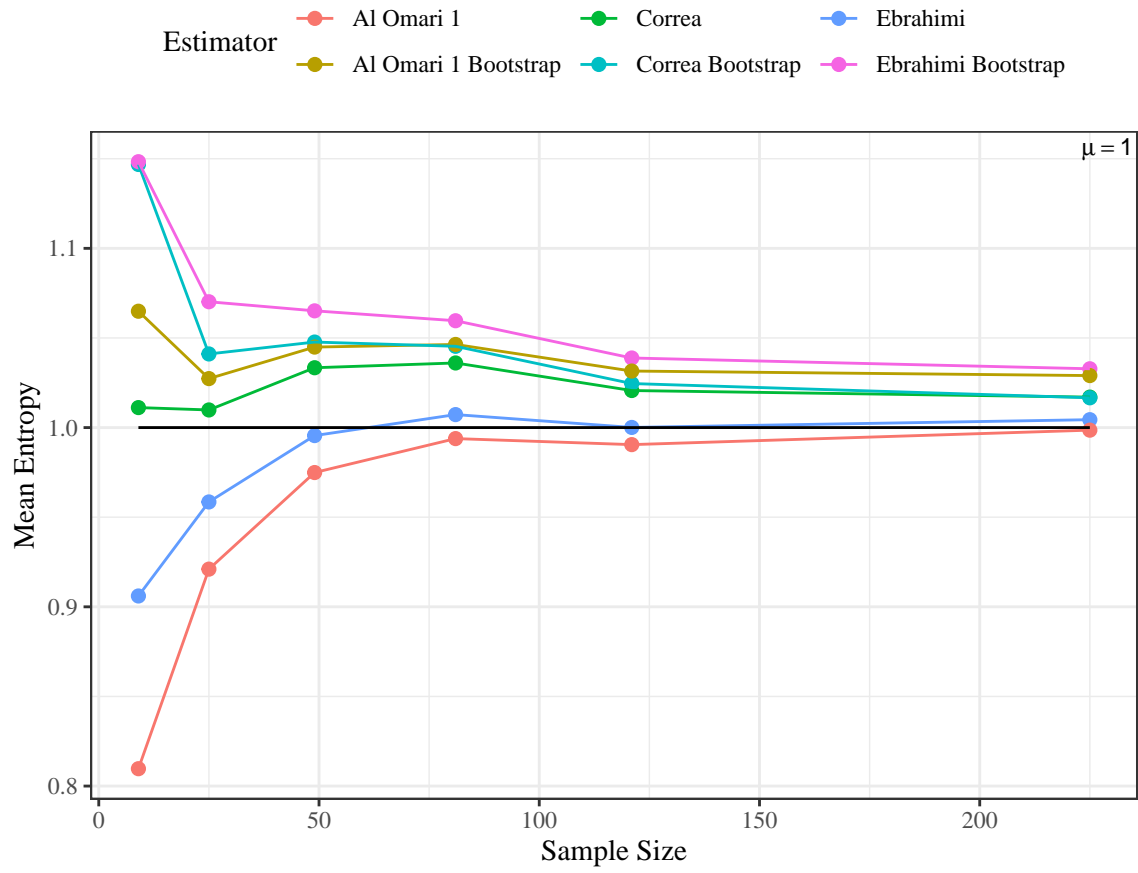Figure 18: Mean entropy for $L = 5$ and $\mu = 10$.
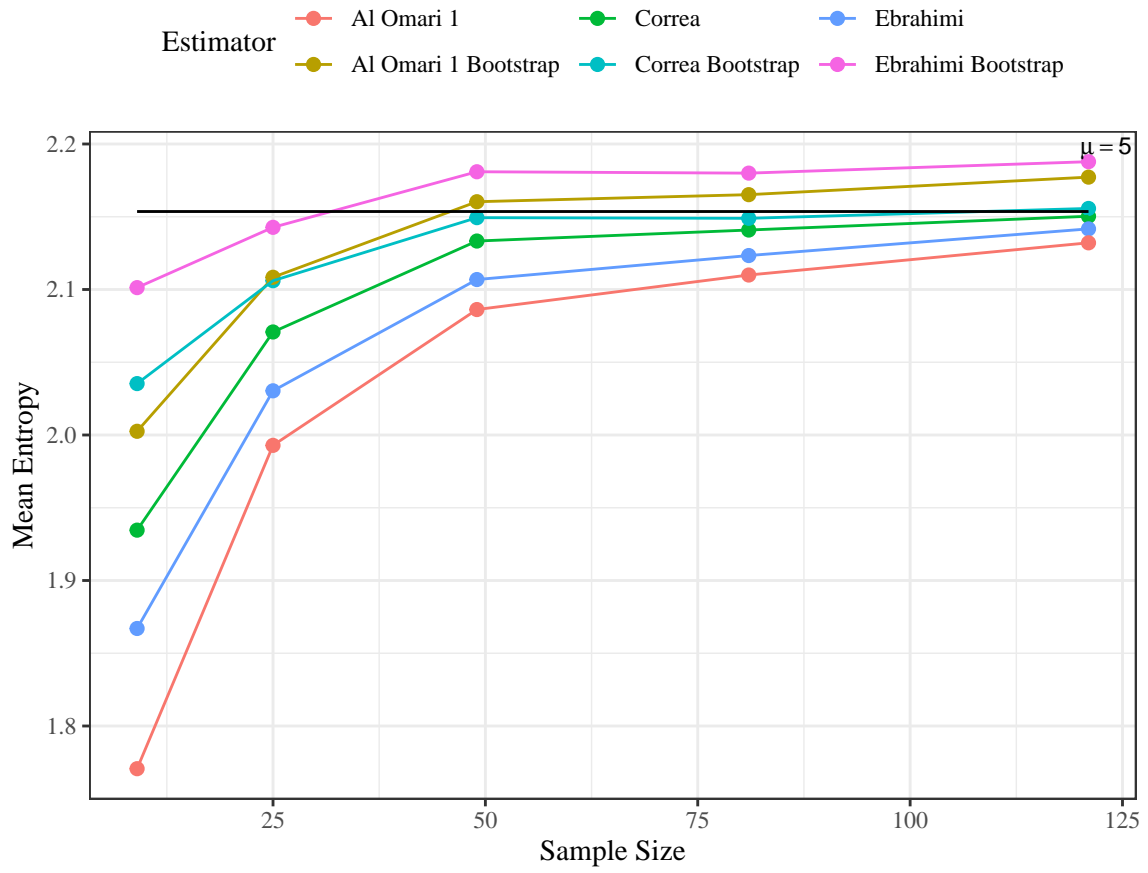
Figure 19: Mean entropy for $L = 1$ and $\mu = 1$.

