Chapter 1

Monte-Carlo Simulations

In this chapter I will explore simulations of the bias of estimator (**??**) in comparison to the size of the sample estimated from, with respect to different values of k; by exploring 1-dimensional distributions and then progressing onto 2-dimensional. Firstly, the distributions considered will be analysed to determine if they satisfy the conditions **??**, **??** and **??** stated for Theorems **??** and **??** to hold. Then, I will explore the estimator of entropy for simulations of samples from certain distributions, for different values of k.

The motivation for these simulations is to explore the consistency of this estimator for different values of k; the relationship between the size of the bias of the estimator $\hat{H}_{N,k}$, $Bias(\hat{H}_{N,k})$, and the sample size, N. Throughout this analysis we will be considering the absolute value of this bias, since when taking its logarithm, we need a positive value. Using Theorem **??**, we can write that the bias of the estimator approaches 0 as $N \to \infty$. This is because we can write $Bias(\hat{H}_{N,k}) = \mathbb{E}(\hat{H}_{N,k}) - H$, which in equation (**??**) implies $Bias(\hat{H}_{N,k}) \to 0$ as $N \to \infty$. Thus, there must be a type of inverse relationship between the modulus of the bias of the estimator, $|Bias(\hat{H}_{N,k})|$, and N. We believe this relationship is of the form;

$$|Bias(\hat{H}_{N,k})| = \frac{c}{N^a} \tag{1.1}$$

for $a, c > 0$ [**?**, **?**]. By taking the logarithm of this, we can generate a linear relationship, which is easier to analyse, and is given by;

$$\begin{aligned} log|Bias(\hat{H}_{N,k})| &\approx log(c) - a[log(N)] + \epsilon \\ &\approx \zeta - a[log(N)] \end{aligned} \tag{1.2}$$

where $\epsilon > 0$ is some small error term. I will investigate the consistency of this estimator for a sample from a specified distribution, dependent on the value of k, this mean finding the optimum value of k for which $|Bias(\hat{H}_{N,k})| \to 0$ for $N \to \infty$. For the relationship in equation (1.1), this will happen for larger values of a and relatively small c, as $N \to \infty$. As previously mentioned, there is evidence supporting that the bias becomes either of order $(\frac{1}{N})^a$ (equation (**??**)) or $(\frac{k}{N})^a$ (equation (**??**)). This leads to also examining the dependence of c / ζ on the value of k.

As I wish to consider the difference in accuracy of the estimator when using different values of k, let us denote the approximate values for a and c dependent on k as a_k and c_k.

I will conduct a range of analysis, for each distribution, to consider how this estimator acts in reality, the process of analysis will be as follows;

1. Create a summary table of the mean absolute value of the bias of the estimator for $N = 100, 25000$ and 50000 for all values of k that satisfy Condition **??**. I could also consider the variance of the bias at the values of N stated above, for all applicable values of k. However, we will find that the $Var|Bias(\hat{H}_{50000,k})| \to 0$ for $k \to 10$, by the definition of the estimator using the nearest neighbour method. Taking a larger k in the nearest neighbour method will produce less varied results, this is because more smoothing takes place for a larger k, eventually - if k is made large enough - the output will be constant and the variance negligible regardless of the inputted values. Thus, considering the variance of the bias of the estimator in comparison to k is not necessarily informative.

2. Graphical representations of the linear relationship shown in equation 1.2, of $log(N)$ against $log|Bias(\hat{H}_{N,k})|$ for sample sizes $N = 100, 200, 300, ..., 50000$ (which are taken 500 times and averaged), for each value of k.

3. Tabulate the results from the regression analysis; I will first discuss the coefficient of determination (R^2), this is a measure of how well the regression model describes the observed data [**?**]. Next I will consider the standard error/deviation of the model (σ^2), this is a measure of accuracy of predictions. Lastly, I will go onto consider the values of a_k and c_k from relationship shown in equation 1.1, for each k, which is the regression line that minimizes the sum of squared deviations (σ^2) of prediction.

4. Graphically compare the values of a_k and c_k for each k.

1.1 1-dimensional Gaussian/Normal Distribution

I will begin by exploring entropy of samples from the normal distribution $N(0, \sigma^2)$, where without loss of generality we can use the mean $\mu = 0$ and change the variance σ^2 as needed. The normal distribution has an exact formula to work out the entropy, given the variance σ^2. Using equation (**??**) and the density function for the normal distribution $f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$.

