A Nearest-Neighbour Approach to Estimation of Entropies

Author:
Thomas Maierhofer

Supervisor:
Prof. Nikolai Leonenko

MMATH project
Academic year: 2014/2015

October 19, 2015

Contents

Abstract

The concept of Shannon entropy as a measure of disorder is introduced and the generalisations of the Rényi and Tsallis entropy are motivated and defined. A number of different estimators for Shannon, Rényi and Tsallis entropy are defined in the theoretical part and compared by simulation in the practical part. In this work the nearest neighbour estimator presented in Leonenko and Pronzato (2010) is compared to spacing based estimators presented in Beirlant et al. (1997) and Song (2000) for the Shannon entropy of one-dimensional distributions. For another special case of entropy, the quadratic entropy, the estimator given in Källberg et al. (2014) is compared with the nearest neighbour estimator for multidimensional densities. Comparisons focus on bias and variance for a given sample size and are executed with simulation studies. Based on the simulations, suggestions for which estimator to use under given conditions are derived. Depending on the conditions different estimators perform better than others; one estimator was not found to be universally superior.

1 Introduction

The concept of entropy is one of the most basic and important in natural sciences and information theory. It all started with Claude E. Shannon's fundamental paper, "A Mathematical Theory of Communication", on information theory, see Shannon (1948), where the concepts of entropy and information are mathematically defined. It remains a commonly used measure of uncertainty and has found its way into many other more or less related areas of science. Examples from very different areas of application include the Boltzmann entropy in thermodynamics, Shannon entropy in coding theory, entropy as a measure of disorder in brain activity as measured by neuroimaging, see Carhart-Harris et al. (2014), entropy as a measure of heart rate variability in order to detect cardiac abnormalities, see Cornforth et al. (2014), entropy as a descriptive measure in linguistics, see Borgwaldt et al. (2005), and entropy based tests for normality, see Vasicek (1976).

This work introduces a reader, familiar with the fundamentals of mathematical statistics, to the concept of entropy and its estimation. Many sources are compactly summarised in order to give a brief but comprehensive overview of the topic. The goal of this work is to compare the asymptotic properties of different estimators for entropy. This will be done using simulations, giving new insights into how conditions affect the performance of the nearest neighbour estimator compared to other estimators. Additionally, code for the software package R, R Core Team (2013b), is in the appendix,

that can be used easily for future research and simulating more parameter and distribution combinations of interest to the reader.

In the literature, consistency and asymptotic unbiasedness were shown for a number of estimators including the nearest neighbour estimator, which are important theoretical results. (Källberg et al.; 2014; Leonenko and Pronzato; 2010; Wang et al.; 2006) Results for the rate of convergence of the bias of the estimators are hard to gain by calculus. There are some results for the nearest neighbour estimator without clear preconditions under which they are valid and no results for the other estimators. In this thesis the convergence of the bias is investigated and checked by means of simulation. For the variance of the estimators there are only asymptotic results in literature that are not valid for small sample sizes. The simulations conducted in this work enlighten the relationship of the sample size and the variance of the estimator for a number of important distributions. The estimators are compared in order to see under which conditions which estimator performs best.

The reminder of the thesis is as follows; In Section 2 a brief introduction to information theory is given and the concepts of information and entropy are established. This work focuses exclusively on the Shannon, Rényi and Tsallis entropy which are the most important in application. The one- and multidimensional entropies are defined for the discrete (Section 2.4) and continuous (Section 2.5) cases. The discrete Boltzmann entropy is introduced due to its importance in natural sciences and its close relation to the Shannon entropy. In Section 3, the maximum entropy principle is presented. The concept of

maximum entropy will be applied for the Shannon, Rényi and Tsallis entropy and maximising distributions will be given for discrete and continuous multivariate random variables with different supports and restraints. The Section 4 presents approaches to estimate entropy. It gives theoretical motivation for the estimators and brief ideas of the derivations. Concrete formulae are given on how to estimate the Shannon, Rényi and Tsallis entropy from a multidimensional sample. Section 5 provides an introduction to simulation studies in general and their use in this thesis. Linear and non-linear models are introduced as they will be used to quantify the results of the simulation. This section also summarises the results of the simulation studies, showing the strengths and weaknesses of the different theoretical estimators by means of simulation. The thesis concludes with the discussion (Section 6), giving an overall review. Appendix A gives a brief introduction to the software-package R and Appendix B shows concepts of using it for efficient programming.

2 Concepts of entropies

This section gives a brief introduction to information theory with its most basic concepts, information and entropy. In Section 2.1 the mathematical concept of information is introduced and in Section 2.2 the idea of entropy is introduced based on the previously defined concept of information. Section 2.3 motivates the use of different generalisations of the Shannon entropy and gives examples for their field of application. The Shannon, Boltzmann, Rényi and Tsallis entropies are mathematically defined in the following section for the discrete (2.4) and the continuous (2.5) case.

2.1 Information and entropy

Information theory started with Claude E. Shannon's fundamental paper "A Mathematical Theory of Communication" in July and October 1948 (Shannon; 1948). In information theory the information of a certain message is a measure of the amount of knowledge gained by knowing this specific message has been received. The idea is that the less likely a message or event is, the more information it provides when it occurs. This idea can be quantified with the concept of information, by considering a message as an outcome of a random experiment.

Let A be a set of outcomes of random experiments with probability $P(A)$.

The information of A is denoted by

$$I(A) = -\log_2 P(A) \tag{1}$$

and is expressed in bits.

Example 2.1. Let H, T denote the two possible outcomes of flipping a coin, heads and tails. For a fair coin $P(H) = P(T) = \frac{1}{2}$ the information gained by seeing a head or a tail is equal

$$I(H) = I(T) = -\log_2(\frac{1}{2}) = 1.$$

Consider the case of an unfair coin with $P(H) = \frac{1}{8}$ and $P(T) = \frac{7}{8}$. The information of the less likely event, heads, is greater than the information gained from the more likely event, tails, precisely

$$I(H) = 3 > 1.933 = I(T).$$

Properties of the information $I(A)$:

1. $\log_2(1) = 0$, which means the observation of a certain event does not contain any information.

2. $I(A)$ increases as $P(A)$ decreases $\Rightarrow$ events with lower probability have higher information, as the logarithm $\log(x)$ is strictly monotonic increasing with x.

Example 2.2. This idea is quite intuitive, when the sun rises in the morning our world view does not change at all. However, in the very unlikely event of the sun failing to rise, our model of physics would require significant changes (Johnson; 2004).

3. (Additivity) For A, B independent outcomes of random experiments

$$I(A, B) = I(A) + I(B)$$

2.2 Entropy of a source

The entropy of a source is the expected amount of information contained in each message received from the source. In other words, entropy is the average amount of information generated by this source. A source can be considered a random experiment generating a set of messages. This set of outcomes is the alphabet of the source. Entropy gives a numerical measure of how far from deterministic a random variable is. It quantifies the diversity, uncertainty or randomness of a system. An entropy of zero means that the random variable is deterministic, an increasing value means that it is more and more unpredictable (Johnson; 2004).

Let S be a source with a finite alphabet $A = \{x_1, x_2, \ldots, x_m\}$ generating an infinite sequence of random variables $X = \{X_1, X_2, \ldots, \}$. The entropy of

the source $H(S)$ is defined as

$$H(S) = \lim_{n\to\infty} \frac{1}{n} G_n \tag{2}$$

where G_n is defined as

$$-\sum_{i_1=1}^{m}, \ldots, \sum_{i_n=1}^{m} P(X_1 = x_{i_1}, \ldots, X_n = x_{i_n}) \log_2 P(X_1 = x_{i_1}, \ldots, X_n = x_{i_n}).$$

For the special case that all $X_i \in X = \{X_1, X_2, \ldots, \}$ are independent and identically distributed (i.i.d.), Equation (2) simplifies to

$$G_n = -n \sum_{i_1=1}^{m} P(X_1 = x_{i_1}) \log_2 P(X_i = x_{i_1})$$

and thus

$$H(S) = -\sum_{i_1=1}^{m} P(X_1 = x_{i_1}) \log_2 P(X_i = x_{i_1})$$

is the first order entropy of the source S, which equates to the expected value of the information of S (Johnson; 2004).

Example 2.3. It is common to speak of the entropy, or rate, of a language. The entropy of a source of information is related to its redundancy in characters. In Shamilov and Yolacan (2006) the entropy of a number of languages is given in Table 1.

It can be seen that the entropies of Russian and Turkish are higher than those of western European languages like English, German and Spanish. A

Table 1: Shannon entropy of a selection of languages

Language	Shannon entropy
Turkish	4.3299
English	4.1489
French	4.0193
German	4.0796
Spanish	4.0142
Russian	4.3452

higher entropy means that the occurrence of the letters in a text written in the given language is more random, meaning the letters occur more evenly.

2.3 Motivation for different entropies

In information theory, the Shannon entropy is used to describe the average amount of information contained in each message received. Known as the Shannon-Wiener Index, it was originally used to quantify the uncertainty of strings in a text (Shannon; 1948).

In thermodynamics, the entropy is standardised with the Boltzmann constant $k_B = R/N_A$, with gas constant R and the Avogadro constant N_A, and referred to as the Gibbs-Boltzmann entropy. (Johnson; 2004)

For a given set of macroscopic variables including temperature, volume and pressure, the entropy measures the degree to which the probability of the system is spread out over different possible microstates. The internal entropy of a molecule depends on random fluctuations in its internal coordinates. The extent of these fluctuations determines the thermodynamic functions and

shapes of the molecule. Therefore the estimation of entropy is an important problem in the fields of molecular biology, chemistry and physics Misra et al. (2010).

The Rényi entropy is a generalisation of the Shannon entropy and important in ecology and statistics as an index of diversity. It still satisfies the property of additivity and reaches its maximum for a uniformly distributed random variable (Conrad; 2013).

The parameter q of the Rényi entropy H_q^* changes the way the average information is calculated. It designates the norm being used to measure the distance of the probability mass function (p.m.f.) to the origin and is used to make the entropy more or less sensitive to the shape of the probability distributions. For $q \to 1$ the Rényi entropy tends to the Shannon entropy.

The Tsallis entropy is another generalisation of the standard Boltzmann-Gibbs or Shannon entropy that is especially useful in cases where there are strong correlations between the different microstates in a system. The parameter q of the Tsallis entropy H_q can be seen as a measure of how strong the correlations are. In a system with weak correlations q tends to one and the Tsallis entropy reduces to the Shannon entropy. But if the correlations in a system are strong, q becomes more distinct from one, meaning more ore less than one, in order to bias the probabilities of certain microstates occurring (Cartwright; 2014).

In the following section, the entropies mentioned above are mathematically defined for discrete and continuous random variables. Due to their

importance for the following section and the remainder of the thesis, let us introduce the discrete and continuous probability distributions.

Definition 2.1. *Discrete probability distribution.* A discrete probability distribution is characterised by a probability mass function. Thus, the distribution of a random variable X is discrete, and X is then called a discrete random variable, if

$$\sum_{x} P(X = x) = 1$$

as x runs through the set of all possible values of X.

It follows that such a random variable can assume only a finite or countably infinite number of values. For the number of potential values to be countably infinite even though their probabilities sum to 1 requires that the probabilities decline to zero fast enough. An example of this is the Poisson distribution, which can be used to express the probability of a given number of events occurring in a fixed time interval if these events occur with fixed rate and independently of each other. Theoretically, an infinite amount of events could happen, but the probability of that goes to zero (Fahrmeir et al.; 1997).

Definition 2.2. *Continuous probability distribution.* A continuous probability distribution is a probability distribution that has a probability density function. Such a distribution is called absolutely continuous, since its cumulative distribution function is absolutely continuous with respect to the Lebesgue measure λ. If the distribution of X is continuous, then X is called a

continuous random variable. Well-known examples of continuous probability distributions are the normal, uniform and chi-squared distributions. Continuous random variables can take a continuous range of values, thus infinitely many (Fahrmeir et al.; 1997).

2.4 Entropy of a discrete distribution

Discrete entropies $H(X)$ are used if the random variable X is discrete, meaning it only takes a set of finite or countably infinite number of values $\{x_1, x_2, \ldots, x_n\}$. This is the case for a die having the numbers $\{1, \ldots, 6\}$, a coin with two sides, heads and tails, or a language with the letters of its alphabet. In this section the discrete Shannon, Rényi and Tsallis entropy are defined.

2.4.1 Shannon entropy of a discrete distribution

Let X be a discrete random variable taking values in the set $\{x_1, x_2, \ldots, x_n\}$ with probabilities $\{p_1, p_2, \ldots, p_n\}$. The Shannon entropy H of X is defined as the expected amount of information we gain upon learning the value of X

$$H(X) = H_1(X) = E[I(X = x_i)] = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{3}$$

Considerations of continuity lead to the adoption of the convention, $0 \log 0 = 0$, see Figure 1.

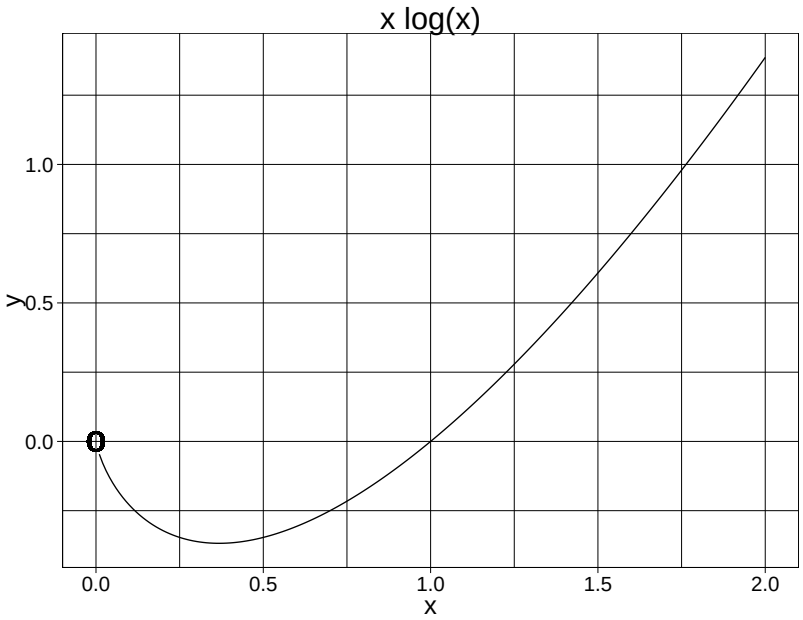


Figure 1: Graph of x against $x \log x$. The function converges to 0 for $x \to 0$.

Some of the most important properties of the Shannon entropy, Equation (2), of a discrete random variable X are (Conrad; 2013; Johnson; 2004; Sabuncu; 2006):

1. $H(X)$ is greater or equal to zero with equality if and only if X is deterministic, meaning $p(x_i) = 1$ for some i.

2. For X taking n values, the entropy is maximised by the discrete uniform distribution with $p_i \equiv 1/n$, so that $0 \leq H(X) \leq \log n$. More details are discussed in Section 3.1.1, where the maximum entropy principle will be presented.

3. For two random variables X and Y the joint entropy is defined as

$$H(X,Y) = -\sum_{x,y} P((X,Y)=(x,y)) \log\{P((X,Y)=(x,y))\}.$$

Then

$$H(X,Y) \leq H(X) + H(Y)$$

with equality for independent X and Y.

Proof.

$$\begin{aligned}
H(X,Y) &= -\sum_{x,y} P((X,Y)=(x,y)) \log\{P((X,Y)=(x,y))\} \\
&= -\sum_{x,y} P((X|Y)=x)P(Y=y) \log\{P((X|Y)=x)P(Y=y)\} \\
&\leq -\sum_{x,y} P(X=x)P(Y=y) \log\{P(X=x)P(Y=y)\} \\
&= -\sum_{x} P(X=x) \log\{P(X=x)\} \\
&\quad -\sum_{y} P(Y=y) \log\{P(Y=y)\} \\
&= H(X) + H(Y)
\end{aligned}$$

With equality for independent X, Y. This important special case shall be kept in mind for comparison with Rényi and Tsallis entropy. □

4. $H(X)$ is independent of the values $x_i, i = 1, \ldots, n$, but solely depends on its distribution $p(x_i)$. For any bijective mapping $f : \Omega_X \to \Omega_X$, note

that the domain is the codomain, the following holds:

$$H(f(x)) = H(X)$$

The probability distribution of a random variable is not affected by such a transformation. Thus entropy is shift and scale invariant.

5. $H(X)$ is a concave function of the probability distribution, p. For example,

$$H(\beta p_1 + (1-\beta)p_2) \geq \beta H(p_1) + (1-\beta)H(p_2),\ \forall \beta \in [0,1].$$

Example 2.4. Consider a random variable X to be the outcome of throwing a fair six-sided die. It has six possible outcomes $\{1,2,3,4,5,6\}$ that are all equally likely, thus $p_i = \frac{1}{6}, i = 1,\ldots,6$. The Shannon entropy of this random experiment is

$$H_1(X) = -\sum_{i=1}^{n} p_i \log_2 p_i = -6\frac{1}{6}\log_2 \frac{1}{6} = \log_2 6 \approx 2.58.$$

2.4.2 Boltzmann entropy of a discrete distribution

Suppose there are n microstates x_i, $i = 1,\ldots,n$, occurring with probability p_i corresponding to one macrostate. The Boltzmann entropy H_B of a system X is defined as

$$H_B(X) = -k_B \sum_{i=1}^{n} p_i \log_e p_i, \tag{4}$$

which is, up to the constant k_B and a different base of the logarithm, the definition of the Shannon entropy (Section 2.4.1) that is used in information theory. The Boltzmann entropy is a strictly monotonic function in the Shannon entropy H_1. We do not particularly worry about the constant, as it passes simply through the analysis. The Boltzmann entropy satisfies all properties of the Shannon entropy stated at the end of Section 2.4.1. Other than the entropies used in information theory, the Boltzmann entropy H_B has the dimension of energy divided by temperature. It is measured in the unit joules per kelvin $\frac{J}{K}$ (Johnson; 2004).

If the occupation of any microstate is assumed to be equally probable (i.e. $p_i \equiv 1/n$, where n is the number of microstates), the entropy simplifies to its maximum (Johnson; 2004)

$$H_B = k_B \log_e n.$$

This assumption, referred to as the fundamental postulate of statistical thermodynamics, is usually justified for an isolated system in equilibrium.

2.4.3 Rényi entropy of a discrete distribution

The Rényi entropy of order q, $q \geq 0$ and $q \neq 1$, of a discrete random variable X taking n values with probabilities $p_1, \ldots, p_n$, is defined as, (Leonenko and

Pronzato; 2010)

$$H_q^*(X) = \frac{1}{1-q} \log \left(\sum_{i=1}^{n} p_i^q \right) = \frac{1}{1-q} \log \|p\|_q^q, \tag{5}$$

using the p-norm of order q of $x \in \mathbb{R}^n$ which is defined as

$$\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{1/q}, \text{ for } q \geq 1 \in \mathbb{R}.$$

The most important special cases are the Manhattan norm $q = 1$ being used in the Shannon entropy, and the Euclidean norm $q = 2$ being used in the so called quadratic entropy. This means that the parameter q changes the way the average information is being calculated by designating the norm being used. It makes the entropy more or less sensitive to the shape of the probability distributions.

The discrete Rényi entropy satisfies the following properties:

1. *Non-negativity.* This means that $H_q^*(X) \geq 0 \; \forall \; X$.

2. The Rényi entropy is concave for $q > 0$.

3. *Additivity.* As for the Shannon entropy, the property of additivity for independent random variables holds. Let X and Y independent discrete random variables with joint probability distribution $f_{X,Y}$, then,

$$H_q^*(X, Y) = H_q^*(X) + H_q^*(Y).$$

The special case of the quadratic entropy is discussed in the appendix of Källberg et al. (2014). In this case, the Rényi entropy simplifies to

$$H_q^*(X) = -\log\left(\sum_{i=1}^{n} p_i^2\right). \tag{6}$$

Lemma 2.1. The Rényi entropy $H_q^*(X)$ of a random variable X can be written as,

$$H_q^*(X) = \frac{1}{1-q}\log(I_q(X)),$$

which leads to the following expression for the information I_q,

$$I_q(X) = \exp((1-q)H_q^*(X)).$$

Note the convention that $I_1 = 1$. (Leonenko and Pronzato; 2010)

2.4.4 Tsallis entropy of a discrete distribution

The Tsallis entropy of order q, $q \geq 0$ and $q \neq 1$, of a discrete random variable X taking n values with probabilities $p_1, \ldots, p_n$, is defined as (Leonenko and Pronzato; 2010)

$$H_q(x) = \frac{1}{1-q}\left(1 - \sum_{i=1}^{n} p_i^q\right). \tag{7}$$

The parameter q changes the way the average information is being calculated analogously to the Rényi entropy.

Properties of Tsallis entropy:

1. Unlike the Rényi entropy it does not keep the property of additivity. Instead it has to be corrected as follows:

$$H_q(X,Y) = H_q(X) + H_q(Y) + (1-q)H_q(X)H_q(Y)$$

for two independent random variables X and Y.

2. The Tsallis entropy is a monotonic conversion of the Rényi entropy.

2.5 Entropy of a continuous distribution

If the random variable X is continuous, with probability density function f, then $H(X)$ denotes the continuous entropy. Well known examples for continuous distributions are the Gaussian and the exponential distribution, see Definitions 2.3 and 2.4. In this section the continuous Shannon, Rényi and Tsallis entropy are defined for the m-dimensional case. This contains the one-dimensional case for $m = 1$.

2.5.1 Shannon entropy of a continuous distribution

The differential Shannon entropy of a continuous random variable $X \in \mathbb{R}^m$ with density f is

$$H_1(X) = -\int_{\mathbb{R}^m} f(x) \log f(x) dx \tag{8}$$

with $0 \log 0 := 0$ and $H(X) = \infty$ if X does not have a continuous distribution function.

Properties of the continuous Shannon entropy (Zografos and Nadarajah; 2005):

1. The continuous Shannon entropy is a concave function in f.

2. It satisfies the same additivity property as the discrete Shannon entropy, $H(X,Y) = H(X) + H(Y)$ for two independent random variables X and Y.

3. It should be noted that, unlike in the discrete case, the Shannon entropy of a continuous random variable does not satisfy the property of non-negativity for arbitrary densities f.

Some of the continuous densities that will be considered throughout this work will be introduced subsequently.

Definition 2.3. *Gaussian Distribution.* The Gaussian distribution or normal distribution is one of the most important distributions. One of the reasons for its special significance is based on of the central limit theorem, according to which averages based on a great number of samples from an arbitrary distribution are approximately Gaussian distributed. A multivariate Gaussian distributed m-dimensional random vector $X = [X_1, \ldots, X_m]^\top$ is denoted by

$$X \sim N(\mu, \Sigma)$$

with mean $\mu = [E(X_1), E(X_2), \ldots, E(X_m)]^\top \in \mathbb{R}^m$ and covariance matrix $\Sigma = \mathrm{Cov}[X_i, X_j]$, $i, j = 1, 2, \ldots, m; \in \mathbb{R}^{m \times m}$. The p.d.f of the m-dimensional

Gaussian is given by

$$\frac{1}{\sqrt{(2\pi)^m|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right).$$

For a one-dimensional Gaussian ($m = 1$) with mean μ and variance σ^2 the p.d.f. simplifies to

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

The standard normal distribution is the commonly used special case, where μ is an all-zero vector of length m and Σ a unit-matrix of dimension $m\times m$. The uni-variate standard normal distribution has parameters $\mu = 0$ and $\sigma^2 = 1$ (Conrad; 2013; Cramér; 1999).

Definition 2.4. *Exponential Distribution.* In probability theory and statistics, the exponential distribution is used to describe the waiting time between Poisson distributed events, which means the events occur continuously and independently at a constant rate $1/\lambda$. It is used to model the length of random time intervals. Examples include the modelling of life expectancy of radioactive atoms and technical equipment but is also used in insurance mathematics.

The p.d.f is given by

$$f(x) = \lambda\exp(-\lambda x)I_{[0,\infty]}(x).$$

The expected value of an exponentially distributed random variable is $1/\lambda$ and its variance is $1/\lambda^2$ (Conrad; 2013). The Shannon entropy of an exponentially distributed random variable is given in 2.6.

Example 2.5. Let X a one-dimensional Gaussian distributed random variable with expected value μ and variance σ^2, see Definition 2.3. The Shannon entropy of X can be calculated to be

$$\begin{aligned}
H_1(X) &= -\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}((x-\mu)/\sigma)^2\right) \\
&\quad \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}((x-\mu)/\sigma)^2\right)\right\} dx \\
&= -\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}((x-\mu)/\sigma)^2\right) \\
&\quad \left\{-\log\{\sqrt{2\pi\sigma^2}\} - \frac{1}{2}((x-\mu)/\sigma)^2\right\} dx \\
&= \frac{1}{2}\log\{2\pi\sigma^2\} \underbrace{\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}((x-\mu)/\sigma)^2\right) dx}_{=1} \\
&\quad + \frac{1}{2\sigma^2} \underbrace{\int (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}((x-\mu)/\sigma)^2\right) dx}_{=E[(x-\mu)^2]=\sigma^2} \\
&= \frac{1}{2}\log\{2\pi\sigma^2\} + \frac{1}{2} \\
&= \frac{1}{2}\log\{2\pi\sigma^2\} + \frac{1}{2}\log\{e\} \\
&= \frac{1}{2}\log\{2\pi\sigma^2 e\}.
\end{aligned}$$

Note that the mean μ is not in the final formula. This makes sense, as the entropy is shift-invariant. For small variances σ ($\sigma < \sqrt{1/(2\pi e)}$) the entropy

takes negative values. This is not a problem, it simply means that the p.d.f. has values greater than one in a comparatively big interval. In this sections it is more concentrated than a uniformely distributed random variable on $[0, 1]$. Thus, the entropy is less than zero, see Figure 2.