Figure 20: Mean entropy for $L = 5$ and $\mu = 5$.
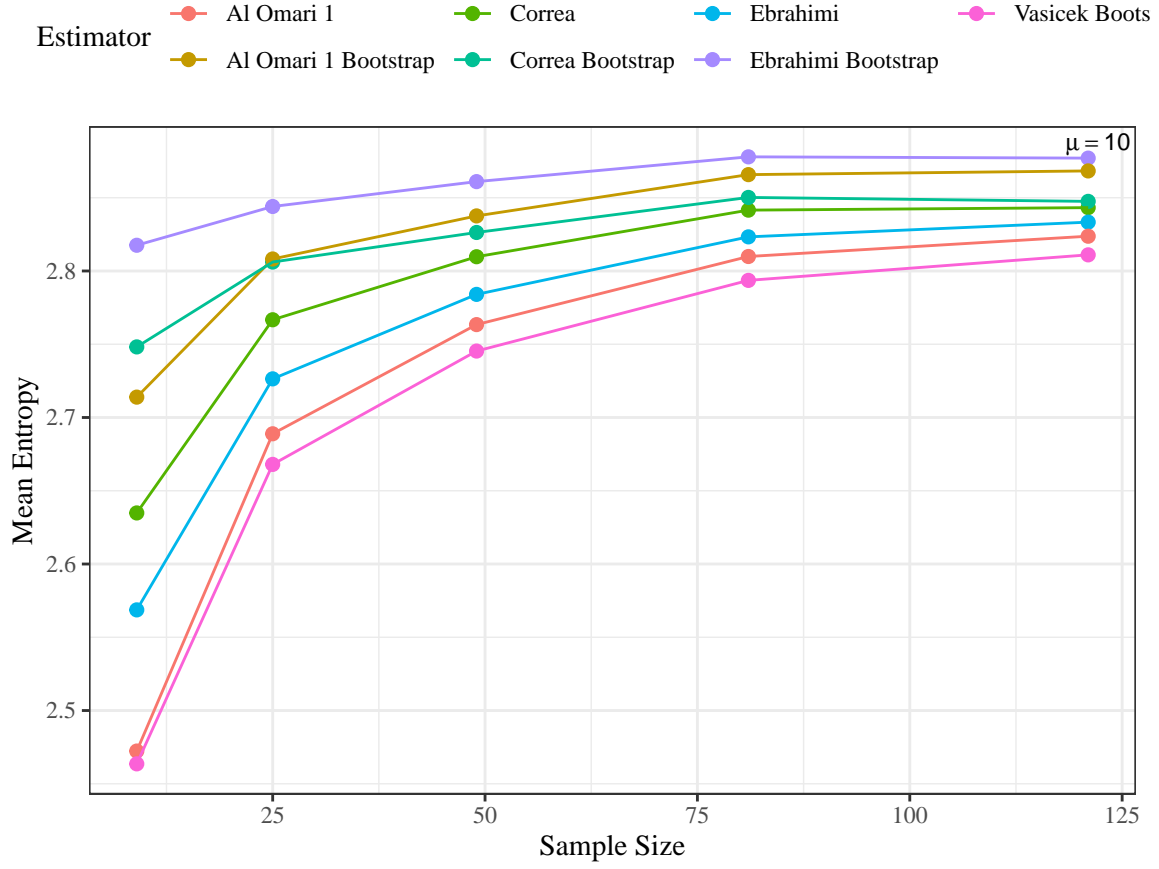
Results for Nonparametric Estimators with $\mathcal{G}_I^0$

Figure 21: Mean entropy for $L = 5$, $\alpha = -1000$ and $\mu = 10$.

In the figure we can observe that when the parameter α goes to $-\infty$, the entropy of $\mathcal{G}_I^0$ is close to the entropy of Γ_{SAR}.