We can write the exact entropy for the normal distribution, using equation (**??**);

$$
\begin{aligned}
H &= -\int_{x:f(x)>0} f(x)log(f(x))dx \\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) log\left[\frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right] dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) \left(log(\sqrt{(2\pi)}\sigma) + \frac{x^2}{2\sigma^2}\right) \\
&= \frac{log(\sqrt{(2\pi)}\sigma)}{\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx + \frac{1}{2\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \frac{x^2}{2\sigma^2} \exp\left(\frac{-x^2}{\sigma^2}\right) dx \\
&= log(\sqrt{(2\pi)}\sigma) + \frac{1}{2}
\end{aligned}
$$

Thus the exact entropy for the normal distribution is given by

$$H = log(\sqrt{(2\pi e)}\sigma) \tag{1.3}$$

I will first explore samples from 1-dimensional standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, $N(0, 1)$, to consider the behavior of the Kozachenko-Leonenko estimator. The exact entropy of this distribution is given by equation (1.3), with $\sigma^2 = 1$;

$$H = log(\sqrt{(2\pi e)}) \approx 1.418939 \tag{1.4}$$

Since, I am first considering the 1-dimensional normal distribution, the estimator takes the form in equation (**??**), which is given by;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right]$$

1.1.1 Estimator Conditions

The density of the normal distribution satisfies Conditions **??**, **??** and **??**, due to the below analysis. Firstly, to satisfy Condition **??**, for density function $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$ and $\sigma^2 = 1$, it must be such that;

- f is bounded - obvious, since for any probability distribution we always have $f(x) \geq 0$, additionally for the normal distribution we have that $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right) < 0.4$, $\forall x \in \mathbb{R}$. Hence, f is bounded above and below; so bounded.
- f is m-times differentiable - using Hermite polynomials, defined as;

$$H_m(x) = (-1)^m e^{\frac{x^2}{2}} \frac{d^m}{dx^m}\left(e^{\frac{-x^2}{2}}\right)$$

multiplying this by the coefficient in the distribution of $f(x)$, $\frac{1}{\sqrt{(2\pi}}$, we then get;

$$\begin{aligned}\frac{d^m}{dx^m} f(x) &= \frac{H_m(x)}{(-1)^m} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \\ &= \frac{H_m(x)}{(-1)^m} f(x)\end{aligned}$$

where $\frac{H_m(x)}{(-1)^m}$ is a polynomial; thus f is m-times differentiable.

- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y-x\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y-x\|^\eta$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$.

  Since we are considering a 1-dimensional distribution, we can write the norms $\|\cdot\|$ as $|\cdot|$. Moreover, considering that for Theorems **??** and **??**, we have the value of $\beta \geq 2$; thus choosing $\beta = 2$, and since $m = \lfloor\beta\rfloor = \lfloor 2 \rfloor = 2 = \beta$ and $\eta = \beta - m$, we have that $\eta = 0$. Thus we need $|f^{(t)}(x)| \leq g_*(x)f(x)$, which is obvious by above, in view of writing $|\frac{d^t}{dx^t} f(x)| = g_*(x)f(x)$, where we choose $g_*(x) = |\frac{H_t(x)}{(-1)^t}| = |H_t(x)|$, for $t = 1, 2, ..., m$, and $|f(x)| = f(x)$, since $f(x) > 0$. Also, g_* is a polynomial and is hence Borel measurable over $\mathbb{R}$, and for any polynomial we obviously have $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$. Additionally, we need $|f^{(m)}(x) - f^{(m)}(x)| \leq g_*(x)f(x)|y-x|^0 = g_*(x)f(x)$. We currently have;

$$\begin{aligned}|f^{(m)}(x) - f^{(m)}(x)| &= \left|\frac{H_m(x)}{(-1)^m} f(x) - \frac{H_m(y)}{(-1)^m} f(y)\right| \\ &\leq \left|\frac{H_m(x)}{(-1)^m} f(x)\right| + \left|\frac{H_m(y)}{(-1)^m} f(y)\right| \\ &= g_*(x)f(x) + g_*(y)f(y) \\ &\leq g_*(x)f(x)\end{aligned}$$

  since we know that $f(x) > 0$ for all $x \in \mathbb{R}$, and $g_*(x) = |H_m(x)| > 0$, which is similar to the g_* before; thus satisfies the conditions for it.

Next, to satisfy Condition **??**, for the density function f of the normal distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$ - this is always true for the normal distribution, all of its moments are finite, since they are defined with respect to σ^n, for some n, and $\sigma < \infty$.

