Quadratic sample entropy as a measure of burstiness

A study in how well Rényi entropy rate and quadratic sample entropy can capture the presence of spikes in time-series data

Kira Huselius Gylling

Supervisor: Sara Sjöstedt de Luna
Abstract

Requests to internet servers do not in general behave in a manner which can be easily modelled and forecast with typical time-series methods, but often have a significant presence of spikes in the data, a property we call “burstiness”. In this thesis we study various entropy measures and their properties for different distributions, both theoretically and via simulation, in order to better find out how these measures could be used to characterise the predictability and burstiness of time series. We find that a low entropy can indicate a heavy-tailed distribution, which for time series corresponds to a high burstiness. Using a previous result that connects the quadratic sample entropy for a time series with the Rényi entropy rate of order 2, we suggest a way of detecting burstiness by comparing the quadratic sample entropy of the time series with the Rényi entropy rate of order 2 for a symmetric and a heavy-tailed distribution.

## Contents

1 Introduction ................................. 1  
1.1 Background .................................. 1  
1.2 Purpose ..................................... 1  
1.3 Scope and limitations ..................... 1  
1.4 Approach and outline .................... 1  

2 Theory ........................................ 2  
2.1 Stationary processes and Markov chains ........................................ 2  
2.2 Entropy measures .......................... 4  
2.2.1 Shannon entropy ........................ 4  
2.2.2 Rényi entropy of order $q$ ............ 4  
2.3 Data-driven entropy measures ............ 7  
2.3.1 Approximate entropy .................... 7  
2.3.2 Sample entropy ........................... 8  
2.3.3 Quadratic sample entropy ............... 9  
2.3.4 Choice of parameters .................... 9  

3 Examples of Rényi entropies and their properties for some distributions .... 10  
3.1 Rényi entropy for some univariate distributions .......................... 10  
3.1.1 The normal distribution .................. 10  
3.1.2 The uniform distribution ................. 11  
3.1.3 The exponential distribution .......... 12  
3.1.4 The Laplace distribution .................. 12  
3.1.5 Order of distributions with respect to Rényi entropies ................. 13  
3.2 Rényi entropy for multivariate normal distributions ...................... 15  
3.2.1 Some properties .......................... 15  
3.2.2 Example: MA(1) Gaussian process .... 17  
3.2.3 Example: AR(1) Gaussian process ....... 17  

4 Numerical illustrations ....................... 20  
4.1 Simulated non-bursty data .................. 20  
4.1.1 Independent and 1-dependent data .... 20  
4.1.2 Autoregressive process .................... 23  
4.1.3 Process with deterministic trend ......... 24  
4.2 Simulated bursty data ..................... 25  
4.2.1 Generation of the series ................. 25  
4.2.2 Characterising the predictability of time series ............... 29  
4.2.3 Sensitivity to choice of $r$ and $m$ . . . . 32  
4.2.4 Comparison of various measures of burstiness .... 35  

5 Discussion and conclusions .................. 36  

6 Acknowledgements ............................ 36  

Bibliography .................................... 37
1 Introduction

1.1 Background

Requests to internet servers do not in general behave in a manner which can be easily modelled and forecast with typical time-series methods, due to their non-stationary behaviour with extreme outliers (spikes), jumps, and changing patterns. Examples of this are news websites at major events such as the recent American presidential election, university admissions websites on the last date of application, and whenever content on a website goes viral on social media. Often big data centres with many servers handle these websites, so at these points in time it is important to have enough servers to keep the website running, so the website does not go down due to server overload. At the same time, in order to conserve electricity, it is important not to have too many servers running at times when a spike is not occurring or about to occur.

In order to deal with this, we want to find summarizing measures that can characterise different predictive properties of time series, and that can be used to classify time series in different predictive classes in an automated manner. Then, given a class, the appropriate predictive model is adapted in an automated manner and predictions are made, once again automated from the adapted model, in order to be able to adjust the number of servers that are needed to handle the future requests.

Properties that are important to capture with such a measure, especially when dealing with cloud data, social media websites, data centers, electricity usage, etc., are variance, stochastic and deterministic dependence, and “burstiness”, i.e. the presence of spikes and extreme outliers.

In connection to this, we find it interesting to study closer various entropy measures, and which type of information they can give us with respect to these needs. In this thesis we therefore study entropy measures and their properties for different distributions, both theoretically and via simulation, in order to better find out how these measures could be used for these purposes.

1.2 Purpose

The main purpose of the thesis is to study various entropy measures for distributions and time series, and the possibilities for them to characterise various features such as burstiness. We will mainly focus on the Rényi entropy of order 2, and the Rényi entropy rate of order 2 for dependent sequences of data. Via theoretical calculations and simulated data, we will study and illustrate the Rényi entropy rate of order 2, as well as properties of some of its estimates, based upon the quadratic sample entropy (QSE), and the focus will be on stationary series. Via simulation we will also study quadratic sample entropy for some possibly non-stationary series, and reflect over how it can be used as a measure to capture burstiness.

1.3 Scope and limitations

The theory used is mainly for stationary processes, although we will also show how QSE behaves for non-stationary, or perhaps non-trivial stationary, cases. It is outside the scope of this thesis to show whether the studied series are stationary or not.

1.4 Approach and outline

In Section 2.1 we will begin with a necessary description of stationary processes and Markov chains. Next, in Section 2.2 we present various entropy measures for univariate and multivariate distributions. Firstly we present Shannon entropy and normalised entropy to give a background on the concept of entropy. After this we present the Rényi entropy and differential Rényi entropy rate of order $q$, which are central to this thesis. We then go on in Section 2.3 to describe various data-driven entropy measures: Approximate entropy (ApEn) and sample entropy (SampEn). The latter is then tied to the Rényi entropy rate of order 2 via the quadratic sample entropy (QSE), which is the measure that is mainly studied in this thesis.

After this, in Section 3 we examine and compute the Rényi entropy for some univariate distributions, as well as for some multivariate normal distributions. The latter is exemplified with moving-average and
autoregressive processes.

In Section 4 we then describe some numerical illustrations that were performed in order to see how well
the QSE estimates the Rényi entropy rate of order 2. We begin in Section 4.1 by studying relatively regular
time series, over which we have a degree of control and for which the theoretical Rényi entropy rate can
be computed. Firstly we study independent data from the normal and the exponential distributions. After
this we study an autoregressive process, and finally a process with a deterministic linear trend. In each case
we study the original series \(X_t\) and the differenced series \(\nabla X_t = X_t - X_{t-1}\), and in each case the QSE is
computed and compared with the Rényi entropy rate of order 2.

In Section 4.2, we study seven bursty and irregular series with spikes, in order to see how well the theory
can be applied to bursty series. The original series \(X_t\), the differenced series \(\nabla X_t = X_t - X_{t-1}\), and the
relative difference \(D_t = (X_t - X_{t-1})/X_{t-1}\) are studied through QSE as well as SampEn. We study the
histograms of the series and compare their QSE’s to the QSE’s of white noise, and comment on how this can
be used to classify different series. Next we study the sensitivity to the choice of various parameters used in
the algorithms. Finally we compare different ways of measuring the burstiness of these seven series, to see
how well these measures measure up to each other.

Finally in Section 5, we present discussions and conclusions.

2 Theory

This chapter begins with an introduction to stationary processes and Markov chains. After this we go on to
describing various entropy measures, and next we describe some data-driven entropy measures.

2.1 Stationary processes and Markov chains

A stochastic process \(\{X_t, t \in \{0, \pm 1, \ldots\}\}\), such as a time series, is said to be stationary if its statistical
properties are, in some sense, similar to those of the shifted series \(\{X_{t+h}, t \in \{0, \pm 1, \ldots\}\}\) for all integers \(h\).
For a more precise definition of stationarity, we need to define what we mean by statistical properties of a
stochastic process. Restricting ourselves to first- and second-order moments, we define the following:

Let \(\{X_t\}\) be a stochastic process such that \(E[X_t^2] < \infty\). Then the mean function \(\mu_X(t)\) of \(\{X_t\}\) is

\[
\mu_X(t) = E[X_t].
\]

The covariance function \(\gamma_X(r, s)\) of \(\{X_t\}\) is

\[
\gamma_X(r, s) = \text{Cov}(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))] = E[X_r X_s] - \mu_X(r)\mu_X(s),
\]

for all integers \(r, s\). Now we can proceed to define what we mean by stationarity. There are two types of
stationarity: weak and strict stationarity. A stochastic process \(\{X_t\}\) is weakly stationary if

1. \(\mu_X(t)\) is independent of \(t\), and
2. \(\gamma_X(t + h, t)\) is independent of \(t\) for all integers \(h\).

A stochastic process \(\{X_t\}\) is strictly stationary if each subsequence \(\{X_t\}_{t=1}^n\) of it has the same distribution
as the shifted sequence \(\{X_{t+h}\}_{t=1}^n\), i.e. if

\[
(X_1, \ldots, X_n) \overset{d}{=} (X_{1+h}, \ldots, X_{n+h}),
\]

for all integers \(h\) and \(n \geq 1\). For stationary processes, we can define the autocovariance function (ACVF).
The autocovariance function \(\gamma_X(h)\) of a stationary process \(\{X_t\}\) at lag \(h\) is

\[
\gamma_X(h) = \text{Cov}(X_{t+h}, X_t).
\]
The autocorrelation function (ACF) $\rho(h)$ of $\{X_t\}$ at lag $h$ is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \text{Corr}(X_{t+h}, X_t).$$

To estimate the ACVF and ACF one uses, respectively, the sample autocovariance function and the sample autocorrelation function. The sample autocovariance function $\hat{\gamma}(h)$ of an observation $x_1, \ldots, x_n$ of a time series is defined as

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n,$$

where $\bar{x}$ is the sample mean of $x_1, \ldots, x_n$. The sample autocorrelation function $\hat{\rho}(h)$ of $x_1, \ldots, x_n$ is defined as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n.$$

The two estimators above are both biased, however they can be made less so by changing the denominator in (1) from $n$ to $n - |h|$.

An important class of stationary processes is so called autoregressive moving-average, or ARMA, processes, defined by

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q},$$

where $\{Z_t\}$ is white noise, i.e. uncorrelated random variables, with mean 0 and variance $\sigma^2$. If $p = 0$ then the process is said to be a MA($q$) process, and if $q = 0$ then it is said to be an AR($p$) process. If both $p > 0$ and $q > 0$ then it is said to be an ARMA($p, q$) process.

Not all processes of the form defined above are stationary. For such a process to be stationary, it is necessary that the polynomial $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ is nonzero for all complex $z$ with $|z| \leq 1$. Furthermore, $\phi(z)$ and $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ must not have any common factors. For a more thorough discussion on stationary processes, see, for example, Brockwell and Davis [2006].

For some of the processes studied in this thesis, Markov chains need to be introduced. Here the basics are introduced; for a more thorough discussion on Markov chains, see, for example, Stirzaker [2005]. A Markov chain is a discrete-time discrete stochastic process which possesses the Markov property, which means that given the current value of the process, the next value is conditionally independent of all past values. Formally,

$$P(X_t = i_t | X_0 = i_0, \ldots, X_{t-1} = i_{t-1}) = P(X_t = i_t | X_{t-1} = i_{t-1}) \forall t_k.$$

The set $S$ of possible values of the process is called the state space, and the probabilities of moving from one state in the state space to another are called transition probabilities. If the transition probabilities are the same for each time step, i.e. if $P(X_n = j | X_{n-1} = i) = P(X_1 = j | X_0 = i) = p_{ij}$, then the Markov chain is said to be homogeneous. For a homogeneous Markov chain, the matrix $P = (p_{ij}), i, j \in S$ is called the transition matrix. For an $n$-state Markov chain,

$$P = (p_{ij}) = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1n} 
p_{21} & p_{22} & \cdots & p_{2n} 
\vdots & \vdots & \ddots & \vdots 
p_{n1} & p_{n2} & \cdots & p_{nn}
\end{pmatrix}.$$

Each row of $P$ necessarily sums to 1, since each row $i$ contains the probabilities of moving from state $i$ to any other state. When in a state $i$, the expected time to stay in that state is equal to $1/(1 - p_{ii})$ if $p_{ii} < 1$.

Processes with occasional bursts can be implemented via two-state Markov chains, the states called normal and exceptional. First, a stationary process is generated in one way or another. Next, in the normal state, the original process is left as it is. With a certain low probability $p_{12}$, the process moves to the exceptional state, where instead, for example, $N(\mu, \sigma^2)$ numbers are generated, which replace the points in the original series. With a relatively high probability $p_{21}$, the process moves back to the normal state.
2.2 Entropy measures

We will now go on to present and describe various entropy measures. Entropy basically measures the degree of uncertainty in the value of a random variable or the outcome of a stochastic process, and was first defined by Shannon [1948] as a way of measuring the information content in communication signals. The information content in the random variable, and hence the entropy, is high when the uncertainty is high and low when the uncertainty is low. In other words, predictability provides a low amount of information and variability provides a high amount of information. As an example one can consider the roll of a die. If the die is biased, and therefore relatively predictable, then identifying the outcome provides a lower amount of information than if the die is fair.

