UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MATHEMATIK

UNIVERSITÄT ZU LÜBECK
GRADUATE SCHOOL FOR COMPUTING
IN MEDICINE AND LIFE SCIENCES

From the Institute of Mathematics
of the University of Lübeck
Director: Prof. Dr. Prestin

Ordinal-patterns-based segmentation and discrimination of time series with applications to EEG data

Dissertation
for Fulfillment of Requirements
for the Doctoral Degree of the University of Lübeck
- from the Department of Computer Science -

Submitted by
Anton M. Unakafov
from Smolensk, Russia

Lübeck 2015

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MATHEMATIK

UNIVERSITÄT ZU LÜBECK
GRADUATE SCHOOL FOR COMPUTING
IN MEDICINE AND LIFE SCIENCES

First referee: Prof. Dr. Karsten Keller

Second referee: Prof. Dr. José M. Amigó

Chairman: Prof. Dr. Heinz Handels

Date of oral examination: 06.03.2015

Approved for printing. Lübeck, 09.03.2015.

Zusammenfassung

Diese Arbeit ist der Zeitreihenanalyse gewidmet, insbesondere der Segmentierung von Zeitreihen und der Diskriminierung von Zeitreihensegmenten. Außerdem wird das Problem der Messung der Komplexität von Zeitreihen angesprochen. Als theoretische Modelle für Zeitreihen betrachten wir zeitdiskrete dynamische Systeme und stochastische Prozesse. Um die oben erwähnten Fragestellungen zu behandeln, verwenden wir die ordinale Musteranalyse (oMA). Der Grundgedanke der oMA besteht darin, nicht die Werte einer Zeitreihe, sondern die Ordnungsrelation zwischen ihren Zeitpunkten zu betrachten. Zentraler Gegenstand der oMA sind ordinale Muster einer Ordnung d, die die Relationen zwischen $(d + 1)$ aufeinanderfolgenden Zeitpunkten einer Zeitreihe beschreiben.

Die wichtigsten Ergebnisse dieser Arbeit sind die folgenden.

- Wir führen ein neues ordinale-Muster-basiertes Komplexitätsmaß, die bedingte Entropie ordinaler Muster, ein und untersuchen die Eigenschaften dieser Charakteristik.

- Wir schlagen eine neue Methode zur Segmentierung von Zeitreihen auf der Grundlage der bedingten Entropie ordinaler Muster vor.

- Wir entwickeln eine Clustering-basierte Methode zur Diskriminierung von Zeitreihensegmenten und wenden diese Methode auf EEG-Zeitreihen erfolgreich an.

Kapitel 1 gibt eine kurze Einführung in die Probleme, die in dieser Dissertation besprochen werden.

In Kapitel 2 wiederholen wir grundlegende Fakten über maßerhaltende dynamische Systeme und stochastische Prozesse. Besonderes Augenmerk wird dabei auf die Kolmogorov-Sinai-Entropie gelegt, die ein traditionelles Komplexitätsmaß für dynamische Systeme darstellt. Wir gehen auch auf die wichtigsten Begriffe der oMA ein und fassen die wichtigsten bekannten Beziehungen zwischen den ordinale-Muster-basierten Komplexitätsmaßen und der Kolmogorov-Sinai-Entropie zusammen.

In Kapitel 3 führen wir die bedingte Entropie ordinaler Muster ein. Diese Charakteristik beschreibt die durchschnittliche Vielfalt ordinaler Muster, die einem bestimmten ordinalen Muster nachfolgen. Wie wir zeigen, liefert die bedingte Entropie ordinaler

Muster in vielen Fällen eine gute Schätzung der Kolmogorov-Sinai-Entropie. Außerdem beweisen wir, dass für Markovshifts über einem binären Alphabet die bedingte Entropie ordinaler Muster einer endlichen Ordnung d und die Kolmogorov-Sinai-Entropie gleich sind. Wir leiten auch das empirische Gegenstück der bedingten Entropie ordinaler Muster her, das ein Maß für die Komplexität von Zeitreihen bietet.

In Kapitel 4 führen wir ordinale-Muster-basierte Methoden zur Aufdeckung von Change-Points in Zeitreihen ein. Wenn man diese Methoden verwendet, erhält man eine Segmentation der Zeitreihe in quasistationäre Segmente. Eine der vorgestellten Methoden basiert auf einer Statistik, die von der bedingten Entropie ordinaler Muster abgeleitet wird. Wie die Ergebnisse der empirischen Untersuchungen zeigen, ist diese Methode besser als andere ordinale-Muster-basierte Methoden und ihre Qualität vergleichbar mit der von klassischen Methoden. Außerdem erfordert die neue Methode im Gegensatz zu den klassischen Methoden keine a-priori-Kenntnis von den Charakteristiken der Zeitreihe, die sich verändern.

Schließlich schlagen wir in Kapitel 5 vor, eine Kombination eines ordinale-Muster-basierte Clusterings und einer Segmentierung von Zeitreihen zu Diskriminierung von Zeitreihensegmenten zu verwenden. Wir untersuchen verschiedene Clustering-Algorithmen empirisch, wählen die, die für das ordinale-Muster-basierte Clustering am besten geeignet sind, aus und wenden sie auf epileptische EEG's und Schlaf-EEG's an.

Abstract

This thesis is devoted to time series analysis, in particular, to segmentation of time series and to discrimination of time series segments. The problem of measuring time series complexity is also addressed. As theoretical models for time series we consider discrete-time dynamical systems and stochastic processes. To solve the above-mentioned problems we use ordinal pattern analysis, a novel approach based on considering order relations between values of time series instead of the values themselves. The central objects of ordinal pattern analysis are ordinal patterns of order d that describe order relations between $(d + 1)$ successive points of a time series.

The main results of this thesis are the following.

- We establish a new ordinal-patterns-based complexity measure, the conditional entropy of ordinal patterns, and investigate the properties of this quantity.
- We suggest a new method for segmentation of time series on the basis of the conditional entropy of ordinal patterns.
- We develop a method for discrimination of time series segments based on clustering and successfully apply this method to EEG time series.

Chapter 1 gives a brief introduction to the problems addressed in this thesis.

In Chapter 2 we recall basic facts about measure-preserving dynamical systems and stochastic processes. Special attention is paid to the Kolmogorov-Sinai entropy, which is a traditional measure of systems complexity. We also recall the main notions from ordinal pattern analysis and review the relationship between ordinal-patterns-based complexity measures and the Kolmogorov-Sinai entropy.

In Chapter 3 we introduce the conditional entropy of ordinal patterns that describes the average diversity of the ordinal patterns succeeding a given ordinal pattern. We demonstrate that the conditional entropy of ordinal patterns provides a good estimation of the Kolmogorov-Sinai entropy in many cases. Besides, we prove that for Markov shifts over a binary alphabet the conditional entropy of ordinal patterns for a finite order d coincides with the Kolmogorov-Sinai entropy. We also discuss the empirical counterpart of the conditional entropy of ordinal patterns, which provides a complexity measure for time series.

In Chapter 4 we introduce several ordinal-patterns-based methods for detecting change-points in time series, which provides a segmentation of time series into pseudo-stationary pieces. One of the introduced methods is based on a statistic strongly related to the conditional entropy of ordinal patterns. Results of the empirical studies show that this method has a better performance than other ordinal-patterns-based methods and a comparable performance to classical methods. Moreover, in contrast to classical methods, this new method does not require a priori knowledge of what characteristic of the time series changes in time.

Finally, in Chapter 5 we suggest to use ordinal-pattern-distributions clustering in combination with ordinal-patterns-based segmentation for discrimination of time series segments. We empirically investigate different clustering algorithms, choose the most suitable for ordinal-pattern-distributions clustering and apply them to sleep and epileptic EEG.

Acknowledgement

Though doctoral thesis is supposed to be an individual work, this work would have been hardly accomplished without participation of the people I'd like to thank here. First of all, I want to thank my supervisor Prof. Karsten Keller, whose help, enthusiasm and constant patience helped me a lot during all stages of my work at this project. I wish to express my gratitude to my co-supervisor Prof. Rolf Verleger, who guided me through the mazes of the EEG analysis. I am very grateful to Prof. Christoph Bandt and PD Dr. Bernd Pompe for their valuable comments and for the inspiring discussions. I would like to thank Prof. Vasil Kolev for placing at my disposal the sleep EEG dataset used for experiments in Chapter 5. I very much appreciate my colleagues from the Institute for Mathematics and all the officers of the Graduate school of University of Lübeck who have helped me during my PhD study.

This work would not have been possible without my Teachers, Yuri Victorovich Chernukhin, Boris Ivanovich Orekhov, Alexey Fedorovich Olkhovoi and Alexander Nikolaevich Karkishenko.

I would like to thank all my friends, whose support I've been feeling during these years, especially Igor Syalev, Oleg Semenov, Alexandru Barbu and Foti Coleca. Finally, I want to thank my Wife and my Parents, to whom I owe all my successes.