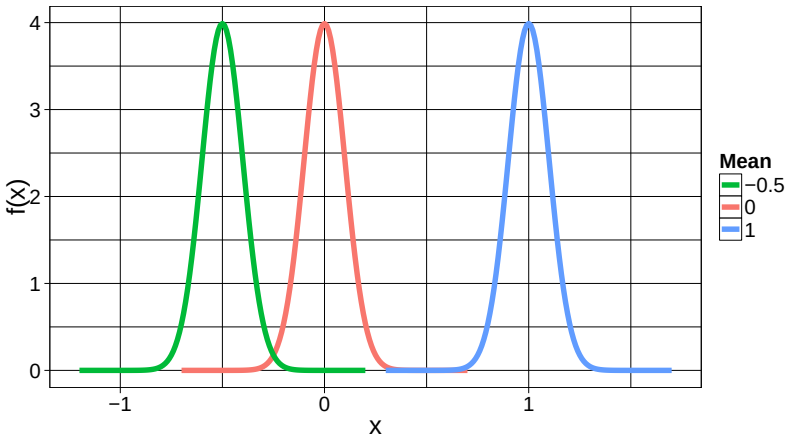


Figure 2: One-dimensional Gaussian with different mean μ but same variance $\sigma^2 = 0.01$ and, thus, same entropy $H_1(X) \approx -0.884$. The p.d.f $f(x)$ is greater than one close to the mean of the distribution.

Example 2.6. Let X be an exponentially distributed random variable with expected value λ, thus rate $1/\lambda$, see definition 2.4. The Shannon entropy is

$$\begin{aligned} H_1(X) &= \int_0^\infty \frac{1}{\lambda} \exp(-\frac{x}{\lambda}) \log \left\{ \frac{1}{\lambda} \exp(-\frac{x}{\lambda}) \right\} dx \\ &= 1 + \log(\lambda). \end{aligned}$$

This means that the smaller the rate $1/\lambda$, the greater the entropy of the exponential distribution. In other words, the longer the expected waiting

period, the greater is the uncertainty of the waiting time. For small $\lambda$the entropy takes negative values.

Lemma 2.2. For random variables from an exponential family the Shannon entropy

$$H_1(f(X;\theta)) = F(\theta) - \langle \theta, \nabla F(\theta) \rangle,$$

where θ denotes the natural parameter of the exponential family, $\langle x, y \rangle$ the inner product $x^\top y$ and ∇ the Gâteaux-differential, is a closed-form expression of the Shannon entropy (Nielsen and Nock; 2011).

Theorem 2.1. For the m-dimensional normal distribution

$$H_1(X) = \frac{1}{2} \log\{(2\pi \exp)^m\}|\Sigma| \tag{9}$$

is a closed form of the Shanon entropy (Nielsen and Nock; 2011).

2.5.2 Rényi entropy of a continuous distribution

The Rényi entropy of a random variable $X \in \mathbb{R}^m$ with probability measure μ which has density f is defined as

$$H_q^* = \frac{1}{1-q} \log \int_{\mathbb{R}^m} f^q(x) dx,\ q \neq 1\,. \tag{10}$$

The parameter q changes the way the average information is being calculated. For $q \to 1$ the Rényi entropy tends to the Shannon entropy (Nielsen and Nock; 2011).

The Rényi entropy of a continuous random variable satisfies the following properties (Zografos and Nadarajah; 2005):

1. The Rényi entropy is concave for $q > 0$ and convex for $q < 0$.

2. As for the Shannon entropy, the property of additivity for independent random variables holds. Let X and Y independent discrete random variables with joint probability distribution $f_{X,Y}$, then

$$H_q^*(X,Y) = H_q^*(X) + H_q^*(Y).$$

For the special case of the quadratic entropy, $q = 2$, the Rényi entropy simplifies to

$$H_q^*(X) = -\log\left(\int_{\mathbb{R}^m} p_i^2 dx\right). \tag{11}$$

For the estimation of the quadratic entropy there are several estimators proposed in literature. The consistent estimator that satisfies normality conditions developed in Källberg et al. (2014) will be presented in Section 4.3.

Definition 2.5. *Exponential Families.* A random variable X belongs to an exponential family if and only if its p.d.f. $f(x)$ can be written in such a form that

$$f(x;\theta) = \exp\left(\langle t(x), \theta\rangle F(\theta) + k(x)\right), \tag{12}$$

where $\langle t(x), \theta\rangle = x^\top\theta$ denotes the inner product of x and θ, $t(x)$ the sufficient statistics, θ the natural parameters, $F(\theta)$ a C^∞ differentiable real-valued

convex function and $k(x)$ a support measure. (Nielsen and Nock; 2011)

Some well known examples of exponential families with support measure $k(x) = 0$ are the normal distribution (see Definition 2.3), the exponential distribution (see Definition 2.4), the gammma distribution, and the Bernoulli distribution. It is useful to speak of them as exponential families as that gives a more general approach to those distributions and their common properties.

Example 2.7. In order to illustrate this Definition 12 consider the decomposition of the univariate Gaussian distribution as introduced in Definition 2.3.

$$\begin{aligned} f(x, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2}\log 2\pi\sigma^2\right) \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right) \\ &= \exp\left(x\frac{\mu}{\sigma^2} - x^2\frac{1}{2\sigma^2} - \frac{\mu}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2\right) \\ &= \exp\left(\underbrace{\langle(x, x^2)}_{t(x)}, \underbrace{(\frac{\mu}{\sigma^2}, -\frac{1}{\sigma^2})\rangle}_{\theta=(\theta_1,\theta_2)} \underbrace{-\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2}_{F(\theta)=-\frac{\theta_1^2}{4\theta_2}+\frac{1}{2}\log\frac{2\pi}{-\theta_2}} + \underbrace{0}_{k(x)}\right), \end{aligned}$$

where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. This decompositions proofs that the univariate Gaussian is a distribution of the exponential family, and thus satisfies all properties of an exponential family.

Lemma 2.3. For random variables from the exponential family with base support measure $k(x) = 0$,

$$H_q^*(f(X;\theta)) = \frac{1}{1-q}\left[F(q\theta) - qF(\theta)\right],$$

with θ the natural parameter of the exponential family, is a closed-form expression of the Rényi entropy.

Proof. Let us recall the definition of the Rényi entropy in Equation (10),

$$H_q^* = \frac{1}{1-q}\log\int_{\mathbb{R}^m} f^q(x)dx\,,\ q \neq 1\,,$$

which can be written as

$$H_q^* = \frac{1}{1-q}\log I_q(f)\,,\ q \neq 1,$$

where $I_q(f) = \int_{\mathbb{R}^m} f^q(x)dx$.

$$
\begin{aligned}
I_q(f) &= \int_{R^m} f^q(x)dx \\
&= \int_{\mathbb{R}^m} \exp\{q(\langle t(x), \theta\rangle - F(\theta) + k(x))\}dx \\
&= \int_{\mathbb{R}^m} \exp\{\langle t(x), q\theta\rangle - qF(\theta) + qk(x) + \\
&\quad \underbrace{(1-q)k(x) - (1-q)k(x)}_{=0} + \underbrace{F(q\theta) - F(q\theta)}_{=0}\}dx \\
&= \int_{\mathbb{R}^m} \exp\{F(q\theta) - qF(\theta)\} f(x; q\theta) \exp\{(q-1)k(x)\}dx \\
&= \exp\{F(q\theta) - qF(\theta)\} \int_{\mathbb{R}^m} \exp\{(q-1)k(x)\} f(x; q\theta)dx \\
&= \exp\{F(q\theta) - qF(\theta)\} E[\exp(q-1)k(x)]
\end{aligned}
$$

For base support measure $k(x) = 0$ this simplifies to

$$
I_q(f) = \exp\{F(q\theta) - qF(\theta)\}. \tag{13}
$$

This means that the Rényi entropy

$$
\begin{aligned}
H_q^* &= \frac{1}{1-q} \log I_q(f) \\
&= \frac{1}{1-q} \log \exp\{F(q\theta) - qF(\theta)\} \\
&= \frac{1}{1-q} \{F(q\theta) - qF(\theta)\},
\end{aligned}
$$

as stated in Lemma 2.3 (Nielsen and Nock; 2011).

□

Theorem 2.2. For the m-dimensional normal distribution the Reńyi entropy is

$$H_q^*(X) = \frac{m}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| + \frac{m\log q}{2(q-1)}. \tag{14}$$

This is a closed form of the Rényi entropy that is implemented in the R-Code Nielsen and Nock (2011).

Example 2.8. The normal distribution can be seen as a special case of the Student distribution, see Definition 3.1, where the degrees of freedom ν tend to infinity. The Rényi entropy $H_q^*(f)$ of an m-variate Student distributed random variable with density

$$f(x) = \frac{1}{(\nu\pi)^{\frac{m}{2}}}\frac{\Gamma(\frac{m+\nu}{2})}{\Gamma(\frac{\nu}{2})}\frac{1}{|\Sigma|^{\frac{1}{2}}[1+(x-\mu)^\top[\nu\Sigma]^{-1}(x-\mu)]^{\frac{m+\nu}{2}}}$$

is given by

$$\frac{1}{1-q}\log\frac{B\left(\frac{q(m+\nu)}{2}-\frac{m}{2},\frac{m}{2}\right)}{B^q(\frac{\nu}{2},\frac{m}{2})} + \log[(\pi\nu)^m|\Sigma|] - \log\Gamma\left(\frac{m}{2}\right), \text{ for } q > \frac{m}{m+\nu}.$$

Consistently this entropy converges to the Rényi entropy of the normal distribution, denoted in this equation by g, when the degrees of freedom ν of f

tend to ∞,

$$\begin{aligned}
\lim_{\nu\to\infty} H_q^*(f) &= \lim_{\nu\to\infty} \frac{1}{1-q} \log \frac{B\left(\frac{q(m+\nu)}{2} - \frac{m}{2}, \frac{m}{2}\right)}{B^q(\frac{\nu}{2}, \frac{m}{2})} \\
&\quad + \log[(\pi\nu)^m |\Sigma|] - \log\Gamma\left(\frac{m}{2}\right) \\
&= \log[(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}] - \frac{m}{2(1-q)} \log q \\
&= H_q^*(g) \\
&= H_1(g) - \frac{m}{2}\left(1 + \frac{\log q}{1-q}\right)
\end{aligned}$$

If additionally q tends to zero, this converges to the Shannon entropy of the normal distribution,

$$\lim_{q\to\infty} H_q^* = \log[(2\pi\exp)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}] = H_1(g)$$

according to the property of the Rényi entropy, that if converges to the the Shannon entropy for $q \to 1$ (Leonenko and Pronzato; 2010).

Theorem 2.3. With Theorem 2.2 and Lemma 2.1 the information I_q of an

m-dimensionally Gaussian distributed variable X is given as

$$\begin{aligned} I_q(X) &= \exp((1-q)H_q^*(X)) \\ &= \exp\left((1-q)\left[\frac{m}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{m}{2(q-1)}\log(q)\right]\right) \\ &= \exp\left((1-q)\frac{m}{2}\log(2\pi) + (1-q)\frac{1}{2}\log|\Sigma| - \frac{m}{2}\log(q)\right) \\ &= \frac{(2\pi)^{\frac{(1-q)m}{2}}|\Sigma|^{\frac{(1-q)}{2}}}{(q)^{\frac{m}{2}}} \\ &= \sqrt{\frac{(2\pi)^{(1-q)m}|\Sigma|^{(1-q)}}{(q)^m}} \end{aligned}$$

and $I_1 = 1$. If X is the standard normal distributed, thus the covariance matrix Σ is the identity matrix, this simplifies to

$$I_q(X) = \sqrt{\frac{(2\pi)^{(1-q)m}}{(q)^m}}$$

as the determinant of the covariance matrix $|\Sigma|$ equals one.

Definition 2.6. The beta distribution is a family of continuous probability distributions defined on $[0,1]$ with shape parameters $\alpha, \beta \in \mathbb{R}^+$. It is defined to be

$$f_{\alpha,\beta}(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(x-1)^{\beta-1}$$

where $\text{Beta}(\alpha,\beta)$ denotes the Betafunction $\text{Beta}(\alpha,\beta) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$. A beta distributed variable X has mean $E[X] = \alpha/\alpha+\beta$ and variance $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$

Definition 2.7. The one-dimensional uniform distribution with support $[a, b]$ has the p.d.f.

$$f(x) = \frac{1}{b-a} \cdot I_{[a,b]}(x)$$

. The case that $a = 0$ and $b = 1$, thus f has support $[0, 1]$ can be seen as a special case of the beta distribution where $\alpha = \beta = 1$. A uniformly distributed variable X on $[a, b]$ has mean $E[X] = (a + b)/2$ and variance $(b - a)^2/12$.

Theorem 2.4. The Rényi entropy $H_q^*(f)$ of a one-dimensional beta distributed variable X with p.d.f. (see Definition 2.6)

$$f(X) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(x-1)^{\beta-1}$$

is given in Nadarajah and Zografos (2003) as

$$\frac{1}{1-q} \log \left(\frac{\text{Beta}(q\alpha - q + 1, q\beta - q + 1)}{\text{Beta}^q(\alpha, \beta)} \right)$$

where $\text{Beta}(\alpha, \beta)$ denotes the beta-function $\text{Beta}(\alpha, \beta) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$.

Theorem 2.5. The uniform distribution with support $[0, 1]$ with p.d.f. (see Definition 2.7)

$$f(x) = 1 \cdot I_{[0,1]}(x)$$

can be seen as the special case of the beta distribution where $\alpha = \beta = 1$.

For the entropy it follows that

$$\begin{aligned} H_q^* &= \frac{1}{1-q}\log\left(\frac{\text{Beta}(q\alpha - q + 1, q\beta - q + 1)}{\text{Beta}^q(\alpha,\beta)}\right) \\ &= \frac{1}{1-q}\log\left(\frac{\text{Beta}(q - q + 1, q - q + 1)}{\text{Beta}^q(1,1)}\right) \\ &= \frac{1}{1-q}\log(1) \\ &= 0, \end{aligned}$$

which is the maximum entropy of a beta distribution with support $[0, 1]$.

2.5.3 Tsallis entropy of a continuous distribution

The Tsallis entropy of a continuos random variable is defined as

$$H_q = \frac{1}{1-q}\,(\int_{R^m} f^q(x)dx - 1),\ q \neq 1. \tag{15}$$

The parameter q changes the way the average information is being calculated. For $q \to 1$ the Tsallis entropy tends to the Shannon entropy.

Unlike the Shannon and Rényi entropy it does not keep the property of additivity, but instead

$$H_q(X, Y) = H_q(X) + H_q(Y) + (1 - q)H_q(X)H_q(Y)$$

for two independent random variables X and Y.

Lemma 2.4. The Tsallis entropy $H_q(X)$ of a random variable X can be written as

$$H_q(X) = \frac{1}{1-q}(I_q(X) - 1),$$

which leads to the following expression for the information I_q,

$$I_q(X) = (1-q)H_q(X) + 1.$$

Note that $I_1 = 1$.

Theorem 2.6. The Rényi and Tsallis entropy are continuous monotonic conversions of one another. This relationship can be concluded from Lemma 2.1 and Lemma 2.4 to be

$$\exp((1-q)H_q^*(X)) = (1-q)H_q(X) + 1$$

. Algebraic transformations leads to the following equations:

$$\begin{aligned} H_q^*(X) &= \frac{1}{1-q}\log((1-q)H_q(X) + 1) \\ H_q(X) &= \frac{1}{1-q}(\exp((1-q)H_q^*(X)) - 1) \end{aligned}$$

Lemma 2.5. For random variables from the exponential family

$$H_q(f(X;\theta)) = \frac{1}{1-q}\{\exp F(q\theta) - qF(\theta) - 1\}$$

with θ the natural parameter of the exponential family, is a closed-form expression of the Tsallis entropy (Nielsen and Nock; 2011).

Proof. Recall the Tsallis entropy as stated in Equation (15)

$$H_q = \frac{1}{1-q}\left(\int_{R^m} f^q(x)dx - 1\right), \; q \neq 1$$

which can be written as

$$H_q = \frac{1}{1-q}\left(I_\alpha(q) - 1\right),$$

with $I_\alpha(q) = \int_{\mathbb{R}^m} f^q(x)dx$ as in Proof 2.5.2. With Equation (13) the Tsallis entropy for an exponential family with base support measure $k(x) = 0$ is given by

$$\begin{aligned} H_q &= \frac{1}{1-q}\left(I_\alpha(q) - 1\right) \\ &= \frac{1}{1-q}\left(\exp\{F(\alpha\theta) - \alpha F(\theta)\} - 1\right), \end{aligned}$$

completing the proof (Nielsen and Nock; 2011). □

Example 2.9. For the m-dimensional normal distribution it follows that

$$H_q(X) = \frac{1}{1-q}((2\pi)^{(1-q)\frac{m}{2}} \cdot |\Sigma|^{\frac{1-q}{2}} \cdot q^{\frac{m}{2}} - 1) \tag{16}$$

is a closed form of the Tsallis entropy that is implemented in the R-Code. This follows from the closed form of H_q^* in Nielsen and Nock (2011) and the

relation $H_q^* = \log(1-(q-1)H_q)/(1-q)$ in Leonenko and Pronzato (2010).

Proof.

$$\begin{aligned}
H_q^* &= \log(1-(q-1)H_q)/(1-q) \\
\Rightarrow H_q &= \frac{1}{q-1}(\exp\{(1-q)H_q^*\}-1) \\
&= \frac{1}{q-1}(\exp\{(1-q)[\frac{m}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| + \frac{m\log q}{2(q-1)}]\}-1) \\
&= \frac{1}{q-1}(\exp\{\frac{(1-q)m}{2}\log 2\pi + \frac{1-q}{2}\log|\Sigma| - \frac{m\log q}{2}\}-1) \\
&= \frac{1}{q-1}(\exp^{\frac{(1-q)m}{2}\log 2\pi}\cdot\exp^{\frac{1-q}{2}\log|\Sigma|}\cdot\exp^{-\frac{m\log q}{2}}-1) \\
&= \frac{1}{1-q}((2\pi)^{(1-q)\frac{m}{2}}\cdot|\Sigma|^{\frac{1-q}{2}}\cdot q^{\frac{m}{2}}-1)
\end{aligned}$$

□

2.6 Kullback-Leibler divergence

In information theory the Kullback-Leibler divergence D_{KL}, also known as information divergence, information gain or relative entropy, quantifies the difference between two probability distributions P and Q on $\mathbb{R}^m$. The Kullback-Leibler divergence of Q from P is a measure of the information lost when Q is used to approximate P. The idea is to use the expected value of the logarithmic difference of the probabilities P and Q, where the expectation is taken using the probabilities P, to measure the discrepancy between two distributions. Note that $D_{\mathrm{KL}} \geq 0$ for all P and Q. The Kullback-Leibler di-

vergence is defined for discrete (Section 2.6.1) and continuous (Section 2.6.2) random variables (Wang et al.; 2006).

2.6.1 Discrete Kullback-Leibler divergence

Let Q, P discrete probability distributions on $\mathbb{R}^m$ and P absolutely continuous with respect to Q, meaning $Q(x) = 0 \Rightarrow P(x) = 0 \; \forall x \in \mathbb{R}^m$. Let $0 \log \frac{0}{0} \equiv 0$, then the Kullback-Leibler divergence is defined as (Wang et al.; 2006)

$$D_{\mathrm{KL}}(P||Q) = \sum_{x \in \mathbb{R}^m} P(x) \log \frac{P(x)}{Q(x)}.$$

2.6.2 Continuous Kullback-Leibler divergence

Let Q, P continuous probability distributions on $\mathbb{R}^m$ and P absolutely continuous with respect to Q. Denote the densities of Q and P with respect to the Lebesgue measure by $p(x)$ and $q(x)$. As P is absolutely continuous with respect to Q, $p(x) = 0$ for almost every x such that $q(x) = 0$ and $0 \log \frac{0}{0} \equiv 0$. Under this condition the Kullback-Leibler divergence is defined in Wang et al. (2006) as:

$$D_{\mathrm{KL}}(P||Q) = D_{\mathrm{KL}}(p(x)||q(x)) = \int_{x \in \mathbb{R}^m} p(x) \log \frac{p(x)}{q(x)} dx. \tag{17}$$

Note that for both the continuous and the discrete case the Kullback-Leibler divergence $D_{\mathrm{KL}}(Q||P)$ is greater or equal than zero with equality if and only if $p(x) = q(x)$. In general, the Kullback-Leibler divergence is

not symmetric, meaning $D_{\mathrm{KL}}(Q||P) \neq D_{\mathrm{KL}}(P||Q)$ and does not satisfy the triangle inequality. Thus it is not a proper distance metric. Nonetheless it is very useful for a number of applications (Sabuncu; 2006).

2.7 Bregman divergence

Let f be real-valued strictly convex function defined on the convex set $S = \mathrm{dom}(f) \in \mathbb{R}$, the domain of f, such that f is differentiable on $int(S)$, the interior of S. The Bregman divergence or Bregman distance of two points z_1, z_2 is defined as

$$D_q(z_1||z_2) = f(z_1) - f(z_2) - \langle z_1 - z_2; \nabla(z2)\rangle,$$

where ∇ is the gradient of f (Banerjee et al.; 2004). Note that the Bregman divergence is non-negative, thus $D_F(z_1, z_2) \geq 0$ for all z_1, z_2, as f is a convex function.

3 Maximum entropy principle

In thermodynamics Lagrangian methods are used to show that the entropy S is maximised subject to an energy constraint by the Gibbs states. The maximum of

$$-\sum_r p_r \log p_r$$

subject to $\sum_r p_r = 1$ and $\sum_r p_r E_r = E$ is reached for $p_i = \exp(-\beta E_i)/Z_\beta$, for some β determined by the total energy E and where the partition function $Z_\beta = \sum_i \exp(-\beta E_i)$. The parameter β can be found with knowledge of Z_β, since

$$-\frac{d}{d\beta}\log Z_\beta = -\frac{Z'_\beta}{Z_\beta} = \frac{\sum_i E_i \exp(-\beta E_i)}{\sum_i \exp(-\beta E_i)} = \sum_i E_i p_i = E$$

According to the second law of thermodynamics S converges to its maximum, the Gibbs state. The total energy E in the systems remains always constant due to energy conservation, that is why it is known from start which Gibbs state will be the limit. In our case this is not of further interest, but it shall be mentioned here that the second law of thermodynamics is still debated controversially, even though it is such a long established principle in natural science (Johnson; 2004).

Analogously, the Central Limit Theorem states that the information-theoretic entropy H_1 increases to its limit as we take convolutions, implying convergence to the Gaussian. As the variance remains constant during convolutions we can tell from start which Gaussian will be the limit. This

similarity between those two of the most basic principles in statistics and physics is quite striking (Johnson; 2004).

The principle of maximum entropy states that, subject to certain constraints as a given mean or variance, the probability density function which best represents the current state of knowledge is the one with largest entropy (Conrad; 2013). In other words, the principle of maximum entropy expresses a claim of epistemic modesty. The distribution selected is the one that makes the least claim to being informed beyond the stated prior data. One admits to the most ignorance beyond the stated prior data. Any probability function satisfying the constraints that has a smaller entropy will contain less uncertainty, therefore more information and thus says something stronger than what we are assuming.

3.1 Shannon entropy maximisation

The Shannon entropy, contained as a special case in the Rényi and Tsallis entropy, is the most basic concept of entropy. The maximisation of the Shannon entropy is a well known problem and can, amongst other things, be used as a way of justifying a set of exponential families. In this section the maximising distributions for a discrete random variable (Section 3.1.1), a positive random variable with known mean (Section 3.1.2) and a rational random variable with known variance (Section 3.1.3) are deduced.

3.1.1 Discrete Shannon entropy maximisation

For a discrete probability function p on a finite set $x_1, \ldots, x_n \in \mathbb{R}^m$ the Shannon entropy is maximised by the uniform distribution. Therefore

$$H_1(p) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \leq \log n \tag{18}$$

with equality if and only if $p(x_i) = 1/n$, $i = 1, \ldots, n$ (Conrad; 2013).

In order to prove this, a number of lemmas will be introduced. The actual proof will be given on page 48.