2.2.1 Shannon entropy

We begin this section with defining the classical Shannon entropy. The Shannon entropy for a discrete random variable \(X\) with probability mass function (p.m.f.) \(P(X)\) is defined as

\[
H(X) = E[-\ln(P(X))] = -\sum_{i=1}^{K} P(x_i) \ln(P(x_i)),
\]

The Shannon entropy for discrete distributions is maximised for the uniform distribution, \(P(x_i) = 1/K, i = 1, \ldots, K\), in which case \(H(X) = \ln K\). It is minimised for the one-point distribution, \(P(x_{i_0}) = 1, P(x_i) = 0 \forall i \neq i_0\), in which case \(H(X) = 0\). In other words, the Shannon entropy for discrete distributions is bounded by \(0 \leq H(X) \leq \ln K\). This upper bound leads to a way of defining normalised entropy when \(K\) is finite.

Normalised entropy for a discrete random variable \(X\) with p.m.f. \(P(X)\) is defined as

\[
H_{NE}(X) = -\sum_{i=1}^{K} P(x_i) \ln(P(x_i))/\ln K,
\]

Considering that \(H_{NE}\) is simply Shannon entropy divided by its maximum value, we directly get that it achieves its maximum value 1 for the uniform distribution, and its minimum value 0 for the one-point distribution.

The definition of Shannon entropy can be extended to continuous random variables in a natural way. The Shannon entropy, or differential entropy, of a continuous random variable \(X\) having probability density function (p.d.f.) \(f(x)\) is defined as

\[
H(X) = E[-\ln f(X)] = -\int_{-\infty}^{\infty} \ln(f(x)) f(x) dx.
\]

Among all continuous distributions with \(\text{Var}(X) = \sigma^2\), the Shannon entropy is maximised for the normal distribution. For proof of this, see, for example, Marsh [2013].

2.2.2 Rényi entropy of order \(q\)

A more general form of entropy is the Rényi entropy of order \(q\), defined by Rényi [1961]. In this chapter we will summarise definitions and known properties of Rényi entropies, e.g. covered in Lake [2006], Golshani et al. [2009], Golshani and Pasha [2010]. For a continuous random variable \(X\) with p.d.f. \(f(x)\), the Rényi entropy of order \(q\) is defined as

\[
R_q(X) = \frac{1}{1-q} \ln(E[f(X)^{q-1}]) = \frac{1}{1-q} \ln \left( \int_{-\infty}^{\infty} f(x)^q dx \right),
\]

4
where \( q > 0 \). A special case, relevant to this thesis, is when \( q = 2 \), in which case the quadratic Rényi entropy
\[
R_2(X) = -\ln \left( \int_{-\infty}^{\infty} f(x)^2 dx \right)
\]
is obtained. Note that it can be shown that as \( q \) tends to 1, the Rényi entropy approaches the Shannon entropy; for proof of this, see, for example, [Bromley et al., 2004].

There exists a simple relationship between Rényi entropy for the standardised and the non-standardised random variable.

If \( X \) is a random variable such that \( E[X] = \mu \) and \( \text{Var}[X] = \sigma^2 \), then the Rényi entropy of order \( q \) for the standardised random variable \( W = (X - \mu)/\sigma \) is
\[
R_q(W) = R_q(X) - \frac{1}{2} \ln(\sigma^2) = R_q(X) - \ln(\sigma).
\]
This relationship follows from the transformation theorem. We have that the density function \( g(w) \) of \( W \) is
\[
g(w) = f(x(w)) \left| \frac{dx}{dw} \right| = f(\sigma w + \mu) |\sigma| = f(\sigma w + \mu)\sigma
\]
where \( f(x) \) is the density function of \( X \), and where the last equality holds since by necessity \( \sigma > 0 \). From this we get that
\[
R_q(W) = \frac{1}{1-q} \ln \left( E \left[ g(W)^{q-1} \right] \right)
= \frac{1}{1-q} \ln \left( E \left[ f(\sigma W + \mu)^{q-1} \sigma^{q-1} \right] \right)
= \frac{1}{1-q} \ln \left( E \left[ f(\sigma W + \mu)^{q-1} \right] \right) + \frac{q-1}{1-q} \ln(\sigma)
= \frac{1}{1-q} \ln \left( E \left[ f(\sigma W + \mu)^{q-1} \right] \right) - \ln(\sigma).
\]
Finally we transform back according to \( X = \sigma W + \mu \) and obtain
\[
R_q(W) = \frac{1}{1-q} \ln \left( E \left[ f(X)^{q-1} \right] \right) - \ln(\sigma) = R_q(X) - \ln(\sigma).
\]
For an \( m \)-dimensional continuous random variable \( \mathbf{X}_m = (X_1, \ldots, X_m) \), the Rényi entropy of order \( q \) is
\[
R_q(\mathbf{X}_m) = \frac{1}{1-q} \ln \left( E \left[ f(\mathbf{X}_m)^{q-1} \right] \right)
= \frac{1}{1-q} \ln \left( \int_{\mathbb{R}^m} f(\mathbf{x}_m)^{q} d\mathbf{x}_m \right),
\]
where \( f(\mathbf{x}_m) \) is the joint p.d.f. of \( X_1, \ldots, X_m \). In particular, the Rényi entropy of order 2 of \( \mathbf{X}_m \) is
\[
R_2(\mathbf{X}_m) = -\ln \left( E \left[ f(\mathbf{X}_m) \right] \right)
= -\ln \left( \int_{\mathbb{R}^m} f(\mathbf{x}_m)^2 d\mathbf{x}_m \right).
\]
A concept related to Rényi entropy is the Rényi entropy rate. For \( X_1, \ldots, X_{m+1} \) the Rényi entropy rate of order \( q \) is defined as
\[
D_{q,m}(\mathbf{X}_{m+1}) = R_q(\mathbf{X}_{m+1}) - R_q(\mathbf{X}_m).
\]
If \( q = 2 \), it is called the quadratic entropy rate. The Rényi entropy rate can be interpreted as how much more information is obtained by adding another point \( X_{m+1} \), compared to the sequence \( X_1, \ldots, X_m \). The differential Rényi entropy rate is defined as

\[
R_q = \lim_{m \to \infty} D_{q,m}(X_{m+1}).
\]

For independent data, we have the following relationship between the Rényi entropy rate and the Rényi entropy.

If \( X_1, \ldots, X_{m+1} \) are independent random variables, then

\[
R_q(X_{m+1}) - R_q(X_m) = R_q(X_{m+1}).
\]  

(4)

This result can be motivated as follows. We have that

\[
R_q(X_{m+1}) = \frac{1}{1-q} \ln \left( \int_{\mathbb{R}^{m+1}} f(x_{m+1})^q dx_{m+1} \right)
\]

\[
= \frac{1}{1-q} \ln \left( \int_{\mathbb{R}} f(x_{m+1})^q dx_{m+1} \int_{\mathbb{R}^m} f(x_m)^q dx_m \right)
\]

\[
= \frac{1}{1-q} \ln \left( \int_{\mathbb{R}} f(x_{m+1})^q dx_{m+1} \right) + \frac{1}{1-q} \ln \left( \int_{\mathbb{R}^m} f(x_m)^q dx_m \right)
\]

\[
= R_q(X_{m+1}) + R_q(X_m),
\]

the second following from the independence of the random variables. Hence,

\[
R_q(X_{m+1}) - R_q(X_m) = R_q(X_{m+1}).
\]

Hence, for independent data, adding a new observation \( X_{m+1} \) to the sequence \( (X_1, \ldots, X_m) \) brings completely new full information, and has no information gained from the previous \( X_1, \ldots, X_m \).

For continuous conditional distributions, the Rényi entropy of order \( q \) for \( X_{m+1} \) given \( X_m = x_m \) is defined as

\[
R_q(X_{m+1}|X_m = x_m) = \frac{1}{1-q} \ln \left( \int_{\mathbb{R}^m} f(x_{m+1}|x_m)^q dx_{m+1} \right)
\]

\[
= \frac{1}{1-q} \ln \left( \int_{\mathbb{R}^m} f(x_{m+1}|x_m)^q dx_{m+1} \right),
\]

(5)

where \( f(x_{m+1}|x_m) \) is the p.d.f. of \( X_{m+1} \) given \( X_m = x_m \). Note that if \( X_{m+1} \) is independent of \( X_m \) then

\[
R_q(X_{m+1}|X_m = x_m) = R_q(X_{m+1}).
\]

The conditional Rényi entropy rate is defined as

\[
R^*_q = \lim_{m \to \infty} C_{q,m}
\]

where

\[
C_{q,m} = R_q(X_{m+1}|X_m).
\]

For multivariate normal distributions there exists a simple relationship between the differential Rényi entropy rate and the conditional Rényi entropy rate, see Section 3.2.1. In general, the following relationship holds between the Rényi entropy for \( X_{m+1} \) and the conditional Rényi entropy for \( X_{m+1} \) given \( X_m \).

For the random variables \( X_1, X_2, \ldots, X_{m+1} \), the Rényi entropy of order \( q \) is

\[
R_q(X_{m+1}) = \frac{1}{1-q} \ln \left( \int_{\mathbb{R}^m} \exp \left\{ \{-1\} R_q(X_{m+1}|X_m = x_m) \right\} f(x_m)^q dx_m \right). 
\]

(5)
This can be seen by first noting that the p.d.f. of $X_{m+1}$ satisfies $f(x_{m+1}) = f(x_{m+1}|x_m)f(x_m)$. Then the Rényi entropy of order $q$ for $X_{m+1}$ equals

$$R_q(X_{m+1}) = \frac{1}{1-q} \ln \int_{\mathbb{R}^m} \left( \int_{\mathbb{R}} f(x_{m+1}|x_m)^q dx_{m+1} \right) f(x_m)^q dx_m.$$ 

From the definition of conditional Rényi entropy, it can be seen that the inner integral equals

$$\exp \left\{ (1-q)R_q(X_{m+1}|X_m=x_m) \right\}$$

and hence the entropy $R_q(X_{m+1})$ gets the desired form.

### 2.3 Data-driven entropy measures

We have so far presented entropy measures in terms of probability density functions or probability mass functions of random variables. Data-driven entropy measures, such as approximate entropy (ApEn), sample entropy (SampEn), and quadratic sample entropy (QSE), have been introduced to quantify the amount of regularity and the unpredictability over time-series data. In this section, we present ApEn, SampEn, and QSE in more detail. Suppose that you want to estimate the entropy from a set of independent and identically distributed (i.i.d.) observations from some unknown distribution. How would you find such estimators, and are they consistently estimating the corresponding theoretical entropy measures? What happens with these estimators if you no longer have i.i.d. observations? There are so far no complete answers to these questions. Below we will focus on two such data-driven estimators, the approximate entropy and the sample entropy.

#### 2.3.1 Approximate entropy

When estimating entropy from experimental data, the results are greatly influenced by noise in the system, and in order to be accurate, vast amounts of data are needed. In order to deal with this, Pincus [1991] introduced approximate entropy. Approximate entropy is computed based on an observed sequence of random variables $x_1, \ldots, x_n$, according to the algorithm described in Algorithm 1.

**Algorithm 1** Approximate entropy

0: Input: Sequence $x_1, \ldots, x_N$ of length $N$
1: Specify a tolerance $r$, a template length $m$, and a distance function $d(\cdot, \cdot)$.
2: Let $x_i(m) = (x_i, \ldots, x_{i+m-1})$ := the $i$th template of length $m$ of the $N$-length sequence.
3: For each template $x_i(m)$, $i = 1, \ldots, N-m$:
   - Let $B_i :=$ number of templates $x_j(m)$, $j = 1, \ldots, N-m + 1$ for which $d(x_i(m), x_j(m)) \leq r$.
4: Let $x_i(m+1) = (x_i, \ldots, x_{i+m})$ := the $i$th template of length $m + 1$ of the $N$-length sequence.
5: For each template $x_i(m+1)$, $i = 1, \ldots, N-m$:
   - Let $A_i :=$ number of templates $x_j(m+1)$, $j = 1, \ldots, N-m$ for which $d(x_i(m+1), x_j(m+1)) \leq r$.
6: Output: Approximate entropy, computed as

$$\text{ApEn}(m, r, N) = -\frac{1}{N-m} \sum_{i=1}^{N-m} \ln \left( \frac{A_i}{B_i} \right).$$

As distance function $d(\cdot, \cdot)$ can for example the Euclidean distance be used. However, the usual distance function is the Chebyshev distance, defined as

$$d(x_i(m), x_j(m)) = \max_{0 \leq k \leq m-1} |x_{i+k} - x_{j+k}|.$$
When \( d(x_i(m), x_j(m)) \leq r \), \( x_i(m) \) and \( x_j(m) \) are said to be similar. For an illustration of approximate entropy, see Figure 1.

As we see in Algorithm 1, self-matches are in general allowed, i.e. templates are allowed to be seen as similar to themselves, in order to avoid expressions of the form \( 0/0 \) when no other matches are found.

For stationary series, given that two similar templates of length \( m \) have been found, the ratios \( A_i/B_i \) could be considered as an estimate of the conditional probability that they will remain similar when one more point is added to each of them. ApEn is the negative average logarithm of this conditional probability, being close to zero if the conditional probability is large, indicating a strong regularity, and a large positive value if the conditional probability is low. ApEn could thus be thought of as the negative logarithm of the probability that similar patterns will be followed by additional similar observations. A time series with many repetitive patterns will thus have a relatively small ApEn, while a less predictable process has a higher ApEn.

It can be shown that as \( N \to \infty \) and \( r \to 0 \), \( \text{ApEn}(m, r, N) \) converges to the Shannon entropy rate if the sequence of observations is i.i.d., which is the same as the Rényi entropy rate of order 1, see [Pincus 1991] for details on the convergence of ApEn.