Lastly, to satisfy Condition **??**, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. To do this we must have, for some $\alpha > d = 1$, let k_0^* and k_1^* denote two deterministic sequences of positive integers with $k_0^* \leq k_1^*$. Taking $\alpha := 2$, we must have;

- $k_1^* = O(N^\tau)$, where $\tau < \min\left\{\frac{2\alpha}{5\alpha+3d}, \frac{\alpha-d}{2\alpha}, \frac{4}{4+3d}\right\} = \min\left\{\frac{4}{13}, \frac{1}{4}, \frac{4}{7}\right\} = \frac{1}{4}$, so we can choose $\tau := \frac{2}{9} < \frac{1}{4}$ so that we have $k_1^* = O(N^{\frac{2}{9}})$

- $\frac{k_0^*}{\log^5 N} \to \infty$ - for this to be true we need to choose $k_0^* := N^A$ for some $A > 0$. Considering that $k_0^* \leq k_1^*$ and $k_1^* = O(N^{\frac{2}{9}})$, thus $A \in (0, \frac{2}{9})$. So we can choose $A := \frac{1}{\eta}$ for some large η, which gives that $k_0^* = O(N^{\frac{1}{\eta}}) \approx 1$.

Thus, on account of the values of N being considered in the simulations; $N = 100, 200, ..., 50000$, we have that for the smallest $N = 100$, the values of k for which Theorem **??** and **??** both hold, are $k \in \{k_0^*, ..., k_1^*\} = \{1, ..., 100^{\frac{2}{9}}\} = \{1, ..., 2.782\} \approx \{1, 2\}$. Also, for the middle value $N = 25,000$, we have the values of k to be in $\{k_0^*, ..., k_1^*\}$, where $k_1^* \approx 25000^{\frac{2}{9}} = 9.491 \approx 9$, thus $k \in \{1, ..., 9\}$. Moreover, for the largest $N = 50,000$, we must consider $k \in \{1, ..., k_1^*\} = \{1, ..., 50000^{\frac{2}{9}}\} = \{1, ..., 11.072\} \approx \{1, 2, ..., 11\}$.

Overall, due to Conditions **??**, **??** and **??** being met, we can say that for the normal distribution, Theorems **??** and **??** hold; henceforth, we can say that the Kozachenko-Leonenko estimator, of a sample from the 1-dimensional normal distribution is an asymptotically unbiased and consistent estimator for entropy, for some values of $k \in \{1, 2, ..., 11\}$, depending on the sample size N.

1.1.2 Simulation Results

I will now conduct some simulations to consider this for each value of k separately, each time considering 500 samples of size N from this distribution, finding the estimator in each case and take the average of these estimators to find our entropy estimator. I will then consider the relationship show in equation (1.2) for each sample and work out the average for the values of a and c, for each $k \in \{1, 2, ..., 11\}$.

For $N = 100$, $N = 25,000$ and $N = 50,000$, using the results from **??**, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered.

The results shown in table 1.1 show that for a larger N, the modulus of the bias of the estimator is smaller, this is true for all values of k except when $k = 2, 3, 7, 8$, for which the bias is smaller when $N = 25,000$ in comparison to the larger value of N. There are a number of reasons why this could be; however, it is first important to notice that when finding the values of k that satisfy condition**??**, we found that for $N = 100$, we must have $k \in \{1, 2\}$, for $N = 25,000$ we have $k \in \{1, 2, ..., 9\}$ and for $N = 50,000$ we have $k \in \{1, 2, ..., 11\}$.

For the smallest values of $N = 100$, we expect the best value of k to be either 1 or 2; and the table agrees with this showing that the smallest bias occurs at $k = 1$ for a small sample size.

When $N = 25,000$ we have that for $k \in \{2, ..., 8\}$ that the bias is very small, especially for the values of $k = 3, 4, 7, 8$ with the smallest bias appearing when $k = 3$; which fits with the previous analysis that the best value of k will lie within 1 and 9.