Lemma 3.1. For arbitrary $x > 0, y \geq 0 \in \mathbb{R}^m$ with $0 \log 0 = 0$ being adopted, see Figure 1,

$$y \log y \leq x - \log x$$

Proof. For $y = 0$ the proof is trivial, as the left side equals zero and $x > 0$. For $y > 0$ the inequality can be transformed as

$$\begin{aligned} y \log y &\leq x - \log x \\ y - x &\leq -y \log x + y \log x \\ 1 - \frac{x}{y} &\leq \log \frac{y}{x} \\ \log \frac{x}{y} &\leq \frac{x}{y} \end{aligned}$$

The proof will be completed graphically in Figure 3 in order to visualise the inequality. For an easier acces substitue $\frac{x}{y}$ by t, $t > 0$ as $x, y > 0$ (Conrad;

2013). □

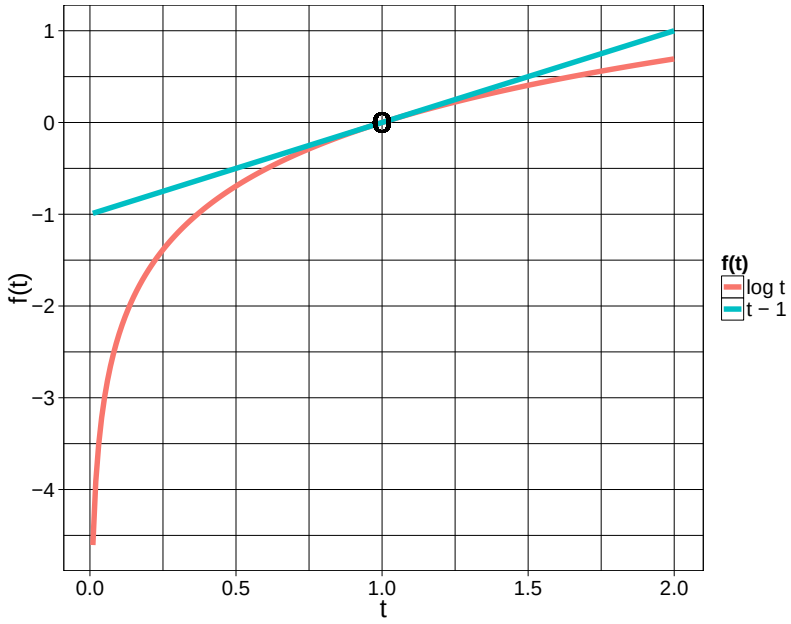


Figure 3: The inequality stated in the proof to Lemma 3.1 is vizualised. It can be seen that $\log t$ is smaller than $t-1$ with equality for $t = 1$ in $f(t) = 0$.

Lemma 3.2. For any continuous p.d.f.s $p(x)$ and $q(x)$ on an interval I with $q > 0$ on I the following holds

$$-\int_I p(x) \log p(x) dx \leq -\int_I p(x) \log q(x) dx$$

with equality if and only if $p(x) = q(x)$ for all $x \in I$. Analogously for two discrete p.m.f.s $p(x)$ and $q(x)$ on a set $\{x_1, x_2, \ldots, \}$, with $q(x_i) \geq 0$ for all i,

$$-\sum_i p(x_i) \log p(x_i) dx \leq -\sum_i p(x_i) \log q(x_i) dx$$

if both sums converge. As in the continuous case there is equality if and only if $p(x_i) = q(x_i)$ for all x_i. Note that the left side matches the definition of the Shannon entropy, thus an upper-bound is given for it (Conrad; 2013).

Proof. By Lemma 3.1 for any $x \in I$ and continuous $p(x), q(x)$

$$p(x) - p(x)\log p(x) \leq q(x) - p(x)\log q(x). \tag{19}$$

By integrating over both sides of equation (19) one gets

$$\begin{aligned} \underbrace{\int_I p(x)dx}_{1} - \int_I p(x)\log p(x)dx &\leq \underbrace{\int_I q(x)dx}_{1} - \int_I p(x)\log q(x)dx \\ -\int_I p(x)\log p(x)dx &\leq -\int_I p(x)\log q(x)dx \end{aligned}$$

For the case of equality of theses integrals, the continuous function

$$q(x) - p(x)\log q(x) - \{p(x) - p(x)\log p(x)\}$$

has integral zero over I. With its property of nonnegativity follows, that this function equals zero. Thus Equation (19) is an equality for all $x \in I$ if and only if $p(x) = q(x)$ for all $x \in I$, by Lemma 3.1. The proof for the discrete case is the same with integrals replaced by sums and is thus not written out again (Conrad; 2013). □

Theorem 3.1. Consider $p(x)$, $q(x)$ continuous probability density functions

(p.d.f.) with finite entropy H_1 on an interval I and $q(x) > 0$ for $x \in I$. From

$$-\int_I p(x) \log q(x) dx = H_1(q)$$

follows that $H_1(p) \leq H_1(q)$ with equality if and only if $p(x) = q(x)$ on I.

For the discrete case let $p(x_i)$, $q(x_i)$ discrete p.m.f.s on the discrete set $\{x_1, x_2, \dots, \}$ and $q(x_i) > 0$ for all i. Further assume that the Shannon entropy H_1 of p and q is finite. Then

$$-\sum_i p(x_i) \log q(x_i) dx = H_1(q)$$

implies $H_1(p) \leq H_1(q)$ with equality if and only if $p(x_i) = q(x_i)$ for all i.

Proof. By Lemma 3.2 the Shannon entropy $H_1(p)$ is bounded from above by $-\int_I p(x) \log q(x) dx$ for continuous $p(x)$, $q(x)$respectively $-\sum_i p(x_i) \log q(x_i)$ in the discrete case. This bound is assumed to equal $H_1(q)$. If it is the case that $H_1(p) = H(q)$, then $H_1(p)$ equals the bound, thus by Lemma 3.2 $p(x)$ is $q(x)$. □

Finally we have all the means necessary to proof Equation (18) in Section 3.1.1.

Proof. This is the proof for the discrete entropy maximisation in Section 3.1.1 on page 45. Let $p(x_i)$ a p.m.f. on $\{x_1, x_2, \dots, x_n\}$. Let $q(x_i) = \frac{1}{n}$ for all

i a uniform distribution, then

$$\begin{aligned} H_1(p) &= -\sum_{i=1}^{n} p_i \log p_i \\ &\leq -\sum_{i=1}^{n} p_i \log q_i \\ &= \sum_{i=1}^{n} p_i \log n \\ &= \log n \\ &= \sum_{i=1}^{n} q_i \log n \\ &= H_1(q), \end{aligned}$$

with equality if and only if $p(x_i) = q(x_i)$ for all x_i, by Lemma 3.2. This means that the Shannon entropy of an arbitrary p.m.f. can only equal the entropy of the uniform distribution, if it is uniformly distributed as well, completing the proof that the uniform distribution maximises the Shannon entropy in the discrete case. □

3.1.2 Continuous Shannon entropy maximisation on $\mathbb{R}^+$

For a continuous probability function f on $\mathbb{R}^+$ with given mean λ the Shannon entropy is maximised by the exponential distribution with mean λ. Therefore

$$H_1(f) \leq 1 + \log \lambda, \tag{20}$$

with equality for

$$f(x) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right) \tag{21}$$

the exponential distribution with mean λ, respectively rate $\frac{1}{\lambda}$, see Definition 2.4 (Conrad; 2013). This is an important case to be considered, as in application many random variables are restricted to be greater than zero. One can think of many examples as time, height, length and weight. None of these can take values smaller than zero. If this is not the case, meaning the random variable can take values in all $\mathbb{R}$, Section 3.1.3 derives the normal distribution to be of maximum entropy.

Proof. Consider an arbitrary continuous p.d.f. p on $\mathbb{R}^+$ with known mean λ. Let q be the exponential distribution with mean λ. Similarly to the proof of the discrete maximum entropy distribution above it holds that

$$\begin{aligned}
H_1(p) &= -\int_0^\infty p(x) \log p(x) dx \\
&\leq -\int_0^\infty p(x) \log q(x) dx \\
&= -\int_0^\infty p(x) \log\left(\frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right)\right) dx \\
&= \int_0^\infty p(x) \left(\log \lambda + \frac{x}{\lambda}\right) dx \\
&= \log \lambda \underbrace{\int_0^\infty p(x) dx}_{1} + \frac{1}{\lambda} \underbrace{\int_0^\infty x p(x) dx}_{\lambda} \\
&= \log \lambda + 1 \\
&= H_1(q).
\end{aligned}$$

□

3.1.3 Continuous Shannon entropy maximisation on $\mathbb{R}$

For a continuous probability density function f on $\mathbb{R}$ with known variance $\sigma^2 \in \mathbb{R}^+$ the Shannon entropy is

$$H_1(f) \leq \frac{1}{2}(1 + \log(2\pi\sigma^2)), \tag{22}$$

with equality if f is Gaussian with variance σ^2 and some mean μ, see Definition 2.3, therefore

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

This is the most general case of an unrestricted random variable. Examples are time differences, economic growth, acceleration and many log-transformed random variables. An interesting application is an entropy based test of normality of a random variable, see Vasicek (1976). It is based on the quality of the normal distribution that its entropy exceeds that of any other distribution with the same variance, giving a non-parametric test for normality of great power. In Conrad (2013) some other probability density functions with different constraints are characterised with maximum entropy.

Proof. Consider an arbitrary continuous p.d.f. p on $\mathbb{R}$ with known variance σ^2 and mean μ. Let q be the normal distribution with variance σ^2 and mean

μ. Similarly to the proof of the discrete maximum entropy distribution on page 48 it holds that

$$\begin{aligned}
H_1(p) &= -\int_{\mathbb{R}} p(x)\log p(x)dx \\
&\leq -\int_{\mathbb{R}} p(x)\log q(x)dx \\
&= -\int_{\mathbb{R}} p(x)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right) \\
&= \int_{\mathbb{R}} p(x)\left(\frac{1}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}(x-\mu)^2\right)dx \\
&= \frac{1}{2}\log(2\pi\sigma^2)\underbrace{\int_{\mathbb{R}} p(x)dx}_{1}+\frac{1}{2\sigma^2}\underbrace{\int_{\mathbb{R}}(x-\mu)^2p(x)dx}_{\sigma^2} \\
&= \frac{1}{2}\log(2\pi\sigma^2)+\frac{1}{2} \\
&= H_1(q),
\end{aligned}$$

with equality if and only if $p(x) = q(x)$ on all $\mathbb{R}$, by Lemma 3.2. Said in words, the Shannon entropy of an arbitrary p.d.f. $p(x)$ is lesser or equal than the entropy of the Gaussian distribution, with equality if and only if $p(x)$ is Gaussian distributed with the same variance σ^2. □

3.1.4 Continuous Shannon entropy maximisation on $\mathbb{R}^m$

In this section a generalisation of Section 3.1.3 for multidimensional random variables is given. Consider a continuous p.d.f. f on $\mathbb{R}^m$ with fixed covariance matrix Σ, then

$$H_1(f) \leq \frac{1}{2}\left(m+\log((2\pi)^m)\det\Sigma\right), \tag{23}$$

with equality if and only if f is m-dimensional Gaussian distributed with covariance matrix Σ, therefore

$$f(x) = \frac{1}{(2\pi)^m \det \Sigma} \exp(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)),$$

where $\mu \in \mathbb{R}^n$ denotes the means (Conrad; 2013). Note that the covariance matrix Σ is symmetric and positive-semidefinit. The variances of the m components of the random variable are on the main diagonal of covariance matrix.

In the following section dealing with Rényi entropy maximisiation the Student distribution is needed as a generalisation of the normal distribution, giving the limiting maximum entropy distribution.

Definition 3.1. *Multivariate Student distribution.* An m-dimensionally Student distributed m-variate random variable X with mean $\mu \in \mathbb{R}^m$, correlation matrix $\Sigma \in \mathbb{R}^{m \times m}$ respectively covariance matrix $C = \nu\Sigma/(\nu-2)$ and ν degrees of freedom is denoted by

$$X \sim T(\nu, \Sigma, \mu).$$

Its p.d.f. $f_\nu(x)$ is given by

$$f_\nu(x) = \frac{1}{(\nu\pi)^{\frac{m}{2}}} \frac{\Gamma(\frac{m+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{|\Sigma|^{\frac{1}{2}}[1+(x-\mu)^\top[\nu\Sigma]^{-1}(x-\mu)]^{\frac{m+\nu}{2}}},$$

with the gammfunction $\Gamma(t) = \int_0^\infty x^{t-1}\exp(x)dx$.

Note that the limiting form for $\nu \to \infty$ is the m-variate Gaussian $N(\mu, \Sigma)$ as given in Definition 2.3. For $\nu = 1$ the m-dimensional Student distribution simplifies to the density of the m-dimensional Cauchy distribution

$$f(x; \mu, \Sigma, k) = \frac{\Gamma\left(\frac{1+m}{2}\right)}{\Gamma(\frac{1}{2})\pi^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}} \left[1 + (x-\mu)^\top \Sigma^{-1}(x-\mu)\right]^{\frac{1+m}{2}}}.$$

The multivariate Student distribution contains the one-dimensional Student distribution as special case for $m = 1$. Due to its importance, the p.d.f. of the one-dimensional Student distribution is written out:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

3.2 Rényi entropy maximisation

In this section distributions that maximise the Rényi entropy under certain constraints will be introduced. One shall be reminded, that the Rényi entropy contains the Shannon entropy as the special case for $q = 1$. This means that this section contains parts of the previous Section 3.1 on Shannon entropy maximisation as special cases.

3.2.1 Continuous Rényi entropy maximisation on $\mathbb{R}^m$

For a continuous probability density function f on $\mathbb{R}^m, m \geq 1$, and $\frac{m}{m+2} < q < 1$ the m-dimensional Student distribution maximises the entropy. The distribution maximising the Rényi entropy is uniquely defined, as H_q^* of a

probability density function f is a concave function of q for $q \geq 0$ and convex for $q \leq 0$. For the case that $\frac{m}{m+2} < q < 1$ the entropy maximising distribution under the variance constraint $E[(X-\mu)(X-\mu)^\top] = C$ is the Student distribution

$$T\left(\nu, \frac{(\nu-2)C}{\nu}, 0\right),$$

with $\nu = \frac{2}{1-q} - m > 2$.

If $q > 1$ and the distribution is subject to the same variance constraint, then the q-entropy maximising distribution $f_p(x)$ is

$$\begin{cases} \frac{\Gamma\left(\frac{p}{2}\right)}{|C|^{\frac{1}{2}}[\pi(p+2)C]^{\frac{m}{2}}\Gamma\left(\frac{p-m}{2}+1\right)} \left[1-(x-\mu)^\top[(p+2)C]^{-1}(x-\mu)\right]^{\frac{1}{q-1}}, & \text{if } x \text{ in } \Omega_q, \\ 0, & \text{otherwise} \end{cases}$$

with finite support

$$\Omega_q = \{x \in \mathbb{R}^m : (x-\mu)^\top[(p+2)C]^{-1}(x-\mu) \leq 1\}.$$

A proof is given in Leonenko and Pronzato (2010).

3.3 Tsallis entropy maximisation

In this section Tsallis distributions will be introduced that maximise the Tsallis entropy under certain constraints. As the Tsallis entropy contains the Shannon entropy as the special case $q = 1$, this section contains generalisations of parts of the foregone Section 3.1 on Shannon entropy maximisation.

3.3.1 Continuous Tsallis entropy maximisation on $\mathbb{R}^m$

The problem of maximisation of the Tsallis entropy under energy (covariance) constraint is solved by the so called Tsallis distribution. Without loss of generality, we consider only the centred case $\mu = 0$. Then the m-variate Tsallis distribution with covariance matrix $K = E[(x-\mu)(x-\mu)^\top]$ is defined for $\frac{m}{m+2} < q$ as

$$f_q(x) = A_q \left(1 - (q-1)\beta x^\top K^{-1} x\right)_+^{\frac{1}{q-1}}, \tag{24}$$

where $x_+ = \max(0; x)$ and $\beta = \frac{1}{2}q - n(1-q)$. The problem of maximisation of the Tsallis entropy under energy (covariance) constraint is solved by f_q. The energy constraint can be interpreted as a covariance constraint, leading to the following expression: (Vignat et al.; 2004)

$$f_q(x) = arg f : \underset{max}{E[xx^\top]} = KH_q(f)$$

Proof. Define a Bregman divergence D_q of two continuous distributions f and g as

$$D_q(f||g) = \text{sign}(q-1) \int \frac{f^q}{q} + \frac{q-1}{q} g^q - f g^{q-1}.$$

The positivity of $D_q(f||g)$ with equality to zero if and only if $f = g$ pointwise is a consequence of the convexity of function $x \mapsto \text{sign}(q-1)\frac{x^q}{q}$. Consider the case $q > 1$: the fact that distribution f has the same covariance

K as f_q a Tsallis distribution defined in Equation (24) can be expressed by

$$\int f_q^q = \int f_q^{q-1} f$$

so that

$$\begin{aligned}
0 &\leq D_q(f||f_q) \\
&= \text{sign}(q-1) \int \frac{f^q}{q} + \frac{q-1}{q} f_q^q - f f_q^{q-1} \\
&= \int \frac{f^q}{q} + \frac{q-1}{q} f_q^q - f_q^q \\
&= \frac{1}{q} \int f^q + (q-1) f_q^q - q f_q^q \\
&= \frac{1}{q} \int f^q - f_q^q \\
&= \frac{1}{q} \{(1 - (q-1) H_q(f)) - (1 - (q-1) H_q(f_q))\} \\
&= \frac{q-1}{q} (H_q(f_q) - H_q(f))
\end{aligned}$$

This implies that $H_q(f_q) \geq H_q(f)$, thus the Tsallis distribution has higher Tsallis entropy than any other distribution with the same covariance matrix K. The proof for the case $q < 1$ follows accordingly (Vignat et al.; 2004).

□

Tsallis distributions have the following properties (Vignat et al.; 2004):

1. *Stochastic representation.* If X is Tsallis distributed with parameter

$q < 1$ and covariance matrix K then

$$X \stackrel{\mathrm{d}}{=} \frac{CN}{A},$$

where A is a chi random variable with $d_f = -m + \frac{2}{1-q}$ degrees of freedom, independent of the Gaussian vector N, with $E[NN^\top] = I$, and $C = (m-2)K$. If Y is Tsallis distributed with parameter $q > 1$, then

$$Y \stackrel{\mathrm{d}}{=} \frac{CN}{\sqrt{A + ||N||_2^2}},$$

where A is a chi random variable with $\frac{2}{q-1} + 2$ degrees of freedom. Note that the denominator is again a chi random variable, but that contrary to the case $q < 1$, it is now dependent on the numerator.

2. *Orthogonal invariance.* This follows as a direct consequence of the invariance under orthogonal transformation of the covariance constraint. Thus Tsallis distributions can be written as

$$f_q(x) = \phi_q(x^\top K^{-1} x).$$

3. *Duality.* There is a natural bijection between the cases $q < 1$ and $q > 1$. Let X Tsallis distributed with parameter $q < 1$, $d_f = -m + \frac{2}{1-q}$ and $C = (d_f - 2)K$. Then the random variable Y which is defined to be

$$Y = \frac{X}{\sqrt{1 - X^\top C^{-1} X}}$$

is Tsallis distributed with covariance matrix $\frac{d_f-2}{d_f+2}K$ and parameter $q' > 1$ such that $\frac{1}{q'-1} = \frac{1}{1-q} - \frac{m}{2} - 1$.

After giving the entropy maximising distributions, theoretical estimators for entropy of a given sample are derived in the following section. Note that the estimated entropy can not exceed the entropy of the respective maximum entropy distribution.

4 Theoretical estimators of entropy

There is a variety of different approaches of estimating entropy in literature. The focus of this thesis is on the nearest neighbour approach of entropy estimation that will be presented in Section 4.1, but other estimators will be discussed as well. In Section 4.2 two estimators based on an m-spacing density estimate will be presented. Section 4.3 presents an estimator for quadratic entropy.

In order to define nearest neighbouring observations in a sample a notion of distance has to be introduced. The nearest neighbour estimator presented in Leonenko and Pronzato (2010) uses the Euclidean distance.

Definition 4.1. *Euclidean distance.* The Euclidean distance is the most commonly used distance metric. Denote with $p = (p_1, p_2, \ldots, p_m)^\top$ and $q = (q_1, q_2, \ldots, q_m)^\top$ vectors in $\mathbb{R}^m$. The Euclidean distance ρ of p and q is defined as

$$\begin{aligned} \rho(p,q) &= \rho(q,p) \\ &= \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + , \ldots, + (q_m-p_m)^2} \\ &= \sqrt{\sum_{i=1}^{m} (q_i - p_i)^2}. \end{aligned}$$

In two and three dimensional space this can be interpreted as the length of a line with end points p and q.

With this definition of distance the nearest neighbour estimator is derived in the following section.

4.1 Nearest neighbour estimator

This section introduces the nearest neighbour approach to estimation of entropies given in Leonenko and Pronzato (2010). The nearest neighbour estimator for information is defined in Section 4.1.1. Properties of this estimator are summarised in Section 4.1.2 for densities with bounded support and in Section 4.1.3 for densities with unbounded support. In Section 4.1.4 theoretical results concerning the convergence of the bias are presented. Finally, the theoretical estimators for the Shannon, Rényi and Tsallis entropy are given in Section 4.1.5.

4.1.1 Estimation of information

This section introduces the nearest neighbour estimator for information. As the Shannon, Rényi and Tsallis entropies are only different transformations of information, this is the essential part of the approach. Let $\rho(p, q)$ denote the Euclidean distance, see Definition 4.1, between two points $p, q \in \mathbb{R}^m$. Let $X_1, \ldots, X_N$ be a sample of size N of a random variable X. For any X_i in the sample compute the $N-1$ distances $\rho(X_i, X_j)$, $j = 1, \ldots, N$, $j \neq i$, to every other random variable in the sample. From these $N-1$ distances we form the order statistics $\rho_1^{(i)} \leq \rho_2^{(i)} \leq, \ldots, \leq \rho_{N-1}^{(i)}$. This includes that $\rho_1^{(i)}$ is the nearest neighbour distance from X_i to any other X_j, $j \neq i$, in the sample

and $\rho_k^{(i)}$ is the kth nearest neighbour distance from X_i to any other X_j, $j \neq i$, in the sample.

The information I_q, $q \neq 1$, can be estimated by (Leonenko and Pronzato; 2010)

$$\hat{I}_{N,k,q} = \frac{1}{N} \sum_{i=1}^{N} (\zeta_{N,i,k})^{1-q}, \tag{25}$$

with

$$\zeta_{N,i,k} = (N-1) C_k V_m \left(\rho_k^{(i)}\right)^m, \tag{26}$$

where $V_m = \frac{\pi^{m/2}}{\Gamma(m/2+1)}$ is the volume of the unit ball $\mathcal{B}(0,1)$ in $\mathbb{R}^m$ and

$$C_k = \left[\frac{\Gamma(k)}{\Gamma(k+1-q)}\right]^{\frac{1}{1-q}}.$$

The following lemma gives outlines the general idea of the derivation of the nearest neighbour estimator for the information.

Lemma 4.1. A Monte Carlo estimator for a known function f based on the sample $X_1, \ldots, X_N$ is

$$\hat{f} = \frac{1}{N} \sum_{i=1}^{N} f(X_i).$$

The estimator for I_q is a plug-in estimator into the Monte Carlo estimator of the form

$$\hat{I}_{N,k,q} = \sum_{i=1}^{N} \left[\hat{f}_{N,k}(X_i)\right]^{q-1},$$

with

$$\hat{f}_{N,k}(x) = 1/\left\{(N-1) C_k V_m [\rho_{k+1}(x)]^m\right\}$$

the function to be estimated. This function can be deduced via nearest neighbour estimation.

4.1.2 Densities with bounded support

For densities with bounded support the following theorems can be derived.

Theorem 4.1. *Asymptotical unbiasedness of* $\hat{I}_{N,k,q}$ *for* $q > 1$. The estimator $\hat{I}_{N,k,q}$ satisfies

$$E[\hat{I}_{N,k,q}] \to I_{N,k,q} \text{ for } N \to \infty$$

for any $q \in]1, k+1[$ for bounded densities f, see Leonenko and Pronzato (2010).

Theorem 4.2. *Asymptotical unbiasedness of* $\hat{I}_{N,k,q}$ *for* $q < 1$. Let f a continuous function on $\mathcal{X}$ and $f(x) = 0$ for $x \notin \mathcal{X}$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^m$. Additionally let the following assumption on the geometry of $\mathcal{X}$ hold

$$\text{vol}\left[\mathcal{B}(x, \theta^{\frac{1}{m}}) \bigcap \mathcal{X}\right] > \alpha \,\text{vol}\left[\mathcal{B}(x, \theta^{\frac{1}{m}})\right] = \alpha V_m \theta$$

for all $x \in \mathcal{X}$, for all $\theta \in [0, B_x]$ and for some $\alpha > 0$. Then the bias of $\hat{I}_{N,k,q}$ tends to zero for $N \to \infty$ if f is bounded from below, see Leonenko and Pronzato (2010). If f is not bounded from below $\hat{I}_{N,k,q}$ is asymptotically unbiased for any $q \in (-1, 1)$ and any k when f is continuously differentiable on $\mathcal{X}$ and $p < 1$. For larger values of p an upper bound for admissible values is given by $(1 + qp^2)p(p-1)$.

Take $q = \frac{1}{2}$ as an example. Then $k = 2$ can be used for $p < \frac{\sqrt{10}+2}{3} \approx 1.72$ to gain an asymptotically unbiased estimator. If p is any larger than that, $k = 1$ is necessary.

Theorem 4.3. *Consistency of* $\hat{I}_{N,k,q}$. The estimator $\hat{I}_{N,k,q}$ satisfies

$$\hat{I}_{N,k,q} \xrightarrow{L_2} I_q \text{ for } N \to \infty$$

and thus

$$\hat{I}_{N,k,q} \xrightarrow{p} I_q \text{ for } N \to \infty$$

for any $q \in \left]1, \frac{k+1}{2}\right[$ for bounded f when $k \geq 2$, or $k = 1$ for $q \in \left]1, \frac{3}{2}\right[$ (Leonenko and Pronzato; 2010).

In order to prove Theorem 4.1 and Theorem 4.3 the following lemmas are needed.

Lemma 4.2. The following properties of I_q are shown in Leonenko and Pronzato (2010)

1. $I_q < \infty$ for $q > 1$ and bounded f.
2. $I_q < \infty$ for some $q < 1$ implies $I_{q'} < \infty$ for any $q' \in]q, 1[$.
3. $I_q < \infty$ for any $q \in [0, 1[$ if f is of finite support.