### 2.3.2 Sample entropy

There are a few problems with approximate entropy, mostly owing to the self-matches. Comparing templates to themselves lowers the approximate entropy, and thus the series is shown to be more regular than it actually is. To alleviate this, [Richman and Moorman 2000] introduced a modification of approximate entropy, called sample entropy, which does not include self-matches. Proceeding in the same manner as when computing approximate entropy, excluding the self-matches in \( A_i \) and \( B_i \), sample entropy is defined as

\[
\text{SampEn}(m, r, N) = - \ln \left( \frac{1}{N-m} \sum_{i=1}^{N-m} \frac{A_i}{B_i} \right) = - \ln \left( \frac{A}{B} \right).
\]

SampEn attempts to quantify the negative logarithm of the conditional probability that similar templates remain similar when adding another point to each of them. For regular, repeating data, \( A/B \) tends to be near 1, and thus the sample entropy is near 0. For independent data, \( A \) tends to be low in relation to \( B \) and thus the sample entropy is high. For an illustration of sample entropy, see Figure 1.

Figure 1: An illustration of the computation of sample entropy and approximate entropy. The time series begins with the \( i \)th template. Here \( m \) equals 2, and \( r(\text{SD}) \) means that \( r \) equals a factor times the standard deviation, and is indicated by the error bars. The template is matched by the 11th and 12th points, marked by a solid box. The \((m+1)\)st points also match, marked by a dashed box. Thus both \( A_i \) and \( B_i \) are incremented by 1. The figure is reproduced with permission from [Richman et al. 2004].
2.3.3 Quadratic sample entropy

Lake [2006] introduced quadratic sample entropy, defined as

\[ \text{QSE}(m, r, N) = \text{SampEn}(m, r, N) + \ln(2r). \]

The motivation behind introducing QSE is that if one is trying to classify different time series via SampEn, the results may vary depending on which \( m \) and \( r \) are used. This raises the question of which \( m \) and \( r \) to choose, and how to interpret the results. QSE was therefore introduced because it is a more stable estimator, and especially because for small \( r \)'s, QSE converges to the quadratic entropy rate.

**Theorem 1.** [Källberg et al., 2014] Under some mild regularity conditions for the stationary \( m \)-dependent sequence \( X_1, \ldots, X_N \), the QSE will converge to the quadratic entropy rate \( R_2(X_{m+1}) - R_2(X_m) \) as \( N \to \infty \) and \( r \to 0 \).

2.3.4 Choice of parameters

A challenge when using ApEn or SampEn for comparing different time series is, as it is already mentioned, that of choosing \( m \) and \( r \). If \( r \) is too large, then all templates are similar, and SampEn = \(-\ln 1 = 0 \). On the other hand, if \( r \) is too small, none of the templates are similar and SampEn = \(-\ln 0 = \infty \). Moreover, \( m \) should be sufficiently small to find similar sequences, and \( r \) should be sufficiently small to be able to distinguish between time series, but not too small, in order to find similar sequences. Often \( m \) is chosen to be 1, 2, or 3, and \( r \) is chosen to be a factor of the standard deviation of the marginal distribution of the series, often \( r = 0.2\sigma \). The sample standard deviation can be used as an estimate of the standard deviation, but it is sensitive to spikes and outliers in the series. It might therefore be wise to replace it by a more robust estimate of the standard deviation. One such estimate is a constant scale factor times the median absolute deviation (MAD), which is the median of the absolute deviations of the observations from its median:

\[ \hat{\sigma} = K \cdot \text{MAD} = K \cdot \text{median}_i(|X_i - \text{median}_j(X_j)|). \]

For i.i.d. normally distributed data, a consistent estimator of the standard deviation \( \sigma \) is

\[ \hat{\sigma} = \frac{\text{MAD}}{\Phi^{-1}(3/4)} \approx 1.4826 \cdot \text{MAD}, \tag{7} \]

where \( \Phi^{-1} \) is the inverse distribution function for the standard normal distribution.

This can be seen from the following reasoning. The theoretical median absolute deviation, mad, is given by

\[ \frac{1}{2} = P(|X - \mu| \leq \text{mad}) = P \left( \left| \frac{X - \mu}{\sigma} \right| \leq \frac{\text{mad}}{\sigma} \right) = P \left( |Z| \leq \frac{\text{mad}}{\sigma} \right), \]

where \( Z \sim \mathcal{N}(0, 1) \). This means that

\[ \Phi \left( \frac{\text{mad}}{\sigma} \right) - \Phi \left( -\frac{\text{mad}}{\sigma} \right) = \frac{1}{2}. \]

Now, since

\[ \Phi \left( -\frac{\text{mad}}{\sigma} \right) = 1 - \Phi \left( \frac{\text{mad}}{\sigma} \right), \]

\[ \Phi \left( \frac{\text{mad}}{\sigma} \right) = \frac{1}{2}, \]

\[ \Phi \left( -\frac{\text{mad}}{\sigma} \right) = \frac{1}{2}, \]

\[ \Phi \left( \frac{\text{mad}}{\sigma} \right) = \frac{1}{2}, \]
we obtain

$$\Phi \left( \frac{\text{mad}}{\sigma} \right) = \frac{3}{4} \Leftrightarrow \sigma = \frac{\text{mad}}{\Phi^{-1}(3/4)}.$$  

Since MAD is a consistent estimator of mad, the result follows.

Another robust estimate of the standard deviation is a constant scale factor times the inter-quartile range (IQR), which is the difference between the 75th percentile $Q_3$ and the 25th percentile $Q_1$ of the data:

$$\hat{\sigma} = K \cdot (Q_3 - Q_1).$$  

(8)

For i.i.d normally distributed data, a consistent estimator of the standard deviation $\sigma$ is

$$\hat{\sigma} = \frac{Q_3 - Q_1}{2\sqrt{2}\text{erf}^{-1}(1/2)} \approx \frac{Q_3 - Q_1}{1.349},$$  

(9)

where erf$^{-1}$ is the inverse of the error function, which is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt.$$  

For more details and other ways of estimating the standard deviation from the IQR, see Wan et al. [2014].

3 Examples of Rényi entropies and their properties for some distributions

To get an idea about how the shape of a distribution affects the size of the Rényi entropy, in this section we will give examples of Rényi entropies and their properties for some distributions with different shapes. In Section 3.1 we consider four univariate continuous distributions: the normal, uniform, exponential, and Laplace distribution. We also study how the Rényi entropy changes given that the distributions have the same variance. In Section 3.2 we consider examples of multivariate normal distributions, with different dependence structures.

3.1 Rényi entropy for some univariate distributions

We here derive the Rényi entropies of order $q$ for the normal, uniform, exponential, and Laplace distributions. Most of these can be found in general forms in Nadarajah and Zografos [2003].

3.1.1 The normal distribution

A random variable $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$ if it has the probability density function (p.d.f.)

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

If $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$, the Rényi entropy of order $q$ for $X$ is

$$R_q(X) = \frac{\ln(q)}{2(q-1)} + \ln \left( \sqrt{2\pi\sigma} \right).$$  

(10)
It can be motivated by the following reasoning. First note that
\[ \int_{-\infty}^{\infty} f(x)^q dx = \frac{1}{\sqrt{2\pi}} \sigma q^{-1/2} \sqrt{2\pi} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma q^{-1/2}} \exp \left\{ -(x - \mu)^2 / 2\sigma^2 q^{-1} \right\} dx \]
\[ = \frac{1}{\sqrt{2\pi}} \sigma^{-1/2} q^{-1/2} \]
\[ = \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^{-1} q^{-1/2}, \]

since the integrand on the right hand side (RHS) is the p.d.f. of a normally distributed random variable with mean \( \mu \) and variance \( \sigma^2 / q \). Hence by the definition (3)
\[ R_q(X) = \frac{1}{1 - q} \ln \left( \int_{-\infty}^{\infty} f(x)^q dx \right) \]
\[ = \frac{1}{1 - q} \ln \left( \frac{1}{\sqrt{2\pi} \sigma} \right) + \frac{1}{1 - q} \left( -\frac{1}{2} \ln(q) \right) \]
\[ = \frac{\ln(q)}{2(q - 1)} + \ln \left( \sqrt{2\pi} \sigma \right). \]

Note that, as \( q \to 1 \), we obtain the Shannon entropy
\[ H(X) = \frac{1}{2} + \ln \left( \sqrt{2\pi} \sigma \right). \]

3.1.2 The uniform distribution
A random variable \( X \) is uniformly distributed on \([a, b]\) if it has the p.d.f.
\[ f(x) = \frac{1}{b - a}, \quad a \leq x \leq b. \]

If \( X \) is uniformly distributed on \([a, b]\), the Rényi entropy of order \( q \) for \( X \) is
\[ R_q(X) = \ln(b - a). \]

This can be seen by using definition (3) and noting that
\[ R_q(X) = \frac{1}{1 - q} \ln \left( \int_{a}^{b} \left( \frac{1}{b - a} \right)^q dx \right) \]
\[ = \frac{1}{1 - q} \ln \left( \frac{1}{b - a} \right)^q (b - a) \]
\[ = \frac{1}{1 - q} (1 - q) \ln(b - a) \]
\[ = \ln(b - a). \]

Since \( R_q(X) \) does not depend on \( q \), we also have that the Shannon entropy is
\[ H(X) = \ln(b - a). \]
3.1.3 The exponential distribution

A random variable \( X \) is exponentially distributed with parameter \( \lambda \) if it has the p.d.f.

\[
f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x < \infty.
\]

If \( X \) is exponentially distributed with parameter \( \lambda \), the Rényi entropy of order \( q \) for \( X \) is

\[
R_q(X) = \frac{\ln(q)}{q-1} + \ln(\lambda).
\]

This can be shown by the following reasoning. From definition (3), we have that

\[
R_q(X) = \frac{1}{1-q} \ln \left( \int_0^\infty \left( \frac{1}{\lambda} e^{-x/\lambda} \right)^q dx \right)
\]

\[
= \frac{1}{1-q} \ln \left( \frac{1}{\lambda^q} \int_0^\infty e^{-qx/\lambda} dx \right)
\]

\[
= \frac{1}{1-q} \ln \left( \frac{1}{\lambda^q} \left[ -\frac{\lambda}{q} e^{-qx/\lambda} \right]_0^\infty \right)
\]

\[
= \frac{1}{1-q} \ln \left( \frac{1}{\lambda^q} \right)
\]

\[
= \frac{1}{1-q} (\ln(\lambda^{1-q}) - \ln(q))
\]

\[
= \frac{\ln(q)}{q-1} + \ln(\lambda).
\]

Letting \( q \to 1 \), we obtain the Shannon entropy

\[
H(X) = 1 + \ln(\lambda).
\]

3.1.4 The Laplace distribution

A random variable \( X \) is Laplace distributed with parameters \( \mu \) and \( b \) if it has the p.d.f.

\[
f(x) = \frac{1}{2b} e^{-|x-\mu|/b}, \quad -\infty < x < \infty.
\]

If \( X \) is Laplace distributed with parameters \( \mu \) and \( b \), the Rényi entropy of order \( q \) for \( X \) is

\[
R_q(X) = \frac{\ln(q)}{q-1} + \ln(2b).
\]

Without loss of generality, we will restrict the motivation to the case where the location parameter \( \mu = 0 \), in which case the p.d.f. reduces to

\[
f(x) = \frac{1}{2b} e^{-|x|/b}.
\]
Using definition (3), we obtain the Rényi entropy of order $q$

\[
R_q(X) = \frac{1}{1-q} \ln \left( \int_{-\infty}^{\infty} \left( \frac{1}{2b} e^{-|x|/b} \right)^q dx \right)
\]

\[
= \frac{1}{1-q} \ln \left( \frac{1}{(2b)^q} \left( \int_{-\infty}^{0} e^{qx/b} dx + \int_{0}^{\infty} e^{-qx/b} dx \right) \right)
\]

\[
= \frac{1}{1-q} \ln \left( \frac{1}{(2b)^q} \left[ \frac{b}{q} e^{qx/b} \right]_{0}^{\infty} + \left[ -\frac{b}{q} e^{-qx/b} \right]_{-\infty}^{0} \right)
\]

\[
= \frac{1}{1-q} \ln \left( \frac{(2b)^{1-q}}{q} \right)
\]

\[
= \frac{\ln(q)}{q-1} + \ln(2b).
\]

Letting $q \to 1$, we obtain the Shannon entropy

\[H(X) = 1 + \ln(2b)\].

### 3.1.5 Order of distributions with respect to Rényi entropies

The Rényi entropies of order $q$ for the univariate normal, exponential, Laplace, and exponential distributions derived in the previous section are summarised in Table 1, along with their respective variances.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Variance</th>
<th>$R_q(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$\sigma^2$</td>
<td>$\ln(q) \over 2(q-1) + \ln(\sqrt{2\pi} \sigma)$</td>
</tr>
<tr>
<td>Uniform</td>
<td>$1/2(b-a)^2$</td>
<td>$\ln(b-a)$</td>
</tr>
<tr>
<td>Laplace</td>
<td>$2b^2$</td>
<td>$\ln(q) \over q-1 + \ln(2b)$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\lambda^2$</td>
<td>$\ln(q) \over q-1 + \ln(\lambda)$</td>
</tr>
</tbody>
</table>

Note how each entropy contains a term which basically is the logarithm of the standard deviation, implying that the entropy grows logarithmically with respect to the standard deviation for these distributions. Note also how none of the entropies is a function of the mean value. Hence the Rényi entropy depends on the scale but not the location of the distribution.