Contents

1 Introduction 13

1.1 Segmentation, discrimination and measuring complexity of time series . . . 13

1.2 Time series, stochastic processes and dynamical systems 15

1.3 Ordinal-patterns-based methods for time series analysis 17

1.4 Outline of the thesis 19

2 Preliminaries 22

2.1 Symbolic dynamics and Kolmogorov-Sinai entropy 22

2.1.1 Basic facts from ergodic theory 22

2.1.2 Symbolic dynamics, Markov shifts 25

2.1.3 Markov property of a partition 26

2.1.4 Kolmogorov-Sinai entropy 29

2.2 Stochastic processes 31

2.3 Ordinal pattern analysis 33

2.3.1 Basic notions 33

2.3.2 Permutation entropy and sorting entropy 38

2.3.3 Empirical permutation entropy and sorting entropy 39

2.3.4 Topological permutation entropy and sorting entropy 40

2.3.5 Order isomorphisms 42

2.4 Proofs 44

3 Conditional entropy of ordinal patterns 49

3.1 Physical meaning of conditional entropy and a motivating example: measuring complexity of logistic maps 50

3.2 Interrelationship between conditional entropy of ordinal patterns, permutation and sorting entropy 52

3.3 Conditional entropy of ordinal patterns and Kolmogorov-Sinai entropy . 53

3.3.1 Relationship in the general case 53

3.3.2 Periodic case 55

3.3.3 Markov property of ordinal partition 57

3.3.4 Ordinal partitions for unimodal maps 60

3.4 Empirical conditional entropy of ordinal patterns 64

3.4.1 Sensitivity to the size of a sample 65

3.4.2 Robustness with respect to noise 68

3.5 Conclusions 69

3.6 Proofs 70

4 Ordinal change-point detection 77

4.1 General framework and problem statement 78

4.1.1 Three problems of change-point detection 79

4.1.2 Structural change-points 83

4.1.3 Ordinal change-point detection 85

4.1.4 Notation for change-point detection 86

4.2 Methods for ordinal change-point detection 88

4.2.1 Detection of changes in distributions of ordinal patterns 88

4.2.2 Change-point detection via the CEofOP statistic 94

4.3 Change-point detection for the noisy logistic and autoregressive processes: simulation studies 96

4.3.1 Performance for Problem 1 96

4.3.2 Performance for Problem 2 99

4.3.3 Performance for Problem 3 101

4.4 Properties of the CEofOP statistic 103

4.4.1 CEofOP statistic: a toy example 103

4.4.2 Relation between the CEofOP and the likelihood ratio statistic . 104

4.4.3 Asymptotic properties of the CEofOP statistic 106

4.5 Implementation details 107

4.5.1 Implementation of detecting changes in distributions of ordinal patterns 107

4.5.2 Implementation of change-point detection via the CEofOP statistic 113

5 Ordinal-pattern-distributions clustering of time series 114

5.1 Ordinal-pattern-distributions clustering and basic facts from cluster analysis 115

5.1.1 Evaluation of clustering results 116

5.1.2 Dissimilarity measures and clustering algorithms 118

5.2 Investigation of algorithms for ordinal-pattern-distributions clustering . 119

5.2.1 Artificial time series for the experiments 120

5.2.2 Ordinal-pattern-distributions clustering of artificial time series . 121

5.3 Applications to biomedical time series 123

5.3.1 Applications to epileptic EEG recordings 123

5.3.2 Applications to sleep data 125

5.3.3 An algorithm for ordinal-patterns-based segmentation of a multi-channel EEG recording 131

5.4 Conclusions 131

Bibliography 133

Nomenclature

Throughout the thesis we use the following conventions and notation:

- $\mathbb{N}_0 := \mathbb{N} \cup 0$.
- $\binom{n}{k} := \frac{n!}{(n-k)!\,k!}$ for $n, k \in \mathbb{N}_0$ with $n \geq k$, $0! = 1$.
- $\lfloor x \rfloor$ is the largest integer not exceeding x.
- $x \mod 1 := x - \lfloor x \rfloor$.
- $\overline{a_1 a_2 \dots a_n} := (a_1, a_2, \dots, a_n, a_1, a_2, \dots, a_n, \dots)$ for $a_1, a_2, \dots, a_n \in \mathbb{N}_0$, $n \in \mathbb{N}$.
- $|x|$ is the absolute value of a number $x \in \mathbb{R}$; $|A|$ is the cardinality of a set A.
- Ω is a non-empty topological space, usually – the state space of a dynamical system.
- $\mathbb{B}(\Omega)$ is the Borel sigma-algebra on the space Ω.
- $\mu : \mathbb{B}(\Omega) \to [0, 1]$ is a probability measure.
- λ is the Lebesgue measure.
- $A^{\mathbb{N}}$ is the set of all one-sided sequences over a finite set (alphabet) $A = \{0, 1, \dots, l\}$ for $l \in \mathbb{N}$.
- $\mathbb{B}_{\Pi}(A^{\mathbb{N}})$ is the Borel sigma-algebra on $A^{\mathbb{N}}$ generated by the topology given by the cylinder sets.
- $\mathrm{id} : \Omega \hookleftarrow$ is the identity map on Ω: for all $\omega \in \Omega$ it holds $\mathrm{id}(\omega) = \omega$.
- The set $\mathbb{T}$ describes "time" for stochastic processes and time series and is either finite ($\mathbb{T} = \{0, 1, \dots, L\}$ for some $L \in \mathbb{N}_0$) or infinite ($\mathbb{T} = \mathbb{N}_0$).

Chapter 1

Introduction

The present thesis is devoted to time series analysis. In particular, we are interested in the segmentation of time series and discrimination of their segments, the question of measuring complexity of time series is also addressed. In this chapter we fix the aims and the object of study (Sections 1.1 and 1.2, respectively), and briefly describe the techniques used to achieve these aims (Section 1.3). Finally, we outline the structure of the thesis in Section 1.4.

1.1 Segmentation, discrimination and measuring complexity of time series

In many fields of research information about the system of interest is provided by sequences of observations, such as stock indices in economics and measurements of brain electrical activity (electroencephalogram) in medicine. These sequences of observations are in general called *time series*; analysis of time series allows to extract information about the underlying system, to model the system, and to predict its future evolution.

Three problems of time series analysis are addressed in this thesis, namely

- segmentation of time series,
- discrimination of time series segments,
- measuring complexity of a system, possibly underlying the time series.

Segmentation means splitting time series into segments in a meaningful way, such that certain characteristics of the time series are constant inside the segments, while the boundaries of the segments correspond to changes in these characteristics. By *discrimination* we understand partitioning segments of time series into classes; segments of one class should correspond to the same state of the system of interest, while segments from different classes should correspond to different states. Segmentation of time series and discrimination of time series segments are problems of general interest and have many applications in medicine, biology, physics, economics, engineering, etc.

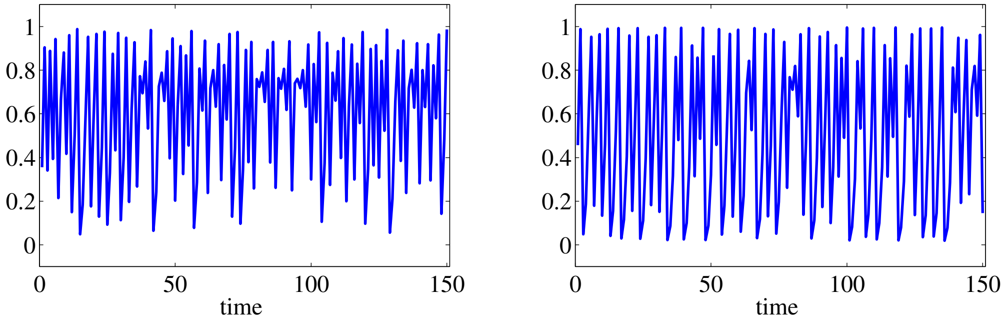
The question how one can quantify the complexity of a system arises in different contexts. When little is known about the system and it is problematic to construct a reliable model of it, one can study the system by assessing its complexity and tracking changes of complexity in time. Though measuring complexity is not directly related to the issues of segmentation and discrimination, in this thesis we introduce a measure of complexity, which is also useful for solving these two problems.

There exists a certain gap between theoretical and empirical measures of complexity. On the one hand, theoretical measures of complexity like the Kolmogorov-Sinai (KS) entropy [Wal00, Cho05] or the Lyapunov exponent [Cho05] are not easy to estimate from data (see, for instance, [ER92, Par98, CE07]). On the other hand, empirical measures of complexity often lack of a theoretical foundation and are not well interpretable. Moreover, most of complexity measures are reliable only for stationary time series, that is in case when characteristics of a time series are constant. This fact links the problem of measuring complexity with segmentation of time series, we provide an example to illustrate it.

Example 1.1. Consider a time series generated by a logistic map:

$$x(t+1) = rx(t)\big(1 - x(t)\big).$$

with $t = 0, 1, \ldots, L-1$ for some positive integer L, for $x(0) \in [0,1]$ being a random point, and $r \in [1,4]$. Figure 1.1 shows logistic time series for $r = 3.95$ and $r = 3.98$. The question is: which of them is more complex?



(a) (b)

Figure 1.1: Logistic time series for $r = 3.95$ (a) and $r = 3.98$ (b)

A theoretical answer to this question is provided by the KS entropy, it shows that the second time series is more complex [Spr03, Subsection 5.1.3]. The values of several empirical measures suggested by different authors agree with this answer (see Table 1.1).

Suppose now that the parameter r of the logistic time series varies with time, and consider a time series x_{v} given by

$$x_{\mathrm{v}}(t+1) = r(t)x_{\mathrm{v}}(t)\big(1 - x_{\mathrm{v}}(t)\big)$$

with $r(t) = 3.95$ for $t < L/2$ and $r(t) = 3.98$ for $t \geq L/2$. Is this time series more complex or less complex than the former two?

Intuitively, two different answers are possible. On the one hand, x_{v} represents two "glued" time series, therefore its complexity should be approximately equal to the mean of the segments complexity. On the other hand, x_{v} may be regarded as more complex than the first two time series since it has a variable parameter. In this case it is not clear how to define theoretical measures of complexity; the first two empirical measures of complexity support the first answer, the other two – the second one (see Table 1.1).

Complexity measure	$r = 3.95$	$r = 3.98$	variable r
KS entropy	0.581	0.602	—
estimate of Lyapunov exponent [Par98, MPW^{+}09]	0.487	0.575	0.556
permutation entropy [BP02, UK13]	0.773	0.808	0.802
conditional entropy of ordinal patterns [UK14, Una15]	0.545	0.598	0.604
approximate entropy [Pin91, Lee12]	0.565	0.592	0.604

Table 1.1: Values of complexity measures for logistic time series. For empirical complexity measures we refer first to the definition and then to the MATLAB realization. The empirical complexity measures are calculated for the length of time series $L = 2000$. Parameters of the Lyapunov exponent estimator are calculated according to the guidelines in [Par98]. Permutation entropy and conditional entropy of ordinal patterns are computed for order 4 and delay 1. Approximate entropy is computed for the embedded dimension estimated by Cao's method [Cao97], and for the tolerance calculated as 20% of the time series standard deviation.