Lemma 4.3. Let $g \in L_1(\mathbb{R}^m)$. For any sequence of open balls $\mathcal{B}(x, R_k)$ with radius $R_k \to 0$ as $k \to \infty$ and for μ-almost any $x \in \mathbb{R}^m$,

$$\lim_{k\to\infty} \frac{1}{V_m R_k^m} \int_{\mathcal{B}(x,R_k)} g(t)dt = g(x)$$

Lemma 4.4. For any $\beta > 0$,

$$\begin{aligned} \int_0^\infty x^\beta F(dx) &= \beta \int_0^\infty x^{\beta-1}[1 - F(x)]dx \\ &\text{and} \\ \int_0^\infty x^{-\beta} F(dx) &= \beta \int_0^\infty x^{-\beta-1} F(x)dx, \end{aligned}$$

in the sense that if one side converges so does the other.

Proof. Proof of Theorem 4.1. As all X_i, $i = 1, \ldots, N$ are i.i.d.,

$$E\left[\hat{I}_{N,k,q}\right] = E\left[\zeta_{N,i,k}^{1-q}\right] = E\left[E\left[\zeta_{N,i,k}^{1-q} | X_i = x\right]\right],$$

with $\zeta_{N,i,k}$ defined as in Equation (26). The corresponding distribution function conditional to $X_i = x$ is

$$F_{n,x,k}(u) = P\left(\zeta_{N,i,k} < u | X_i = x\right) = P\left(\rho_k^{(i)} < R_N(u) | X_i = x\right),$$

where $R_N(u) = \left(\frac{u}{(N-1)V_m C_k}\right)^{\frac{1}{m}}$ and V_m and C_k defined as for Equation (26).

Let $\mathcal{B}(x,r)$ be the open ball with center x and radius r. Then

$$\begin{aligned} F_{N,x,k}(u) &= P\{k \text{ elements or more in } \mathcal{B}(x, R_N(u))\} \\ &= \sum_{j=k}^{N-1} \binom{N-1}{j} p_{N,u}^j (1-p_{N,u})^{N-1-j} \\ &= 1 - \sum_{j=0}^{k-1} \binom{N-1}{j} p_{N,u}^j (1-p_{N,u})^{N-1-j}, \end{aligned}$$

where $p_{N,u} = \int_{\mathcal{B}(x,R_N(u))} f(t)dt$. From the Poisson approximation of the binomial distribution, Lemma 4.3 gives

$$F_{n,x,k}(u) \to F_{x,k}(u) = 1 - \exp(\lambda u) \sum_{j=0}^{k-1} \frac{(\lambda u)^j}{j!}, \text{ as } N \to \infty$$

for μ-almost any x, with $\lambda = \frac{f(x)}{C_k}$. This means that $F_{n,x,k}(u)$ tend to the Erlang-distribution $F_{x,k}(u)$ with p.d.f.

$$f_{x,k} = \frac{\lambda^k u^{k-1} \exp(-\lambda u)}{\Gamma(k)}.$$

With straightforward calculation it can be shown that for the case that $q < 1$ (Leonenko and Pronzato; 2010)

$$\int_0^\infty u^{1-q} f_{x,k}(u)du = \frac{\Gamma(k+1-q)}{\lambda^{1-q}\Gamma(k)} = f^{q-1}(x)$$

for any $q < k+1$.

For the case $1 < q < k+1$ recall that due to Lemma 4.2.1 $I_q < \infty$. Let

$$J_N = \int_0^\infty u^{(1-q)(1+\delta)} F_{N,x,k}(du).$$

With Theorem 2.5.1 of Bierens (1996), page 34,

$$\begin{aligned} z_{N,k}(x) &= \int_0^\infty u^{1-q} F_{N,x,k}(du) \\ \text{converges to } z_k(x) &= \int_0^\infty u^{1-q} F_{x,k}(du) = f^{q-1}, \text{ as } N \to \infty \end{aligned} \tag{27}$$

for μ-almost any x in $\mathbb{R}^m$.

Define $\beta = (1-q)(1+\delta)$ so that $\beta < 0$ and take $\delta < \frac{k+1-q}{q-1}$ so that $\beta + k > 0$. With Lemma 4.4 one can deduce that

$$\begin{aligned} J_N &= -\beta \int_0^\infty u^{\beta-1} F_{N,x,k}(u)du \\ &= -\beta \int_0^1 u^{\beta-1} F_{N,x,k}(u)du - \beta \int_1^\infty u^{\beta-1} F_{N,x,k}(u)du \\ &\leq -\beta \int_0^1 u^{\beta-1} F_{N,x,k}(u)du - \beta \int_1^\infty u^{\beta-1} du \\ &= 1 - \beta \int_0^1 u^{\beta-1} F_{N,x,k}(u)du. \end{aligned} \tag{28}$$

Let $\bar{f}(x)$ (or shortened $\bar{f}$) be a boundary of $f(x)$, then $\forall x \in \mathbb{R}^m, \forall u \in \mathbb{R}, \forall N$,

$$\bar{f}(x) V_m R_N(u)^m = \frac{\bar{f}(x) u}{(N-1) C_k},$$

implying

$$\begin{aligned}
\frac{F_{N,x,k(u)}}{u^k} &\leq \sum_{j=k}^{N-1} \binom{N-1}{j} \frac{\bar{f}^j u^{j-k}}{C_k^j (N-1)^j} \\
&\leq \sum_{j=k}^{N-1} \frac{\bar{f}^j u^{j-k}}{C_k^j j!} \\
&= \frac{\bar{f}^k}{C_k^k k!} + \sum_{j=k+1}^{N-1} \frac{\bar{f}^j u^{j-k}}{C_k^j j!} \\
&\leq \frac{\bar{f}^k}{C_k^k k!} + \frac{\bar{f}^k}{C_k^k} \sum_{j=1}^{N-k-1} \frac{\bar{f}^j u^j}{C_k^j j!} \\
&\leq \frac{\bar{f}^k}{C_k^k k!} + \frac{\bar{f}^k}{C_k^k} \sum_{j=1}^{\infty} \frac{\bar{f}^j u^j}{C_k^j j!} \\
&= \frac{\bar{f}^k}{C_k^k k!} + \frac{\bar{f}^k}{C_k^k} \left\{ \exp\left(\frac{\bar{f} u}{C_k}\right) - 1 \right\}
\end{aligned}$$

and for $u < 1$

$$\frac{F_{N,x,k(u)}}{u^k} < U_k = \frac{\bar{f}^k}{C_k^k k!} + \frac{\bar{f}^k}{C_k^k} \left\{ \exp\left(\frac{\bar{f} u}{C_k}\right) - 1 \right\}.$$

Using Equation (28), it follows that

$$J_N \leq 1 - \beta U_k \int_0^1 u^{k+\beta-1} du = 1 - \frac{\beta U_k}{k+\beta},$$

implying Equation (27). The convergence of

$$\int_{\mathbb{R}^m} z_{N,k}(x) f(x) dx \to \int_{\mathbb{R}^m} z_k(x) f(x) dx = I_q, \text{ as } N \to \infty$$

follows from Lebegues's bounded convergence theorem, because $z_{N,k}(x)$ is bounded, take $\delta = 0$ in J_N, completing the proof of Theorem 4.1.

□

4.1.3 Densities with unbounded support

The case of f being of unbounded support is an important case to consider for $q < 1$. Define

$$r_c(f) = \sup \left\{ \int_{\mathbb{R}^m} |x|^r f(x)dx < \infty \right\}$$

so that $E[|X_i|^r] < \infty$ for $r < r_c(f)$ and $E[|X_i|^r] = \infty$ for $r > r_c(f)$. Using this notation one can obtain the following statements concerning asymptotic unbiasedness.

Theorem 4.4. If $0 < q < 1$

1. If $I_q < \infty$ and $r_c(f) > m\frac{1-q}{1}$, then

$$E[\hat{I}_{N,k,q}] \to I_q, \text{ as } N \to \infty$$

2. If $I_q < \infty, q > \frac{1}{2}$ and $r_c(f) > 2m\frac{1-q}{2q-1}$, then

$$E[\hat{I}_{N,k,q} - I_q]^2 \to 0, \text{ as } N \to \infty$$

4.1.4 Theoretical approximations of the bias

This section summarises the results concerning the convergence of the bias of the nearest neighbour estimator given in Leonenko and Pronzato (2010).

Lemma 4.5. With the assumption of f being three times continuously differentiable $\mu_{\mathcal{L}}$-almost everywhere, the following equation can be shown,

$$\frac{1}{V_m R^m} \int_{\mathcal{B}(x,R)} f(z)dz = f(x) + \frac{R^2}{2(m+2)} \sum_{i=1}^{m} \frac{\partial^2 f(x)}{\partial x_i^2} + o(R^2), R \to 0,$$

which can be used for approximating $F_{N,x,k}(u) - F_{x,k}(u)$ in the proof of Theorem 4.1.

Theorem 4.5. In Leonenko and Pronzato (2010) some approximations of the bias

$$\begin{aligned} \hat{B}_{N,k,q} &= E\left[\hat{I}_{N,k,q}\right] - I_q \\ &= E\left[\zeta_{N,k,q}^{1-q}\right] - I_q \end{aligned}$$

are given under the conditions of Lemma 4.5 as:

$$\hat{B}_{N,k,q} = \begin{cases} \frac{(q-1)(2-q)I_q}{2N} & + \quad \mathcal{O}\left(\frac{1}{N^2}\right), \quad \text{for } m = 1, \\ \frac{(q-1)}{N}\left[\frac{(k+1-q)J_{q-2}}{8\pi} + \frac{(2-q)I_q}{2}\right] & + \quad \mathcal{O}\left(\frac{1}{N^{\frac{3}{2}}}\right), \quad \text{for } m = 2, \\ \frac{(q-1)}{N^{\frac{2}{m}}} \frac{\Gamma\left(k+1+\frac{2}{m}-q\right)}{D_m \Gamma(k+1-q)} J_{q-1-2/m} & + \quad \mathcal{O}\left(\frac{1}{N^{\frac{3}{m}}}\right), \quad \text{for } m \geq 3, \end{cases}$$

where

$$\begin{aligned} J_\beta &= f^\beta(x) \sum_{i=1}^{m} \frac{\partial f(x)}{\partial x_i^2} dx \text{ and} \\ D_m &= 2(m+2) V_m^{\frac{2}{m}} \end{aligned}$$

Lemma 4.6. Some of the most important properties of I_q are (Leonenko and Pronzato; 2010)

1. $I_q < \infty$ for $q > 1$ if f is bounded.
2. $I_{q'} < \infty$ for any $q' \in]q, 1[$ if $I_q < \infty$ for some $q < 1$.
3. $I_q < \infty$ for any $q \in [0, 1[$ if f is of finite support.

4.1.5 Estimation of entropy

With the nearest neighbour estimator for the information given in Equation (25) the following theorems can be derived.

Theorem 4.6. *Nearest neighbour estimator of Rényi entropy.* Under the conditions of Theorem 4.3

$$\hat{H}^*_{N,k,q} = \log(\hat{I}_{N,k,q})/(1-q) \xrightarrow{L_2} H^*_q \tag{29}$$

is an estimator for the Rényi entropy using the estimator of the information I_q proposed in equation (25).

Theorem 4.7. *Nearest neighbour estimator of Tsallis entropy.* Similarly to the Rényi entropy the estimator for the Tsallis entropy

$$\hat{H}_{N,k,q} = (1 - \hat{I}_{N,k,q})/(q-1) \xrightarrow{L_2} H_q \tag{30}$$

is another straight-forward transformation of the nearest neighbour estimator of I_q in equation (25) under the conditions of Theorem 4.3 (Leonenko and Pronzato; 2010).

Theorem 4.8. *Nearest neighbour estimator of Shannon entropy.* For the estimation of H_1, we take the limit of the Rényi entropy $\hat{H}^*_{N,k,q}$ as $q \to 1$ which gives

$$\hat{H}_{N,k,1} = \frac{1}{N} \sum_{i=1}^{N} \log \xi_{N,i,k}, \tag{31}$$

with

$$\xi_{N,i,k} = (N-1) \exp\left[-\Psi(k)\right] V_m \left(\rho^{(i)}_{k,N-1}\right)^m,$$

where $\Psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function, $\Psi(1) = -\mu$ with $\mu = 0.5772$ the Euler-Mascheroni constant and for $k \in \mathcal{N}$, $\Psi(k) = -\mu + A_{k-1}$ with $A_0 = 0$ and $A_j = \sum_{i=1}^{j} 1/j$ (Leonenko and Pronzato; 2010).

Another way of estimating the Shannon entropy for one-dimensional random variables using m-spacings is presented in the following section.

4.2 Spacing estimators

These estimators are based on a density estimate using sample-spacings. As sample spacings are only defined for the one-dimensional case, spacing estimators only work for one-dimensional random variables. The estimators for the Shannon entropy presented in Beirlant et al. (1997) and Song (2000) will be presented in Section 4.2.1 and 4.2.2 respectively. For the Rényi entropy biased estimators exist in literature, see e.g. Hegde et al. (2005), but we will focus on the estimation of the most important special case, the Shannon entropy.

Let $\{X_1, \ldots, X_N\}$ an i.i.d. sample of a real valued random variable X. The corresponding order statistics is denoted by $\{X_{(1)}, \ldots, X_{(N)}\}$. An m-spacing is defined as $[X_{(i+m)} - X_{(i)}]$, for $(1 \leq i < i + m \leq N)$. The corresponding density estimate is

$$f_n(x) = \frac{m}{N} \frac{1}{X_{(im)} - X_{(im-m)}}$$

if $x \in [X_{(im-m)}, X_{(im)}]$. A consistent density estimator can be achieved if m is chosen dependent on the sample size N with the properties $m_N \to \infty$ and $\frac{m_N}{N} \to 0$ as $N \to \infty$.

An estimator for Shannon entropy based on this spacing density estimator can be constructed as a plug-in estimator. The following m-spacing based estimate presented in this section is constructed from an inconsistent spacing density estimator, but is nonetheless consistent (Beirlant et al.; 1997).

4.2.1 m-spacing estimator

Let us consider the m-spacing estimator for a fixed m: (Beirlant et al.; 1997)

$$H_{m,N} = \frac{1}{N}\sum_{i=1}^{N-m} \log\Big(\frac{n}{m}(X_{(i+m)} - X_{(i)})\Big) - \Psi(m) + log(m),$$

where $\Psi(x) = \frac{d-(\log(\Gamma(x))}{dx}$ is the digamma function. This implies that the corresponding density estimate is inconsistent. Consequential, an additional term correcting the asymptotic bias is included in the formula for $H_{m,N}$. In Cressie (1976) root-n consistency of the form of asymptotic normality,

$$\lim_{n\to\infty} \sqrt{n}(H_{m,N} - H(f)) \xrightarrow{D} N(0, \sigma^2),$$

is proven for bounded distributions f under the tail condition $\inf_{f(x)>0} f(x) > 0$. The asymptotic variance is given as

$$\sigma^2 = (2m^2 - 2m + 1)\Psi'(m) - 2m + 1 + V[\log(f(x))].$$

For the special case $m = 1$ this simplifies to

$$\sigma^2 = \frac{\pi^2}{6} - 1 + V[\log(f(x))].$$

4.2.2 m_n-spacing estimator

An improved estimator based on sample spacings can be gained by choosing the the difference between two sample quantiles $2m$ dependent of the sample size n. Such an estimator is given in Song (2000) as

$$H_{m_n} := \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\right).$$

Let us introduce several notations that will be used throughout this section. Let $R_m := \sum_{j=1}^{m} \frac{1}{j}$ and $\gamma := \lim_{n\to\infty}(R_n - \log n)$ be the Euler-Mascheroni constant. Let $\phi(F) := \sup\{x : F(x) = 0\}$ define the lower end point of F and $\psi(F) := \inf\{x : F(x) = 0\}$ define the upper end point of F.

Let us define the following assumptions:

(I) $E[\log^2 f(X) < \infty]$

(II) $\sup_{\phi(F)<x<\psi(F)} F(x)(1 - F(x))\frac{|f'(x)|}{f^2(x)} < \infty$,
where $f'(x)$ denotes the first derivative with respect to x.

(III)

$$\begin{aligned} \frac{m}{\log(n)} &\to \infty \\ \frac{m\log(n)}{n} &\to 0 \text{ as } n \to \infty \end{aligned}$$

(IV)

$$\begin{aligned} \frac{m}{\log(n)} &\to \infty \\ \frac{m(\log(n)^{2/3}}{n^{1/3}} &\to 0 \text{ as } n \to \infty \end{aligned}$$

Note that choosing $m = n^{1/4}$ satisfies properties (III) and (IV). This will be used for the simulation in Section 5.6. This will be used to implement the estimator in R.

Let us now summarise the most important properties in the following theorems.

Theorem 4.9. Under the assumptions (I), (II), and (IV) the following limit theorem can be shown,

$$\sqrt{(}n)(H_{m_n} - H(F) + \log(2m) + \gamma - R_{2m_1}) \xrightarrow{D} N(0, \sigma^2(F),$$

where $\sigma^2(F) = V[\log(f(x))]$, see Song (2000).

This central limit theorem is of great importance, as it makes it possible not only to estimate the Shannon entropy nonparametrically, but also to construct confidence intervals. This allows hypothesis testing as well.

Theorem 4.10. Under the slightly different set of assumptions (II) and (III) and $E[\log^- f(X) < \infty]$ the convergence in probability is shown in Song

(2000),

$$H_{m_n} \xrightarrow{P} H(F).$$

A special case of the Rényi and Tsallis entropy being discussed in literature is the quadratic entropy. Consistent estimators can be found using another approach which is presented in the following section.

4.3 Estimation of quadratic entropy

For the estimation of quadratic entropy a variety of asymptotic properties can be obtained. In Leonenko and Seleznjev (2010) and Källberg et al. (2014) a U-statistic estimator is presented that is based on the number of ϵ-close vectors. It is used to estimate the entropy of the marginal distribution of a stationary d-dependent series, which is a weaker assumption than interdependency of the series. The following notations will be used. The dimension is denoted by d, in order to not be confused with m-dependency. The Lebesgue space or L_p-space of real valued functions in $\mathbb{R}^d$ is denoted by $L_a(\mathbb{R}^d),\ a \geq 1$. Let P the distribution of a random variable X with density $p(\cdot) \in L_2(\mathbb{R}^d)$ and quadratic Rényi entropy $H_2^*(X)$. For $x, y \in \mathbb{R}^d$ let $d(x, y) = ||x - y||$ the Euclidean distance in $\mathbb{R}^m$, see Definition 4.1 . Using this, two vectors x and y are ϵ-close if $d(x, y) \leq \epsilon$. Denote the ϵ-ball with center x with

$$B_\epsilon(x) := \{y : d(x, y) \leq \epsilon\}.$$

Its volume $b_\epsilon(m)$ is given as $b_\epsilon(m) = \epsilon^m b_1(m)$ where $b_1(m)$ denotes the volume of the m-dimensional unit ball $b_1(m) = \frac{2\pi^{m/2}}{(m\Gamma(d/2))}$. The ϵ-coincidence probability of independent X and Y with common distribution P is defined as

$$q_{2,\epsilon} := P(d(X,Y) < \epsilon) = E\left[p_{X,\epsilon}(Y)\right]$$

with the ϵ-ball probability $p_{X,\epsilon} := P(X \in B_\epsilon(x)),\ x \in \mathbb{R}^m$.

The following assumptions are made about the distribution of the random vector $\{X_i\}, i = 1, \ldots, \infty$:

1. *Finite-dimensionality.* The distribution of all X_i is d-dimensional with $d < \infty$.

2. *Stationary sequence.* The $\{X_i\}$ are a random sequence whose joint probability distribution is invariant over time.

3. *m-dependency.* In a series of random vectors $X_1, X_2, \ldots$, that is taken from the random variable, the vectors X_i and X_j are independent if $|i - j| > m$.

This set of assumptions on the distribution of the $X_i, i = 1, \ldots, \infty$ are valid throughout the entire Section 4.3.

The following assumptions will be referred to as $\mathcal{A}_1$ in this section:

1. Let the marginal distribution $p(\cdot)$ fulfil $p(\cdot) \in L_3(\mathbb{R}^d)$.

2. Each four-tuple of positive and unique integers $t = (t_1, t_2, t_3, t_4)$, the random vector $(X_{t_1}, X_{t_2}, X_{t_3}, X_{t_4})$ has distribution P with density

$p_t(x_1, x_2, x_3, x_4)$ satisfying

$$g_t(x_1, x_2) = \left(\int_{\mathbb{R}^{2m}} p_t(x_1, x_2, x_3, x_4)^2 dx_3 dx_4 \right)^{\frac{1}{2}}.$$

for $g_t(\cdot, \cdot) \in L_1(\mathbb{R}^{2m})$.

Let $\epsilon = \epsilon(n) \to 0$ as $n \to \infty$ and $|C|$ denote the cardinality of the finite set C. Introduce the random variable N_n counting the ϵ-close observations in the sample $X_1, \ldots, X_n$ be defined as

$$\begin{aligned} N_n = N_{n,\epsilon} &= \binom{n}{2} Q_n \\ &:= |\{d(X_i, X_j) \leq \epsilon,\ i, j = 1, \ldots, n,\ (i < j)\}| \\ &= \sum_{i<j} I(d(X_i, X_j) \leq \epsilon), \end{aligned}$$

where $I(D)$ is an indicator for an event D. Here Q_n is a U-statistic of Hoeffding with varying kernel. Write $U_n = O_p(1)$ as $n \to \infty$ for a sequence of random variables $U_n, n \in \mathbb{N}$ if for any $\delta > 0$ and large enough $n \geq 1$, there exists $C > 0$ such that $P(|U_n| > C) \leq \delta)$. For a numerical sequence $\nu_n, n \geq 1$, let $U_n = O_p(\nu_n)$ as $n \to \infty$.

Some asymptotic properties for estimation of the quadratic Rényi entropy under m-dependence and the distributional assumptions $\mathcal{A}_1$, are now presented.

Let us look into the asymptotic distribution of the number of small interpoint distances N_n. This is random variable with expectation value $\mu_n =$

$\mu_{n,\epsilon} := E[N_n]$ and variance $\sigma_n^2 = \sigma_{n,\epsilon}^2 := V[N_n]$. For $h = 0, 1, \ldots,$ we introduce the characteristic $\sigma_{1,h,\epsilon}^2 := \mathrm{Cov}\left[p_{X,\epsilon}(X_1), p_{X,\epsilon}(X_{1+h})\right]$. Define

$$\zeta_{1,m} = \lim_{n\to\infty} \frac{1}{n} V\left[\sum_{i=1}^{n} p(X_i)\right] = V\left[p(X_1)\right] + 2\sum_{h=1}^{m} \mathrm{Cov}\left[p(X_1), p(X_{1+h})\right]. \quad (32)$$

Lemma 4.7. Under the set of assumptions $\mathcal{A}_1$ the following holds:

1. The expectation value μ_n and variance σ_n^2 of N_n fulfill

$$\begin{aligned} \mu_n &= \binom{n}{2} q_{2,\epsilon} + o(n\epsilon^{d/2}) \\ \sigma_n^2 &= \frac{n^2}{2} q_{2,\epsilon} n^3 \left(\sigma_{1,0,\epsilon}^2 + 2\sum_{h=1}^{m} \sigma_{1,h,\epsilon}^2\right) + o(n\epsilon^{d/2}) + o(n^2\epsilon^d) \text{ as } n \to \infty \end{aligned}$$

2. For $n^2\epsilon^2 \to a,\ 0 < a \leq \infty$, and $\zeta_{1,m} > 0$ when $\sup_{n\geq 1}\{n\epsilon^d\} = \infty$, then

$$\begin{aligned} \mu_n &\sim \frac{1}{2} b_1(d) q_2 n^2 \epsilon^d, \\ \sigma_n^2 &\sim \frac{1}{2} b_1(d) q_2 n^2 \epsilon^d + b_1(d)^2 \zeta_{1,m} n^3 \epsilon^{2d} \text{ as } n \to \infty. \end{aligned}$$

Theorem 4.11. Under $\mathcal{A}_1$ the following holds:

1. If $n^2\epsilon^d \to 0$, then $N_n \xrightarrow{D} 0$ as $n \to \infty$.

2. If $n^2\epsilon^d \to a$, $0 < a \leq \infty$, then $\mu = \lim_{n\to\infty} \mu_n$ and

$$N_n \xrightarrow{D} Po(\mu) \text{ as } n \to \infty.$$

3. If $n^2\epsilon^d \to \infty$ and $n\epsilon^d \to a$, $0 < a \leq \infty$, and $\zeta_{1,m} > 0$ when $a = \infty$, then

$$\frac{N_n - \mu_n}{\sigma_n} \xrightarrow{D} N(0,1) \text{ as } n \to \infty.$$

Let us now design an estimator of the quadratic Rényi entropy based on N_n. To begin with, we estimate the quadratic functional $q_2 = q_2(P) := \int_{\mathbb{R}^d} p(x)^2 dx$ with

$$\tilde{Q}_n = \tilde{Q}_{n,\epsilon} = \frac{\binom{n}{2}^{-1} N_n}{b_\epsilon(d)}.$$

The corresponding plug-in estimator of the quadratic Rényi entropy H_2^* is

$$\begin{aligned} \tilde{H}_n^* &= \min\left\{-\log\left(\tilde{Q}_n\right), -\log\left(\frac{1}{n}\right)\right\} \\ &= -\log\left(\max\left\{\tilde{Q}_n, \frac{1}{n}\right\}\right), \end{aligned}$$

with varying asymptotic behaviour depending on the rate of decreasing of $\epsilon(n)$. Two examples for different rates of decreasing of $\epsilon(n)$ are given here.