Assuming equal variances, the Rényi entropies of order $q$ for the normal, uniform, Laplace, and exponential distributions in Table 1 can be rewritten in terms of their common standard deviation $\sigma$, as shown in Table 2.
Table 2: Rényi entropy of order $q$ and order 2 for the normal, uniform, Laplace, and exponential distributions with common variance $\sigma^2$.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$R_q(X)$</th>
<th>$R_2(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$\frac{\ln(q)}{2(q-1)} + \ln (\sqrt{2\pi}\sigma)$</td>
<td>$\ln(2\sqrt{\pi}) + \ln (\sigma)$</td>
</tr>
<tr>
<td>Uniform</td>
<td>$\ln (\sqrt{12}\sigma)$</td>
<td>$\ln (2\sqrt{3}) + \ln (\sigma)$</td>
</tr>
<tr>
<td>Laplace</td>
<td>$\frac{\ln(q)}{q-1} + \ln (\sqrt{2}\sigma)$</td>
<td>$\ln (2\sqrt{2}) + \ln (\sigma)$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\frac{\ln(\sigma)}{q-1} + \ln (\sigma)$</td>
<td>$\ln(2) + \ln (\sigma)$</td>
</tr>
</tbody>
</table>

Since the Rényi entropies still are different for equal variances, the differences must be explained by other features of the distributions. For the special case $q = 2$, Table 2 gives that $R_2(X)$ tends to be smaller for more heavy-tailed distributions. Here we see more clearly that each entropy grows logarithmically with respect to the standard deviation. For standardised variables, $\sigma = 1$ and thus the term $\ln(\sigma)$ equals zero.

Figure 2 shows the normal, exponential, Laplace, and uniform probability density functions, centered around 0 and with common variance $\sigma^2 = 1$. We see that the Laplace and exponential distributions are more heavy-tailed than the normal and uniform distributions, thus having larger probability of extreme values.

Figure 2: The normal, exponential, Laplace, and uniform probability density functions with common variance 1.

For other values of $q$, the order of the distributions, given equal variances or equivalently on standardised variables, according to the Rényi entropy may change as shown in Table 3. The distributions are denoted by their initial letters, and the boundaries are approximated and were found via Wolfram|Alpha. For the special case $q = 1$, the Shannon entropy is used.
Table 3: Order of the normal, uniform, Laplace, and exponential distributions according to \( R_q(X) \), depending on the order \( q \).

<table>
<thead>
<tr>
<th>Interval</th>
<th>Ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 &lt; q &lt; 0.2537 )</td>
<td>( U &lt; N &lt; E &lt; L )</td>
</tr>
<tr>
<td>( 0.2537 &lt; q &lt; 0.6369 )</td>
<td>( U &lt; E &lt; N &lt; L )</td>
</tr>
<tr>
<td>( 0.6369 &lt; q &lt; 0.7583 )</td>
<td>( E &lt; U &lt; N &lt; L )</td>
</tr>
<tr>
<td>( 0.7583 &lt; q &lt; 1.2411 )</td>
<td>( E &lt; U &lt; L &lt; N )</td>
</tr>
<tr>
<td>( 1.2411 &lt; q &lt; 2.2605 )</td>
<td>( E &lt; L &lt; U &lt; N )</td>
</tr>
<tr>
<td>( q &gt; 2.2605 )</td>
<td>( E &lt; L &lt; N &lt; U )</td>
</tr>
</tbody>
</table>

We see here that for \( q > 2.26 \), the distributions are ordered according to their heavy-tailedness. When considering empirical data, a heavy-tailed distribution indicates a significant number of spikes and outliers. Hence we would expect low entropy to capture what we call burstiness. On the other hand, high entropy could indicate high variability in the data, seeing how the entropy of each of the four considered distributions is a function of the standard deviation.

### 3.2 Rényi entropy for multivariate normal distributions

We now go on to consider the Rényi entropy for multivariate normal distributions with various dependence structures, starting with some general properties of multivariate normal distributions.

#### 3.2.1 Some properties

A vector-valued random variable \( X_m = (X_1, \ldots, X_m)^T \in \mathbb{R}^m \) has a multivariate normal distribution with mean value

\[
\mathbb{E}[X_m] = \bar{\mu}_m = (\mu_1, \ldots, \mu_m)^T \in \mathbb{R}^m,
\]

and covariance matrix

\[
\Sigma_m = \begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2
\end{pmatrix} \in \mathbb{R}^{m \times m},
\]

if it has the probability density function

\[
f(x_m) = (2\pi)^{-m/2} |\Sigma_m|^{-1/2} \exp \left\{ -\frac{1}{2} (x_m - \bar{\mu}_m)^T \Sigma_m^{-1} (x_m - \bar{\mu}_m) \right\},
\]

where \( |\Sigma_m| \) denotes the determinant of \( \Sigma_m \). A nice property of multivariate normal distributions is that the conditional distribution of a subset of its variables given the others will again be normally distributed. More specifically, suppose that \( X_{m+1} = (X_1, \ldots, X_{m+1})^T \in \mathbb{R}^{m+1} \) is a multivariate normally distributed random variable with mean value \( \bar{\mu}_{m+1} \) and covariance matrix

\[
\Sigma_{m+1} = \begin{pmatrix}
\Sigma_m \\
\mathbf{c}_m \\
\mathbf{c}_m^T \\
\sigma_{m+1}^2
\end{pmatrix},
\]

where \( \mathbf{c}_m = \text{Cov}(X_m, X_{m+1}) \). The conditional distribution of \( X_{m+1} \) given \( X_m \) is then again normally distributed with mean value

\[
\mathbb{E}[X_{m+1} | X_m] = \mu_{m+1} + \mathbf{c}_m^T (X_m - \bar{\mu}_m),
\]
and variance
\[
\text{Var}[X_{m+1}|X_m] = \sigma^2_{m+1} - c_m^T \Sigma_m^{-1} c_m.
\] (11)

Note that the conditional mean, but not the conditional variance, depends on \(X_m\). For \(m = 1\) the conditional variance reduces to
\[
\text{Var}[X_{m+1}|X_m] = \sigma^2_{m+1} - \text{Cov}^2(X_{m+1}, X_m) = \sigma^2_{m+1}(1 - \text{Corr}^2(X_{m+1}, X_m)).
\] (12)

For more details and proofs, see, for example, Gut [2009].

For multivariate conditional distributions, we have the following relationship between unconditional and conditional Rényi entropies.

Let \(X_{m+1} = (X_1, X_2, \ldots, X_{m+1})^T\) be multivariate normally distributed. Then the Rényi entropy of order \(q\) for \(X_{m+1}\) satisfies
\[
R_q(X_{m+1}) = \sum_{k=1}^m R_q(X_{k+1}|X_k) + R_q(X_1).
\] (13)

Moreover, the Rényi entropy rate of order \(q\) for \(X_{m+1}\) equals the conditional Rényi entropy rate of \(X_{m+1}\) given \(X_m\), i.e.
\[
R_q(X_{m+1}) - R_q(X_m) = R_q(X_{m+1}|X_m).
\] (14)

Equations (13) and (14) are motivated by the following reasoning. For normally distributed data and by (10),
\[
R_q(X_{m+1}|X_m) = \frac{\ln(q)}{2(q-1)} + \ln(\sqrt{2\pi}) + \frac{1}{2} \ln(\sigma^2_{m+1} - c_m^T \Sigma_m^{-1} c_m),
\]
which does not depend on \(X_m\). Hence by (5) we have that
\[
R_q(X_{m+1}) = \frac{1}{1-q} \ln \left( \int \exp \{(1-q)R_q(X_{m+1}|X_m)\} f(x_m)^q dx_m \right)
= \frac{1}{1-q} \ln \left( \exp \{(1-q)R_q(X_{m+1}|X_m)\} \right) + \frac{1}{1-q} \ln \left( \int f(x_m)^q dx_m \right)
= R_q(X_{m+1}|X_m) + R_q(X_m).
\] (15)

By iteration, we finally end up with
\[
R_q(X_{m+1}) = \sum_{k=1}^m R_q(X_{k+1}|X_k) + \frac{1}{1-q} \ln \left( \int f(x_1)^q dx_1 \right)
= \sum_{k=1}^m R_q(X_{k+1}|X_k) + R_q(X_1),
\]
which is the desired form. From (15), (14) immediately follows.

In this section we have examined properties of the multivariate normal distribution, and found that these properties lead to the result that the differential Rényi entropy rate of order \(q\) for \(X_{m+1}\) equals the conditional Rényi entropy rate of \(X_{m+1}\) given \(X_m\). This will now be used to study Rényi entropies for some subsequences of Gaussian processes with specific dependence structures.
3.2.2 Example: MA(1) Gaussian process

Consider the MA(1) process

\[ X_t = \varepsilon_t + \theta \varepsilon_{t-1} \]

where \( \varepsilon_t \in \mathcal{N}(\mu, \sigma^2) \) i.i.d. The autocovariance function of this process is given by

\[ \gamma_k = \text{Cov}(X_t, X_{t-k}) = \begin{cases} (1 + \theta^2)\sigma^2 & k = 0 \\ \theta \sigma^2 & |k| = 1 \\ 0 & \text{otherwise} \end{cases} \]

meaning that \( X_{m+1} \) is independent of \( X_{m+1-k} \) if \( k > 1 \), which is called 1-dependence. This implies that

\[ \text{Var}[X_{m+1}|X_m] = \text{Var}[X_{m+1}|X_m] = \sigma^2(1 + \theta^2) \left( 1 - \frac{\theta^2}{(1 + \theta^2)^2} \right), \quad (16) \]

where the second equality follows from (12). Hence we get from (14) and (10) that the Rényi entropy of order \( q \) for the marginal distribution of the 1-dependent MA(1) process is, due to stationarity, for all \( m \),

\[ R_q(X_{m+1}) = R_q(X_1) = \frac{\ln(q)}{2(q-1)} + \ln \left( \frac{\sqrt{2\pi}}{\theta^2(1 + \theta^2)^2} \right) + \frac{1}{2} \ln \left( 1 - \frac{\theta^2}{(1 + \theta^2)^2} \right), \]

and the Rényi entropy rate of order \( q \) is, by (14) and (16),

\[ R_q(X_{m+1}) - R_q(X_m) = R_q(X_{m+1}|X_m) = R_q(X_{m+1}) + \frac{1}{2} \ln \left( 1 - \frac{\theta^2}{(1 + \theta^2)^2} \right). \quad (17) \]

Note that the Rényi entropy rate is smaller than the Rényi entropy due to the dependence structure of the MA(1) process. The Rényi entropy rate does not depend on \( m \) due to the 1-dependent structure of the MA(1) process. Furthermore, from (13) we have that

\[ R_q(X_{m+1}) = \sum_{k=1}^{m} R_q(X_{k+1}|X_k) + R_q(X_1) = (m + 1)R_q(X_1) + \frac{m}{2} \ln \left( 1 - \frac{\theta^2}{(1 + \theta^2)^2} \right), \]

showing that the Rényi entropy of the \( m \)-dimensional random variable is lower than what it had been had the data been independent.

3.2.3 Example: AR(1) Gaussian process

Consider the AR(1) process

\[ X_t = \theta X_{t-1} + \varepsilon_t, \quad t = 1, 2, 3, \ldots, \]

where \( \varepsilon_t \in \mathcal{N}(0, \sigma^2) \) i.i.d. If the process is stationary, i.e. if \( |\theta| < 1 \), then

\[ X_t \in \mathcal{N} \left( 0, \frac{\sigma^2}{1 - \theta^2} \right) \]
and the autocovariance function of the process is given by
\[
\gamma_k = \text{Cov}(X_t, X_{t-k}) = \begin{cases} 
\frac{\sigma^2}{1-\theta^2} & k = 0 \\
\frac{\sigma^2}{1-\theta^2} |k| & k \geq 1
\end{cases}.
\]

Note that the AR(1) process has a decaying dependence proportional to the rate \(|\theta|k|\). The Rényi entropy of order \(q\) is also here, due to stationarity, independent of \(m\) and equals
\[
R_q(X_{m+1}) = R_q(X_1) = \frac{\ln(q)}{2(q-1)} + \ln\left(\frac{\sqrt{2\pi}}{\sqrt{1-\theta^2}\sigma}\right).
\]

The conditional variance of \(X_{m+1}\) given \(X_m\) is independent of \(m\) and equals
\[
\text{Var}[X_{m+1}|X_m] = \text{Var}[X_{m+1}X_m] = \sigma^2.
\]

The Rényi entropy rate of order \(q\) is therefore, for all \(m\),
\[
R_q(X_{m+1}) - R_q(X_m) = R_q(X_{m+1}|X_m)
\]
\[
= \frac{\ln(q)}{2(q-1)} + \ln(\sqrt{2\pi}\sigma).
\]

Again we see that the Rényi entropy rate is smaller than the Rényi entropy \(R_q(X_{m+1})\). The Rényi entropy of order \(q\) for \(X_{m+1}\) is
\[
R_q(X_{m+1}) = \sum_{k=1}^{m} R_q(X_{k+1}|X_k) + R_q(X_1)
\]
\[
= (m + 1) \frac{\ln(q)}{2(q-1)} + m \ln(\sqrt{2\pi}\sigma) + \ln\left(\frac{\sqrt{2\pi}}{\sqrt{1-\theta^2}\sigma}\right).
\]

Note that \(R_q(X_{m+1})\) can be considered as the Rényi entropy rate of an i.i.d. normal process with variance \(\text{Var}[X_{m+1}]\).