Table 1.1: *1-dimensional normal distribution, comparison of* k

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0031912	0.0006312	0.0004428
2	0.0195347	0.0000092	0.0003632
3	0.0167902	0.0000056	0.0002278
4	0.0264708	0.0001657	0.0001196
5	0.0238265	0.0002138	0.0000003
6	0.0311576	0.0001546	0.0001471
7	0.0356302	0.0000217	0.0003024
8	0.0396299	0.0000984	0.0001021
9	0.0460706	0.0003620	0.0002070
10	0.0458648	0.0002752	0.0002611
11	0.0387339	0.0003332	0.0002458

This table is comparing the values of $|Bias(\hat{H}_{N,k})|$ *for the values of* k *with* $N = 100$*,* $N = 25,000$ *and* $N = 50,000$*, when the estimator is taken over* 500 *samples*

Now considering the largest sample size $N = 50,000$, the bias when $k = 5$ sticks out since it is $\approx 10^{-3}$ smaller than the other bias values in the table. However, for all other values of k the bias is still extremely small in comparison to the bias for $N = 100$ and even in comparison to $N = 25,000$ in some places. This extreme difference could be an outlier in my data; thus in table 1.2 I have shown the values for the modulus of the bias, when $k = 5$, for different, also large values of N. This table does indeed show that $|Bias(\hat{H}_{50000,5})| \approx 0.0000003$ is an anomaly in the data, and that $k = 5$ is not necessarily the best value of k for $N = 50,000$. Thus, we cannot yet draw any major conclusions about the best value of k for the estimator of a sample this size.

I now wish to consider the equation 1.2 and plot the simulated data, to fit a regression line for each value of k separately, these are shown in Figures 1.1 and 1.2. All of these graphs agree with the relationship previously stated between the sample size and the bias of the estimator; they all show that the logarithm of this equations gives a negative linear relationship - with relatively small error bars.

Moreover, I would like to consider the coefficient of determination (R^2) for each of the above regression lines, this value provides an estimate of the strength of the relationship between the model and the response variable. Also, I would like to consider the standard error/deviation (σ^2), for each of the different graphs, which shows a measure of the predictions' accuracy. These are all depicted for each value of k in table 1.3.

Both columns of this table essentially point to the same conclusion; the

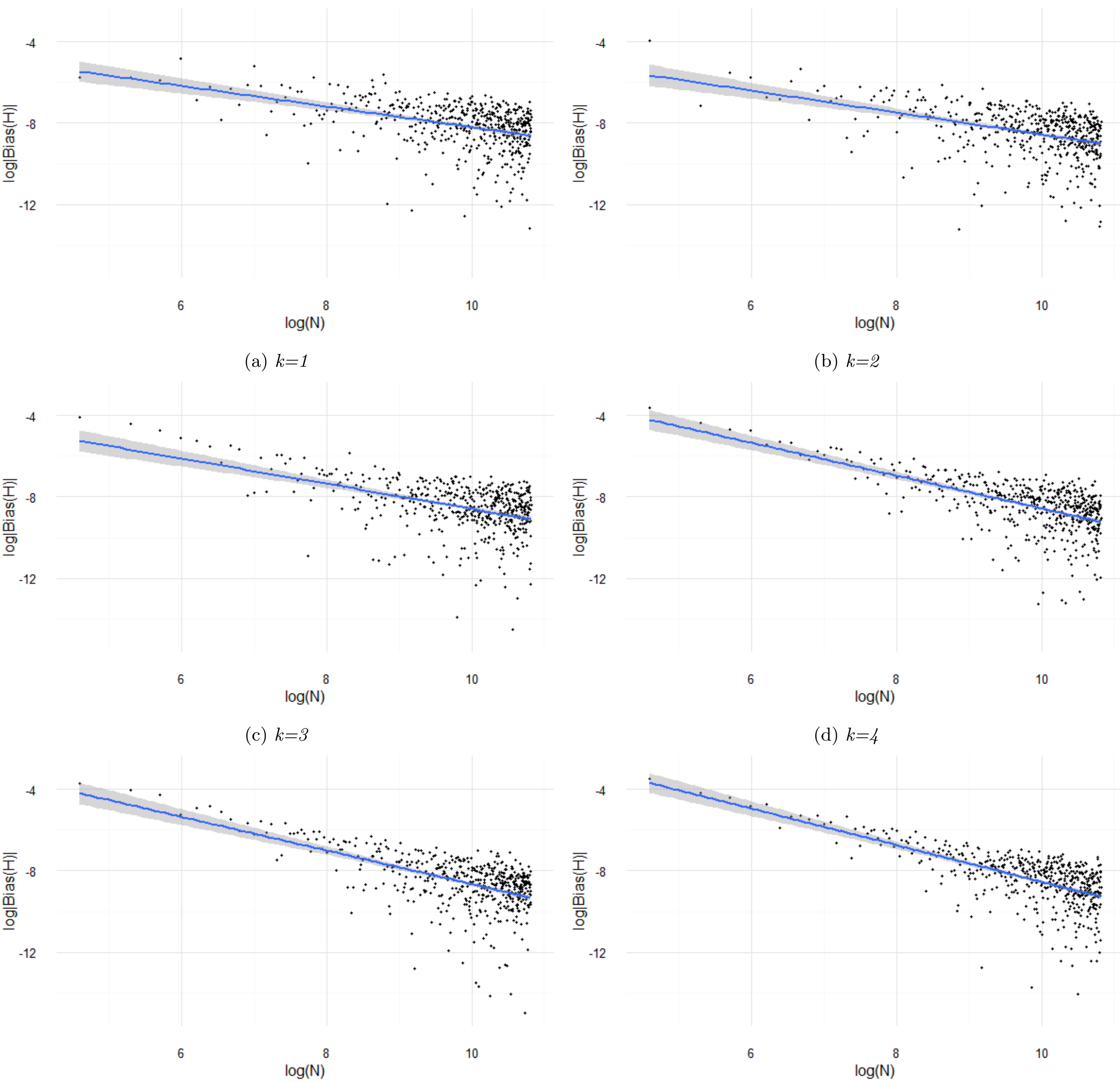
Table 1.2: *1-dimensional normal distribution, $k = 5$ for large N*