At this point we just apply complexity measures without giving definitions. For a discussion of the KS entropy and the Lyapunov exponent see Subsection 2.1.4, permutation entropy is considered in Subsection 2.3.2, and conditional entropy of ordinal patterns is introduced in Chapter 3.

Example 1.1 motivates us to consider measuring complexity together with segmentation of time series.

1.2 Time series, stochastic processes and dynamical systems

An investigation of a time series requires a reliable model of it. A model can be either deterministic or stochastic (having some random component). In order to support prediction, it is good to have a deterministic model of time series provided by a *dynamical system.* It consists of a space of possible states of the system and of an evolution rule describing dynamics of the system in time. This evolution rule can be given by a differential equation (then the system is a continuous-time dynamical system) or by an evolution map that determines the next state for the given current state. The last case is called discrete-time dynamical system; there the notion of time is provided by applications of the evolution map: a single time slice corresponds to one application.

Example 1.2. Consider the interval $\Omega = [0, 1)$ and the dyadic map $T_2 : \Omega \hookleftarrow$ given by

$$T_2(\omega) = \begin{cases} 2\omega, & 0 \le \omega < \frac{1}{2}, \\ 2\omega - 1, & \frac{1}{2} \le \omega < 1. \end{cases}$$

Then the couple (Ω, T_2) provides a simple example of a discrete-time dynamical system.

A sequence of states generated by repeated application of an evolution rule is called an *orbit*. For an initial state ω and an evolution rule T, an orbit is given by $\big(\omega, T(\omega), T^2(\omega), \ldots\big)$ and is completely determined by ω. For instance, an orbit of the dynamical system from Example 1.2 starting with $\frac{1}{3}$ is given by $\big(\frac{1}{3}, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, \ldots\big)$. Dynamical systems provide deterministic models for time series: given a function X that associates a real number with each state from the space of dynamical system, one gets an artificial time series, the sequence of real numbers $\big(X(\omega), X\big(T(\omega)\big), X\big(T^2(\omega)\big), \ldots\big)$.

However, the construction of purely deterministic models for time series is possible only if there is enough information about the system of interest. This is sometimes not the case due to measurement errors and noises. For this reason models for time series are in many cases provided by *stochastic processes*, that is sequences of random variables. In contrast to dynamical systems, stochastic processes are unpredictable by nature, therefore one cannot expect from such models an accurate prediction of a time series evolution. Despite of this obvious difference, stochastic processes are closely related to dynamical systems. If dynamical system possesses certain properties and its initial state is known only approximately, with some finite precision, then after several applications of T the state of the dynamical system becomes unpredictable (that is, random from the viewpoint of an external observer).

Example 1.3. A sequence of independent random variables taking values from the set $\{0, 1\}$ is, perhaps, the most simple stochastic process. It is called a Bernoulli process and can be thought of as a mathematical description of successive tossing a coin. Consider an orbit of a dynamical system from Example 1.2 for some initial state $\omega \in [0, 1)$ and suppose that only the first n binary digits of ω are determined:

$$\omega = 0.a_0 a_1 \ldots a_{n-1} a_n \ldots,$$

where $a_0, a_1, \ldots, a_{n-1}$ are given and $a_n, a_{n+1}, \ldots$ are unknown. Then for the k-th iterate $T_2^k(\omega)$ of the dyadic map it holds

$$T_2^k(\omega) = 0.a_k a_{k+1} a_{k+2} \ldots,$$

which means that already for $T_2^n(\omega)$ all digits describing the current state are unknown. Then for $X(\omega)$ given by

$$X(\omega) = \begin{cases} 0, & 0 \le \omega < \frac{1}{2}, \\ 1, & \frac{1}{2} \le \omega < 1, \end{cases}$$

the sequence $\big(X\big(T_2^n(\omega)\big), X\big(T_2^{(n+1)}(\omega)\big), \ldots\big)$ models a Bernoulli process equivalent to the successive tossing a fair coin.

So dynamical systems provide both deterministic and stochastic models for time series. For this reason they are the main object of our interest, though sometimes it will be more convenient for us to speak about stochastic processes.

1.3 Ordinal-patterns-based methods for time series analysis

Ordinal pattern analysis [BP02, KSE07, Ami10] is a promising and effective approach to time series analysis. The idea behind ordinal pattern analysis is to consider order relations between values of time series instead of the values themselves. The interest to order relations between values of a time series is not new, see [SSH99] for the general discussion of rank tests and order statistics. This interest is motivated by the fact that order relations between values are invariant under translation and scaling [KSE07] and are usually more robust to noise than the values themselves [BP02], [Ami10, Subsection 3.4.3]. Ordinal-patterns-based methods are also computationally simple [UK13].

An ordinal pattern of an order d describes order relations between $(d+1)$ successive points of a time series [KL03]. The original time series is converted to a sequence of ordinal patterns, as demonstrated in Figure 1.2 for order $d = 3$.

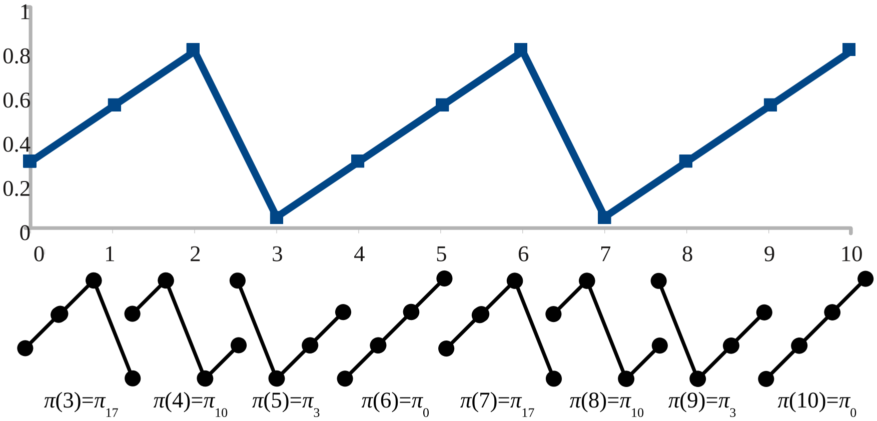


Figure 1.2: Ordinal patterns of order $d = 3$ for a periodic time series, four different patterns (π_{17}, π_{10}, π_3 and π_0) occur with period 4. An ordinal pattern $\pi(t)$ characterizes order relations between $\big(x(t-d), x(t-d+1), \ldots, x(t)\big)$. See Example 1.4 for further details

In the current thesis we use ordinal pattern analysis to solve the three above-mentioned problems: segmentation, discrimination and measuring complexity of time series. The idea of using ordinal pattern analysis for time series discrimination is not new. For instance, Keller and Lauffer provide examples demonstrating that frequencies of ordinal patterns significantly differs for EEG in normal state and during an epileptic seizure [KL03]. However, most of the existing contributions concentrate on showing that ordinal-patterns-based quantities reflect changes in complexity of a time series and there are only few works suggesting ordinal-patterns-based approaches for segmentation of time series and discrimination of their segments [Bra11, SGK12, SKC13].

Meanwhile, ordinal-patterns-based methods for measuring complexity have been developing starting from the seminal paper of Bandt and Pompe [BP02]. There are several ordinal-patterns-based measures of complexity (see, for instance, [HN11, MAAB13, Pom13]), the most renowned is the *permutation entropy.*

The more complex a time series is, the more diverse ordinal patterns occur in it, and this diversity is just what the permutation entropy measures. This concept provides a theoretically justified and simple approach to measuring complexity. Permutation entropy for order d tending to infinity is connected to the KS entropy, the central theoretical measure of complexity for dynamical systems; permutation entropy for finite d is often used as a practical complexity measure; we refer to [Ami10, AK13] for a review of applications. However, the value of permutation entropy strongly depends on the order d, and permutation entropy for finite d can be either much higher or much lower than the limit of permutation entropy as order d tends to infinity [UK14]. Let us provide an example.

Example 1.4. Consider the interval $\Omega = [0, 1)$ with an interval map $T(\omega) = (\omega + 0.25) \bmod 1$. Figure 1.2 shows a part of an orbit of this dynamical system and corresponding ordinal patterns of order $d = 3$.

Map T is periodic with period 4 (that is all points of T are periodic with this period), so dynamics provided by this map is very simple, and the KS entropy for this dynamical system is equal to zero [KAH$^+$06]. Meanwhile, the permutation entropy of order $d = 3$ is equal to $\frac{1}{3} \ln 4 > 0$ since there are four different ordinal patterns occurring with the same frequency (see Subsection 2.3.2 for the general formula of permutation entropy).

We propose to consider the *conditional entropy of ordinal patterns of order d*: it characterizes the average diversity of ordinal patterns succeeding a given one and, as we demonstrate in Chapter 3, in many cases it provides a much better practical estimation of the KS entropy than the permutation entropy. For instance for the dynamical system in Example 1.4, the conditional entropy of ordinal patterns of order $d = 3$ is equal to zero that is coincides with the KS entropy. Indeed, consider the orbit in Figure 1.2: for each ordinal pattern only one successive ordinal pattern occurs (π_{10} is the only successive ordinal pattern for π_{17}, π_3 is the only successive ordinal pattern for π_{10} and so on).

Moreover, conditional entropy of ordinal patterns appears to be rather useful for time series segmentation (see Chapter 4).

The conditional entropy of ordinal patterns is the cornerstone of this thesis; we illustrate application of this quantity to segmentation, discrimination and measuring complexity of real-world time series by the following example.