Consider now the Rényi entropies for the differenced process
\[
\nabla X_t = X_t - X_{t-1} = (\theta - 1)X_{t-1} + \varepsilon_t.
\]

The Rényi entropies depend on the variance and covariances of the process, so we will first derive them. First note that this process is also stationary for \(0 < \theta < 1\) and has variance
\[
\text{Var}[\nabla X_t] = (\theta - 1)^2 \text{Var}[X_{t-1}] + \text{Var}[\varepsilon_t]
\]
\[
= \frac{(\theta - 1)^2}{1-\theta^2}\sigma^2 + \sigma^2
\]
\[
= \frac{(\theta^2 + 2\theta + 1)(1-\theta^2)}{(1+\theta)(1-\theta)}\sigma^2
\]
\[
= \frac{2}{1+\theta}\sigma^2.
\]

For \(k \geq 1\), we get the covariances
\[
\gamma_k = \text{Cov}(\nabla X_t, \nabla X_{t-k})
\]
\[
= \text{Cov}((\theta - 1)X_{t-1} + \varepsilon_t, (\theta - 1)X_{t-k-1} + \varepsilon_{t-k})
\]
\[
= (\theta - 1)^2 \text{Cov}(X_{t-1}, X_{t-k-1}) + (\theta - 1)\text{Cov}(X_{t-1}, \varepsilon_{t-k})
\]
\[
= (\theta - 1)^2 \frac{\theta^k}{1-\theta^2}\sigma^2 + (\theta - 1)\theta^{k-1}\sigma^2
\]
\[
= \frac{\theta^{k-1}(\theta - 1)}{1+\theta}\sigma^2.
\]
For $m = 1$ the conditional variance of $\nabla X_{m+1}$ given $\nabla X_m$ is

$$\text{Var}[\nabla X_{m+1}|\nabla X_m] = \text{Var}[\nabla X_{m+1}] - \frac{\text{Cov}^2(\nabla X_{m+1}, \nabla X_m)}{\text{Var}[\nabla X_m]}$$

$$= \frac{2}{1 + \theta} \sigma^2 - \left[\frac{\theta - 1}{1 + \theta} \sigma^2\right]^2 \cdot \frac{1 + \theta}{2 \sigma^2}$$

$$= \frac{2\sigma^2}{1 + \theta} \left(1 - \frac{(\theta - 1)^2}{4}\right).$$

For $m = 2$ the covariance matrix of $(\nabla X_{m+1}, \nabla X_m)^T$ is

$$\Sigma_m = \sigma^2 \begin{pmatrix} 2 & \theta - 1 \\ \theta - 1 & 2 \end{pmatrix},$$

with inverse

$$\Sigma^{-1}_m = \frac{1 + \theta}{\sigma^2} \begin{pmatrix} 2 & -(\theta - 1) \\ -(\theta - 1) & 2 \end{pmatrix}, \frac{1}{4 - (\theta - 1)^2}.$$  

Furthermore, we have

$$c_m = \begin{pmatrix} \text{Cov}(\nabla X_1, \nabla X_3) \\ \text{Cov}(\nabla X_1, \nabla X_2) \end{pmatrix} = \begin{pmatrix} \theta(\theta - 1) \\ (\theta - 1) \end{pmatrix} \cdot \frac{\sigma^2}{1 + \theta}.$$  

Therefore the conditional variance of $\nabla X_{m+1}$ given $(\nabla X_m, \nabla X_{m-1})$ is, by (11),

$$\text{Var}[\nabla X_{m+1}|\nabla X_m, \nabla X_{m-1}] = \text{Var}[\nabla X_{m+1}] - c_m^T \Sigma^{-1}_m c_m$$

$$= \frac{2\sigma^2}{1 + \theta} - \frac{1}{4 - (\theta - 1)^2} \sigma^2 \begin{pmatrix} \theta(\theta - 1) & \theta - 1 \end{pmatrix} \begin{pmatrix} \theta(\theta - 1) & \theta - 1 \end{pmatrix}$$

$$= \frac{2\sigma^2}{1 + \theta} - \frac{1}{4 - (\theta - 1)^2} \sigma^2 \left(2\theta^2(\theta - 1)^2 - 2\theta(\theta - 1)^3 + 4(\theta - 1)^2\right)$$

$$= \frac{2\sigma^2}{1 + \theta} \left(1 - \frac{(\theta - 1)^2(\theta^2 - \theta(\theta - 1) + 1)}{4 - (\theta - 1)^2}\right)$$

$$= \frac{2\sigma^2}{1 + \theta} \left(1 - \frac{(\theta - 1)^2(\theta + 1)}{4 - (\theta - 1)^2}\right).$$

We are now ready to compute the Rényi entropies. The Rényi entropy of order $q$ for $\nabla X_{m+1}$ is independent of $m$ and equals, by (10),

$$R_q(\nabla X_{m+1}) = \frac{\ln(q)}{2(q - 1)} + \ln \left(\sqrt{2\pi} \sqrt{\frac{2\sigma^2}{1 + \theta}}\right).$$

For $m = 1$ the Rényi entropy rate of order $q$ is

$$R_q(\nabla X_{m+1}) - R_q(\nabla X_m) = R_q(\nabla X_{m+1}|\nabla X_m)$$

$$= \frac{\ln(q)}{2(q - 1)} + \ln \left(\sqrt{2\pi} \sqrt{\frac{2\sigma^2}{1 + \theta} \left(1 - \frac{(\theta - 1)^2}{4}\right)}\right)$$

$$= R_q(\nabla X_{m+1}) + \frac{1}{2} \ln \left(1 - \frac{(\theta - 1)^2}{4}\right),$$

(19)
and for \( m = 2 \) the Rényi entropy rate of order \( q \) is

\[
R_q(\nabla X_{m+1}) - R_q(\nabla X_m) = R_q(\nabla X_{m+1} | \nabla X_m, \nabla X_{m-1})
\]

\[
= \frac{\ln(q)}{2(q-1)} + \ln \left( \sqrt{2\pi} \sqrt{\frac{2\sigma^2}{1+\theta} \left( 1 - \frac{(\theta - 1)^2(\theta + 1)}{4 - (\theta - 1)^2} \right)} \right)
\]

\[
= R_q(\nabla X_{m+1}) + \frac{1}{2} \ln \left( 1 - \frac{(\theta - 1)^2(\theta + 1)}{4 - (\theta - 1)^2} \right).
\]

Note that for \(|\theta| < 1\), we have that \(4\theta + 4 > 4 - (\theta - 1)^2\) and thus

\[
\frac{\theta + 1}{4 - (\theta - 1)^2} > \frac{1}{4},
\]

implying that the Rényi entropy rate of order \( q \) for \( m = 1 \) is larger than that for \( m = 2 \) when \(|\theta| < 1\), and both of them are smaller than the Rényi entropy rate of order \( q \) for i.i.d. normally distributed variables with variance \( \text{Var}[\nabla X_{m+1}] \). Hence there is more information in the distribution of \( \nabla X_{m+1} \) given \( \nabla X_m \) than if one conditions on both \( \nabla X_m \) and \( \nabla X_{m-1} \). This is natural since there is a strong dependence. Furthermore, for \( m = 1 \) we have that

\[
\text{Var}[\nabla X_{m+1} | \nabla X_m] = \text{Var}[X_{m+1} | X_m] \cdot \frac{2}{1+\theta} \left( 1 - \frac{(\theta - 1)^2}{4} \right) > \text{Var}[X_{m+1} | X_m],
\]

meaning that the Rényi entropy rate of order \( q \) for \( X_{m+1} \) is smaller than that of \( \nabla X_{m+1} \), and moreover saying that \( X_{m+1} \) given \( X_m \) is more predictable compared to \( \nabla X_{m+1} \) given \( \nabla X_m \).

4 Numerical illustrations

In this section numerical illustrations are performed aiming at illustrating and drawing inference on different properties of entropy measures. In Section 4.1 QSE\((m, r, N)\) is studied as an estimator of the Rényi entropy rate of order 2 for independent and dependent, mainly stationary sequences. In Section 4.2 we study the performance of QSE\((m, r, N)\), approximate entropy, and sample entropy for bursty, non-stationary sequences of random variables, with the aim of finding ways of making inferences about the predictability of the original series.

4.1 Simulated non-bursty data

From Theorem 1 by Källberg et al. [2014], we know that, at least for \( m \)-dependent stationary sequences, the QSE\((m, r, N)\) converges to the quadratic entropy rate \( R_2(X_{m+1}) - R_2(X_m) \) as \( r \to 0 \) and the sample size \( N \to \infty \). However, since QSE\((m, r, N)\) = SampEn\((m, r, N)\) + \( \ln(2r) \), we cannot evaluate it at \( r = 0 \).

Here we will study various properties of QSE\((m, r, N)\) for some simulated time series data, in order to see how close QSE\((m, r, N)\) is to the quadratic entropy rate for sufficiently large sample sizes \((N = 1000)\), and varying values of \( r \) and \( m \). We will see how QSE behaves when it is computed on both the original and the differenced time series, and try to identify whether it can capture certain features in the time series that are related to the time series predictability, including burstiness. We will investigate this for \( m = 1, 2 \) and for \( r = 0.01\sigma, 0.1\sigma, 0.2\sigma \), where \( \sigma \) is both the true value and a robust estimate \( \hat{\sigma} = 1.4826 \cdot \text{MAD} \), see equation [7].

4.1.1 Independent and 1-dependent data

Here we will first study two sequences \( X_t, t = 1, \ldots, 1000 \) of i.i.d. observations with the same variance \( \sigma^2 = 1 \) but with different distributions, being normally (with mean \( \mu = 0 \)) and exponentially distributed. We will also study their first-order differences \( \nabla X_t, t = 2, \ldots, 1000 \) which are 1-dependent with mean \( \mu = 0 \)
and variance $\sigma^2 = 2$. Note that when $X_t$ is normally distributed, so is $\nabla X_t$, but when $X_t$ is exponentially distributed $\nabla X_t$ is Laplace distributed with parameters $\mu = 0$ and $b = 1$.

For all four series $\text{QSE}(m, r, N)$ is computed for $m = 1, 2$ and $r = 0.01 \sigma, 0.1 \sigma, 0.2 \sigma$, where $\sigma$ both corresponds to the true standard deviation of each process, i.e. $\sigma = 1$ for $X_t$ and $\sigma = \sqrt{2}$ for $\nabla X_t$, and computed from the data as $\hat{\sigma} = 1.4826 \cdot \text{MAD}$ being a robust estimator of $\sigma$, which is less sensitive to outliers than the sample standard deviation.

In Table 4 is summarised these standard deviations together with the Rényi entropy rates of order 2.

A few things are noteworthy here: Firstly, recall that $1.4826 \cdot \text{MAD}$ is only a robust consistent estimator for i.i.d. normally distributed data. Therefore we should not expect it to consistently estimate the standard deviation for data that is not normally distributed. We can see from the Table 4 that $\hat{\sigma}$ is close to the true value for 1-dependent normally distributed data. Note that it underestimates the standard deviation for exponentially and Laplace distributed data. This is likely because it tends to disregard outliers and extreme values, and the exponential and Laplace distributions are relatively heavy-tailed compared to the normal distribution.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Standard deviation</th>
<th>Rényi entropy rate of order 2</th>
<th>$R_2(\cdot)$</th>
<th>$R_2(\cdot)$ if $\mathcal{N}(\cdot, \hat{\sigma}^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_t \in \mathcal{N}(0, 1)$ indep.</td>
<td>$\sigma = 1$</td>
<td>0.9586</td>
<td>1.266</td>
<td>1.266</td>
</tr>
<tr>
<td>$\nabla X_t \in \mathcal{N}(0, 2)$ 1-dep.</td>
<td>$\sqrt{2} \approx 1.414$</td>
<td>1.445</td>
<td>1.468</td>
<td>1.612</td>
</tr>
<tr>
<td>$Y_t \in \text{Exp}(1)$ indep.</td>
<td>$\sigma = 1$</td>
<td>0.7241</td>
<td>0.6931</td>
<td>0.6931</td>
</tr>
<tr>
<td>$\nabla Y_t \in \mathcal{L}(0, 1)$ 1-dep.</td>
<td>$\sqrt{2} \approx 1.414$</td>
<td>1.121</td>
<td>Unknown</td>
<td>1.386</td>
</tr>
</tbody>
</table>

The Rényi entropies of order 2, which as we have seen in Section 2.2.2 coincide with the Rényi entropy rates $R_2(\cdot)$ for the independent data, are computed in Table 1. The Rényi entropy rate of order 2 for the normally distributed data is computed in Section 3.2.2, with $\theta = -1$. The Rényi entropy rates and entropies are given in Table 2. The last column of Table 4 corresponds to the Rényi entropy of order 2 for a normally distributed random variable with standard deviation $\hat{\sigma}$.

Note, for the normally distributed $\nabla X_t$, that the Rényi entropy rate is lower than the Rényi entropy for the normal distribution, which would be the Rényi entropy rate had the data been independent. This difference is clearly due to the dependence structure.