N	$\lvert Bias(\hat{H}_{N,5}) \rvert$
49100	0.0000639
49200	0.0001463
49300	0.0001700
49400	0.0001037
49500	0.0000711
49600	0.0003221
49700	0.0001047
49800	0.0000644
49900	0.0001240
50000	0.0000003

This table is comparing the values of $Var\lvert Bias(\hat{H}_{N,5})\rvert$ for the large values of N.

Table 1.3: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional normal distribution*

k	R^2	σ^2
1	0.1766	1.0661
2	0.1793	1.1477
3	0.2292	1.1053
4	0.3556	1.0759
5	0.3322	1.1752
6	0.4260	1.0180
7	0.4532	1.0155
8	0.4623	1.0088
9	0.4962	0.9730
10	0.5227	0.9759
11	0.5839	0.8566

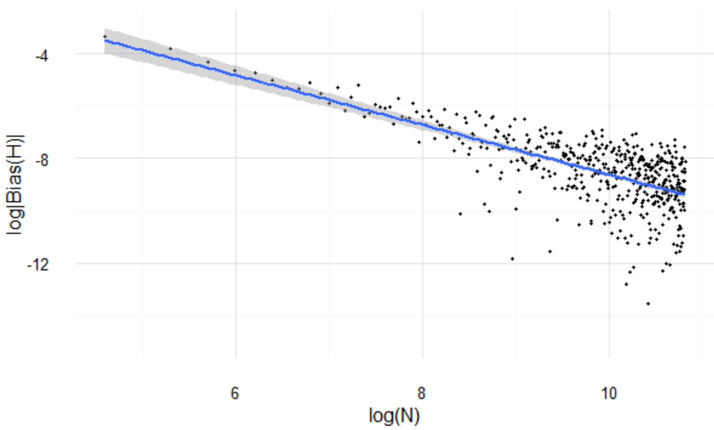


(a) *k=1* (b) *k=2*

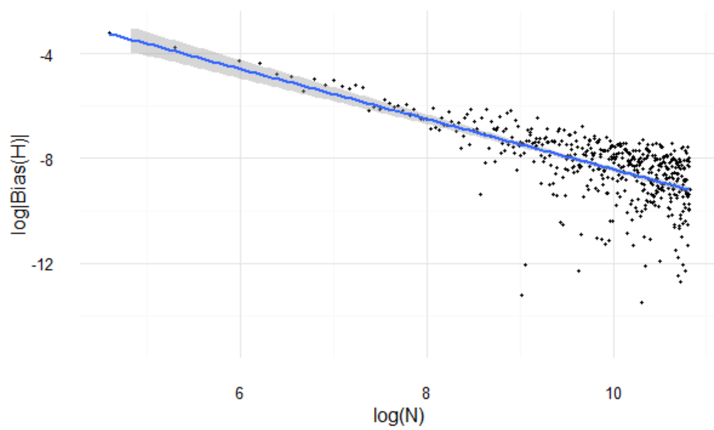
(c) *k=3* (d) *k=4*

(e) *k=5* (f) *k=6*

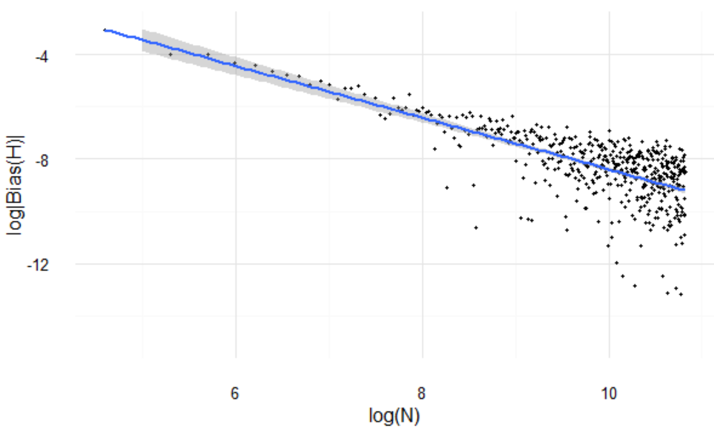
Figure 1.1: *1-dimensional normal distribution with different* $k = 1, ..., 6$



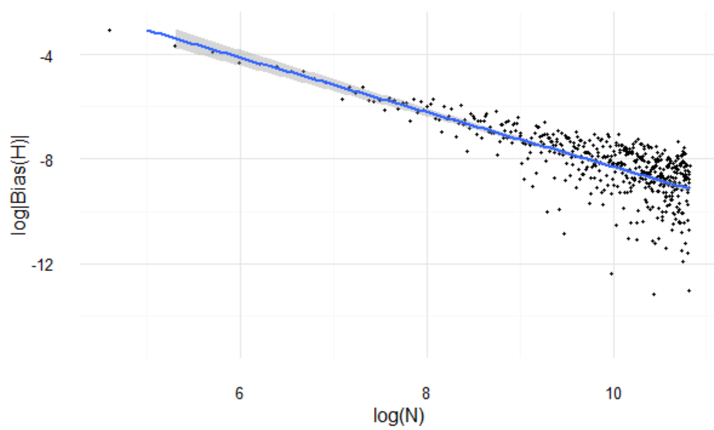
(a) *k=7*



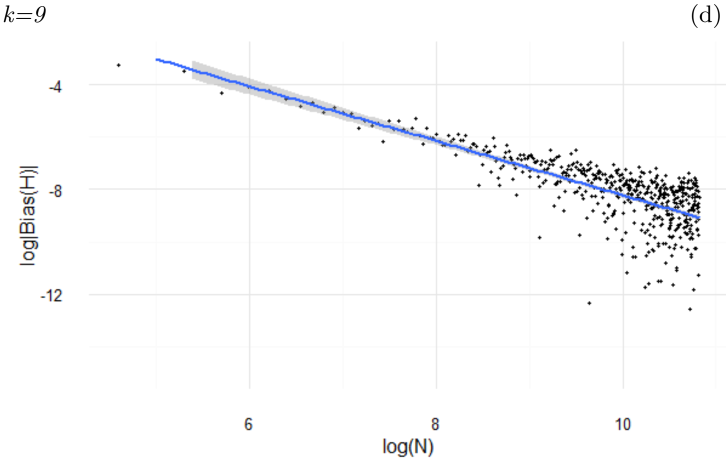
(b) *k=8*



(c) *k=9*



(d) *k=10*



(e) *k=11*

Figure 1.2: *1-dimensional normal distribution with different* $k = 7, ..., 11$

Table 1.4: *Comparison of coefficients of regression a_k and c_k from equation 1.1, for 1-dimensional normal distribution*

k	a_k	c_k
1	0.5054	0.0433
2	0.5490	0.0459
3	0.6169	0.0894
4	0.8181	0.6690
5	0.8486	0.8235
6	0.8976	1.5514
7	0.9464	2.3576
8	0.9574	3.2021
9	0.9883	4.4558
10	1.0454	8.5402
11	1.0386	8.7457

larger the value of k, the more accurate the linear model is to fitting the data. This is shown by the R^2 value increasing towards 1 and the σ^2 values decreasing positively.

The R^2 is very small for $k \leq 3$, which points towards the line being a poor fir to the data; however, due the the standard deviation being $\sigma^2 \approx 1.1$, we cannot say that these lines are poorly fitting; since the majoring of the data is within a very small range of the line.

The most important information found from the regression analysis is shown in table 1.4; where the values of a_k and c_k are given for each value of k.