1. Let $\mathcal{A}_1$ hold and $n^2\epsilon^d \to \infty$, then

$$\begin{aligned} \tilde{Q}_n &\xrightarrow{m.s} q_2 \text{ and} \\ \tilde{H}_n^* &\xrightarrow{P} H_2^* \text{ as } n \to \infty. \end{aligned}$$

2. Let $p(\cdot) \in L_3(\mathbb{R}^d)$ and $X_i \neq X_j \forall i \neq j$. For $n\epsilon^d \to a,\ 0 < a \leq \infty$, then

$$\begin{aligned} \tilde{Q}_n &\xrightarrow{m.s} q_2 \text{ and} \\ \tilde{H}_n^* &\xrightarrow{P} H_2^* \text{ as } n \to \infty. \end{aligned}$$

Theorem 4.12. Define $\tilde{q}_{2,\epsilon} = q_{2,\epsilon}/b_\epsilon(d)$ and $\tilde{H}_2^* = -\log \tilde{q}_{2,\epsilon} + \log(b_\epsilon(d))$. Let $\nu = 2q_2/b_1(d)$ and $\zeta_{1,m}$ as in Equation (32). Let $\mathcal{A}_1$ hold and $n^2\epsilon^d \to \infty$.

1. For $n\epsilon^d \to a,\ 0 < a \leq \infty$, and $\zeta_{1,m} > 0$ when $a = \infty$, then

$$\begin{aligned} \sqrt{n}(\tilde{Q}_n - \tilde{q}_{2,n}) &\xrightarrow{D} N(0, \frac{\nu}{a} + 4\zeta_{1,m}) \text{ and} \\ \sqrt{n}\tilde{Q}_n(\tilde{H}_n - \tilde{H}_{2,\epsilon}) &\xrightarrow{D} N(0, \frac{\nu}{a} + 4\zeta_{1,m}) \text{ as } n \to \infty. \end{aligned}$$

2. If $n\epsilon^d \to 0$, then

$$\begin{aligned} n\epsilon^{\frac{d}{2}}(\tilde{Q}_n - \tilde{q}_{2,\epsilon}) &\xrightarrow{D} N(0, \nu) \text{ and} \\ n\epsilon^{\frac{d}{2}}\tilde{Q}_n(\tilde{H}_n - \tilde{H}_{2,\epsilon}) &\xrightarrow{D} N(0, \nu) \text{ as } n \to \infty. \end{aligned}$$

Define

$$\begin{aligned} U_{h,n} &= U_{h,n,\epsilon_0} \\ &:= M_{h,n}^{-1} b_{\epsilon_0}(d)^{-2} \sum_{(i,j,k)\in\Xi_{h,n}} I(d(X_i, X_j) \leq \epsilon_0, d(X_{i+h}, X_k \leq \epsilon_0),\ \epsilon_0 > 0 \end{aligned}$$

where $\Xi_{h,n} := (i,j,k) : 1 \leq i \leq n-(h+1), \quad j,k \neq i, \quad i+h, j \neq k$ and the number of summands $M_{h,n} := |\Xi_{h,n}| = (n-(h=1))(n-2)(n-3)$. Let $\epsilon = \epsilon_0(n) \to 0$ as $n \to \infty$, then $z_{1,r,n}$ denotes a consistent plug-in estimator for $\zeta_{1,r,n}$ where

$$z_{1,r,n} := U_{0,n} - \tilde{Q}_n^2 + 2\sum_{h+1}^{r} \left(U_{h,n} - \tilde{Q}_n^2 \right),$$

under the assumption that the sequence $\epsilon_0 = \epsilon_0(n)$ satisfies $n\epsilon_0^{3d} \to c, \ 0 < c \leq \infty$.

Theorem 4.13. Define a consistent estimator $w_{r,n}^2$ for $\nu/a + 4\zeta_{1,m}$ for $n\epsilon^d \to a, 0 \leq \infty$ as

$$w_{r,n}^2 := \frac{2\tilde{Q}_n}{nb_\epsilon(d)} = 4\max\left(z_{1,r,n}, \frac{1}{n} \right).$$

Denote with $H_2^{(\alpha)}(K), 0 < \alpha \leq 1, K > 0$ a space of functions in $\mathbb{R}^d$ satisfying an α Hölder condition in L_2-norm with constant K, which is a smoothness condition.

Suppose $\mathbb{A}_1$ holds and $p(\cdot) \in H_2^{(\alpha)}(K), \ \alpha > \frac{d}{4}$, and $r \geq m$. If $\epsilon \sim L(n)n^{-\frac{1}{d}}$ and $n\epsilon^d \to a, \ 0 < a \leq \infty$, and $\zeta_{1,m} > 0$ when $a = \infty$, then

$$\begin{aligned} \frac{\sqrt{n}(\tilde{Q}_n - q_2)}{w_{r,n}} &\xrightarrow{D} N(0,1) \text{ and} \\ \frac{\sqrt{n}\tilde{Q}_n(\tilde{H}_n - \tilde{H}_{2,\epsilon})}{w_{r,n}} &\xrightarrow{D} N(0,1) \end{aligned}$$

is a consistent estimator estimator of $\frac{\nu}{a} + 4\zeta_{1,m}$ when $n\epsilon^d \to a, \ 0 < a \leq \infty$.

Theorem 4.14. Under the set of assumptions $\mathbb{A}_1$ and the assumption $p \in H_2^{(\alpha)}(K)$ and $n^2\epsilon^d \to \infty$:

1. If $\alpha > (d/4)C_\beta$ for some $0 < \beta < 1$ and $\epsilon \sim cn^{-(2-\beta)/d}, c > 0$, then

$$\begin{aligned} \frac{n^{\beta/2}c^{d/2}(\tilde{Q}_n - q_2)}{u_n} &\xrightarrow{D} N(0,1) \text{ and} \\ \frac{n^{\beta/2}c^{d/2}\tilde{Q}_n(\tilde{H}_n - H_2^*)}{u_n} &\xrightarrow{D} N(0,1) \text{ as } n \to \infty \end{aligned}$$

2. If $\epsilon \sim L(n)^{2/d}n^{-2/d}$, then

$$\begin{aligned} \frac{L(n)(\hat{Q}_n - q_2)}{u_n} &\xrightarrow{D} N(0,1) \text{ and} \\ \frac{L(n)\hat{Q}_n(\hat{H}_n - H_2^*)}{u_n} &\xrightarrow{D} N(0,1) \text{ as } n \to \infty \end{aligned}$$

Note that the practical applicability strongly depends on the choice of ϵ for a given sample size n. An optimal choice of this parameter has not been found yet.

4.4 Overview of the theoretical estimators

In Table 2 a summary of the type of entropy (Shannon, Rényi and Tsallis entropy) that can be estimated with the estimators presented above is given.

Note that $m \in \mathbb{N}$ and the Rényi and Tsallis entropy simplify to the Shannon entropy for $q = 1$. For $q = 2$ one speaks of the quadratic Rényi or Tsallis entropy.

Table 2: Overview of the applicability of the estimators

Distribution	Shannon	Rényi	Tsallis
Nearest neighbour	m-dim	m-dim for $q \in \mathbb{R}^+$	m-dim for $q \in \mathbb{R}^+$
Spacing estimator	one-dim	one-dim for $q = 1$	one-dim for $q = 1$
U-statistic estimator	–	m-dim for $q = 2$	m-dim for $q = 2$

5 Simulation study

In this thesis simulation studies are utilised to affirm and expand theoretical results and findings presented earlier in this thesis. The most important qualities concerning expectation value, bias and variance are verified. Additionally new insights can be gained by simulation that could not be proven theoretically so far, for example the rate of convergence of the bias.

In statistics a simulation study usually implies a Monte Carlo simulation. These are a broad class of computational algorithms based on repeated random sampling in order to gain numerical results. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use calculus or other analytical methods. A typical Monte Carlo simulation as done in this thesis is composed of the following steps (Davidian; 2005):

1. Generate N independent data sets from a distribution of choice with a set of fixed parameters.

2. Compute the numerical value of the estimator $T(data)$ for each set of samples $T_1, \ldots, T_N$.

3. For large N, summary statistics of $T_1, \ldots, T_N$ are good approximations of the true sampling properties of the estimator for the distribution of choice with given parameters.

Example 5.1. Let us estimate the expected value of the estimator T. Let T_i the value of T from the i^{th} data set, $i = 1, \ldots, N$. The sample mean over all N data sets is an estimate of the true mean of the sampling distribution of the estimator,

$$\hat{E}[T] = \frac{1}{N} \sum_{i=1}^{N} T_i.$$

An estimator for the variance and approximate confidence intervals for the estimator can be gained analogously.

Due to the vast number of calculations and the generation of (pseudo) random samples that have to be conducted in order to get accurate results, Monte Carlo simulation studies are sensibly only executed by computers. Here the software package R is used, see Appendix A. The summary statistics of the random experiments are described using linear models (Section 5.1) or non-linear models (Section 5.2). The Sections 5.5 and 5.4 explain how the bias and the variance of an estimator are modelled from the results of a simulation. These methods will be used in Section 5.6 in order to compare the nearest neighbour estimator with the estimator defined in Section 4.2 for the one-dimensional Shannon entropy as well as in Section 5.7 where the nearest neighbour estimator is compared with the estimator presented in Section 4.3 for the quadratic entropy. In Section 5.8 the theoretical findings concerning the bias of the nearest neighbour estimator are surveyed.

5.1 Linear model

The linear model is a tool that will be used to describe the convergence of the estimators. Suppose we are given a variable of primary interest y and we aim to model the relationship between this response variable y and a set of explanatory variables $\underline{x} = x_1, \ldots, x_p$. In general, we model the relationship between y and $x_1, \ldots, x_p$ with a function $f(x_1, \ldots, x_p, \underline{\beta})$. This relationship is not exact, as it is affected by random noise ε that is usually assumed to be additive errors. Thus the model is

$$y_i = f(\underline{x}_i, \underline{\beta}) + \varepsilon_i, \ i = 1, \ldots, N.$$

Our goal is to estimate the unknown function f, that means to separate the systematic component f from random noise. Within the framework of linear models, the following specific assumptions regarding the unknown function f and the noise are made (Fahrmeir et al.; 2007):

1. *The systematic component f is a linear combination of covariates.* The unknown function $f(x_1, \ldots, x_p)$ is modelled as a linear combination of covariates, i.e.,

$$f(x_1, \ldots, x_p, \underline{\beta}) = \beta_0 + \beta_1 x_1 +, \ldots, +\beta_p x_p$$

The parameters $\underline{\beta} = \beta_0, \ldots, \beta_p$ are unknown and need to be estimated. The parameter β_0 represents the intercept. In the linear model $\underline{\beta}$ is

chosen so that the squared sum of the residuals in minimised, thus

$$\sum_{i=1}^{N}(y_i - f(x_i, \underline{\beta})) \to \min_{\underline{\beta}}.$$

For a one-dimensional example see Figure 4.

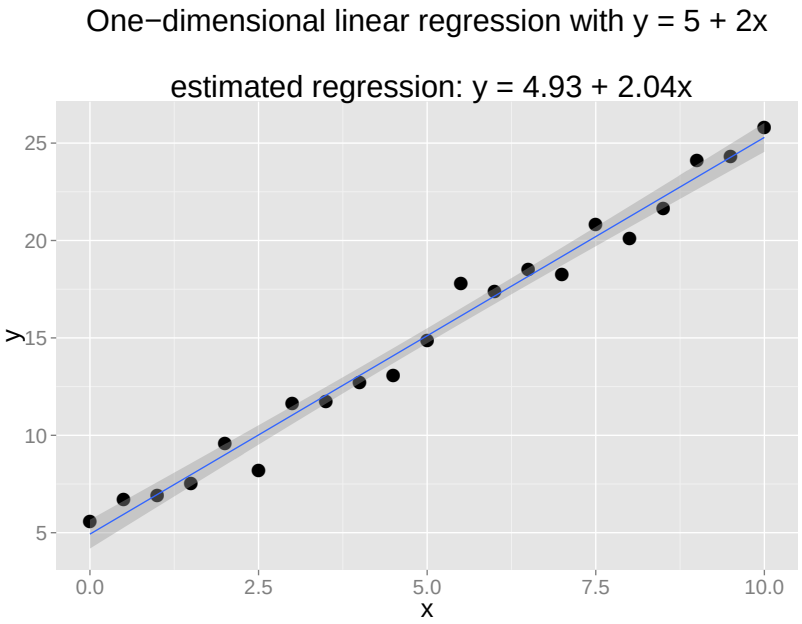


Figure 4: Linear regression model with 95 % confidence interval in grey.

The assumption of a linear relation between y and $x_1, \ldots, x_p$, appears to be very restrictive, but nonlinear relationships can also be modelled within the framework of linear models (Fahrmeir et al.; 2007).

Example 5.2. Consider two random variables X and Y with a relationship of the form

$$Y = \frac{C}{X^a}\varepsilon$$

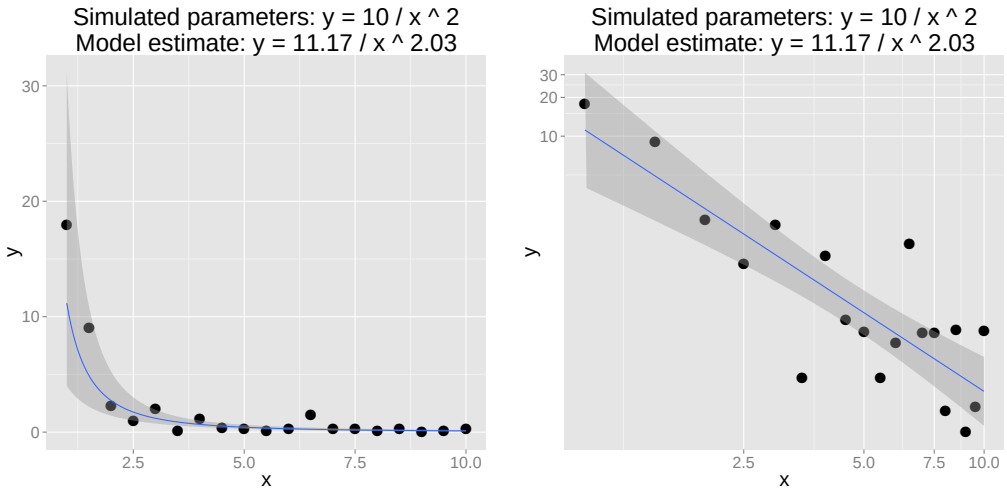


Figure 5: Visualisation of Example 5.2, in the right hand figure the $x-$ and $y-$axes are log-scaled in order to see the linear relation of $\log x$ and $\log y$.

with constants C, $a \in \mathbb{R}$ and multiplicative error $\varepsilon \sim \log N(0, \sigma^2)$, see Figure 5. In our case $\log \varepsilon \sim N(0, \sigma^2)$, $E[\varepsilon] = \exp(\frac{\sigma^2}{2})$, $V[\varepsilon] = \exp(\sigma^2)\exp(\sigma^2 - 1)$. After a log-transformation of the equation we get the following form

$$\log Y = \log C - a \log X + \log \varepsilon,$$

where the parameters C and a can be estimated with a normal linear model (Fahrmeir et al.; 1997, p. 301).

2. *Assumptions on the Errors.* Another basic assumption of the linear model is additivity of errors, which implies

$$y_i = \beta_0 + \beta_1 x_{1,i} + , \ldots , + \beta_k x_{k,i} + \varepsilon_i, \ i = 1, \ldots, N.$$

Even though this appears to be very restrictive, this assumption is reasonable for many practical applications. Moreover, problems, which at first do not show additive error structure, can be specified by models with additive errors after a transformation of the response variable y (Fahrmeir et al.; 2007). The errors are assumed to have expectation zero, $E[\varepsilon_i = 0],\ i = 1, \ldots, N$ and we assume a constant error variance $V[\varepsilon_i] = \sigma^2,\ i = 1, \ldots, N$ to exist across observations. In addition to homoscedastic variances, we assume that errors are uncorrelated, meaning $Cov[\varepsilon_i, \varepsilon_j] = 0$, for $i \neq j$. Additionally, errors and stochastic covariates are assumed to be independent, which can be a problem in real data settings (Fahrmeir et al.; 2007).

3. *Gaussian errors.* To construct confidence intervals and hypothesis tests for the regression coefficients, in the classical normal regression case we assume a normal distribution for the errors (at least approximately). Together with assumptions 1 and 2, we obtain $\varepsilon_i \sim N(0, \sigma^2)\ i = 1, \ldots, N$ (Fahrmeir et al.; 2007). In order to verify the normal distribution of the errors a common tool are Q-Q plots ("Q" stands for quantile). They are a graphical method for comparing two probability distributions by plotting their quantiles against each other. In the case of linear regression, the quantiles of the residuals are plotted against the quantiles of a Gaussian distribution (Definition 2.3). This shows if the residuals follow approximately a normal distribution. Let $x_{(1)}, \ldots, x_{(N)}$ be an ordered sample. For $i = 1, \ldots, N$ let z_i the $(i - 0.5)/N$- quantiles

of the standard normal distribution. The points $(z_1, x_{(1)}), \ldots, (z_N, x_{(N)})$ form the Q-Q plot (Fahrmeir et al.; 1997). The Q-Q plot for the regression model seen in Figure 4 is shown in Figure 6.

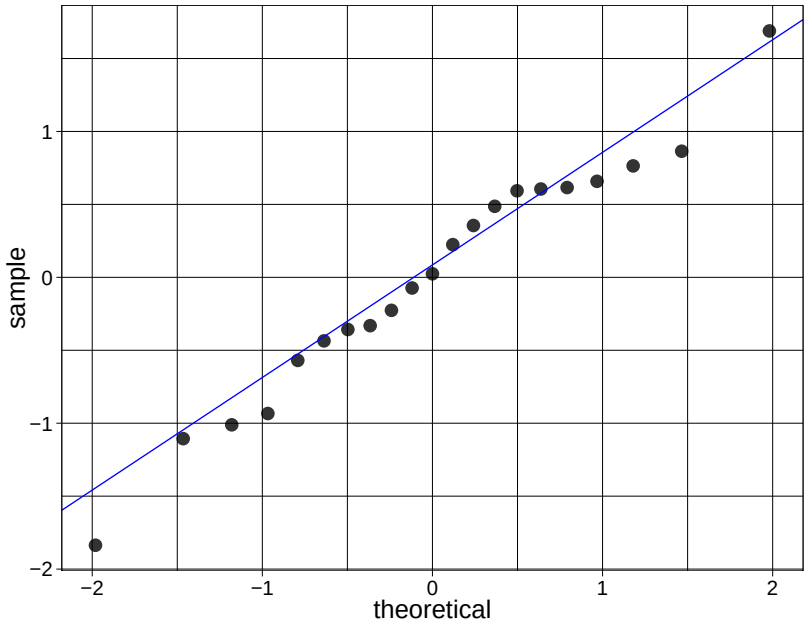


Figure 6: Example for Q-Q plot of normally distributed errors as stated in assumption 3 of the linear model in Section 5.1.

Confidence intervals for the regression parameters $\underline{\beta}$ can be constructed, due to the assumption of normally distributed errors. This is a precondition for the construction of exact tests and confidence intervals, that can in application be replaced by a large sample size.

As the model parameters $\underline{\hat{\beta}}$ are normally distributed with expectation value $\underline{\beta}$ and covariance matrix $\sigma^2(X^\top X)^{-1}$ we obtain the following confidence

intervals for our model parameters $\underline{\beta}$ with level $1-\alpha$:

$$\left[\hat{\beta}_j - t_{N-p}\left(1-\frac{\alpha}{2}\right)\mathrm{se}_j,\ \hat{\beta}_j - t_{N-p}\left(1-\frac{\alpha}{2}\right)\mathrm{se}_j\right],$$

where $t_{N-p}(1-\alpha/2)$ denotes the $(1-\alpha/2)$ quantile of the Student distribution with $N-p$ degrees of freedom (Definition 3.1)

$$\mathrm{se}_j = \sqrt{\hat{Var}[\beta_j]} = \frac{\hat{\sigma}^2}{(1-R_j^2)\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)}$$

denotes the estimated standard deviation of $\hat{\beta}_j$. We define R_j as the coefficient of determination for the regression between x_j as response variable and all other explanatory variables $x_l, l \neq j$. In general the coefficient of determination R for a regression with response y and predicted response $\hat{y}$ is defined by $R = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})/\sum_{i=1}^{N}(y_i - \bar{y})$.

5.2 Nonlinear model

In some cases where the linear model is not applicable, a more general nonlinear regression model is used, where the function f relating the response to the predictors is not necessarily linear:

$$y_i = f(x_{1,i}, \ldots, x_{p,i}, \underline{\beta}) + \varepsilon_i,\ i = 1, \ldots, N$$

As in the linear model, $\underline{\beta}$ is a vector of parameters and $\underline{x}_i = x_{1,i}, \ldots, x_{p,i}$ is a vector of predictors, and $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, N$.

The likelihood for the nonlinear regression model is

$$L(\underline{\beta}, \sigma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{\sum_{i=1}^{N}[y_i - f(\underline{x}_i, \underline{\beta})]^2}{2\sigma^2} \right\}$$

This likelihood is maximised when the sum of squared residuals, which is proportional to the negative log-likelihood $l(\beta)$,

$$-l(\underline{\beta}) = -\log L(\underline{\beta}) \propto \sum_{i=1}^{N}[y_i - f(\underline{x}_i, \underline{\beta})]^2$$

is minimised. This is done by differentiating the negative log-likelihood resulting in the so-called score function $s(\underline{\beta})$,

$$\begin{aligned} s(\underline{\beta}) &= -\frac{\partial l(\underline{\beta})}{\partial \underline{\beta}} \\ &= -2\sum_{i=1}^{N}[y_i - f(\underline{\beta}, \underline{x}_i)]\frac{\partial f(\underline{\beta}, \underline{x}_i)}{\partial \underline{\beta}} \end{aligned}$$

and setting the partial derivatives to zero, producing estimating equations for the regression coefficients $\underline{\beta}$. Because these equations are in general nonlinear, they require solution by numerical optimisation. As in a linear model, it is usual to estimate the error variance by dividing the residual sum of squares for the model by the number of observations less the number of parameters. Coefficient variances may be estimated from a linearised version of the model.

Let the Fisher information F be defined as the matrix $F = [F_{ij}]$ with

$$F_{ij} = \frac{\partial f(\underline{x}_i, \underline{\beta})}{\partial \beta_i \beta_j}.$$

Then the estimated asymptotic covariance matrix of the estimated regression coefficients is

$$V(\underline{\hat{\beta}}) = \sigma^2 (F^\top F)^{-1}$$

where σ^2 is the error variance, estimated by $\hat{\sigma}^2 = \frac{1}{N-p} \sum_{i=1}^{N} [y_i - f(\underline{\beta}, \underline{x}_i)]$ with p the dimension of the parameter vector $\underline{\beta}$. For large sample sizes the parameters $\underline{\beta}$ is approximately multivariate normal distributed. Thus, like in linear regression, standard error and confidence intervals can be constructed, or confidence ellipses if several variables are considered at once (Fox; 2002; Gallant; 1975).

Example 5.3. A model for population growth towards an asymptote is the logistic model

$$y_i = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 x_i)} + \varepsilon_i$$

where y_i is the population size at time x_i, β_1 is the asymptote towards which the population grows, β_2 reflects the size of the population at time $x = 0$ (relative to its asymptotic size) and β_3 controls the growth rate of the population (Fox; 2002).

5.3 Setup of the simulation

The setup to simulate the bias and variance of an estimator under given conditions is the following:

1. Define a series of sample sizes N_k, e.g. $N_k \in \{10, 20, \ldots, 1000\}$.

2. Draw r random samples of size N_k of m-dimensional random variables of a chosen distribution of interest.

3. Compute the estimated entropy for all r of the N_k-sized random samples with the estimator to be investigated.

4. Summary statistics like the variance or bias based on the r estimated entropies can be computed for every sample size N_k.

5. Based on those summary statistics, models can be fitted given a relationship between the sample size N_k, the chosen distribution and the entropy estimator being used.

Example 5.4. For every sample size N_k one can compute the mean and the variance of the nearest neighbour estimators for the Shannon, Reńyi and Tsallis entropy. As to be expected from theory, the estimators are basically the same for all three entropies as $q \to 1$. This can be seen in Figure 7, as all estimators plot at the same spot. This is based on the property of the Rényi and Tsallis entropy to converge to the Shannon entropy for $q \to 1$. For $N \to \infty$ the estiamted entropy and thus the mean of the estimated entropy

converges to the theoretical entropy, see Figure 7 while their variance tends to zero, see Figure 8. Again, for a given sample size the variance of the nearest neighbour estimators for the different entropy types is almost the same, as $q \to 1$. For $N \to \infty$ the variance converges obviously to zero for all estimators, see Figure 8.

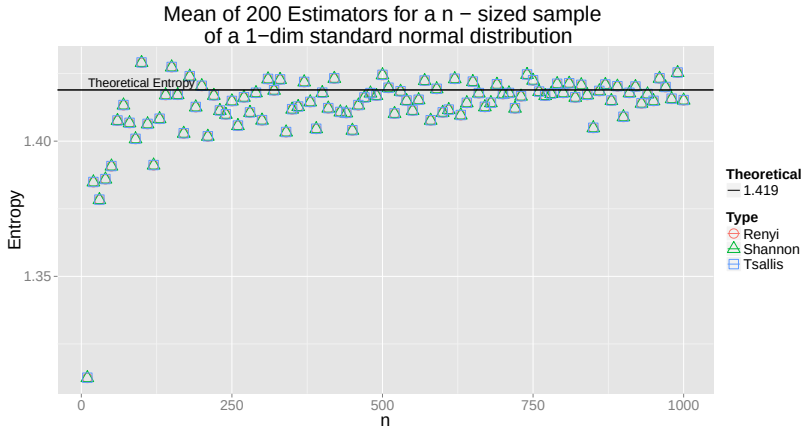


Figure 7: Mean of estimated entropy of 200 repetitions against sample size n , $n = \{10, 20, ..., 1000\}$.