$\text{QSE}(m, r, N)$ computed based on the two sequences of normally and exponentially distributed observations ($N = 1000$) and their differenced series using various $m$’s and $r$’s are shown in Tables 5 and 6.
Table 5: QSE($m, r, N$) for $N = 1000$ i.i.d. normally distributed observations with $\mu = 0$, $\sigma^2 = 1$, computed using various $m$’s and $r$’s, and for the sequence $\nabla X_t \in \mathcal{N}(0, 2)$ being 1-dependent. The QSE’s are computed with both the true $\sigma$ and the robust estimate $\hat{\sigma} = 1.4826 \cdot \text{MAD}$.

| Sequence | \begin{tabular}{|c|c|c|c|c|c|} \hline & \multicolumn{3}{|c|}{$m = 1$} & \multicolumn{2}{|c|}{$m = 2$} \\
 & \multicolumn{2}{|c|}{Exact $\sigma$} & Robust $\sigma$ & \multicolumn{2}{|c|}{Exact $\sigma$} \\
 & $r$ & $\sigma$ & $\hat{\sigma}$ & $\sigma$ & $\hat{\sigma}$ \\
 \hline $X_t$ & 0.01$\sigma$ & 1.594 & 1.594 & NaN & NaN \\
 & 0.1$\sigma$ & 1.262 & 1.259 & 1.371 & 1.323 \\
 & 0.2$\sigma$ & 1.422 & 1.424 & 1.223 & 1.230 \\
 \hline $\nabla X_t$ & 0.01$\sigma$ & 1.651 & 1.626 & NaN & NaN \\
 & 0.1$\sigma$ & 1.438 & 1.441 & 1.412 & 1.407 \\
 & 0.2$\sigma$ & 1.446 & 1.446 & 1.405 & 1.402 \\
 \hline \end{tabular} |

NaN in the tables, and in any table below, stands for “not a number”, and is due to an attempt to take the logarithm of zero. This is because, with $m = 2$, $r = 0.01\sigma$ is a too small interval in order to find any matches of length $m + 1$, and subsequently $A = 0$ in the SampEn algorithm.

Table 6: QSE($m, r, N$) for $N = 1000$ i.i.d. exponentially distributed observations with $\lambda = 1$, computed using various $m$’s and $r$’s, and for the sequence $\nabla X_t \in \mathcal{L}(0, 1)$ being 1-dependent.

| Sequence | \begin{tabular}{|c|c|c|c|c|c|} \hline & \multicolumn{3}{|c|}{$m = 1$} & \multicolumn{2}{|c|}{$m = 2$} \\
 & \multicolumn{2}{|c|}{Exact $\sigma$} & Robust $\sigma$ & \multicolumn{2}{|c|}{Exact $\sigma$} \\
 & $r$ & $\sigma$ & $\hat{\sigma}$ & $\sigma$ & $\hat{\sigma}$ \\
 \hline $Y_t$ & 0.01$\sigma$ & 0.6224 & 0.5388 & NaN & NaN \\
 & 0.1$\sigma$ & 0.7890 & 0.7854 & 0.6719 & 0.6238 \\
 & 0.2$\sigma$ & 0.8322 & 0.8056 & 0.7920 & 0.7508 \\
 \hline $\nabla Y_t$ & 0.01$\sigma$ & 0.9494 & 1.024 & NaN & NaN \\
 & 0.1$\sigma$ & 1.082 & 1.071 & 1.050 & 1.083 \\
 & 0.2$\sigma$ & 1.131 & 1.114 & 1.070 & 1.054 \\
 \hline \end{tabular} |

Recall from Theorem 1 by Källberg et al. [2014] that the QSE($m, r, N$) converges to the Rényi entropy rate of order 2 as $r \to 0$ and $N \to \infty$ since the sequences are independent or 1-dependent and stationary. Comparing the computed QSE’s in Table 5 to the Rényi entropy rates of order 2 in Table 4, we see that with a very low $r$, i.e. $r = 0.01\sigma$, the QSE overshoots the Rényi entropy rate for normally distributed data, independent as well as 1-dependent. This is likely due to the fact that the number of observations is limited. For independent normally distributed data, we can note that $r = 0.1\sigma$ and $m = 1$ seems to work best. There is not a big difference in the QSE for $m = 1$ and $m = 2$, but $m = 1$ tends to work better. Since they are estimating the same quantity, the explanation might be that with a smaller $m$, for the same $r$ it can be easier to find similar templates.

For the independent exponentially distributed data we can see from Table 6 the $m = 2$ and $r = 0.1\sigma$ seems to work best, being closest to the Rényi entropy rate 0.6931. For $m = 1$, $r = 0.01\sigma$ seems to work best, out of the considered $r$-values. It seems important to find a good $r$ given $m$, a suggestion is to use an $r$ that minimises QSE($m, r, N$), seeing how the QSE tends to overshoot the theoretical values for both too small and too large $r$’s.
For the 1-dependent Laplace distributed data we do not know the theoretical Rényi entropy rate of order 2, as indicated in Table 4. However, from the experimental results in Table 6, we can conclude that it is likely lower than the Rényi entropy of independent Laplace distributed data, due to the dependence structure.

### 4.1.2 Autoregressive process

Here we will study a simulated stationary sequence of \( N = 1000 \) normally distributed observations with a decaying dependence structure, following an AR(1) model. More specifically we consider \( X_t = \theta X_{t-1} + \varepsilon_t \) with \( \theta = 0.8 \) and \( \varepsilon_t \sim N(0, \sigma^2) \) where \( \sigma^2 = 1 \). We also consider \( \nabla X_t = X_t - X_{t-1} = (\theta - 1)X_{t-1} + \varepsilon_t \). The true variances are

\[
\text{Var}[X_t] = \frac{\sigma^2}{1 - \theta^2} \approx 2.778,
\]

\[
\text{Var}[\nabla X_t] = (\theta - 1)^2 \text{Var}[X_t] + \sigma^2 \approx 1.111.
\]

Table 7 summarises the corresponding true \( \sigma \) and robust \( \hat{\sigma} = 1.4826 \cdot \text{MAD} \) computed based on the 1000 observations for each of the two series, as well as the Rényi entropy rates of order 2 and the Rényi entropy rates of order 2 if the data were independent normally distributed with variance \( \hat{\sigma} \).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Standard deviation</th>
<th>Rényi entropy rate of order 2</th>
<th>( R_2(\cdot) ) if ( N(\cdot, \hat{\sigma}^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_t )</td>
<td>True ( \sigma ) = 1.667</td>
<td>Robust ( \hat{\sigma} ) = 1.752</td>
<td>( m = 1 ) 1.266 ( m = 2 ) 1.266</td>
</tr>
<tr>
<td>( \nabla X_t )</td>
<td>True ( \sigma ) = 1.054</td>
<td>Robust ( \hat{\sigma} ) = 3.272</td>
<td>( m = 1 ) 1.313 ( m = 2 ) 1.309</td>
</tr>
</tbody>
</table>

The Rényi entropy rates of order 2 in Table 7 are computed in Section 3.2.3 with \( \theta = 0.8 \) and \( \sigma^2 = 1 \), using 19 for \( m = 1 \) and 20 for \( m = 2 \). The Rényi entropies in the last column of Table 7 are computed in Table 1. Note how the estimator \( 1.4826 \cdot \text{MAD} \) overestimates the standard deviation, an overestimation that is greater the stronger the dependence structure is.

Table 8 shows the QSE\((m, r, N)\) computed using the previously specified values of \( m \) and \( r \).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>QSE((m, r, 1000))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_t )</td>
<td>( r ) 0.01(( \sigma ) = 1.289 1.315 NaN NaN) ( 0.1\sigma ) = 1.237 1.241 1.249 1.251) ( 0.2\sigma ) = 1.247 1.250 1.249 1.260)</td>
</tr>
<tr>
<td>( \nabla X_t )</td>
<td>( r ) 0.01(( \sigma ) = 1.410 1.430 NaN NaN) ( 0.1\sigma ) = 1.563 1.570 1.190 1.245) ( 0.2\sigma ) = 1.570 1.575 1.444 1.418)</td>
</tr>
</tbody>
</table>

23
We can see from Table 7, as it is found in Section 3.2.3, that the standard deviations and Rényi entropy rates do indeed increase when differencing the series and are lower for \( m = 2 \) than for \( m = 1 \) for \( \nabla X_t \). This holds for the most part for the QSE’s in Table 8 as well. Comparing the QSE’s to the Rényi entropy rates we see that it is best not to use a too low \( r \) for \( m = 2 \), likely due to the limited \( N \). Furthermore, for \( \nabla X_t \) it is best to use \( m = 2 \), in order to capture the dependence structure, or use a smaller value of \( r \) for \( m = 1 \). For \( X_t \) the size of \( m \) does not make a very big difference in the value of the QSE’s whereas it makes a difference for \( \nabla X_t \), which has different theoretical Rényi entropy rates for \( m = 1 \) and \( m = 2 \).

Once again we see that \( r = 0.01\sigma \) is too small to find any matches of length \( m + 1 \) when \( m = 2 \).

4.1.3 Process with deterministic trend

Finally for this subsection we study a simulated sequence of \( N = 1000 \) normally distributed observations with a linearly growing expected value. Specifically we consider the process \( X_t = t + \varepsilon_t \in \mathcal{N}(t, \sigma^2_{\varepsilon}) \) where \( \sigma^2_{\varepsilon} = 1 \). We also consider \( \nabla X_t = X_t - X_{t-1} \in \mathcal{N}(1, 2) \). Note that the sequence \( \nabla X_t = 1 + \varepsilon_t - \varepsilon_{t-1} \) and is hence a 1-dependent process.

From Table 2 we have that \( R_2(X_t) = \ln(2\sqrt{\pi}) + \ln(\sigma_{\varepsilon}) \), \( R_2(\nabla X_t) = \ln(2\sqrt{\pi}) + \ln(2) \). The Rényi entropy rate of order 2 for \( X_t \) is by (4) equal to \( R_2(X_t) \), and for \( \nabla X_t \) corresponds to \( R_2(\nabla X_t) + \ln (3/4)/2 \) by (17) with \( \theta = -1 \). Note that the Rényi entropy rates are independent of \( m \).

Table 9 summarises the corresponding true standard deviations \( \sigma \) and robust estimates \( \hat{\sigma} = 1.4826 \cdot \text{MAD} \) based on \( N = 1000 \) observations for each of the two series, as well as the Rényi entropy rates of order 2 if the data were independent normally distributed with variance \( \sigma^2 \) and \( \hat{\sigma}^2 \) respectively.

Table 9: Standard deviations, Rényi entropies, and Rényi entropy rates of order 2 for a Gaussian process with deterministic trend \( X_t = t + \varepsilon_t \in \mathcal{N}(t, 1) \), as well as its first-order differences \( \nabla X_t \). The \( \hat{\sigma} \) is computed based on a simulated realisation of the process of length \( N = 1000 \).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Standard deviation</th>
<th>Rényi entropy rate of order 2</th>
<th>( R_2(\cdot) ) if ( \mathcal{N}(\cdot, \sigma^2) )</th>
<th>( R_2(\cdot) ) if ( \mathcal{N}(\cdot, \hat{\sigma}^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_t )</td>
<td>1</td>
<td>370.2</td>
<td>1.266</td>
<td>1.266</td>
</tr>
<tr>
<td>( \nabla X_t )</td>
<td>( \sqrt{2} \approx 1.414 )</td>
<td>1.445</td>
<td>1.468</td>
<td>1.612</td>
</tr>
</tbody>
</table>

The discrepancy between the true \( \sigma \) and the robust \( \hat{\sigma} \) for \( X_t \) is clearly due to the deterministic trend, which makes any variance estimate unreliable unless specifically taking the deterministic trend into account.

Table 10 shows the QSE\((m, r, N)\) computed based on the \( N = 1000 \) observations of each of the sequences \( X_t \) and \( \nabla X_t \) using the previously specified \( r \)'s and \( m \)'s.
Table 10: QSE\((m, r, N)\) computed based on \(N = 1000\) simulated normally distributed observations with a deterministic linearly growing trend \(X_t = t + \varepsilon_t \in \mathcal{N}(t, 1)\) as well as for \(\nabla X_t\).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>QSE((m, r, 1000))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(r) Exact (\sigma) Robust (\hat{\sigma}) Exact (\sigma) Robust (\hat{\sigma})</td>
</tr>
<tr>
<td>(X_t)</td>
<td>(0.01\sigma) NaN 2.278 NaN 2.171</td>
</tr>
<tr>
<td></td>
<td>(0.1\sigma) 1.705 4.326 NaN 4.316</td>
</tr>
<tr>
<td></td>
<td>(0.2\sigma) 2.781 5.009 NaN 5.003</td>
</tr>
<tr>
<td>(\nabla X_t)</td>
<td>(0.01\sigma) 1.651 1.626 NaN NaN</td>
</tr>
<tr>
<td></td>
<td>(0.1\sigma) 1.438 1.441 1.412 1.407</td>
</tr>
<tr>
<td></td>
<td>(0.2\sigma) 1.446 1.446 1.405 1.402</td>
</tr>
</tbody>
</table>

Comparing the Rényi entropy rates in Table 9 to the QSE’s in Table 10, we can see that for the original data \(X_t\), \(r = 0.01\sigma\) seems to work best, using robust \(\hat{\sigma}\). For the differenced data \(m = 2\) and \(r = 0.1\sigma\) works best when using exact \(\sigma\), and \(m = 1\) and \(r = 0.2\sigma\) works best when using robust \(\hat{\sigma}\).

This time we see that, for the original data, \(m = 2\) makes all the studied \(r\)’s too small to find any matches of length \(m + 1\) when using the exact \(\sigma\). We see too that \(0.01\sigma\) is too small to find any matches of length \(m + 1\) even when \(m = 1\). This led to a division-by-zero error when using \(m = 2\) and \(r = 0.01\sigma\), in which case \(A = B = 0\).

To summarise, for the most part we have seen that \(r\) should not be too low, which is to say that \(r = 0.1\sigma\) or \(r = 0.2\sigma\) work well. Furthermore, except in the case of differenced autoregressive data, the size of \(m\) does not make much of a difference and one would be well advised to use \(m = 1\) in order to try to avoid logarithms of zero. A good approach could be to choose \(r\) such that QSE\((m, r, N)\) finds an approximate minimum, since QSE\((m, r, N)\) tends to overestimate the Rényi entropy rate of order 2 for both too low \(r\)’s and too high \(r\)’s.