Example 1.5. The automatic scoring of sleep stages is a relevant problem in biomedical research. According to the classification in [RK68], there are 6 sleep stages:

- the waking state (W);
- two stages of light sleep (S1, S2);
- two stages of deep sleep (S3, S4);
- rapid eye movement (REM), also called paradoxical sleep since activity of neurons at this stage is similar to that during wakefulness [Lib12, p. 20].

To investigate sleep stages one measures electrical activity of the brain by means of electroencephalogram (EEG). Today discrimination of sleep EEG is mainly carried out manually by experts: they assign a sleep stage (W, REM, S1–S4) to every 30-s. epoch of the EEG recording [SAIB$^+$07]; the result is often visualized by a *hypnogram* (see Figure 1.3).

We suggest to apply ordinal-patterns-based methods to discrimination of sleep EEG. Here we consider EEG recording 14 from the dataset kindly provided by Vasil Kolev (details and other results for this dataset are provided in Experiment 5.4, p. 126). First of all, we employ the ordinal-patterns-based segmentation procedure (see Subsection 5.3.3). Then we calculate the conditional entropy of ordinal patterns for every obtained segment of the EEG recording, results are shown in Figure 1.3 in comparison with the manually scored hypnogram. Though the conditional entropy of ordinal patterns does not discriminate between sleep stages, it reflects dynamics of the EEG signal complexity, which decreases with increase of sleep deepness and vice versa.

Figure 1.4 illustrates the outcome of the ordinal-patterns-based discrimination of sleep EEG in comparison with the manual scoring by an expert; the automated identification of a sleep type (waking, REM, light sleep, deep sleep) is correct for 79.6% of 30-second epochs.

Note that ordinal-patterns-based segmentation and discrimination are completely data-driven procedures: we do not use any expert knowledge about the data to obtain the above results. This fact emphasizes potential of ordinal-patterns-based methods.

1.4 Outline of the thesis

In this thesis we discuss segmentation of time series and discrimination of time series segments, we also address the problem of measuring time series complexity. The main results of this thesis are the following:

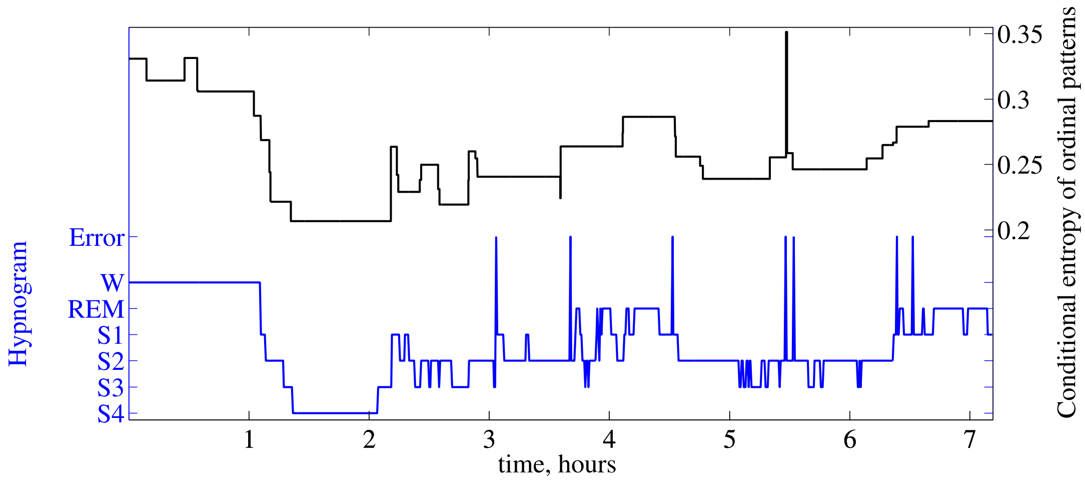


Figure 1.3: Manually scored hypnogram (lower plot) and the conditional entropy of ordinal patterns (upper plot) for an EEG recording (channel C4)

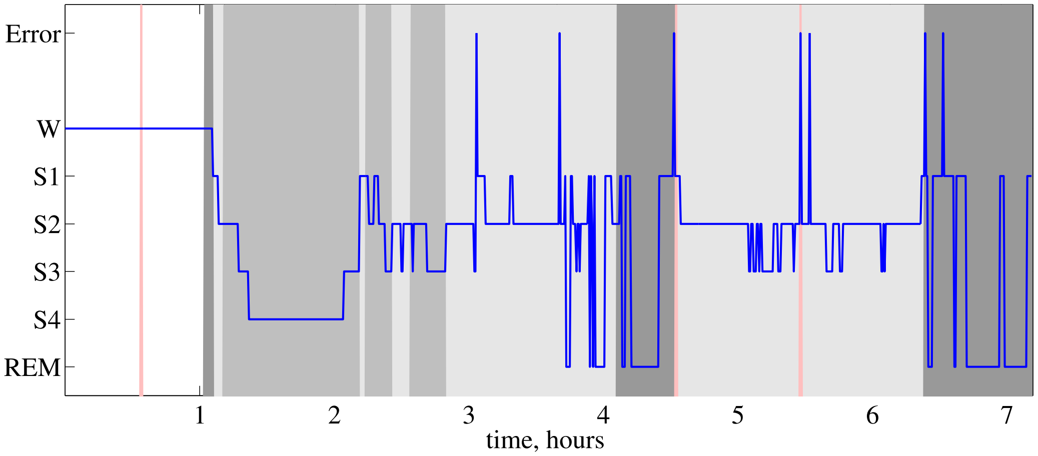


Figure 1.4: Hypnogram (bold curve) and the results of ordinal-patterns-based discrimination of sleep EEG (white color indicates epochs classified as waking state, light gray – as light sleep, gray – as deep sleep, dark gray – as REM, red color indicates unclassified segments)

- a new ordinal-patterns-based complexity measure with interesting theoretical and practical properties, the conditional entropy of ordinal patterns, is established;
- a new ordinal-patterns-based method for detection of change-points and for segmentation of time series on the basis of conditional entropy of ordinal patterns is suggested;
- a method for discrimination of time series segments on the basis of clustering is described and successfully applied to real-world (EEG) data.

The thesis is organized as follows. Chapter 2 introduces the main concepts used in the thesis. We start from a brief introduction to the theory of measure-preserving dynamical systems, which are used as theoretical models for time series throughout Chapters 2-3. Special attention is paid to symbolic dynamics and, in particular, to Markov shifts, since further in Chapter 3 some theoretical results are established for them. We consider the Kolmogorov-Sinai (KS) entropy as a traditional measure of complexity for dynamical systems and discuss some properties of it. Then we recall basic facts about stochastic processes, which are used as models for time series throughout the thesis. The rest of Chapter 2 is devoted to ordinal pattern analysis, which is applied further to theoretical objects such as dynamical systems and stochastic processes, and to real-world data. We define ordinal patterns and ordinal partitions, consider ordinal-patterns-based complexity measures (permutation entropy and sorting entropy), review relationship between ordinal-patterns-based quantities and the KS entropy.

In Chapter 3 we introduce a new ordinal-patterns-based complexity measure, the conditional entropy of ordinal patterns. We demonstrate that under certain assumptions it estimates the KS entropy better than the permutation entropy (Theorem 3.4); this fact can be useful for various applications. Besides, we prove that for periodic dynamics and for Markov shifts over a binary alphabet the conditional entropy of ordinal patterns for a finite order d coincides with the KS entropy (Theorems 3.6, 3.10), while the permutation entropy only asymptotically approaches the KS entropy. We also discuss the empirical counterpart of the conditional entropy of ordinal patterns, which can be applied to real-world time series either directly as a complexity measure or, as one will see from Chapter 4, as an ingredient of a statistic for detecting changes in time series.

Chapter 4 is devoted to segmentation of time series using ordinal-patterns-based methods. We consider there stochastic processes as models for time series and aim to find changes in dynamics (change-points) in order to segment the process into pseudo-stationary pieces. We propose a modification of an existing method for ordinal change-point detection introduced in [SGK12, SKC13] and suggest two new methods based on similar ideas. We also introduce a new method strongly related to the conditional entropy of ordinal patterns; according to the results of empirical studies, this method has better performance than other ordinal-patterns-based methods and comparable performance to classical methods. Moreover, in contrast to them, this new method does not require a priori knowledge of what characteristic of the time series changes in time.

Finally, in Chapter 5 we suggest to use ordinal-pattern-distributions clustering in combination with ordinal-patterns-based segmentation for discrimination of time series segments. We empirically investigate different clustering algorithms, choose the most suitable for ordinal-pattern-distributions clustering and apply them to sleep and epileptic EEG.

Chapter 2

Preliminaries

In this chapter we introduce the main concepts and notions used further in the thesis in order to make it self-containing. Most of results here are either known or trivial, so we go into details only when we believe that our result or interpretation of a concept can be interesting to the reader. Section 2.1 introduces basic notions from ergodic theory related to measure-preserving dynamical systems, in particular, to symbolic dynamics. We define the Kolmogorov-Sinai (KS) entropy, which is used further as a reference measure of theoretical complexity, and discuss some possible approaches to computing this quantity. In Section 2.2 we recall the definition of stochastic processes and their relationship with dynamical systems. In Section 2.3 we discuss ordinal pattern analysis of dynamical systems, stochastic processes and time series. In particular, we discuss an ordinal-patterns-based characterization of the KS entropy, which is of interest since direct computation of the KS entropy is in the general case problematic. In Section 2.4 we provide some technical proofs.

The reader who is familiar with these topics may proceed to Chapter 3.

2.1 Symbolic dynamics and Kolmogorov-Sinai entropy

2.1.1 Basic facts from ergodic theory

In this subsection we summarize relevant material from ergodic theory, for a general reference see [LM95, Cho05, ELW11]. We are mainly interested in measure-preserving dynamical systems.