The number of repetitions r for every sample size influences the precision of the results. The greater r is chosen, the smaller the variance of the summary statistics and the models based on them. This is obviously a favoured quality, but on the other hand, a large number of repetitions for each sample size results in longer computation time. Trying to reduce the computation time for a given number of repetitions r, which corresponds to a given accuracy in estimation, can be achieved by writing more efficient code. Ways of achieving this in R are discussed in the following section.

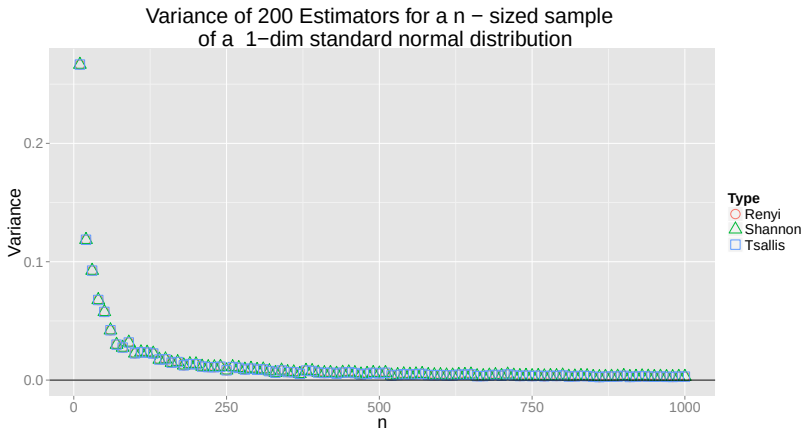


Figure 8: Variance of estimated entropy of 200 repetitions against sample size $n = \{10, 20, ..., 1000\}$.

5.4 Estimation of convergence of variance

In order to estimate the rate of convergence of the variance of the estimated entropy $\hat{H}$ a linear model is used. The assumption is that

$$V[\hat{H}] = C/N^a \varepsilon$$

where C is a constant and a is the power of convergence of N and ε is a log-normally distributed multiplicative error. The constant C may depend on q, m and the distribution of the random samples, but not on N. In order to be able to estimate the unknown parameters one uses a logarithmic transformation on the equation, analogously to Example 5.2, yielding

$$\log V[\hat{H}] = \log C - a \log N + \log \varepsilon,$$

which can be estimated in a simple linear model as presented in Section 5.1. The output of the linear model corresponding to the simulated data used in Figure 7 and Figure 8 suggests that $a = 1$, see Table 3. This induces a conversion rate proportional to $\frac{1}{N}$.

Table 3: Estimated coefficients of the linear model $\log V[\hat{H}] = \log C - a \log N + \log \varepsilon$ with intercept C and slope $-a$, giving lower and upper bounds for 95% confidence intervals.

	lower.bound	estimated	upper.bound
$\hat{C}$	0.704	0.842	0.980
$\hat{a}$	1.003	0.980	0.957

The model fits the sampled data very well, see Figure 9. The line representing the model fits the simulated data throughout all sample sizes. For details of the composition of the parameter C and the effect of different values of the parameters m and k on the power of N further investigation is necessary.

5.5 Estimation of convergence of bias

To estimate the bias of the estimated entropy $\hat{H}$, the residuals of the estimation are being used. The function to be estimated,

$$\text{Bias}[\hat{H}] = C/N^a + \varepsilon,$$

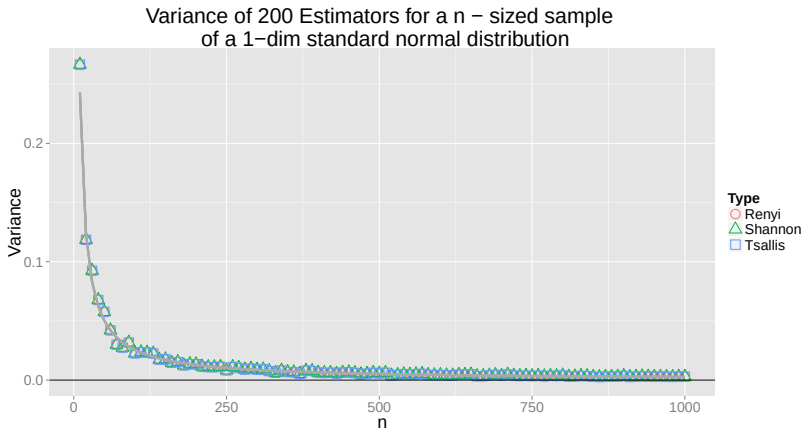


Figure 9: Variance of estimated entropy of 200 repetitions against sample size $n = \{10, 20, ..., 1000\}$. The same sample as in Figure 8 is used. The grey line represents the log-transformed linear model with parameters seen in Table 3.

looks similar to the case of the variance at first glance. Theoretically the same log-transformation would be helpful here as well. As the estimated values converge from below to the theoretical entropy, the transformation

$$\log(-\text{Bias}[\hat{H}]) = \log C - a \log N$$

would be an obvious choice. But unlike the variance that is restricted to be strictly positive, the bias does change its sign when calculated from a sample of a random variable, leading to undefined values for $\log(x)$ for negative x that can not be handled by the linear model. Therefore the more general normal nonlinear regression model is used as proposed in Section 5.2.

The estimated parameters for the bias of the simulation seen in Figure 7 are shown in Table 4. For $q = 1$ all estimators give the same result,

therefore the estimated parameters are all the same. It can be seen that the rate of convergence of the bias is very close to 1 with the confidence interval overlapping it. Thus it can be assumed that the true rate of convergence of the bias is $1/N$.

Table 4: Estimated model coefficients for the bias of the *NearestNeighbour*-estimator of the Shannon entropy for a one-dimensional standard normal distribution with 95% confidence interval.

	lower.bound	estimated	upper.bound
β_1	-2.565	-1.215	-0.648
β_2	0.841	1.076	1.376
β_3	-0.004	-0.001	0.002

The visual fit of the model is very good, see Figure 10, where the model is represented by the grey line.

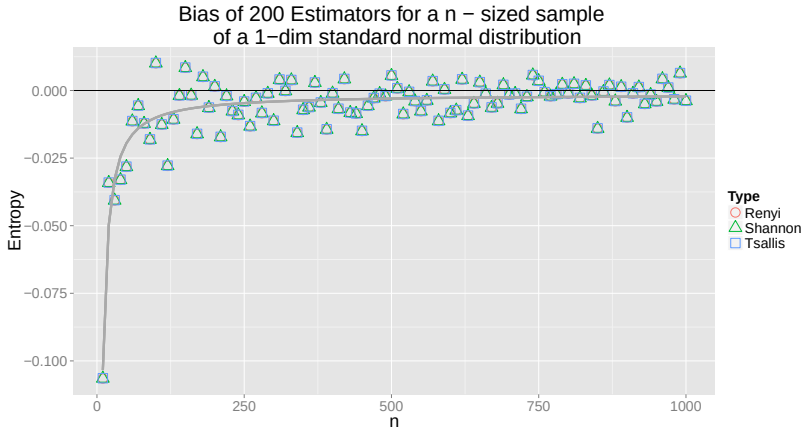


Figure 10: Bias of estimated entropy of 200 repetitions against sample size n, $n = \{10, 20, ..., 1000\}$. The same sample as in Figure 7 is used. The grey line represents the log-transformed linear model with parameters seen in Table 4.

5.6 Comparison of one-dimensional Shannon entropy estimators

In this section the two different estimators based on sample spacings that were presented in Section 4.2, the *m_spacing*-estimator presented in Beirlant et al. (1997), see Section 4.2, and the *mn_spacing*-estimator presented in Song (2000), see Section 4.2.2, will be compared to the nearest neighbour estimator presented in Leonenko and Pronzato (2010), see Section 4.1. All three estimators can be used to obtain asymptotically unbiased estimators of the Shannon entropy for one-dimensional samples. This is the historically oldest definition of entropy and still the most commonly used special case in application. A good estimator is an important real world task of great interest. In applied literature the term entropy is often used equivalently to Shannon entropy. In this section the alias *m_spacing* will be used for the estimator presented in Beirlant et al. (1997), *mn_spacing* for the estimator presented in Song (2000) and for the nearest neighbour estimator presented in Leonenko and Pronzato (2010) we will use the alias *NearestNeighbour*. As the spacing estimators are only capable of estimating the entropy for a one-dimensional random variable, this section exclusively focuses on that. The essential part of the *NearestNeighbour*-estimator is the estimation of the integral over the quadratic function, whereas the spacing estimators are based on a density estimate. A comparison of these completely different approaches is an interesting task, that will be executed by a simulation study.

Estimating the Shannon entropy of a random sample of various conditions, the estimators will be compared in their bias (Section 5.6.1) and variance (Section 5.6.2) for the one-dimensional standard normal distribution.

A selection of other distributions is given in Section 5.6.3, where the estimation of entropy for bounded distributions will be compared for the *NearestNeighbour*-estimator and the *m_spacing*-estimator. The section ends in a brief discussion (Section 5.6.4) of the comparison of the two estimators. We will see that the *m_spacing*-estimator is superior to the *NearestNeighbour*-estimator under some conditions.

5.6.1 Comparison of bias

In this section the *NearestNeighbour*-, *m_spacing*- and *mn_spacing* estimators will be compared concerning their biasedness. The biasedness of the respective estimator will be estimated using the method presented in Section 5.5. All three estimators are asymptotically unbiased. For growing sample sizes $n \to \infty$ the estimators converge to the real Shannon entropy of the underlying distribution from which the samples were drawn. This asymptotic unbiasedness is an important theoretical property for any estimator, but in application it is usually not possible to draw an infinite number of samples from the source. The rate of convergence is thus an important criterion for a good estimator. For $n \to \infty$ the bias converges to zero for all estimators, see Figure 11, but the bias of the *NearestNeighbour*-estimator is smaller than the bias of the spacing-estimators for small sample sizes. For large sample

sizes N the picture is less clear, as the bias of *mn_spacing*-estimator seems to converge faster to zero. As all of the estimators are asymptotically unbiased, the difference for large sample sizes is marginal.

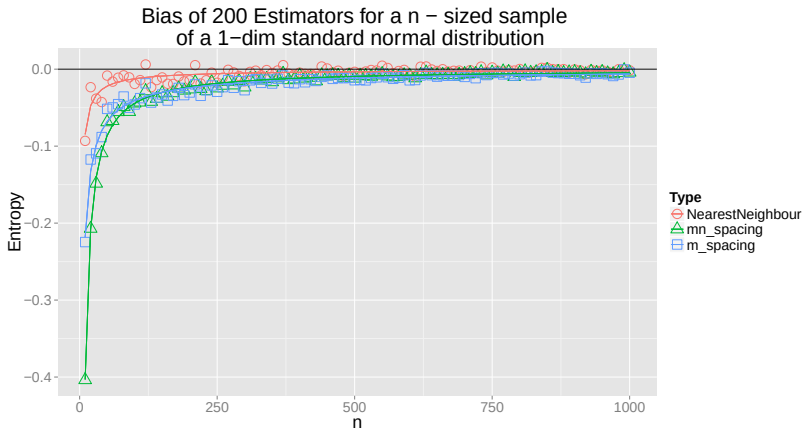


Figure 11: Bias of estimated entropy of 200 repetitions against sample size $n = \{10, 20, ..., 1000\}$.

These statements can be confirmed and quantified with a nonlinear regression model (see Section 5.2). It is employed in order to describe the connection of the bias and the sample size, analogously to Section 5.5. The assumed relationship of bias and sample size n is of the form

$$\text{Bias}[\hat{H}] = \beta_1 n^{-\beta_2} + \beta_3 + \varepsilon.$$

The estimated model coefficients shown in Tables 5, 6 and 7 are estimated from the samples shown in Figure 11. The estimated parameters for the *NearestNeighbour*-estimator are shown in Table 5. For the *m_spacing*-estimator the estimated model parameters are shown in Table 6 and the *mn_spacing*-

estimator in Table 7. The *mn_spacing*-estimator has the largest bias for small sample sizes, but it converges faster to zero than the other estimators.

Table 5: Estimated model coefficients for the bias of the *NearestNeighbour*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	-1.007	-0.587	-0.361
β_2	0.661	0.842	1.051
β_3	-0.003	-0.0003	0.003

Table 6: Estimated model coefficients for the for the bias of the *m_spacing*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	-1.310	-1.137	-0.989
β_2	0.660	0.713	0.766
β_3	-0.001	0.002	0.005

Table 7: Estimated model coefficients for the for the bias of the *mn_spacing*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	-3.912	-3.650	-3.409
β_2	0.931	0.957	0.983
β_3	-0.001	0.0002	0.002

5.6.2 Comparison of variance

As all of the estimators that are being investigated, that is the the *NearestNeighbour*-, the *m_spacing*- and the *mn_spacing*-estimator, are asymptotically unbiased, their variance is an important second criterion to be considered. All three estimators, are consistent, that means that for sample sizes $n \to \infty$ the variance of the estimators converges to 0. A small variance for small sample sizes is an important quality of a good estimator, because this ensures good results even for small sample sizes. For $n \to \infty$ the variance converges to zero for all estimators, see Figure 12. Just by looking at Figure 12 it seems that the variance of the *mn_spacing*-estimator is somewhat smaller than of the *m_spacing*-estimator and *NearestNeighbour*-estimator.

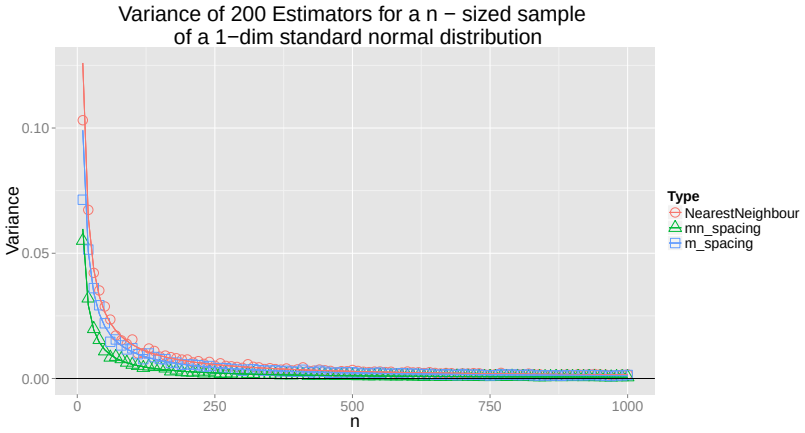


Figure 12: Variance of estimated entropy of 200 repetitions against sample size $n = \{10, 20, ..., 1000\}$.

As there might me a small difference in the variance of the estimators, a log-transformed linear regression model (see Section 5.1, Example 5.2) is

used to describe the relationship of the variance and the sample size. With this tool the rate of convergence of the variance to zero can be estimated for both estimators. The modelled relation of the variance of the estimator and the sample size n is of the form

$$V[\hat{H}] = \beta_0 n^{\beta_1} \varepsilon,$$

that can be estimated after a log-transform in a linear model of the form

$$\log(V[\hat{H}]) = \log(\beta_0) + \beta_1 \log(n) + \varepsilon,$$

analogously to Section 5.4.

The estimated model coefficients based on the samples shown in Figure 12 for the *NearestNeighbour*-estimator are shown in Table 8, for the *m_spacing*-estimator in Table 10 and for the *m_spacing*-estimator in Table 9 respectively.

Table 8: Estimated model coefficients for the variance of the *NearestNeighbour*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
$\log(\beta_0)$	0.029	0.159	0.288
β_1	-0.990	-0.969	-0.947

It can be seen that the *mn_spacing*-estimator has the smallest variance for all sample sizes and it converges faster to zero than the other estimators. This is an important quality. When it comes to comparing the entropy of

Table 9: Estimated model coefficients for the variance of the *m_spacing*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
$\log(\beta_0)$	-0.210	-0.080	0.050
β_1	-0.991	-0.969	-0.948

Table 10: Estimated model coefficients for the variance of the *mn_spacing*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
$\log(\beta_0)$	-0.584	-0.458	-0.332
β_1	-1.046	-1.025	-1.004

random samples of the same sample size the variance is a way more important measure of performance for an estimator than its bias, because the bias for a given sample size is the same in all samples of this size. In application the biasedness of an estimator might not be of great interest as usually same sized samples will be compared. In this case the *mn_spacing*-estimator is superior to the other estimators investigated.

5.6.3 Bounded distributions

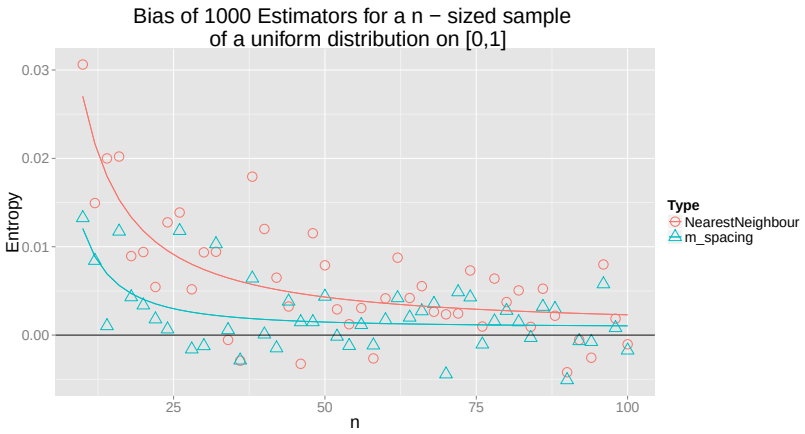
In this section the *NearestNeighbour*- and the *m_spacing*-estimator for the Shannon entropy of a one-dimensional random variable will be compared for bounded distributions. We will have a look at the uniform distribution, see Definition 2.7, and its generalisation the beta distribution, see Definition

2.6. Their theoretical Rényi entropy is given in Theorem 2.4 and Theorem 2.5 respectively.

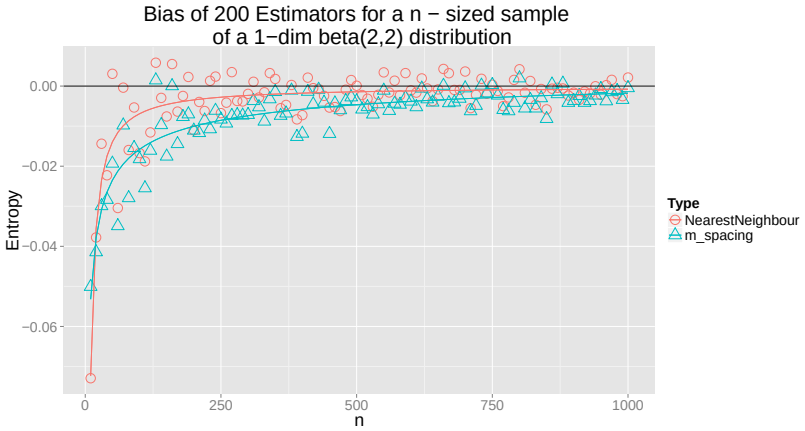
Both the *NearestNeighbour*- and the *m_spacing*-estimator are found to be asymptotically unbiased for the uniform distribution on $[0, 1]$, see Figure 13a, as well as for the beta distribution with parameters $\alpha = \beta = 2$, see Figure 13b, that has the same bounded support.

The observation of asymptotic unbiasedness of the estimators can be verified using the non-linear model approach for estimating the bias presented in Section 5.5. The *NearestNeighbour*-estimator can be assumed to be asymptotically unbiased as the confidence interval of the parameter β_3 includes zero, see Table 11 for the uniform distribution and Table 13 for the beta(2,2) distribution. The same argument holds for the *m_spacing*-estimator, see Tables 12 and 14. Surprisingly the bias of the *NearestNeighbour*-estimator is smaller for the beta(2,2) distribution, but bigger for the uniform distribution. This difference is striking as the uniform distribution can be seen as a beta(1,1) distribution. It would be an interesting topic for future research to find the subsets of the paramater space where one estimator has a smaller bias than the the other.

Technical remark: For the uniform distribution a larger number of repetitions of smaller sample sizes had to be drawn in order to estimate the bias of the *m_spacing*-estimator for a given sample size. For the usual number of repetitions of 200 estimations per given sample size n the algorithm of the estimation of the non-linear model ran into numerical problems. After an in-



(a) Average bias for sample sizes n of the estimation of the Shannon entropy for a one-dimensional uniform distribution on $[0, 1]$.



(b) Average bias for sample sizes n of the estimation of the Shannon entropy for a one-dimensional beta(2,2) distribution with finite support $[0, 1]$.

Figure 13: Average bias for a sample sizes n of the *m_spacing* and *Nearest-Neighbour*-estimator for the Shannon entropy.

Table 11: Estimated model coefficients for the bias of the *NearestNeighbour*-estimator for a uniform distribution on $[0, 1]$.

	lower.bound	estimated	upper.bound
β_1	-0.568	0.474	1.515
β_2	0.321	1.258	2.195
β_3	-0.004	0.001	0.006

Table 12: Estimated model coefficients for the bias of the *m_spacing*-estimator for a uniform distribution on $[0, 1]$.

	lower.bound	estimated	upper.bound
β_1	-3.178	0.740	4.658
β_2	-0.372	1.821	4.014
β_3	-0.002	0.001	0.003

Table 13: Estimated model coefficients for the bias of the *NearestNeighbour*-estimator for a beta distribution with parameters $\alpha = \beta = 2$.

	lower.bound	estimated	upper.bound
β_1	-1.450	-0.770	-0.444
β_2	0.822	1.028	1.279
β_3	-0.002	-0.0001	0.002

Table 14: Estimated model coefficients for the bias of the *Ustatistic*-estimator for a beta distribution with parameters $\alpha = \beta = 2$.

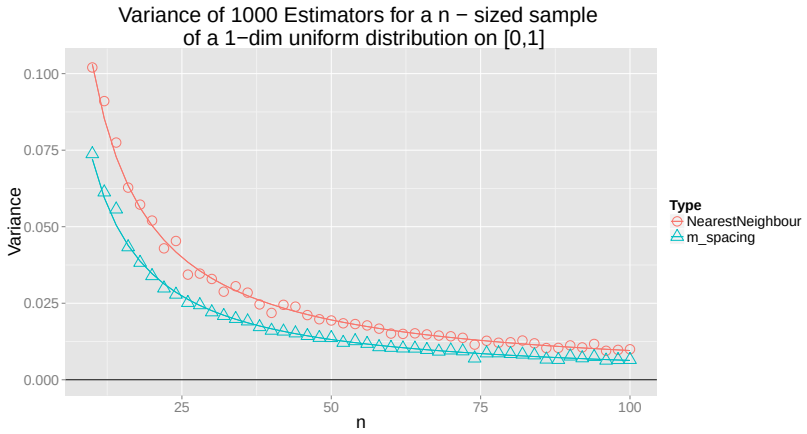
	lower.bound	estimated	upper.bound
β_1	-0.227	-0.160	-0.119
β_2	0.281	0.426	0.571
β_3	0.001	0.007	0.017

creased number of repetitions per sample size in order to get more precision and a shorter sequence of sample size n to keep the computation time within a reasonable limit, the results presented in this section were gained.

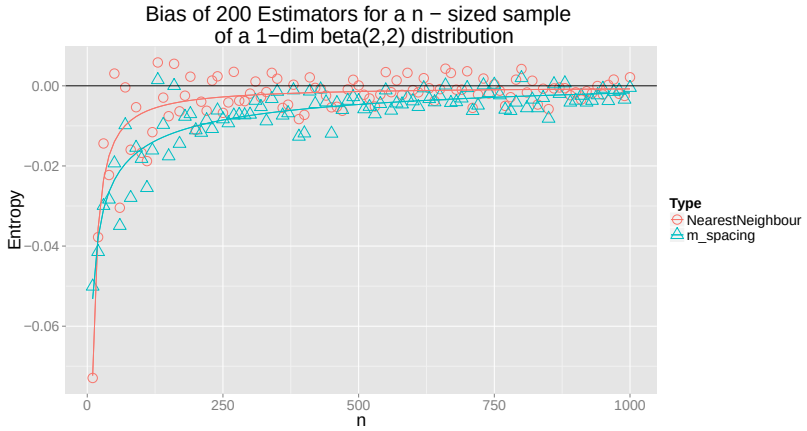
Now the variance of the *NearestNeighbour*-estimator and the *m_spacing*-estimator will be compared under the same conditions. In Figure 14, depicting the variance of the estimators shown in Figure 13, it can be seen that the *m_spacing*-estimator has a smaller variance than the *NearestNeighbour*-estimator. This can be quantified with a log-transformed linear regression model (see Section 5.1, example 5.2). In an analogous manner to Section 5.4 the model is used to describe the rate of convergence of the variance to zero. For a uniform distribution on $[0, 1]$ the estimated model coefficients based on the samples shown in Figure 14a for the *NearestNeighbour*-estimator are shown in Table 15 and for the *m_spacing*-estimator in Table 16 respectively. The estimated model coefficients for a beta(2,2) distribution of the *NearestNeighbour*-estimator are shown in Table 15 and for the *m_spacing*-estimator in Table 16 respectively. The variance shown in Figure 14b is modeled.