4.2 Simulated bursty data

In this section, we consider a set of 7 different more or less bursty time series of sample size \(N = 1000\). The theoretical quadratic Rényi entropy rates are not possible to compute for these series, but we study the behaviour of the sample entropy, conditional probability and QSE for the original data, the differenced data and the relative difference, and try to draw conclusions about how these measures can be used to characterise "burstiness" such as large variance, spikes, unpredictability or other features in the time series. We study both the original data, the differenced data and the relative differences in order to have three classes of series with various properties.

4.2.1 Generation of the series

Here we describe how the seven bursty time series in Figure 3 are constructed and generated. Series 1 is a sine function with random phase, and with added i.i.d. \(\mathcal{N}(0,1)\) random variables in each time point. The entire series is then shifted by ten units so that it does not achieve negative values. Formally, Series 1 is generated from the following model:

\[
y_t = 10 + 2\sin(2\pi t/100 + \phi) + Z_t,
\]

where

\[
\phi \in U(0, 2\pi)
\]
\[
Z_t \in \mathcal{N}(0,1), \text{i.i.d.}
\]
Series 2 is built upon an AR(2) process with AR coefficients 0.9 and -0.75. The absolute values of the process are then computed, and for each value is independently added a uniformly distributed random variable, $U(0, 30)$, with probability 0.005. The entire series is then shifted by 0.1 units to prevent close-to-zero values. The model for Series 2 is thus

$$y_t = 0.9 y_{t-1} - 0.75 y_{t-2} + \varepsilon_t,$$

where $\varepsilon_t \in \mathcal{N}(0, 1)$. Then Series 2 is constructed as

$$Z_t = 0.1 + |y_t| + X_t I_t,$$

where

$$X_t \in U(0, 30),$$

$$I_t = \begin{cases} 1 & \text{with probability 0.005} \\ 0 & \text{with probability 0.995} \end{cases}.$$

The underlying AR(2) process is generated in Python using the software package Statsmodels, using the method described in McKinney et al. [2011].

Series 3 is constructed based on an AR(1) process generated in the same manner as Series 2, with AR coefficient $\theta = 0.6$. It is then split into a uniformly random number of chunks between 1 and 20, with uniformly random lengths. To all values in each chunk, an extra term is added with a certain probability, and finally the entire series is shifted by ten units to prevent negative values. More exactly, let

$$y_t = 0.6 y_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \in \mathcal{N}(0, 1)$. Next, split the time series into $n_s + 1$ chunks, where $n_s \in U\{1, 20\}$. Let the points of splits be defined by $n_s$ discrete uniformly distributed random numbers between 1 and $N$, with $N = 1000$ being the length of the time series. On each chunk $i$, add the value $20 + Z_i$ with a certain probability $p_s$, where $Z_i \in \mathcal{N}(0, 1)$. Finally, add 10 to the entire series. Here, $p_s = 0.1$.

Series 4 is generated by first splitting a time vector $(1, \ldots, 1000)$ into chunks of equal length. On these are computed 1 minus the absolute value of the cosine function. Each chunk is then multiplied by the absolute value of a $\mathcal{N}(0, 2)$ random number, and to each point in the chunk is added a $\mathcal{N}(0, 0.1)$ error term. Formally, split the time vector $[1, \ldots, n]$ into chunks $\{I_i\}_{i=1}^m$, each having length $n/m$. Then on chunk $I_i$, let

$$y_t = |Z_i|(1 - |\cos(t)|) + W_t,$$

where

$$Z_i \in \mathcal{N}(0, 2),$$

$$W_t \in \mathcal{N}(0, 0.1).$$

Each of these four series is generated in Python with NumPy, with seed 1.

The series shown in the bottom row of Figure 3 on page 27, called Series 5 through 7, are all generated via the method described in Section 2.1. The states are implemented as functions that call each other with a certain probability. To start the simulation, the function representing the normal state is called, which takes a stationary process as input. A counter is used to traverse the process. In the normal state, nothing is done except that the counter is incremented. With a given probability $p_{12}$, the function representing the exceptional state is called. In the exceptional state, normal white noise is computed, and the counter is incremented. With a given probability $p_{21}$, the normal state is called. When the counter reaches the length of the process, the simulation stops.

When generating each series, $p_{12} = 0.02$, $p_{21} = 0.4$, and the white noise of the exceptional state has mean value 30 and standard deviation 5. Hence we expect the simulation to remain in the normal state for
$1/(1 - p_{11}) = 1/p_{12} = 50$ steps, and in the exceptional state for $1/(1 - p_{22}) = 1/p_{21} = 2.5$ steps. Finally, to each series, the arbitrarily chosen constant 10 is added, in order to prevent negative numbers.

Series 5 is based upon the AR(1) process $X_t - 0.9X_{t-1} = Z_t$, Series 6 upon the MA(1) process $X_t = Z_t - 0.2Z_{t-1}$, and Series 7 upon the ARMA(1, 1) series $X_t + 0.75X_{t-1} = Z_t - 0.5Z_{t-1}$. In each case, $\text{Var}\{Z_t\} = 1$.

Series 5 through 7 are generated in Python with NumPy, with seed $i$ for the $i$th series, in order to vary the value and location of the bursts. Their underlying stationary processes are generated in Python with StatsModels, once again using the method described in McKinney et al. [2011].

Figure 3 shows the original series $X_t$. Their differences $\nabla X_t = X_t - X_{t-1}$ are shown in Figure 4, and their relative differences $D_t = (X_t - X_{t-1})/X_{t-1}$ are shown in Figure 5.
Figure 4: Differenced series $\nabla X_t = X_t - X_{t-1}$.

Figure 5: Relative differences $D_t = (X_t - X_{t-1})/X_{t-1}$. 
4.2.2 Characterising the predictability of time series

In this section we will investigate some summarizing measures of time series that could assist in characterising them with respect to predictability. One measure is a robust estimate of its standard deviation, here quantified as $s = 1.4826 \cdot \text{MAD}$. Small variability means high predictability. A second measure is

$$r_1 = \frac{\sum_{t=1}^{n-1} (x_{t+1} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2},$$

which for stationary data would correspond to the lag-1 sample autocorrelation, and which indicates the strength of the first neighbour dependence in the time series in terms of how close the data cloud of points $(x_{t+1}, x_t)$ are to a straight line, the sign of $r_1$ indicating if the line has a positive or negative slope. Note that $|r_1| \leq 1$, and that if $r_1$ is far away from 0 the data cloud is more closely gathered around a straight line. A large (in absolute value) value of $r_1$ hence indicates high predictability. Yet another measure is the QSE for the original series and for the standardised series, a summarizing measure taking into account the size of the variability of the processes, the shape of the marginal distribution, and the dependence structure.

Table 11 presents the QSE’s, the $r_1$, and the $s$ for the seven original series in Fig. 3. The histograms of the seven series are shown in Fig. 6. The QSE $= \text{QSE}(m, r, N)$ of the original time series is computed using $m = 1$ and $r = 0.2s$, and on the standardised time series as $\text{QSE}_{\text{STAND}} = \text{QSE}(m, r, N) - \ln(s)$. $\text{QSE}_{\text{STAND}}$ can be compared to known $\text{QSE}_{\text{STAND}}$ for i.i.d. sequences, e.g. for normally distributed i.i.d. data $\text{QSE}_{\text{STAND}} = 1.266$ (indicating no spikes), for exponentially distributed i.i.d. data $\text{QSE}_{\text{STAND}} = 0.6931$ (being heavy-tailed and indicating spikes).

As seen in Section 4.1, the QSE for a time series with dependence structure is smaller than for a time series with i.i.d. data having the same marginal distribution. Now a small QSE indicates a low variance and potentially a strong dependence structure, or more heavy-tailed data. Combining this information with $\text{QSE}_{\text{STAND}}$, $s$, and $r_1$ can help us to identify if the low value of the QSE comes from heavy-tailedness or a strong dependence structure. If $\text{QSE}_{\text{STAND}}$ is much smaller than the $\text{QSE}_{\text{STAND}} = 0.6931$ for i.i.d. exponentially distributed data then the low QSE is due to dependence, which is also indicated by a large $|r_1|$. On the other hand, if $|r_1|$ is low then a lower $\text{QSE}_{\text{STAND}}$ could indicate more spiky data.

For the seven original series in Table 11 Series 4 has the smallest QSE (-1.11), caused by a small $s$, large $r_1$, and negative $\text{QSE}_{\text{STAND}}$, indicating a strong dependence. Hence Series 4 is predictable. Series 1 on the other hand has the largest QSE (1.516), followed by Series 3 and 7 (1.416 and 1.407 respectively). It is interesting to note that Series 3 has about three times larger $s$ than Series 1 and 7 but that $r_1 = 0.99$ for Series 3, compared to 0.659 and 0.146 for Series 1 and 7 respectively. This causes the QSE for Series 3 to be smaller than for Series 1.

Series 2 with small $r_1 = 0.085$ has a $\text{QSE}_{\text{STAND}} = 1.155$ rather close to a Gaussian i.i.d. $\text{QSE}_{\text{STAND}} = 1.266$, indicating a low presence of extreme values. On the other hand, Series 7 also with a small $r_1 = 0.1455$ has $\text{QSE}_{\text{STAND}} = 0.6374$, rather close to the exponential i.i.d. $\text{QSE}_{\text{STAND}} = 0.693$, hence indicating more extreme values.
Figure 6: Histograms of the original series $X_t$ in Fig. 3.

Table 11: $s = 1.4826 \cdot \text{MAD}$, $r_1$, QSE(1, 0.2$s$, 1000), and $\text{QSE}_{\text{STAND}} = \text{QSE} - \ln(s)$ for the original series $X_t$ in Fig. 3.

<table>
<thead>
<tr>
<th>Series</th>
<th>$s$</th>
<th>$r_1$</th>
<th>QSE</th>
<th>QSE standardised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series 1</td>
<td>1.912</td>
<td>0.6590</td>
<td>1.516</td>
<td>0.8679</td>
</tr>
<tr>
<td>Series 2</td>
<td>0.9669</td>
<td>0.08501</td>
<td>1.121</td>
<td>1.155</td>
</tr>
<tr>
<td>Series 3</td>
<td>6.262</td>
<td>0.9907</td>
<td>1.416</td>
<td>-0.4185</td>
</tr>
<tr>
<td>Series 4</td>
<td>0.3547</td>
<td>0.9841</td>
<td>-1.110</td>
<td>-0.07352</td>
</tr>
<tr>
<td>Series 5</td>
<td>2.171</td>
<td>0.5573</td>
<td>1.290</td>
<td>0.5148</td>
</tr>
<tr>
<td>Series 6</td>
<td>1.047</td>
<td>0.5439</td>
<td>1.287</td>
<td>1.241</td>
</tr>
<tr>
<td>Series 7</td>
<td>2.159</td>
<td>0.1455</td>
<td>1.407</td>
<td>0.6374</td>
</tr>
</tbody>
</table>

Figure 7 shows histograms of the seven differenced series $\nabla X_t$ in Fig. 4 and Table 12 shows the $s$, the $r_1$, and the QSE’s of the seven differenced series. All but one $r_1$ are negative and Series 7 has the largest QSE depending on of its large $\hat{\sigma}$, but its $\text{QSE}_{\text{STAND}} = 0.463$ is low and indicates the presence of extreme values. However it also has a relatively strong dependence ($r_1 = -0.569$) that lowers the value of $\text{QSE}_{\text{STAND}}$. Series 4 has the lowest QSE because of a low $s$ and a low $r_1$, and its $\text{QSE}_{\text{STAND}} = 1.232$ indicates few extreme values, which matches the histogram in Fig. 7.
Table 12: $s = 1.4826 \cdot \text{MAD}, r_1, \text{QSE}(1.0, s, 1000)$, and $\text{QSE}_{\text{STAND}} = \text{QSE} - \ln(s)$ for the differenced series $\nabla X_t = X_t - X_{t-1}$ in Fig. 4.

<table>
<thead>
<tr>
<th>Series</th>
<th>$s$</th>
<th>$r_1$</th>
<th>QSE</th>
<th>QSE standardised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series 1</td>
<td>1.462</td>
<td>-0.5430</td>
<td>1.448</td>
<td>1.068</td>
</tr>
<tr>
<td>Series 2</td>
<td>1.236</td>
<td>-0.4596</td>
<td>1.395</td>
<td>1.183</td>
</tr>
<tr>
<td>Series 3</td>
<td>1.039</td>
<td>-0.1034</td>
<td>1.316</td>
<td>1.278</td>
</tr>
<tr>
<td>Series 4</td>
<td>0.1095</td>
<td>0.1403</td>
<td>-0.9795</td>
<td>1.232</td>
</tr>
<tr>
<td>Series 5</td>
<td>1.036</td>
<td>-0.2977</td>
<td>1.340</td>
<td>1.305</td>
</tr>
<tr>
<td>Series 6</td>
<td>1.615</td>
<td>-0.2962</td>
<td>1.536</td>
<td>1.057</td>
</tr>
<tr>
<td>Series 7</td>
<td>4.309</td>
<td>-0.5693</td>
<td>1.897</td>
<td>0.4363</td>
</tr>
</tbody>
</table>

Fig. 8 shows histograms of the seven relative differences $D_t$ in Fig. 5, and Table 13 shows the $s$, the $r_1$, and the QSE's of the seven relative differences. Now all $r_1$ are negative. In general these series have a low $|r_1|$, except for Series 1 and 7 where $r_1$ is lower than -0.5. Series 3 has the lowest QSE because of its low $s$. Series 2 has the largest QSE because of its high $s$, and Series 4 with a more heavy-tailed distribution has $\text{QSE}_{\text{STAND}} = 0.935$. 