Definition 2.1. Here a *measure-preserving dynamical system* is a quadruple $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$, where Ω is a non-empty topological space, $\mathbb{B}(\Omega)$ is the Borel sigma-algebra on it, $\mu : \mathbb{B}(\Omega) \to [0,1]$ is a probability measure, and $T : \Omega \hookleftarrow$ is a $\mathbb{B}(\Omega)$-$\mathbb{B}(\Omega)$-measurable *μ-preserving map*, i.e. $\mu\big(T^{-1}(B)\big) = \mu(B)$ for all $B \in \mathbb{B}(\Omega)$ (we also say that μ is an *invariant measure*).

A *topological dynamical system* is a couple (Ω, T), where Ω is a non-empty topological space and $T : \Omega \hookleftarrow$ is a continuous map.

From now on, we write that some property holds for μ-almost all $\omega \in \Omega$ if it holds for all $\omega \in \Omega_0$, where $\Omega_0 \in \mathbb{B}(\Omega)$ is a set with $\mu(\Omega_0) = 1$.

Definition 2.2. The map T is said to be *ergodic* with respect to μ (and μ is said to be ergodic with respect to T) if for every $B \in \mathbb{B}(\Omega)$ with $T^{-1}(B) = B$ it holds either $\mu(B) = 0$ or $\mu(B) = 1$. We also say that the system $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ is ergodic.

The importance of the ergodic property becomes clear from Birkhoff's Ergodic Theorem [Cho05, Theorem 3.8]: If a measure μ is ergodic then for every μ-integrable $\mathbb{R}$-valued random variable X on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ it holds

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X\big(T^i(\omega)\big) = \int_\Omega X \,\mathrm{d}\mu \quad \text{for } \mu\text{-almost all } \omega \in \Omega. \tag{2.1}$$

This means that for an ergodic dynamical system space averaging can be replaced by averaging over an orbit of μ-almost every point. In particular, the measure $\mu(B)$ of some set B coincides with the relative frequency of visiting B by points from the orbit of μ-almost every point $\omega \in \Omega$.

However, an ergodic measure μ is not necessary informative from the practical point of view. For instance, let the whole measure μ be concentrated in a fixed point $\omega_0 \in \Omega$, that is $\mu\{\omega_0\} = 1$, $T(\omega_0) = \omega_0$. Then μ is ergodic, though it is not possible to extract any information about the system behavior from the points of measure 1. The following property guarantees that the measure μ is related to the dynamics of the system [ER85, You02]. There are different definitions of SRB measure and we use the definition from [MN00] for the case $\Omega \subseteq \mathbb{R}^N$.

Definition 2.3. The measure μ is said to be *Sinai-Ruelle-Bowen (SRB)* on $\big(\Omega, \mathbb{B}(\Omega)\big)$ with respect to T if for Lebesgue-almost all $\omega \in \Omega$ it holds

$$\mu(B) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_B\big(T^i(\omega)\big), \tag{2.2}$$

where 1_B is the characteristic function of the set B.

From Definition 2.3 it follows that for every μ-integrable $\mathbb{R}$-valued random variable X on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ it holds [Mis10]

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} X\big(T^i(\omega)\big) = \int_\Omega X \,\mathrm{d}\mu \quad \text{for Lebesgue-almost all } \omega \in \Omega$$

(cf. (2.1)). The SRB measure is the most natural measure for the system. If μ is absolutely continuous with respect to Lebesgue measure then the SRB property is equivalent to ergodicity, however, in general an ergodic measure is not always SRB (for more information we refer to [You02]). Note that the SRB measure can still be supported on a set of Lebesgue measure zero (for instance, if Lebesgue-almost all orbits are attracted by a finite periodic cycle, see [MN00, Section 1]).

From Birkhoff's Ergodic Theorem it follows that a measure-preserving map T is ergodic if and only if for every $A, B \in \mathbb{B}(\Omega)$ it holds [Cho05, Theorem 3.12]

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mu\big(T^{-i}(A) \cap B\big) = \mu(A)\mu(B).$$

A map T is said to be *strong mixing* if for every $A, B \in \mathbb{B}(\Omega)$

$$\lim_{n\to\infty} \mu\big(T^{-n}(A) \cap B\big) = \mu(A)\mu(B).$$

Example 2.1. Throughout the thesis for illustrating various notions and results we consider the *golden mean map* having many nice properties. It is a particular case of the beta-transformation [Par60], and it is defined on the unit interval $[0, 1]$ by

$$T_{gm}(\omega) = \begin{cases} \varphi\omega, & 0 \le \omega \le \frac{1}{\varphi}, \\ \varphi\omega - 1, & \frac{1}{\varphi} < \omega \le 1, \end{cases}$$

where $\varphi = \frac{1}{2}(\sqrt{5} + 1)$ is the golden ratio. Figure 2.1 shows a graph of the map T_{gm}.

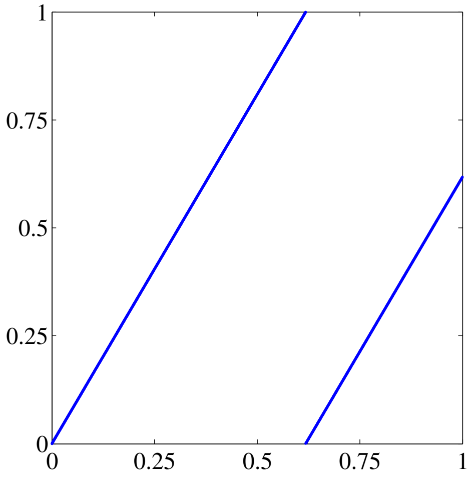


Figure 2.1: The golden mean map

The map T_{gm} preserves the measure μ_{gm} [Cho05] given by $\mu_{gm}(U) = \int_U p(\omega)d\omega$ for all $U \in \mathbb{B}\big([0, 1]\big)$ and for

$$p(\omega) = \begin{cases} \frac{\varphi^3}{1+\varphi^2}, & 0 \le \omega \le \frac{1}{\varphi}, \\ \frac{\varphi^2}{1+\varphi^2}, & \frac{1}{\varphi} < \omega \le 1. \end{cases}$$

The measure-preserving *golden mean dynamical system* $\Big([0, 1], \mathbb{B}\big([0, 1]\big), \mu_{gm}, T_{gm}\Big)$ is ergodic, moreover, it is strong-mixing [Cho05]. Since the measure μ_{gm} is absolutely continuous with respect to the Lebesgue measure, μ_{gm} is also the SRB measure.

2.1.2 Symbolic dynamics, Markov shifts

The idea of symbolic dynamics is to provide a coarse-grained description of the original dynamical system in order to understand its behavior. Here we only sketch the concept and consider some important particular cases; for a general discussion and details we refer to [LM95, Kit98, Cho05, ELW11].

First of all, define a symbolic space. Let $A^{\mathbb{N}}$ be the set of all one-sided sequences over a *finite alphabet* $A = \{0, 1, \ldots, l\}$. For $n \in \mathbb{N}$ an *n-letter word* $a_0 a_1 \ldots a_{n-1}$ over A defines the *cylinder set* $C_{a_0 a_1 \ldots a_{n-1}}$ as

$$C_{a_0 a_1 \ldots a_{n-1}} = \{(s_0, s_1, \ldots) \in A^{\mathbb{N}} \mid s_0 = a_0, s_1 = a_1, \ldots, s_{n-1} = a_{n-1}\};$$

we distinguish the case $n = 1$ and for all $a \in A$ call the set C_a a *cylinder*. The cylinder sets generate a *product topology* on $A^{\mathbb{N}}$ and are both open and closed in this topology [Kit98, Chapter 1]. Therefore the sigma-algebra $\mathbb{B}_\Pi(A^{\mathbb{N}})$ generated by the cylinder sets is the corresponding Borel sigma-algebra [Kit98, Chapter 6]. Dynamics in the symbolic space is given by the *shift map* $\sigma : A^{\mathbb{N}} \hookleftarrow$ such that

$$(\sigma s)_j = s_{j+1} \text{ for all } j \in \mathbb{N}_0, s = (s_0, s_1, \ldots) \in A^{\mathbb{N}}.$$

Now we construct the symbolic dynamics corresponding to a measure-preserving dynamical system $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$. Given a finite partition $\mathcal{P} = \{P_0, P_1, \ldots, P_l\} \subset \mathbb{B}(\Omega)$ of Ω, one assigns to each set $P_a \in \mathcal{P}$ the symbol a from the alphabet $A = \{0, 1, \ldots, l\}$. The n-letter word $a_0 a_1 \ldots a_{n-1}$ is associated with the set $P_{a_0 a_1 \ldots a_{n-1}}$ defined by

$$P_{a_0 a_1 \ldots a_{n-1}} = P_{a_0} \cap T^{-1}(P_{a_1}) \cap \ldots \cap T^{-(n-1)}(P_{a_{n-1}}). \tag{2.3}$$

We define a *coding* via the partition $\mathcal{P}$ as a map $\phi_{\mathcal{P}} : \Omega \to A^{\mathbb{N}}$ such that

$$\phi_{\mathcal{P}}(\omega) = (a_0, a_1, \ldots) \text{ with } T^i(\omega) \in P_{a_i}.$$

In particular, $\phi_{\mathcal{P}}$ maps any set $P_{a_0 a_1 \ldots a_{n-1}} \in \mathbb{B}(\Omega)$ to the corresponding cylinder set $C_{a_0 a_1 \ldots a_{n-1}} \in \mathbb{B}_\Pi(A^{\mathbb{N}})$. It holds $\sigma \circ \phi_{\mathcal{P}} = \phi_{\mathcal{P}} \circ T$, i.e. the dynamics in the symbolic space corresponds to the original dynamics. A measure-preserving symbolic dynamical system $\big(A^{\mathbb{N}}, \mathbb{B}_\Pi(A^{\mathbb{N}}), m_{\mathcal{P}}, \sigma\big)$ arises when the measure $m_{\mathcal{P}}$ is transported by the coding via $\mathcal{P}$ (below we consider only partitions $\mathcal{P} \subset \mathbb{B}(\Omega)$ without mentioning this explicitly):

$$m_{\mathcal{P}}(C_{a_0 a_1 \ldots a_{n-1}}) = \mu(P_{a_0 a_1 \ldots a_{n-1}}) \text{ for all } a_0, a_1, \ldots, a_{n-1} \in A, \text{ and } n \in \mathbb{N}.$$