Table 15: Estimated model coefficients for the variance of the *NearestNeighbour*-estimator for a one-dimensional uniform distribution on [0,1].

	lower.bound	estimated	upper.bound
$\hat{C}$	-0.0001	0.106	0.212
$\hat{a}$	-1.060	-1.033	-1.006



(a) Variance of the estimators of the Shannon entropy for a one-dimensional uniform distribution on $[0, 1]$ for a sample sizes n.



(b) Variance of the estimators of the Shannon entropy for a one-dimensional beta(2,2) distribution with finite support $[0, 1]$ for a sample sizes n.

Figure 14: Variance of the *m_spacing* and *NearestNeighbour*-estimator for the Shannon entropy for a sample sizes n. The same samples are used as in Figure 13.

Table 16: Estimated model coefficients for the variance of the *m_spacing*-estimator for a one-dimensional uniform distribution on [0,1].

	lower.bound	estimated	upper.bound
$\hat{C}$	-0.300	-0.194	-0.088
$\hat{a}$	-1.085	-1.058	-1.030

Table 17: Estimated model coefficients for the variance of the *NearestNeighbour*-estimator for a one-dimensional beta(2,2) distribution.

	lower.bound	estimated	upper.bound
$\hat{C}$	-0.049	0.110	0.269
$\hat{a}$	-1.033	-1.007	-0.980

Table 18: Estimated model coefficients for the variance of the *m_spacing*-estimator for a one-dimensional beta(2,2) distribution.

	lower.bound	estimated	upper.bound
$\hat{C}$	-0.318	-0.193	-0.069
$\hat{a}$	-1.026	-1.006	-0.985

5.6.4 Discussion of the comparison

For the case of a normally distributed random variable, the *mn_spacing*-estimator is superior to the *m_spacing*-estimator and the *NearestNeighbour*-estimator. It has a smaller variance for any given sample size and a smaller bias for large sample sizes. Note that for small sample sizes the *NearestNeighbour*-estimator has a smaller bias. For bounded distributions the results are not as clear. The *m_spacing*-estimator has a smaller bias than the *NearestNeighbour*-estimator for the uniform distribution, but vice versa for the beta(2,2)-distribution. This is especially striking as the beta distribution is a generalisation of the uniform distribution. Further investigation is promising for future research. The key information is summarised in Table 19.

Table 19: Overview of the best performing estimators for the one-dimensional Shannon entropy under the aspects of variance and unbiasedness.

	Bias	Variance
Normal Distribution	*NearestNeighbour*	*mn_spacing*
Uniform distribution	*mn_spacing*	*mn_spacing*
Beta(2,2) distribution	*NearestNeighbour*	*NearestNeighbour*

5.7 Comparison of quadratic entropy estimators

In this section the estimator presented in Källberg et al. (2014), see Section 4.3, will be compared to the nearest neighbour estimator presented in Leonenko and Pronzato (2010), see Section 4.1. Both of them give asymp-

totically unbiased estimators of the quadratic entropy. The alias *Ustatistic* will be used for the estimator given in Källberg et al. (2014) and the alias *NearestNeighbour* for the nearest neighbour estimator given in Leonenko and Pronzato (2010). This section deals exclusively with the estimation of the quadratic Rényi entropy, as the quadratic Tsallis entropy is just a one-to-one transformation of it, see Theorem 2.6. The crucial part is the estimation of the integral over the quadratic function, which is estimated differently by the *NearestNeighbour*-estimator and the *Ustatistic*-estimator. Transforming the quadratic function to get the entropy is exactly the same for both estimators. An investigation of the comparison of the estimators for the quadratic Tsallis entropy might be interesting for future research, but similar results are to be expected. The two approaches to estimate the quadratic Rényi entropy will be compared in their bias (Section 5.7.1), variance (Section 5.7.2) and computation time (Section 5.7.3) for the one-dimensional standard normal distribution. A prospect of higher dimensional distributions is given in Section 5.7.4. Especially for higher dimensional data small sample sizes effect the bias of the estimators differently. This will be illuminated in Section 5.7.5. A prospect of other distributions is given in Section 5.7.6, where the estimation of entropy for bounded distributions will be investigated. The section ends in a short discussion (Section 5.7.7) of the comparison of the two estimators. We will see that the *Ustatistic*-estimator is superior to the *NearestNeighbour*-estimator depending on the conditions.

5.7.1 Comparison of bias

The *Ustatistic*- and the *NearestNeighbour*-estimator are asymptotically unbiased, that means that for sample sizes $n \to \infty$ the estimators converge to the theoretical quadratic Rényi entropy. But often in application only finite samples are at hand, for which a small bias for a small sample size is of great importance. For $n \to \infty$ the bias converges to zero for both estimators, see Figure 15, but the bias of the *Ustatistic*-estimator is consistantly smaller than the *NearestNeighbour*-estimator. Thus, it is a better estimator when in comes to unbiasedness.

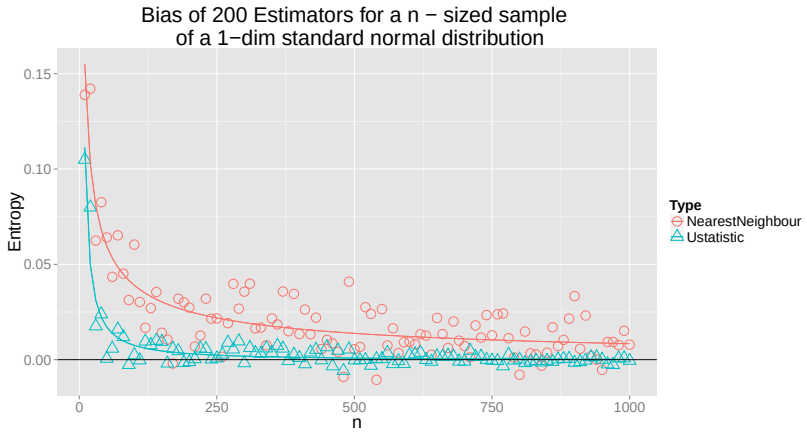


Figure 15: Bias of estimated entropy of 200 repetitions against sample size $n = \{10, 20, ..., 1000\}$.

In order to quantify that difference, a nonlinear regression model (see Section 5.2) is used to describe the relationship of the bias and the sample size, analogously to Section 5.5. The assumed relation between bias and

sample size is

$$\text{Bias}[\hat{H}_2^*] = \beta_1 n^{-\beta_2} + \beta_3 + \varepsilon.$$

The estimated coefficients based on the samples shown in Figure 15 for the *NearestNeighbour*-estimator are shown in Table 20 and for the *Ustatistic*-estimator in Table 21 respectively. The rate of convergence is a lot higher for the *Ustatistic*-estimator.

Table 20: Estimated model coefficients for the expectation value of the *NearestNeighbour*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	0.394	0.596	0.930
β_2	0.414	0.578	0.749
β_3	-0.018	-0.003	0.006

Table 21: Estimated model coefficients for the for the expectation value of the *Ustatistic*-estimator for a one-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	1.005	1.579	2.606
β_2	0.978	1.150	1.351
β_3	-0.002	-0.0004	0.001

5.7.2 Comparison of variance

Both the *Ustatistic*-estimator and the *NearestNeighbour*-estimator are consistent, that means that for sample sizes $n \to \infty$ the variance of the estimators converges to zero. As both estimators are asymptotically unbiased,

their variance is an important second criterion to be considered. A small variance for small sample sizes is an important quality criterion of an estimator, as this ensures good estimation results even for small sample sizes. For $n \to \infty$ the variance converges to zero for both estimators, see Figure 16, but the variance of the *Ustatistic*-estimator is consistently smaller than the *NearestNeighbour*-estimator for a fixed sample size. This makes it superior, especially in the case of small sample sizes.

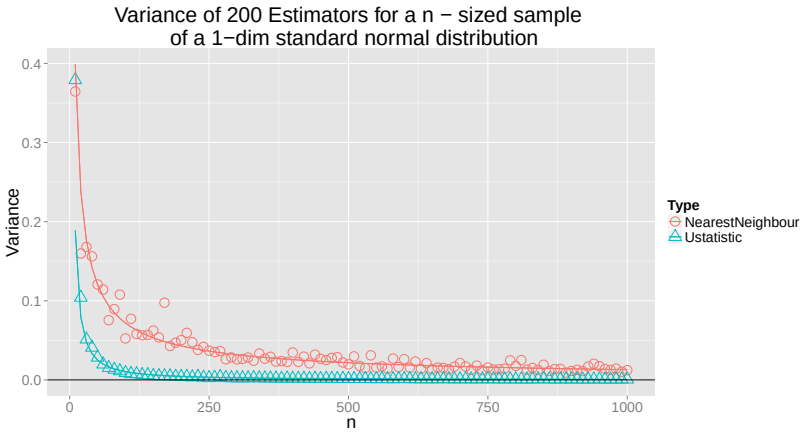


Figure 16: Variance of estimated entropy of 200 repetitions against sample size $n = \{10, 20, ..., 1000\}$.

In order to quantify that difference, a log-transformed linear regression model (see Section 5.1, Example 5.2) is used to describe the relationship of the variance and the sample size, as done in Section 5.4. The modelled relation of the variance of the estimator and the sample size n is of the form

$$V[\hat{H}_2^*] = \beta_0 n^{\beta_1} \varepsilon,$$

which can be estimated after a log-transformation in a linear model as

$$\log(V[\hat{H}_2^*]) = \log(\beta_0) + \beta_1 \log(n) + \varepsilon.$$

The estimated coefficients based on the samples shown in Figure 16 for the *NearestNeighbour*-estimator are shown in Table 22 and for the *Ustatistic*-estimator in Table 23 respectively.

Table 22: Estimated model coefficients for the variance of the *NearestNeighbour*-estimator for a one-dimensional standard normal distribution

	lower.bound	estimated	upper.bound
$\log(\beta_0)$	0.512	0.793	1.075
β_1	-0.791	-0.744	-0.697

Table 23: Estimated model coefficients for the variance of the *Ustatistic*-estimator for a one-dimensional standard normal distribution

	lower.bound	estimated	upper.bound
$\log(\beta_0)$	1.042	1.223	1.404
β_1	-1.285	-1.255	-1.225

5.7.3 Comparison of computation time

From a mathematical point of view the computation time of the simulation is a secondary problem, but in real life applications this is an important issue. When it comes to large data sets a more run-time efficient estimator might

be preferable even if it is not superior in other properties. A shorter runtime can be achieved by writing more efficient code it computation power is limited. Appendix B discusses efficient programming using the software package R .

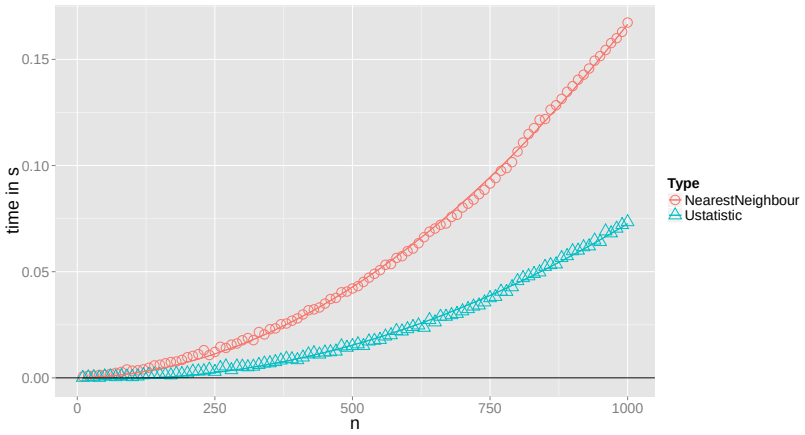


Figure 17: The computing time for both estimators for a given sample size n.

The *Ustatistic*-estimator performs a lot faster than the *NearestNeighbour*-estimator, see Figure 17. Further investigation shows, that the ratio of the compution times converges to 2, see Figure 18, which means that the *Ustatistic*-estimator performs about twice as fast as the *NearestNeighbour*-estimator. A look in the code, see Appendix, gives the clue to this property, in both approaches the distances of all sampled vectors have to be computed. In the *NearestNeighbour*-estimator the $k-th$ nearest neighbour has to be computed which is a rather extensive operation involving finding the minimum of the distances for every sample. For the *Ustatistic*-estimator the

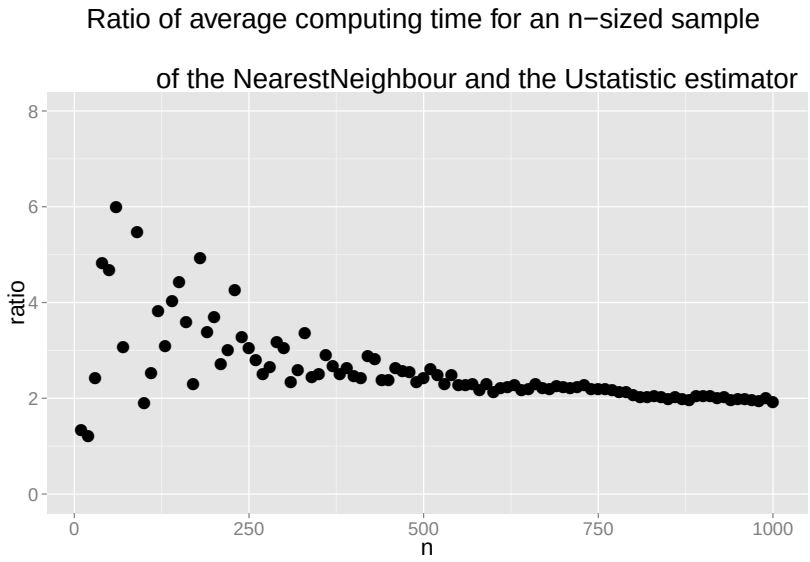


Figure 18: Ratio computing time for both estimators, the *Ustatistic*-estimator performs around twice as fast

distances only have to be compared to one fixed threshold ϵ, which means that a lot less computations have to be done.

5.7.4 Comparison for multidimensional densities

In the simulation study samples of a two- and a three-dimensional standard normal random variable have been drawn and evaluated in the same manner as above in this section. On a big scale the resulting plots for the variance and the residuals look very similar to the one-dimensional case. For small sample sizes N the *Ustatistic*-estimator behaves oddly, see Figure 20, which will be discussed in the following section. The results of the model for two-dimensional standard normal distribution with trimmed sample size N are

given in this section. In order to avoid the problematic small sample sizes, the models tabulated here were only fitted on the data sets for which $N \geq 40$, see Figure 19. Both of the estimators are asymptotically unbiased and their variance tends to zero. The *Ustatistic*-estimator is converging faster and has a smaller variance for a given sample size. The estimated parameters for the non-linear model for the residuals of the *NearestNeighbour*-estimator are given in Table 24 and for the *Ustatistic*-estimator in Table 25. The residuals for a given sample size are smaller for the *Ustatistic*-estimator than for the *NearestNeighbour*-estimator. Based on the simulation it can be concluded that the bias of the *Ustatistic*-estimator is smaller.

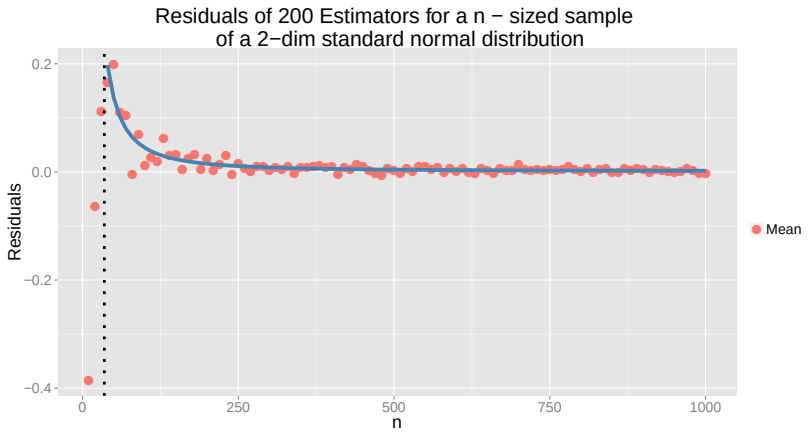


Figure 19: Averaged residuals of *Ustatistic*-estimator for a sample size n for a two-dimensional standard normal distribution. The vertical black line represents the minimal threshold for sample sizes that were being used for calculating the bias and variance models. The blue line is the estimated bias from the model described in Table 24.

The estimated parameters for the log-transformed linear model for the variance of the *NearestNeighbour*-estimator are given in Table 26 and for the

Table 24: Estimated model coefficients for the for the bias of the *Nearest-Neighbour*-estimator for a twodimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	1.421	1.825	2.363
β_2	0.656	0.749	0.847
β_3	-0.010	-0.002	0.005

Table 25: Estimated model coefficients for the for the bias of the *Ustatistic*-estimator for a two-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
β_1	35.010	85.348	225.652
β_2	1.419	1.646	1.897
β_3	-0.002	0.001	0.004

Ustatistic-estimator in Table 27. The variance of the *Ustatistic*-estimator is smaller than the variance of the *NearestNeighbour*-estimator, given the same sample size. The rate in which the variance of the *Ustatistic*-estimator tends to zero is greater than the rate of the *NearestNeighbour*-estimator.

Table 26: Estimated model coefficients for the variance of the *NearestNeighbour*-estimator for a two-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
$\hat{C}$	0.226	0.518	0.811
$\hat{a}$	-0.731	-0.682	-0.633

Table 27: Estimated model coefficients for the variance of the *Ustatistic*-estimator for a two-dimensional standard normal distribution.

	lower.bound	estimated	upper.bound
$\hat{C}$	4.064	4.400	4.736
$\hat{a}$	-1.581	-1.525	-1.469

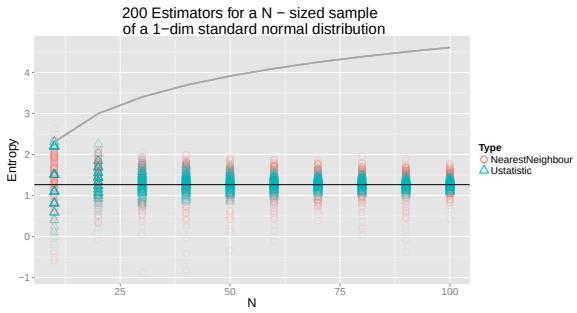
5.7.5 Biasedness for small sample sizes

Both estimators are biased for small sample sizes N, they are only asymptotically unbiased, that means for $N \to \infty$ the bias tends to zero. The *NearestNeighbour*-estimator tends to overestimate the quadratic entropy for small N, whereas the *Ustatistic*-estimator runs into a numerical problem for small sample sizes. The number of ϵ-close observations N_n in the sample is a random number in the finite set of natural numbers $\left\{1, \ldots, \frac{n(n-1)}{2}\right\}$, see Equation (32) in Section 4.3. This results in a discreteness of the possible results of the estimator, which is not really a problem in a simulation study with many repetitions. There is a problem with the choice of ϵ, as for too small ϵ the probability of having any ϵ-close vectors is very low for small sample sizes, which can result in an additional bias for small repetition sizes. More of a problem is an anomaly in the *Ustatistic*-estimator for small sample sizes caused by trimming the estimator for the quadratic Rényi entropy H_2^*, see equation (11),

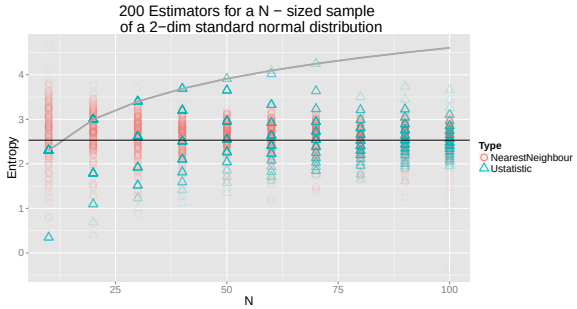
$$\tilde{H}_N^* = -\log\left(\max\left\{\tilde{Q}_N, \frac{1}{N}\right\}\right).$$

For small sample sizes all estimated Q_N are smaller than $\frac{1}{N}$, resulting in the estimated entropy defaulting to $-\log(\frac{1}{N})$, which underestimates the real entropy for small sample sizes, see Figure 20. A suitable choice of ϵ is crucial for small sample sizes, but not trivial to be found as there are no theoretical results in literature. A data based choice of ϵ might be a topic of future investigation that can help minimising the bias for small sample sizes. This cause of bias seems to be more of a problem for higher dimensional data, as the *Ustatistic*-estimator of the one-dimensional standard normal distribution seems to be almost unaffected by trimmimg to the limit $-\log\left(\frac{1}{N}\right)$, see Figure 20a.

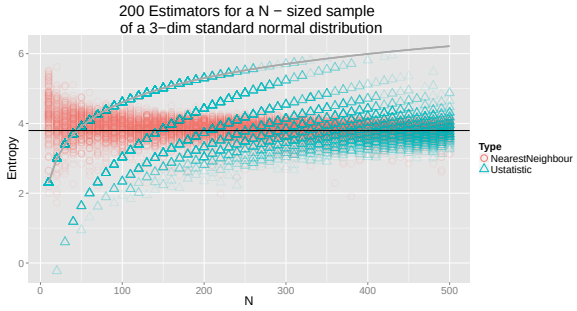
For two-dimensional data (Figure 20b) the effect starts to be noticable for small N and for three-dimensional data a relatively big sample size is required in order to make up for that effect, see Figure 20c. Also the discreteness of the possible outcomes of the *Ustatistic*-estimator is more visible in the higher-dimensional cases. One can see that for higher-dimensional data the sample size needs to be greater for the *Ustatistic*-estimator not to be unbiased.



(a) Estimation of the quadratic entropy for small sample sizes N for a one-dimensional standard normal distribution.



(b) Quadratic entropy for a two-dimensional standard normal distribution.



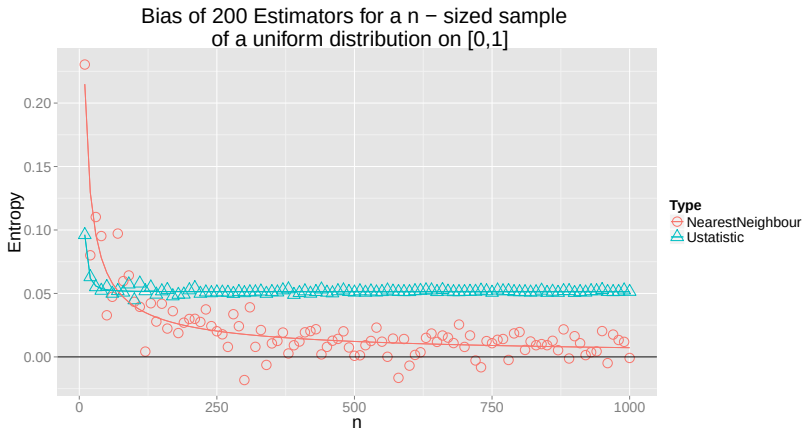
(c) Quadratic entropy for a three-dimensional standard normal distribution. Note the bigger scale of the x-axis.

Figure 20: Quadratic entropy for small sample sizes N for a one- two- and three-dimensional standard normal distribution. The grey line is the theoretical maximum, $-\log\left(\frac{1}{N}\right)$, of the *Ustatistic*-estimator based on the sample size. The horizontal black line is the theoretical entropy of the distribution from which the samples were drawn. In order to reduce the effects of overplotting, the points have been plotted transparently. A more intense colour represents more points in that spot.

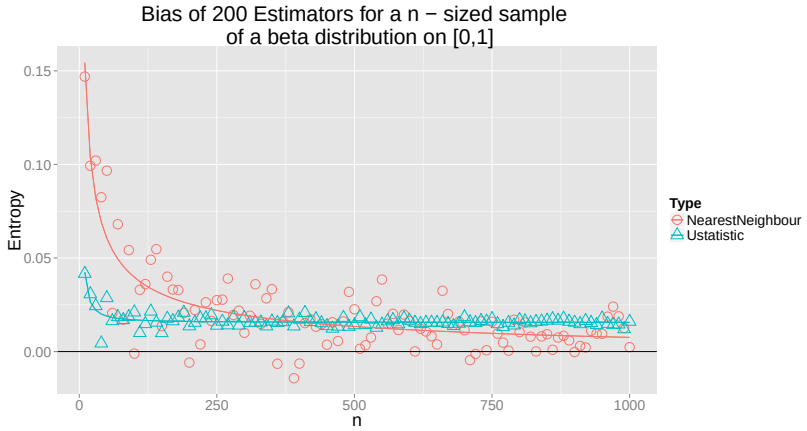
5.7.6 Bounded distributions

In this section the two estimators for quadratic entropy considered in this thesis, the *Ustatistic*-estimator and the *NearestNeighbour*-estimator, will be compared for bounded distributions. We will compare the two estimators for the beta distribution, see Definition 2.6 with its Rényi entropy given in Theorem 2.4, and its special case the uniform distribution, see Definition 2.7 with its Rényi entropy given in Theorem 2.5. Comparing the entropy estimators for bounded distributions is an important case to consider, as stronger convergence of the estimators could be proven theoretically. Surprisingly the *Ustatistic*-estimator is found to be biased for the Uniform distribution on $[0, 1]$, see Figure 21a, as well as for the beta distribution with parameters $\alpha = \beta = 2$, see Figure 21b, that has the same bounded support.

The observation of biasedness of the *Ustatistic*-estimator can be verified by using the non-linear model for the bias presented in Section 5.5. The *NearestNeighbour*-estimator, see Tables 28 and 30, can be assumed to be asymptotically unbiased, as the confidence interval of the parameter β_3 includes zero. The *Ustatistic*-estimator, see Tables 29 and 31, can be assumed to be asymptotically biased, as the confidence interval for the parameter β_3, which represents the bias for $N \to \infty$ does not include zero. On a 95% confidence level this estimator is biased for a beta distribution with parameters $\alpha = \beta = 2$.