As k runs from $1 \to 11$, we have that a_k and c_k both increase, with smooth values of a_k and a large jump, in the value of c_k, between $k = 3$ and 4, and $k = 9$ and 10. The higher the value of a_k, the stronger the negative relationship is between the two variables in question, so for a larger values of a_k, we have that $|Bias(\hat{H}_{N,k})| \to 0$ for large N faster than smaller values of a_k. This is due to the relationship between $|Bias(\hat{H}_{N,k})|$ and a_k shown in equation (1.1)

Recall, from section **??** we have that the bias acts in one of two ways (equations **??** and **??**); it is either of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. Thus we have $|Bias(\hat{H}_{N,k})| \approx \frac{c_k}{N^{a_k}}$ where either c_k is constant or it depends on k and a_k - more specifically is $O(k^{a_k})$. There is evidence here to support the latter claim. If we consider the jump between $k = 3$ and 4 shown in the value of c_k, and consider the results in table 1.5.

This shows that the proportional behaviour between k^{a_k} and c_k also has a large jump when k goes from $3 \to 4$. This agrees with the claim of c_k depending on k in this fashion; however, in table **??** we mentioned another jump between $k = 9$ and $k = 10$, and the evidence here does not show a large jump in the same area. We cannot yet make any conclusions about the dependence of c_k on

Table 1.5: *Considering the dependence of k on c_k*

k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
1	1	0.0433	23.095
2	1.4631	0.0459	31.875
3	1.9694	0.0894	22.029
4	3.1085	0.6690	4.646
5	3.9187	0.8235	4.759
6	4.9942	1.5514	3.219
7	6.3067	2.3576	2.675
8	7.3218	3.2021	2.287
9	8.7716	4.4558	1.969
10	11.1020	8.5402	1.300
11	12.0668	8.7457	1.380

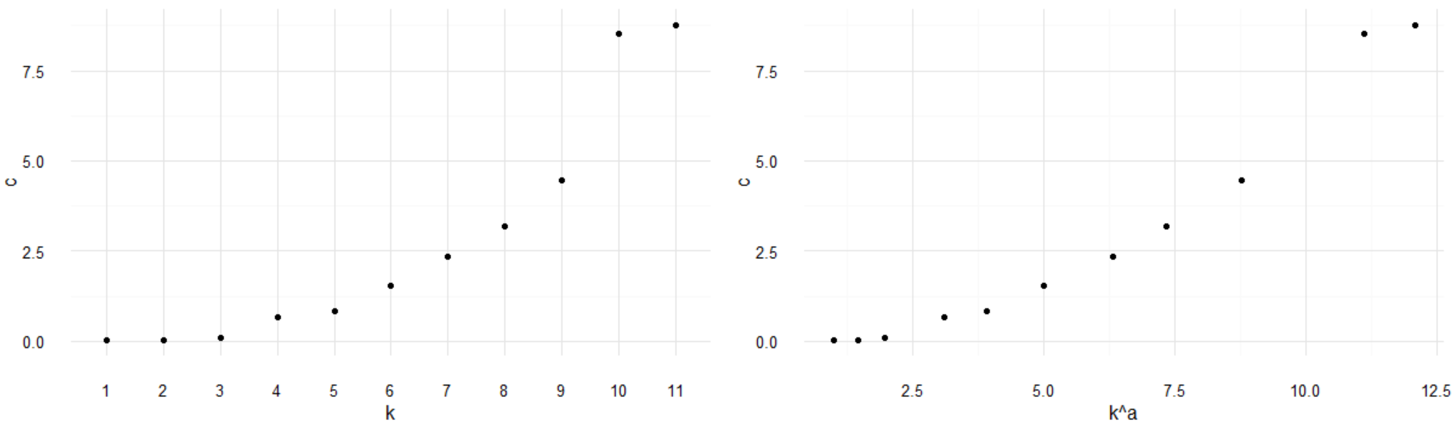
k; this motivates a graphical representation of the value of c_k against k to see if there is any relation, Figure 1.3.

Interestingly, plot 1.3 (a) shows an almost exponential relationship between the values of c_k and the values of k. This leads me to believe that there is some kind of relationship between the two variables, and looking at plot 1.3(b) this shows that there's a strong possibility that the relationship is of the form stated in equation **??**.

To better study the linear relationship between the logarithm of the bias and the logarithm of the sample size, I have generated a comparison plot, shown in Figure 1.4.

From this we can see obviously that for smaller values of N, smaller values of $log(N)$, the smallest bias occurs when $k = 2$, since this line is the lowest for the data up until $log(N) \approx 9$ - i.e. $N \approx 13,000$. For a larger sample size, we cannot accurately see in this graph which line is the best. This motivates us to look at a section of the graph when $9 \leq log(N) \leq 11$ - i.e. $8,000 \leq N \leq 50,000$, which is shown in Figure 1.5.