Definition 2.4. Let $A^{\mathbb{N}}$ be the space of one-sided sequences over $A = \{0, 1, \ldots, l\}$ for $l \in \mathbb{N}$, and $\mathbb{B}_\Pi(A^{\mathbb{N}})$ be the Borel sigma-algebra generated by the cylinder sets. Given an $(l+1) \times (l+1)$ stochastic matrix $Q = (q_{ij})$ and a stationary probability vector $p = (p_0, p_1, \ldots, p_l)$ of Q with $p_0, p_1, \ldots, p_l > 0$, the measure m defined on the cylinder sets $C_{a_0 a_1 \ldots a_{n-1}}$ by

$$m(C_{a_0 a_1 \ldots a_{n-1}}) = p_{a_0} q_{a_0 a_1} q_{a_1 a_2} \cdots q_{a_{n-2} a_{n-1}},$$

is said to be a *Markov measure* on $A^{\mathbb{N}}$. The dynamical system $\big(A^{\mathbb{N}}, \mathbb{B}_\Pi(A^{\mathbb{N}}), m, \sigma\big)$, where $\sigma : A^{\mathbb{N}} \hookleftarrow$ is the shift map, is called a *(one-sided) Markov shift.*

In the particular case when $q_{0a} = q_{1a} = \ldots = q_{la} = p_a$ for all $a \in A$, the measure m_B defined as follows is said to be a *Bernoulli measure*:

$$m_B(C_{a_0 a_1 \ldots a_{n-1}}) = p_{a_0} p_{a_1} \cdots p_{a_{n-1}}.$$

The system $(A^{\mathbb{N}}, \mathbb{B}_\Pi(A^{\mathbb{N}}), m_B, \sigma)$ is then called a *Bernoulli shift.* We use this concept below for illustration purposes.

Example 2.2. Given the golden mean dynamical system $\Big([0,1], \mathbb{B}\big([0,1]\big), \mu_{gm}, T_{gm}\Big)$ defined in Example 2.1, consider the coding of it via the following partition

$$\mathcal{M}_{gm} = \{M_0, M_1\}, \text{ with } M_0 = \left[0, \frac{1}{\varphi}\right],\ M_1 = \left(\frac{1}{\varphi}, 1\right], \tag{2.4}$$

where $\varphi = \frac{1}{2}(\sqrt{5}+1)$. Since $\varphi - 1 = 1/\varphi$, for the measure m_{gm} transported by the coding via the $\mathcal{M}_{gm}$ it holds:

$$\begin{aligned}
m_{gm}(C_0) = \mu_{gm}(M_0) &= \int_{M_0} \frac{\varphi^3}{1+\varphi^2} dx = \frac{1}{\varphi} \frac{\varphi^3}{1+\varphi^2} = \frac{\varphi^2}{1+\varphi^2},\\
m_{gm}(C_1) = \mu_{gm}(M_1) &= \int_{M_1} \frac{\varphi^2}{1+\varphi^2} dx = \frac{\varphi-1}{\varphi} \frac{\varphi^2}{1+\varphi^2} = \frac{1}{1+\varphi^2},\\
m_{gm}(C_{00}) = \mu_{gm}(M_{00}) &= \int_{M_{00}} \frac{\varphi^3}{1+\varphi^2} dx = \frac{1}{\varphi^2} \frac{\varphi^3}{1+\varphi^2} = \frac{\varphi}{1+\varphi^2},\\
m_{gm}(C_{01}) = \mu_{gm}(M_{01}) &= \int_{M_{01}} \frac{\varphi^3}{1+\varphi^2} dx = \frac{\varphi-1}{\varphi^2} \frac{\varphi^3}{1+\varphi^2} = \frac{1}{1+\varphi^2},\\
m_{gm}(C_{10}) = m_{gm}(C_1) &= \frac{1}{1+\varphi^2},\\
m_{gm}(C_{11}) &= 0.
\end{aligned}$$

One can show, that m_{gm} is a Markov measure given by

$$\begin{aligned}
Q_{gm} &= \begin{pmatrix} m_{gm}(C_{00})/m_{gm}(C_0) & m_{gm}(C_{01})/m_{gm}(C_0) \\ m_{gm}(C_{10})/m_{gm}(C_1) & m_{gm}(C_{11})/m_{gm}(C_1) \end{pmatrix} = \begin{pmatrix} 1/\varphi & 1/\varphi^2 \\ 1 & 0 \end{pmatrix},\\
p_{gm} &= \big(m_{gm}(C_0), m_{gm}(C_1)\big) = \left(\frac{\varphi^2}{1+\varphi^2}, \frac{1}{1+\varphi^2}\right).
\end{aligned}$$

Therefore we obtain a Markov shift $\Big(\{0,1\}^{\mathbb{N}}, \mathbb{B}_\Pi\big(\{0,1\}^{\mathbb{N}}\big), m_{gm}, \sigma\Big)$ known as a *golden mean shift.*

2.1.3 Markov property of a partition

The concept of Markov shifts is rather convenient for studying the dynamical systems behavior in special cases. The following property of a partition $\mathcal{M}$ guarantees that the symbolic dynamics induced via $\mathcal{M}$ forms a Markov shift.

Definition 2.5. A finite partition $\mathcal{M} = \{M_0, M_1, \ldots, M_l\} \subset \mathbb{B}(\Omega)$ of Ω has the *(measure-theoretic) Markov property* with respect to T and μ if for all $i_0, i_1, \ldots, i_n \in \{0, 1, \ldots, l\}$ with $n \in \mathbb{N}$ and $\mu\big(M_{i_0} \cap T^{-1}(M_{i_1}) \cap \ldots \cap T^{-(n-1)}(M_{i_{n-1}})\big) > 0$ it holds

$$\frac{\mu\big(M_{i_0} \cap T^{-1}(M_{i_1}) \cap \ldots \cap T^{-n}(M_{i_n})\big)}{\mu\big(M_{i_0} \cap T^{-1}(M_{i_1}) \cap \ldots \cap T^{-(n-1)}(M_{i_{n-1}})\big)} = \frac{\mu\big(M_{i_{n-1}} \cap T^{-1}(M_{i_n})\big)}{\mu(M_{i_{n-1}})}. \tag{2.5}$$

For instance, the partition (2.4) has the Markov property for the golden mean dynamical system (see Example 2.2). Originally in [PW77] a partition with property (2.5) was called Markov partition, but we use another term to avoid confusion with the topological Markov partition. To our knowledge the difference between these two notions is seldom discussed, so it seems worth mentioning here. However, this problem is not essential for the thesis and the reader may omit the rest of this subsection.

Here we follow the definition of Markov partition given in [AKS92] for expanding maps (for other dynamical systems see the definitions in [Adl98, BS02]). Let (Ω, ρ) be a metric space, compact with respect to the induced topology. The map $T : \Omega \hookleftarrow$ is *expanding* if there exists some $c > 0$ such that for all $\omega_1, \omega_2 \in \Omega$ with $\omega_1 \neq \omega_2$ there exists some $n \in \mathbb{N}$ with $\rho\big(T^n(\omega_1), T^n(\omega_2)\big) > c$.

Definition 2.6. For a topological dynamical system (Ω, T) with expanding map T a finite cover $\mathcal{R} = \{R_0, R_1, \ldots, R_l\}$ of Ω is said to be *Markov* if it has the following properties:

(i) each R_i is a closure of its interior $\mathrm{int}R_i$;

(ii) $\mathrm{int}R_i \cap \mathrm{int}R_j = \emptyset$ for $i \neq j$;

(iii) if $T(\mathrm{int}R_i) \cap \mathrm{int}R_j \neq \emptyset$ then every point in $\mathrm{int}R_j$ has one preimage in R_i.

(Note that Ashley, Kitchens and Stafford in [AKS92] used for the cover defined as above the term "Markov partition"). To get a partition instead of a cover, it is sufficient to assign boundaries of the sets $R \in \mathcal{R}$ to certain sets. We call the partition $\mathcal{M} = \{M_0, M_1, \ldots, M_l\}$ *Markov*, if there exist a Markov cover $\mathcal{R} = \{R_0, R_1, \ldots, R_l\}$ with $M_i \subset R_i$ for $i = 0, 1, \ldots, l$. For instance, the golden mean map T_{gm} is obviously expanding and the partition (2.4) is Markov for T_{gm}.

A Markov partition does not necessarily possess the Markov property and vice versa, as is illustrated by the following example. On the one hand, the Markov property describes measure on Ω. On the other hand, the definition of the Markov partition implies that the map T is one-to-one on every element of partition.

Example 2.3. Consider two expanding maps on the interval $[0, 1]$ equipped with the Lebesgue measure λ: the tent map $\hat{T}_2$ (see Figure 2.2a) defined by

$$\hat{T}_2(\omega) = \begin{cases} 2\omega, & 0 \leq \omega \leq \frac{1}{2}, \\ 2 - 2\omega, & \frac{1}{2} < \omega \leq 1, \end{cases}$$

and the map $T_*(\omega)$ (see Figure 2.2b) given by

$$T_*(\omega) = \begin{cases} 4\omega, & 0 \le \omega \le \frac{1}{12}, \\ \frac{1}{5}(8\omega+1), & \frac{1}{12} < \omega \le \frac{1}{2}, \\ \frac{1}{3}(7-8\omega), & \frac{1}{2} < \omega \le \frac{3}{4}, \\ \frac{1}{3}(4-4\omega), & \frac{3}{4} < \omega \le 1. \end{cases} \tag{2.6}$$

Both maps preserve the Lebesgue measure λ.