---

Figure 7: Histograms of the differenced series $\nabla X_t = X_t - X_{t-1}$ in Fig. 4.
Figure 8: Histograms of the relative differences $D_t = (X_t - X_{t-1})/X_{t-1}$ in Fig. 5.

Table 13: $s = 1.4826 \cdot \text{MAD}$, $r_1$, QSE(1, 0.2s, 1000), and QSE_{STAND} = QSE − ln(s) for the relative differences $D_t = (X_t - X_{t-1})/X_{t-1}$ in Fig. 5.

<table>
<thead>
<tr>
<th>Series</th>
<th>$s$</th>
<th>$r_1$</th>
<th>QSE</th>
<th>QSE standardised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series 1</td>
<td>0.1449</td>
<td>-0.5290</td>
<td>-0.8495</td>
<td>1.082</td>
</tr>
<tr>
<td>Series 2</td>
<td>0.9408</td>
<td>-0.1201</td>
<td>1.103</td>
<td>1.158</td>
</tr>
<tr>
<td>Series 3</td>
<td>0.05920</td>
<td>-0.09748</td>
<td>-1.659</td>
<td>1.168</td>
</tr>
<tr>
<td>Series 4</td>
<td>0.2381</td>
<td>-0.1125</td>
<td>-0.5001</td>
<td>0.9350</td>
</tr>
<tr>
<td>Series 5</td>
<td>0.1024</td>
<td>-0.1264</td>
<td>-0.9972</td>
<td>1.282</td>
</tr>
<tr>
<td>Series 6</td>
<td>0.1588</td>
<td>-0.2278</td>
<td>-0.7493</td>
<td>1.091</td>
</tr>
<tr>
<td>Series 7</td>
<td>0.4216</td>
<td>-0.5749</td>
<td>-0.3582</td>
<td>0.5055</td>
</tr>
</tbody>
</table>

4.2.3 Sensitivity to choice of $r$ and $m$

We now proceed to studying how QSE, sample entropy, and the conditional probability $A/B$ behave when varying the tuning parameters $r$ and $m$.

First $r$ is varied with fixed $m = 1$. The $r$-values used are 0.01, 0.03, 0.05, 0.1, 0.3, 0.5, 1, 3. Additionally, $r$-values corresponding to the factor 0.2 multiplied by three estimates of the standard deviation of the series were used: The sample standard deviation, 1.4826 times the median absolute deviation, and the interquartile range divided by 1.349. Discussion of these multiplicative constants can be found in Section 2.3.4. Finally an $r$-value corresponding to 0.3 times the seventieth percentile of the series is used, as proposed by Ali-Eldin et al. [2014]. In Figs. 9 to 11 below, these four special cases are marked: Sample standard deviation as rings, median absolute deviation as stars, inter-quartile range as crosses, seventieth percentile as plusses.

32
The QSE, sample entropy, and conditional probability for \( m = 1 \) and varying \( r \) are given in the top three panels of Figs. 9 to 11 respectively for the seven series, their differences, and their relative differences.

One thing to note in Fig. 9 is that, as \( r \) decreases, the QSE converges to a certain value. However, noting, for example, differenced Series 7 or the relative difference of Series 2, we see that \( r \) shouldn’t be too low, since then the QSE has a tendency to overshoot the value to which it otherwise converges. Looking at the markers, we see that the stars, corresponding to \( 0.2 \cdot 1.4826 \cdot \text{MAD} \), are in general close to hitting the minimum QSE value. As is pointed out in Section 4.1.3, optimal \( r \) to hit the convergence value of the QSE is the one that minimises \( \text{QSE}(m, r, N) \).

It should be noted that as the tolerance interval \( r \) increases, the conditional probability of finding a match of length \( m + 1 \) given that a match of length \( m \) is found approaches 1, as should be expected due to the fact that more templates are seen as similar when the tolerance interval increases. This means that SampEn approaches 0 since it is the negative logarithm of this conditional probability.

Moreover, as \( r \) decreases, SampEn approaches infinity as the conditional probability approaches 0. This suggests that, for small \( r \)’s, QSE is a better measure for comparing time series than SampEn, since the former converges while the latter does not. It is therefore difficult to know which \( r \) to use for comparison of SampEn and conditional probability.

Because of the observation about MAD above, when varying \( m \), \( r \) fixed at \( 0.2 \cdot 1.4826 \cdot \text{MAD} \). The \( m \)’s used are 1, 2, and 3, and the QSE, SampEn, and conditional probability are shown in the bottom three panels of Figs. 9 to 11. It should be noted that, for fixed \( r \), QSE is in general relatively stable with respect to changes in \( m \).

Note also how the conditional probability in general tends to increase when \( m \) increases, for fixed \( r \). This is because it is more probable that one more data point will keep the templates similar if there are initially many points in the template. This of course means that sample entropy decreases, since it is the negative logarithm of the conditional probability.

Figure 9: QSE of the seven series, their differences, and their relative differences when varying \( r \) and \( m \)
Figure 10: Sample entropy of the seven series, their differences, and their relative differences when varying $r$ and $m$.

Figure 11: Conditional probability estimate $A/B$ of the seven series, their differences, and their relative differences when varying $r$ and $m$. 
4.2.4 Comparison of various measures of burstiness

Finally for this section, we look at ApEn, SampEn, and QSE for the studied series. The values of ApEn, SampEn, and QSE for the original series are shown in Table 14, the values for the differences are shown in Table 15 and the values for the relative differences are shown in Table 16. The parameters used are \( m = 1 \) and \( r = 0.2 \cdot 1.4826 \cdot \text{MAD} \). Two things are particularly prominent here: Firstly, ApEn and SampEn are in general close to each other, which is a good thing since they are supposed to estimate the same quantity. However, ApEn is almost consistently higher than SampEn, due to the self-matches. Secondly, the various methods do not in general agree on the ordering of different series, which is something to take into account when comparing different series. However, a recommendation would be to use QSE for comparisons, since ApEn and SampEn are too unstable with respect to changes in the tuning parameters \( m \) and \( r \).

Table 14: ApEn, SampEn, and QSE for the original series \( X_t \) in Fig. 3

<table>
<thead>
<tr>
<th>Series</th>
<th>( s )</th>
<th>( r_1 )</th>
<th>ApEn</th>
<th>SampEn</th>
<th>QSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series 1</td>
<td>1.912</td>
<td>0.6590</td>
<td>1.838</td>
<td>1.784</td>
<td>1.516</td>
</tr>
<tr>
<td>Series 2</td>
<td>0.9669</td>
<td>0.08501</td>
<td>2.148</td>
<td>2.071</td>
<td>1.121</td>
</tr>
<tr>
<td>Series 3</td>
<td>6.262</td>
<td>0.9907</td>
<td>0.5558</td>
<td>0.4983</td>
<td>1.416</td>
</tr>
<tr>
<td>Series 4</td>
<td>0.3547</td>
<td>0.9841</td>
<td>1.050</td>
<td>0.8427</td>
<td>-1.110</td>
</tr>
<tr>
<td>Series 5</td>
<td>2.171</td>
<td>0.5573</td>
<td>1.527</td>
<td>1.431</td>
<td>1.290</td>
</tr>
<tr>
<td>Series 6</td>
<td>1.047</td>
<td>0.5439</td>
<td>2.141</td>
<td>2.158</td>
<td>1.287</td>
</tr>
<tr>
<td>Series 7</td>
<td>2.159</td>
<td>0.1455</td>
<td>1.631</td>
<td>1.554</td>
<td>1.407</td>
</tr>
</tbody>
</table>

Table 15: ApEn, SampEn, and QSE for the differences \( \nabla X_t = X_t - X_{t-1} \) in Fig. 4

<table>
<thead>
<tr>
<th>Series</th>
<th>( s )</th>
<th>( r_1 )</th>
<th>ApEn</th>
<th>SampEn</th>
<th>QSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series 1</td>
<td>1.462</td>
<td>-0.5430</td>
<td>2.034</td>
<td>1.985</td>
<td>1.448</td>
</tr>
<tr>
<td>Series 2</td>
<td>1.236</td>
<td>-0.4596</td>
<td>2.150</td>
<td>2.010</td>
<td>1.395</td>
</tr>
<tr>
<td>Series 3</td>
<td>1.039</td>
<td>-0.1034</td>
<td>2.222</td>
<td>2.193</td>
<td>1.316</td>
</tr>
<tr>
<td>Series 4</td>
<td>0.1095</td>
<td>0.1403</td>
<td>2.248</td>
<td>2.149</td>
<td>-0.9795</td>
</tr>
<tr>
<td>Series 5</td>
<td>1.036</td>
<td>-0.2977</td>
<td>2.178</td>
<td>2.221</td>
<td>1.340</td>
</tr>
<tr>
<td>Series 6</td>
<td>1.615</td>
<td>-0.2962</td>
<td>1.974</td>
<td>1.972</td>
<td>1.536</td>
</tr>
<tr>
<td>Series 7</td>
<td>4.309</td>
<td>-0.5693</td>
<td>1.481</td>
<td>1.353</td>
<td>1.897</td>
</tr>
</tbody>
</table>

Table 16: ApEn, SampEn, and QSE for the relative differences \( D_t = (X_t - X_{t-1})/X_{t-1} \) in Fig. 5

<table>
<thead>
<tr>
<th>Series</th>
<th>( s )</th>
<th>( r_1 )</th>
<th>ApEn</th>
<th>SampEn</th>
<th>QSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series 1</td>
<td>0.1449</td>
<td>-0.5290</td>
<td>2.071</td>
<td>1.998</td>
<td>-0.8495</td>
</tr>
<tr>
<td>Series 2</td>
<td>0.9468</td>
<td>-0.1201</td>
<td>2.081</td>
<td>2.074</td>
<td>1.103</td>
</tr>
<tr>
<td>Series 3</td>
<td>0.05920</td>
<td>-0.09748</td>
<td>2.272</td>
<td>2.084</td>
<td>-1.659</td>
</tr>
<tr>
<td>Series 4</td>
<td>0.2381</td>
<td>-0.1125</td>
<td>2.155</td>
<td>1.851</td>
<td>-0.5001</td>
</tr>
<tr>
<td>Series 5</td>
<td>0.1024</td>
<td>-0.1264</td>
<td>2.211</td>
<td>2.197</td>
<td>-0.9972</td>
</tr>
<tr>
<td>Series 6</td>
<td>0.1588</td>
<td>-0.2278</td>
<td>2.050</td>
<td>2.007</td>
<td>-0.7493</td>
</tr>
<tr>
<td>Series 7</td>
<td>0.4216</td>
<td>-0.5749</td>
<td>1.564</td>
<td>1.422</td>
<td>-0.3582</td>
</tr>
</tbody>
</table>
5 Discussion and conclusions

The main purpose of this thesis has been to study various entropy measures and the possibilities for them to be used as measures of burstiness for time-series data. Studying quadratic sample entropy (QSE) and Rényi entropy rate of order 2, we have found that a low entropy may mean that the marginal distribution of the time series is heavy-tailed, and hence contains a significant number of spikes and outliers.

A method for finding out whether the QSE is to be considered low that has been investigated in this thesis is that of comparing the QSE of the standardised series, $QSE_{STAND}$, to the theoretical Rényi entropy rate of order 2 for i.i.d. normally or exponentially distributed data. If the $QSE_{STAND}$ is close to the Rényi entropy rate of order 2 for i.i.d. exponentially distributed data then it might indicate that the marginal distribution of the data is heavy-tailed and thus that the data contains a significant amount of spikes and outliers. Note that the $QSE_{STAND}$ should be used since it effectively removes the dependence of the QSE on the variance; since the QSE is higher for series with a large variance, studying only the size of the QSE could lead to wrong conclusions. Furthermore, the QSE as well as the $QSE_{STAND}$ decreases if the dependence in the data is strong, or if there are many extreme values. The $QSE_{STAND}$ together with the lag-1 sample autocorrelation can give information about whether a low QSE depends on a strong dependence structure or many extreme values.

Something else that has been found is a way of choosing $r$ when computing quadratic sample entropy: Fix $m$, try various factors of a robust estimate of the standard deviation of the series, preferably $1.4826 \cdot \text{MAD}$, and choose the $r$ that approximately minimises $QSE(m, r, N)$. On real-life data this could however be unfeasible, and should be done on a smaller subset of the data in order to find a suitable $r$ for the entire data set.

It is also worth mentioning that QSE is a better measure of burstiness than SampEn, since the former converges as $r$ decreases while the latter does not. Moreover, while SampEn is an improvement on ApEn, they are nearly equivalent.

The theoretical results show that for a given variance, the Rényi entropy of order 2 has a given order for different distributions, where heavy-tailed distributions have lower entropy than less heavy-tailed ones. Furthermore, the quadratic entropy rate is lower for dependent than for independent data. These results agree with the experimental results.

When studying bursty time-series, we tried to use simulations in order to approximate how real-world data sets could look. However, the application of the theory to real-world data remains outside the scope of this thesis. More work remains to be done, which should apply the theory to real-world data.

6 Acknowledgements

I would like to thank my supervisor Sara Sjöstedt de Luna for extensive help with writing this thesis. I would also like to thank Joshua S. Richman for letting me use Fig. 1.
Bibliography