From this graph we can obviously discount $k = 1$ for large N, since this is the most gradual descent; thus the bias will be largest for this k. Also, both the lines for $k = 2$ and $k = 3$ are more gradual in their descent at larger N, so are probably not the best to choose. Even though, for $k = 9, 10$ and 11, the slope is the steepest - a_k is largest - the intercept is larger so around the biggest sample size considered $N = 50,000$, $log(N) \approx 10.8$, there is not the smallest bias. Actually, for large values of $N \leq 50,000$ we can see from this graph that the best lines appear to be those which are blue/green; $k = 4, 5, 6, 7, 8$. Where the lowest lines around the maximal sample size are those for $k = 5$ and $k = 7$; thus these values of k could possible be the best nearest neighbour value to choose, when looking at a sample of size $N \approx 50,000$ from the normal distribution.



(a) *The values of k against the values of c_k*

(b) *The values of k^{a_k} against the corresponding values of c_k*

Figure 1.3: *Graphically representing the relationship between c_k and k*

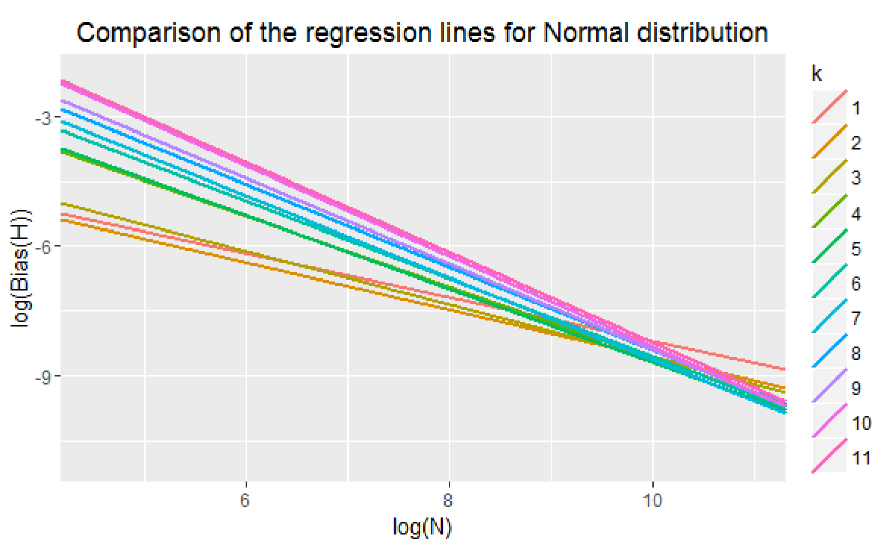


Figure 1.4: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$, *for* $k = 1, 2, ..., 11$, *for samples from the normal distribution*

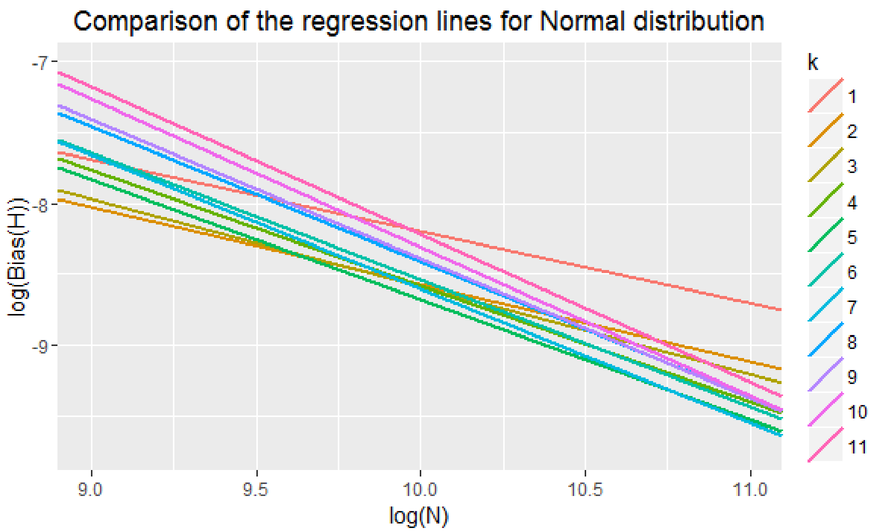


Figure 1.5: *Figure 1.4 zoomed in around large N*

1.2 1-dimensional Uniform Distribution