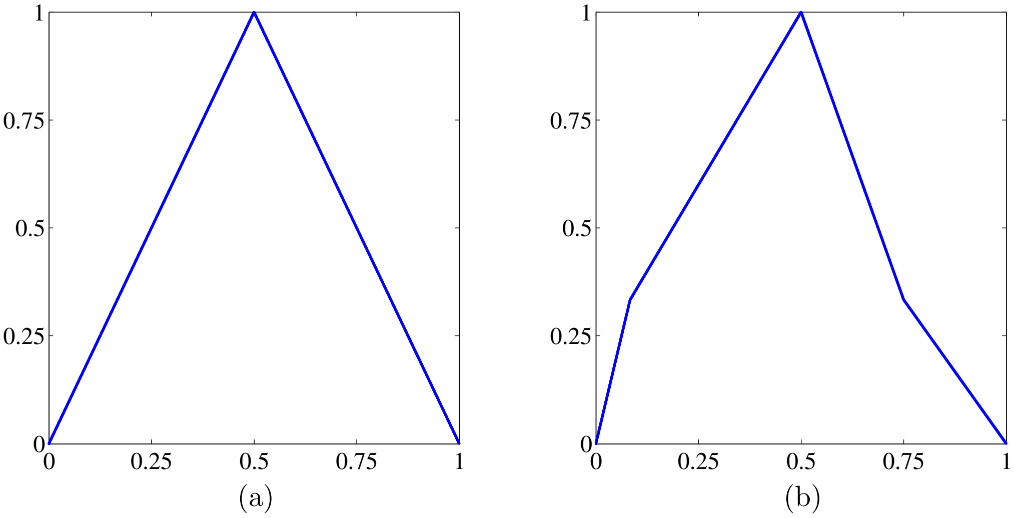


Figure 2.2: The tent map $\hat{T}_2(\omega)$ (a) and the map $T_*(\omega)$ given by (2.6) (b)

The partition $\mathcal{M} = \{[0, \frac{1}{2}], (\frac{1}{2}, 1]\}$ is Markov for both maps, as is easy to check. Though $\mathcal{M}$ possesses the Markov property with respect to $\hat{T}_2$, $\mathcal{M}$ does not possess the Markov property with respect to T_*. To see this, let us take $M_{i_0} = M_{i_1} = M_{i_2} = [0, \frac{1}{2}]$ and show that

$$\frac{\lambda\big(M_{i_0} \cap T_*^{-1}(M_{i_1}) \cap T_*^{-2}(M_{i_2})\big)}{\lambda\big(M_{i_0} \cap T_*^{-1}(M_{i_1})\big)} \ne \frac{\lambda\big(M_{i_1} \cap T_*^{-1}(M_{i_2})\big)}{\lambda(M_{i_1})}.$$

Indeed,

$$\lambda\left(\left[0, \frac{1}{2}\right] \cap T_*^{-1}\left(\left[0, \frac{1}{2}\right]\right) \cap T_*^{-2}\left(\left[0, \frac{1}{2}\right]\right)\right) = \frac{3}{64},$$

$$\lambda\left(\left[0, \frac{1}{2}\right] \cap T_*^{-1}\left(\left[0, \frac{1}{2}\right]\right)\right) = \frac{3}{16},$$

$$\lambda\left(\left[0, \frac{1}{2}\right]\right) = \frac{1}{2},$$

hence $\dfrac{\lambda\Big([0, \frac{1}{2}] \cap T_*^{-1}\big([0, \frac{1}{2}]\big) \cap T_*^{-2}\big([0, \frac{1}{2}]\big)\Big)}{\lambda\Big([0, \frac{1}{2}] \cap T_*^{-1}\big([0, \frac{1}{2}]\big)\Big)} = \dfrac{1}{4} \ne \dfrac{3}{8} = \dfrac{\lambda\Big([0, \frac{1}{2}] \cap T_*^{-1}\big([0, \frac{1}{2}]\big)\Big)}{\lambda\Big([0, \frac{1}{2}]\Big)}.$

Finally, consider a partition $\mathcal{M}' = \{[0, \frac{2}{3}], (\frac{2}{3}, 1]\}$. It is clearly not Markov for the tent map $\hat{T}_2$ since every point from $(\frac{2}{3}, 1]$ has two preimages from $[0, \frac{2}{3}]$. However, $\mathcal{M}'$ possesses the Markov property with respect to $\hat{T}_2$, as one can easily check.

2.1.4 Kolmogorov-Sinai entropy

2.1.4.1 The concept of Kolmogorov-Sinai entropy

In this subsection we discuss the concept of entropy, which allows to study the complexity of a system by means of symbolic dynamics. Given a finite partition $\mathcal{P} = \{P_0, P_1, \ldots, P_l\} \subset \mathbb{B}(\Omega)$ of Ω, the *Shannon entropy* of $\mathcal{P}$ is defined by

$$H(\mathcal{P}) = -\sum_{P_a \in \mathcal{P}} \mu(P_a) \ln \mu(P_a)$$

(with $0 \ln 0 := 0$).

Remark. A concrete base of the logarithm is not essential for the concept of entropy, commonly employed values are 2 and e. We use throughout the natural logarithm.

Consider a partition of Ω generated by the sets $P_{a_0 a_1 \ldots a_{n-1}}$ defined by (2.3):

$$\mathcal{P}_n = \{P_{a_0 a_1 \ldots a_{n-1}} \mid a_0, a_1, \ldots, a_{n-1} \in A\}. \tag{2.7}$$

The partition $\mathcal{P}_1$ coincides with $\mathcal{P}$ and the larger n is the finer is the partition $\mathcal{P}_n$, i.e. each element of $\mathcal{P}_{n+1}$ is a subset of some element of $\mathcal{P}_n$.

The *entropy rate* of T with respect to μ and $\mathcal{P}$ is defined by

$$h_\mu(T, \mathcal{P}) = \lim_{n \to \infty} \frac{H(\mathcal{P}_n)}{n} = \lim_{n \to \infty} \big(H(\mathcal{P}_{n+1}) - H(\mathcal{P}_n)\big). \tag{2.8}$$

Since the partition $\mathcal{P}_{n+1}$ is finer than $\mathcal{P}_n$, the difference $H(\mathcal{P}_{n+1}) - H(\mathcal{P}_n)$ is the *conditional entropy of $\mathcal{P}_{n+1}$ given $\mathcal{P}_n$*. The conditional entropy monotonically decreases with increasing n (for details see [CT06, Section 4.2]).

Finally, the *Kolmogorov-Sinai (KS) entropy* of T with respect to μ is given by

$$h_\mu(T) = \sup_{\mathcal{P} \text{ finite partition}} h_\mu(T, \mathcal{P}).$$

The KS entropy is an important characteristic of a dynamical system. However, computation of it involves taking a supremum over all finite partitions of Ω and is unfeasible in the general case. By this reason, the problem of computation and estimation of the KS entropy is of interest. In the rest of this section we review some possible solutions that allow computation of the KS entropy in particular cases. We will come back to this problem in Subsection 2.3.2, where we consider an ordinal-patterns-based approach to the computation of the KS entropy. In Section 3.3 we introduce a novel ordinal-patterns-based method for the estimation of the KS entropy.

2.1.4.2 Kolmogorov-Sinai entropy and generating partitions

One method for computing the KS entropy is provided by the Kolmogorov-Sinai theorem (for details we refer to [Wal00]): it holds $h_\mu(T) = h_\mu(T, \mathcal{G})$ if $\mathcal{G}$ is a generating partition in the sense of the following definition.

Definition 2.7. A finite partition $\mathcal{G} = \{G_0, G_1, \dots, G_l\} \subset \mathbb{B}(\Omega)$ of Ω is said to be *generating* (under T) if, given the sigma-algebra $\mathbb{A}$ generated by the sets $T^{-n}(G_i)$ with $i = 0, 1, \dots, l$ and $n \in \mathbb{N}_0$, for every $B \in \mathbb{B}(\Omega)$ there exists a set $A \in \mathbb{A}$ such that $\mu(A \bigtriangleup B) = 0$.

Generating partitions are in general unknown or even do not exist. By Krieger's theorem [Kri70] a generating partition exists if the map T is invertible, ergodic and has finite KS entropy (the standardized procedure for constructing the generating partition under these conditions is suggested in [Den74]). In the case of non-invertible maps there is no general method for constructing generating partitions (a comprehensive review of the generating partition problems can be found in [Sch01]).

Let us consider a special case, where Kolmogorov-Sinai entropy can be represented in a particular simple form. For ergodic Markov shifts it holds the following [Kit98, Observation 6.2.10]:

$$h_m(\sigma) = -\sum_{i=0}^{l}\sum_{j=0}^{l} m(C_{ij}) \ln \frac{m(C_{ij})}{m(C_i)} = H(\mathcal{C}_2) - H(\mathcal{C}) \tag{2.9}$$

(with $0/0 := 0$ and $0 \ln 0 := 0$), where C_{ij} are cylinder sets, C_i – cylinders and $\mathcal{C} = \{C_0, C_1, \dots, C_l\}$ is the partition consisting of all cylinders. It follows immediately that the KS entropy of a Bernoulli shift is given by

$$h_{m_B}(\sigma) = \sum_{i=0}^{l} p_i \ln p_i = \sum_{i=0}^{l} m_B(C_i) \ln m_B(C_i) = H(\mathcal{C}). \tag{2.10}$$

Example 2.4. Since the golden mean shift $\Big(\{0,1\}^{\mathbb{N}}, \mathbb{B}_{\Pi}\big(\{0,1\}^{\mathbb{N}}\big), m_{gm}, \sigma\Big)$ is a Markov shift (see Example 2.2), from (2.9) it follows:

$$h_{m_{gm}}(\sigma) = -\frac{\varphi^2}{1+\varphi^2}\frac{1}{\varphi}\ln\frac{1}{\varphi} - \frac{\varphi^2}{1+\varphi^2}\frac{1}{\varphi^2}\ln\frac{1}{\varphi^2} = \frac{\varphi+2}{1+\varphi^2}\ln\varphi = \ln\varphi.$$

The representation (2.9) of the KS entropy can be extended to a more general case.