(a) Average bias for a sample sizes N of the estimation of the quadratic entropy for a one-dimensional uniform distribution on $[0, 1]$



(b) Average bias for a sample sizes N of the estimation of the quadratic entropy for a one-dimensional beta(2,2) distribution with finite support $[0, 1]$

Figure 21: Average bias for a sample sizes N of the *Ustatistic* and *Nearest-Neighbour*-estimator for the quadratic entropy.

Table 28: Estimated model coefficients for the bias of the *NearestNeighbour*-estimator for a uniform distribution on $[0, 1]$.

	lower.bound	estimated	upper.bound
β_1	0.807	1.152	1.679
β_2	0.596	0.729	0.872
β_3	-0.009	-0.0003	0.006

Table 29: Estimated model coefficients for the bias of the *Ustatistic*-estimator for a uniform distribution on $[0, 1]$.

	lower.bound	estimated	upper.bound
β_1	2.135	5.455	35.375
β_2	1.622	2.085	2.890
β_3	0.051	0.051	0.052

Table 30: Estimated model coefficients for the bias of the *NearestNeighbour*-estimator for a beta distribution with parameters $\alpha = \beta = 2$.

	lower.bound	estimated	upper.bound
β_1	0.387	0.568	0.858
β_2	0.397	0.552	0.711
β_3	-0.021	-0.005	0.004

Table 31: Estimated model coefficients for the bias of the *Ustatistic*-estimator for a beta distribution with parameters $\alpha = \beta = 2$.

	lower.bound	estimated	upper.bound
β_1	0.193	0.514	2.517
β_2	0.866	1.282	1.949
β_3	0.014	0.016	0.016

5.7.7 Discussion of the comparison

Comparing the *Ustatistic*-estimator and the *NearestNeighbour*-estimator gets to ambivalent results. For the case of a one-dimensional normally distributed random variable the *Ustatistic*-estimator is superior to the *NearestNeighbour*-estimator concerning the bias as well as the variance. The bias converges faster to zero for growing sample sizes and the variance is smaller for a given sample size, which make the *Ustatistic*-estimator superior in this case. For higher-dimensional normal distributions this has to be put into perspective by the observation that the *Ustatistic*-estimator only has a discrete number of possible outcomes, which can make it impossible to discriminate between the entropy of similar samples for small sample sizes. In practical application this is a great weakness. Additionally, the *Ustatistic*-estimatoris heavily for small sample sizes of high dimensional distributions. Unless the sample size is big enough to outweigh this effect, the *NearestNeighbour*-estimator is a better choice to estimate the quadratic entropy of higher-dimensional samples.

Different again is the case of bounded distributions. On the basis of uniformly distributed and beta distributed samples the *NearestNeighbour*-estimator is asymptotically unbiased, but an asymptotical bias was observed for the *Ustatistic*-estimator. This is usually a disqualifying property for an estimator, but as the bias converges to a fixed value it can still be used to compare the entropy of samples of different sizes from the same distribution, given they have a minimum sample size where the bias can be assumed to be constant. Using the *Ustatistic*-estimator is still attractive, as it has a

smaller variance than the *NearestNeighbour*-estimator. That being said, for a general estimation of the entropy of a bounded distribution the *NearestNeighbour*-estimator is recommended. In Table 32 this information is summarised compactly.

Table 32: Overview of the best performing estimator for the quadratic Rényi entropy under the aspects of variance and unbiasedness. With * labelling the cases where the *Ustatistic*-estimator performs better for large sample sizes.

Distribution	Bias	Variance
Normal Distribution (one-dim)	*Ustatistic*	*Ustatistic*
Normal Distribution (two-dim)	*NearestNeighbour**	*NearestNeighbour**
Normal Distribution (three-dim)	*NearestNeighbour**	*NearestNeighbour**
Uniform distribution	*NearestNeighbour*	*Ustatistic*
Beta(2,2) distribution	*NearestNeighbour*	*Ustatistic*

Note that in this simulation the parameter ϵ of the *Ustatistic*-estimator and the parameter k of the *NearestNeighbour*-estimator are chosen based on values recommendations in literature without further investigating. A suitable choice obviously affects the bias and the variance for a given sample size and distribution. This is left for further investigation. One shall be reminded that the *Ustatistic*-estimator can exclusively be used for estimating the Rényi entropy in the special case of $q = 2$, the quadratic Rényi entropy, whereas the *NearestNeighbour*-estimator can be used to estimate the general case for all q, including the special case of $q = 2$.

5.8 Investigation of biasedness of the nearest neighbour estimator

In this section the theoretical results concerning the bias of the *Nearest-Neighbour*-estimator given in Theorem 4.5 of Section 4.1 are investigated by simulation. The theoretical information of a normally distributed variable X is given in Theorem 2.3 as

$$I_q(X) = \sqrt{\frac{(2\pi)^{(1-q)m}|\Sigma|^{(1-q)}}{(q)^m}},$$

and $I_1 = 1$. This relationship between the Information I_q, its order q and the dimension m is shown in Figure 22.

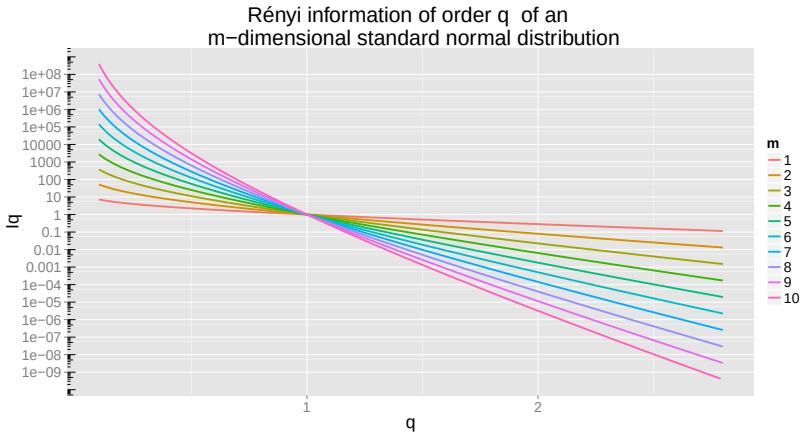


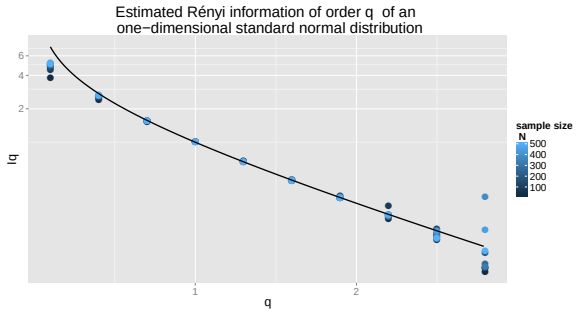
Figure 22: Theoretical Rényi information I_q of an m-dimensional standard normal distribution. Note the log-transformed y-axis.

In order to investigate the behaviour of the bias of the nearest neighbour estimator, the entropy was estimated for a range of parameters. In accor-

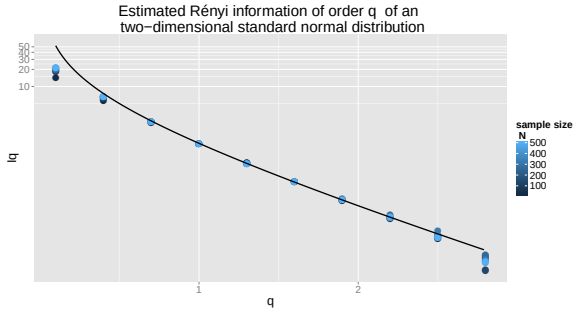
dance with the simulation study executed in Leonenko and Pronzato (2010), the bias of the estimated information grows for small and big values of q. Interestingly the nearest neighbour estimator for information is unbiased for any dimension m and sample size N of the standard normal distributed sample. Additionally, the bias is smaller for growing sample sizes N, which can be seen from Figures 23a, 23b and 23c.

The dimensionality m of the samples does not influence the general trend of the bias. The parameter k for the kth nearest neighbour is set to two. Changing it would only change the picture quantitatively but not qualitatively. The general trend of the bias would stay the same.

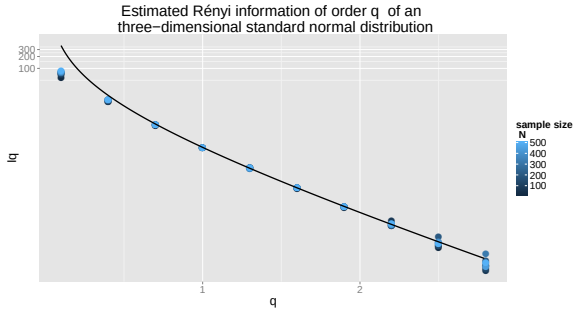
Using non-linear models (Section 5.2) the convergence bias proposed in Theorem 4.5 can be checked for the normal distribution. Due to the limited time of this project, this could not be investigated further.



(a) Estimated Rényi information I_q of an one-dimensional standard normal distribution.



(b) Estimated Rényi information I_q of an two-dimensional standard normal distribution.



(c) Estimated Rényi information I_q of an three-dimensional standard normal distribution.

Figure 23: Estimated Rényi information I_q of an one- two- and three-dimensional standard normal distribution. The black line is the theoretical Rényi entropy, the blue points the average of the estimated information for a given sample size. Note the log-transformed y-axis.

6 Discussion

After introducing the concept of Shannon entropy, and its generalisations the Rényi and Tsallis entropy, as measures of disorder, a number of different estimators for Shannon, Rényi and Tsallis entropy are compared theoretically and by simulation. There is a number of entropy estimators for which, under varying conditions, a number of important theoretical results have been proven in literature, including asymptotic unbiasedness and consistency. Theoretical results on the bias for a given sample size and other performance criteria are hard to quantify theoretically and illuminated in the simulation study. Comparing the nearest neighbour estimator presented in Leonenko and Pronzato (2010) to the other estimators like the spacing based estimators of entropy presented in Beirlant et al. (1997) and Song (2000) for one-dimensional distributions and the estimator for quadratic entropy by Källberg et al. (2014) shows its usefulness. In estimating the Shannon entropy the nearest neighbour estimator performed well, see Table 19, but in some cases not as good as spacing estimators given in Beirlant et al. (1997) and Song (2000). For estimating the quadratic entropy, the estimator presented in Källberg et al. (2014) is superior only for a one-dimensional normal distribution, but for high dimensional distribution it has some severe problems that make the nearest neighbour estimator a better choice. For bounded distributions the estimator given in Källberg et al. (2014) has a smaller variance, but is biased, see Table 32. A major advantage of the nearest neighbour

estimator is that it is applicable for all cases, it can estimate the Shannon, Reńyi and Tsallis entropy of arbitrary order for one- and multidimensional distributions, see Table 2. This is a very useful property as it allows a broad spectrum of appliance. There are some special cases where other estimators perform better, but generally the nearest neighbour estimator performs well. This makes it a useful workhorse for real life applications.

Acknowledgements

I would like to thank Professor Nikolai Leonenko for his assistance and guidance throughout this project. His help and time is much appreciated and I was able to learn a lot from him.

A Introduction to R

For all the simulations and plots R version 3.0.0 is used (R Core Team; 2013b). R is a free software environment for statistical computing and graphics that provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques and is highly extensible. It is often the vehicle of choice for research in statistical methodology, and R provides an open source route to participation in that activity. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. R can be extended via packages. There are some basic packages supplied with the R distribution and numerous are available through the CRAN family of Internet sites covering a very wide range of modern statistics. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS (R Core Team; 2013b).

A number of packages is used in addition to the base package. These are the packages *ggplot2* for the graphical output (Wickham; 2009), *reshape2* for manipulating data frames (Wickham; 2007), and *stargazer* for the presentation of model outputs in tables (Hlavac; 2014).

B Efficient simulating in R

Efficient programming is an important issue for all programming, as efficient code tends to be shorter, simpler, safer and obviously faster. With growing computing power nowadays it may seem less of a pressing subject, but in fact it is even more so, as bigger and bigger data sets are available to be analysed. For a simulation study like in this project, efficient coding pays back directly in more precise results for the same computation time. As R is a high level language, it is not an obvious first choice for a simulation program, but it is designed to connect to high-performance programming languages like *Fortran*, `C` and `C++`. Thus it combines the convenience of a high level language, for example automated memory and data type allocation, with the computational speed of a low level language. In Section B.1 the topic of vectorisation will be introduced and motivated. Section B.2 introduces the concept of parallel computing and how it can be used to achieve more efficiency.

B.1 Vectorisation

Vectorisation in R is a way to achieve better performance. Most of the basic functions in R are actually written in a low-level language and only have a wrapper passing the data on to a low level language where the operation will be executed. But the input of the function still has to be interpreted before passing it on to the compiled code, which can be time consuming.

Example B.1. This is a rather technical example, where efficient programming will be explained on an example from the code used for this project. Efficient code writing will be demonstrated by means of the implementation of the *mn_spacing*-estimator. Consider a sample of size N with the ordered observations $\{X_{(1)}, \ldots, X_{(N)}\}$. First, the indices $i+m, i-m$, of the spacing $X_{(i+m)}, \ldots, X_{(i-m)}$ have to be computed and afterwards bounded to $[1, N]$. The code necessary to do this is timed with the system.time() function, returning the run time in seconds. The following is the output of the console from running a chunk of code from the Appendix.

```
> vec = get_sample(N = 10000000, m = 1,
+                  distribution = "Normal",
+                  parameters = c(0,1))
>
> # Fast vectorised Version
> system.time({
+   N <- length(vec)
+
+   # mn is function of n
+   mn <- (1:N)^(1/3)
+
+   # X_{i+m}
+   upper <- 1:N + mn
+
```

```
+    # X_[i-m]
+    lower <- 1:N - mn
+
+    # Limit to bound [1, N]
+    upper[upper > N] <- N
+    lower[lower < 1] <- 1
+ }
+ )
   user  system elapsed
   1.78    0.16    1.94
>
> # Slow version using loops
> system.time({
+    N <- length(vec)
+
+    # mn is function of n
+    mn <- N^(1/3)
+
+    temp = 0
+
+    for(i in 1:N){
+
+      # X_{i+m}
```

```
+       upper <- i + mn
+
+       # X_[i-m]
+       lower <- i - mn
+
+       # Limit to bound [1, N]
+       if(i + mn > N){
+         upper <- N
+       }
+       if(i - mn < 1){
+         lower <- 1
+       }
+     }
+ }
+ )
    user  system elapsed
 111.30    0.10  111.54
```

It can be seen that for this example just by vectorising the code, it runs about $111.54s/1.94s = 57.49$ times faster than using loops. This is due to a number of causes.

Vectorising the bounding of the vector to $[1, N]$ makes the code run faster. After being handed down to a low level language the command

```
# Limit to bound [1, N]
upper[upper > N] <- N
```

will still be executed using loops, but loop runs are a lot faster in a low level language than in R. How much of a difference this makes can be seen in Figure 24. The regression lines drawn represent a linear model of the form

$$E[\text{computation time}] = \beta_0 + \beta_1 N.$$

The estimated model parameters are shown in Table 33.

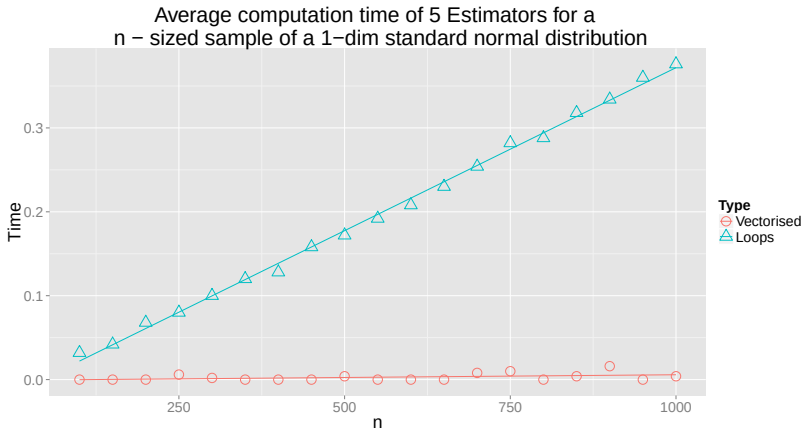


Figure 24: Average computation time of 5 repetitions against sample size n, $n = \{100, 150, 200, ..., 1000\}$. The lines represent linear models fit for the two approaches, see Table 33.

Of special interest is the estimated slope $\hat{\beta}_1$. It can be interpreted as the expected time the computation takes longer if the sample size is increased by one. Obviously this number is very small for both the vectorised code and

Table 33: Estimated model parameters of a linear regression of computation time $\sim N$.

$\hat{\beta}_0$	$\hat{\beta}_1$
$-7.474 \cdot 10^{-4}$	$6.526 \cdot 10^{-6}$

(a) Vectorised Code

$\hat{\beta}_0$	$\hat{\beta}_1$
$-1.680 \cdot 10^{-2}$	$3.886 \cdot 10^{-4}$

(b) Code using loops

using loops. More insightful is the ratio of these parameters, the vectorised code runs about 60 times ($\frac{6.526 \cdot 10^{-6}}{3.886 \cdot 10^{-4}} \approx 59.54$) faster than the version using loops. For large sample sizes and a large number of repetitions this makes a huge difference.

B.2 Parallelisation

Speeding up the run time by vectorisation is limited. Another way of writing more time efficient code is parallelisation. Parallel computing allows carrying out many calculations simultaneously, by splitting up the problem into smaller problems that can be solved independently. These smaller tasks are split among the (multiple) processors of the computer. The result for the problem is gained by putting the results of the small problems back together. This means, that the computations are done simultaneously by the different processor cores at the same time. This does not change the number of calculations to be done nor the speed at which the calculations are executed, but it can reduce the run time by doing the computations at the same time. The downside of parallelisation is that there is a lot of frame work needed to

be set up. The problem has to be divided and the results put back together. Also it involves more lines of code and is harder to read. For small problems there is no point in parallelising, as the extra time the computer need to set up the parallelisation exceeds the time saved by solving the problem faster. This problem is shown Figure 25. The run time is plotted against the number of executions of the same function. It can be seen that for a small number of repetitions the vectorised version is faster, but for larger problems the parallelised code runs faster. The additional time it takes to set up the parallelisation is only made up for when the problem is large. From a

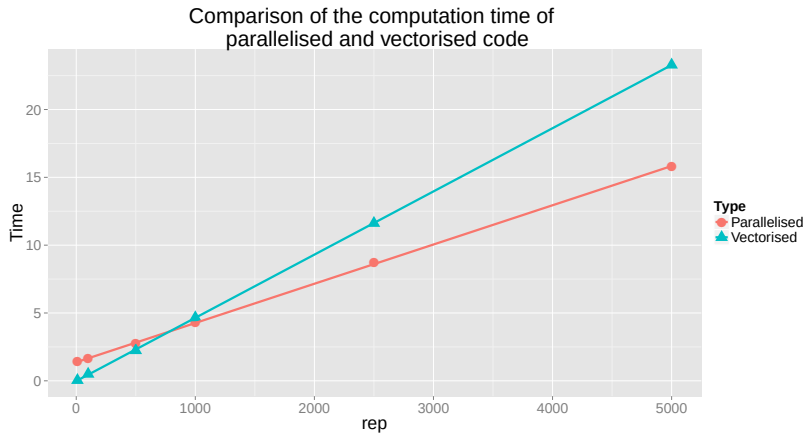


Figure 25: The run time is plotted against the number of times the function is executed.

more practical point of view, problems have to be really large in order for the additional time spent programming being made up for by achieving a shorter run time. In R the progress of the simulation can be monitored using a progress bar from the package utils (R Core Team; 2013c). Unfortunately

this handy tool is incompatible with parallelisation so far. There is a number of R packages that allow parallelisation. For this project the packages *parallel* (R Core Team; 2013a), *doSNOW* (Revolution Analytics and Weston; 2014a), and *foreach* (Revolution Analytics and Weston; 2014b) were used.

References

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S. and Modha, D. S. (2004). A generalized maximum entropy approach to bregman co-clustering and matrix approximation, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 509–514.

Beirlant, J., Dudewicz, E. J., Györfi, L. and Van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview, *International Journal of Mathematical and Statistical Sciences* **6**(1): 17–39.

Bierens, H. J. (1996). *Topics in advanced econometrics*, Cambridge University Press.

Borgwaldt, S. R., Hellwig, F. M. and De Groot, A. M. B. (2005). Onset entropy matters–letter-to-phoneme mappings in seven languages, *Reading and Writing* **18**(3): 211–229.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R. and Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs, *Frontiers in human neuroscience* **8**.

Cartwright, J. (2014). Roll over, Boltzmann, *Physics World* pp. 31–35.

Conrad, K. (2013). Probability distributions and maximum entropy.
URL: *http://www.math.uconn.edu/˜kconrad/blurbs/analysis/entropypost.pdf*

Cornforth, D. J., Tarvainen, M. P. and Jelinek, H. F. (2014). How to calculate Renyi entropy from heart rate variability, and why it matters for detecting cardiac autonomic neuropathy, *Frontiers in bioengineering and biotechnology* **2**.

Cramér, H. (1999). *Mathematical methods of statistics*, Vol. 9, Princeton University Press.

Cressie, N. (1976). On the logarithms of high-order spacings, *Biometrika* **63**(2): 343–355.

Davidian, M. (2005). Simulation studies in statistics.
URL: *http://www4.stat.ncsu.edu/˜davidian/st810a/simulation_handout.pdf*

Fahrmeir, L., Kneib, T. and Lang, S. (2007). *Regression*, Springer Berlin Heidelberg.

Fahrmeir, L., Künstler, R., Piegot, I. and Tutz, G. (1997). *Statistik-der Weg zur Datenanalyse*.

Fox, J. (2002). *An R and S-Plus companion to applied regression*, Sage.

Gallant, A. R. (1975). Nonlinear regression, *The American Statistician* **29**(2): 73–81.

Hegde, A., Lan, T. and Erdogmus, D. (2005). Order statistics based estimator for Renyi's entropy, *2005 IEEE Workshop on Machine Learning for Signal Processing*, IEEE, pp. 335–339.

Hlavac, M. (2014). *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*, Harvard University, Cambridge, USA. R package version 5.1.
URL: *http://CRAN.R-project.org/package=stargazer*

Johnson, O. (2004). *Information Theory and the Central Limit Theorem*, Imperial College Press, London.

Källberg, D., Leonenko, N. and Seleznjev, O. (2014). Estimation of quadratic density functionals under m-dependence, *Journal of Nonparametric Statistics* **26**(2): 385–411.

Leonenko, N. and Pronzato, L. (2010). Correction of "A class of rényi information estimators for multidimensional densities", *The Annals of Statistics* **38**(6): 3837–3838.

Leonenko, N. and Seleznjev, O. (2010). Statistical inference for the ϵ-entropy and the quadratic Rényi entropy, *Journal of Multivariate Analysis* **101**(9): 1981–1994.

Misra, N., Singh, H. and Hnizdo, V. (2010). Nearest neighbor estimates of entropy for multivariate circular distributions, *Entropy* **12**(5): 1125–1144.

Nadarajah, S. and Zografos, K. (2003). Formulas for Rényi information and related measures for univariate distributions, *Information Sciences* **155**(1): 119–138.

Nielsen, F. and Nock, R. (2011). On Rényi and Tsallis entropies and divergences for exponential families, *arXiv preprint arXiv:1105.3259* .

R Core Team (2013a). *parallel: Support for Parallel computation in R.*
URL: *https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf*

R Core Team (2013b). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

R Core Team (2013c). *utils: R utility functions*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://stat.ethz.ch/R-manual/R-devel/library/utils/html/utils-package.html*

Revolution Analytics and Weston, S. (2014a). *doSNOW: Foreach parallel adaptor for the snow package.* R package version 1.0.12.
URL: *http://CRAN.R-project.org/package=doSNOW*

Revolution Analytics and Weston, S. (2014b). *foreach: Foreach looping construct for R.* R package version 1.4.2.
URL: *http://CRAN.R-project.org/package=foreach*

Sabuncu, M. R. (2006). *Entropy-based image registration*, PhD thesis, Princeton University.

Shamilov, A. and Yolacan, S. (2006). Statistical structure of printed Turkish, English, German, French, Russian and Spanish, *WSEAS Transactions on Mathematics* **5**(6): 756.

Shannon, C. E. (1948). Communication theory of secrecy systems, *Bell system technical journal* **27**: 379–423.

Song, K.-S. (2000). Limit theorems for nonparametric sample entropy estimators, *Statistics & probability letters* **49**(1): 9–18.

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

Vignat, C., Hero III, A. O. and Costa, J. A. (2004). About closedness by convolution of the Tsallis maximizers, *Physica A: Statistical Mechanics and its Applications* **340**(1): 147–152.

Wang, Q., Kulkarni, S. R. and Verdú, S. (2006). A nearest-neighbor approach to estimating divergence between continuous random vectors, 2006 IEEE Int. Symp. Information Theory, Seattle, WA, USA.

Wickham, H. (2007). Reshaping data with the reshape package, *Journal of Statistical Software* **21**(12): 1–20.
URL: *http://www.jstatsoft.org/v21/i12/*

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*, Springer New York.
URL: *http://had.co.nz/ggplot2/book*

Zografos, K. and Nadarajah, S. (2005). Expressions for Rényi and Shannon entropies for multivariate distributions, *Statistics & Probability Letters* **71**(1): 71–84.

Eidesstattliche Erklärung

Hiermit versichere ich, Thomas Maierhofer, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

(Ort, Datum)	(Unterschrift)