Theorem 2.1. *If the partition $\mathcal{G} = \{G_0, G_1, \dots, G_l\}$ is generating and has the Markov property, then*

$$h_\mu(T) = H(\mathcal{G}_2) - H(\mathcal{G}) = \sum_{i=0}^{l}\sum_{j=0}^{l} \mu\big(G_i \cap T^{-1}(G_j)\big) \ln \frac{\mu\big(G_i \cap T^{-1}(G_j)\big)}{\mu(G_i)}, \tag{2.11}$$

where $\mathcal{G}_2$ is the partition of the sets corresponding to the two-letter words induced via $\mathcal{G}$ in the sense of (2.7).

For a Markov shift over the alphabet $\{0, 1, \dots, l\}$ the partition $\mathcal{C} = \{C_0, C_1, \dots, C_l\}$ is generating and has the Markov property. Therefore Theorem 2.1 is formally an extension of Observation 6.2.10 from [Kit98], but the idea of the proof does not differ significantly from the original one. However, we provide it in Subsection 2.4.1 for the sake of completeness.

2.1.4.3 Kolmogorov-Sinai entropy and Lyapunov exponent

Another approach to computing the KS entropy is provided by Pesin's formula [You13, Theorem 1], which in certain cases links together KS entropy and Lyapunov exponents. Roughly speaking, Lyapunov exponents characterize the speed of divergence for orbits of the nearby points in the space Ω. For a detailed discussion of Pesin theory we refer to [Pes97, Chapter 8] and [You03], in this thesis we use only a particular case of Pesin's formula for differentiable maps on the space $\Big([0,1], \mathbb{B}\big([0,1]\big), \mu\Big)$.

Definition 2.8 ([Cho05])**.** For a map $T : [0,1] \hookleftarrow$, ergodic with respect to the measure μ and having a piecewise continuous derivative T', the *Lyapunov exponent* is defined by

$$\mathrm{LE}(T) = \int_0^1 \ln |T'(\omega)| \mathrm{d}\mu.$$

The following result allows to compute or at least to estimate the KS entropy in certain cases.

Theorem 2.2. *If μ is the SRB measure on $\Big([0,1], \mathbb{B}\big([0,1]\big)\Big)$, then for every μ-preserving map T with* $\mathrm{LE}(T) > 0$ *it holds*

$$h_\mu(T) = \mathrm{LE}(T) = \int_0^1 \ln |T'(\omega)| \mathrm{d}\mu. \tag{2.12}$$

Moreover, in this case the KS entropy can be estimated from the orbit of λ-almost every point ω:

$$h_\mu(T) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \ln \big|T'\big(T^i(\omega)\big)\big|. \tag{2.13}$$

Proof. Equality (2.12) is a direct consequence of Pesin's formula [You13, Theorems 1], and equation (2.13) follows from (2.12) by Birkhoff's Ergodic Theorem [Cho05]. □

Example 2.5. For the golden mean dynamical system $\Big([0,1], \mathbb{B}\big([0,1]\big), \mu_{gm}, T_{gm}\Big)$ the Lyapunov exponent is given by

$$\mathrm{LE}(T) = \int_0^1 \ln |T'_{gm}(\omega)| \, \mathrm{d}\mu_{gm} = \int_0^1 \ln(\varphi) \, p(\omega) \, \mathrm{d}\omega = \left(\frac{\varphi^2}{1+\varphi^2} + \frac{\varphi^2 - \varphi}{1+\varphi^2} \right) \ln \varphi = \ln \varphi.$$

2.2 Stochastic processes

While measure-preserving dynamical systems provide a mathematical foundation for investigating real-world data, stochastic processes are used to link together theory and practice. In this section we briefly recall some basic definitions related to stochastic processes, for a more substantial introduction we refer to [TK98, GKK$^+$10].

Definition 2.9. Let $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ be a probability space. Then a sequence $\big(\mathbf{Y}(t)\big)_{t\in\mathbb{T}}$ of random vectors with $\mathbf{Y}(t) = \big(Y_1(t), Y_2(t), \ldots, Y_N(t)\big) : \Omega \to \mathbb{R}^N$ for $N \in \mathbb{N}$ is called a *stochastic process* on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$. Moreover, a stochastic process $\big(\mathbf{Y}(t)\big)_{t\in\mathbb{T}}$ is said to be *stationary* if the distributions of $\big(\mathbf{Y}(t_1), \mathbf{Y}(t_2), \ldots, \mathbf{Y}(t_n)\big)$ and $\big(\mathbf{Y}(t_1+t), \mathbf{Y}(t_2+t), \ldots, \mathbf{Y}(t_n+t)\big)$ coincide for all $t_1, t_2, \ldots, t_n, t_1+t, t_2+t, \ldots, t_n+t \in \mathbb{T}$.

By fixing a point $\omega \in \Omega$ one obtains a *realization* $\big(\mathbf{y}(t)\big)_{t\in\mathbb{T}}$ of a stochastic process:

$$\mathbf{y}(t) = \big(\mathbf{Y}(t)(\omega)\big).$$

Stochastic processes are directly linked to measure-preserving dynamical systems. Indeed, given a dynamical system $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ and a random vector $\mathbf{X} : \Omega \to \mathbb{R}^N$ called *observable*, the sequence $\big(\mathbf{Y}(t)\big)_{t\in\mathbb{N}_0}$ with

$$\mathbf{Y}(t) = \mathbf{X} \circ T^{\circ t} \tag{2.14}$$

forms an $\mathbb{R}^N$-valued stationary stochastic process on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$.

To construct a dynamical system corresponding to a stationary stochastic process $\big(\mathbf{Y}(t)\big)$ on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$, let $A \subset \mathbb{R}^N$ be the set of all possible values of $\big(\mathbf{Y}(t)\big)$. Then the process $\big(\mathbf{Y}(t)\big)$ corresponds to the dynamical system $\big(A^{\mathbb{N}}, \mathbb{B}_\Pi(A^{\mathbb{N}}), m_{\mathbf{Y}}, \sigma\big)$, where $\mathbb{B}_\Pi(A^{\mathbb{N}})$ is the sigma-algebra generated by the cylinder sets, $m_{\mathbf{Y}}$ is the measure defined on the cylinder sets $C_{a_0 a_1 \ldots a_{n-1}}$ for $a_0, a_1, \ldots, a_{n-1} \in A$ and $n \in \mathbb{N}$ by

$$m_{\mathbf{Y}}(C_{a_0 a_1 \ldots a_{n-1}}) = \mu(\{\omega \in \Omega \mid \mathbf{Y}(0)(\omega) = a_0, \mathbf{Y}(1)(\omega) = a_1, \ldots, \mathbf{Y}(n-1)(\omega) = a_{n-1}\}),$$

where σ is the shift map. See [CFS82, Chapter 8], [Gra09] for details.

Definition 2.10. A stationary stochastic process is said to be *ergodic* if the dynamical system corresponding to this process is also ergodic.

Two particular types of stochastic processes are used throughout this thesis: sequences of independent and identically distributed random variables (*IID processes*) and *Markov chains.* An example of IID process is the *standard additive white Gaussian noise* $\big(\epsilon(t)\big)_{t\in\mathbb{T}}$ with $\epsilon(t) \sim \mathcal{N}(0,1)$ for all $t \in \mathbb{T}$.

Definition 2.11. We understand a *Markov chain* as a stochastic process $\big(Y(t)\big)$ with values in $A = \{0, 1, \ldots, l\}$ for some $l \in \mathbb{N}$ such that for all $a_0, a_1, \ldots, a_{t+1} \in A$, $t \in \mathbb{T}$ it holds

$$\begin{aligned}\Pr\big(Y(t+1) = a_{t+1} \mid Y(0) = a_0, Y(1) = a_1, \ldots, Y(t) = a_t\big)& \\ = \Pr\big(Y(t+1) = a_{t+1} \mid Y(t) = a_t\big)&.\end{aligned}$$

Remark. The concept of a Markov chain is linked to the concept of a Markov shift and one gets a Markov chain $\big(Y(t)\big)$ from a Markov shift by taking an observable $X : A^{\mathbb{N}} \to A$ (for instance, $X\big((s_0, s_1, \ldots)\big) = s_0$ for all $(s_0, s_1, \ldots) \in A^{\mathbb{N}}$).

A Markov chain is determined by an $(l+1) \times (l+1)$ stochastic matrix Q of transition probabilities and a stationary probability vector p such that for all $i, j \in A$ it holds $q_{i,j} = \Pr\big(Y(1) = j \mid Y(0) = i\big)$, $p_i = \Pr\big(Y(0) = i\big)$.

2.3 Ordinal pattern analysis

In this section we provide an introduction to ordinal pattern analysis. We begin with the definitions of ordinal patterns and ordinal partition in Subsection 2.3.1. Then we introduce ordinal-patterns-based complexity measures, permutation and sorting entropies, in Subsection 2.3.2. We also discuss there the relationship between permutation entropy and KS entropy, and the ordinal characterization of KS entropy in general. In Subsection 2.3.3 we consider the empirical counterparts of permutation and sorting entropy that provide practical measures of complexity. In Subsection 2.3.4 we mention topological permutation and sorting entropy and touch the problem of calculation the number of ordinal patterns realized by a map. There we present some minor new results, since this question is not addressed in the rest of the thesis. An order isomorphism, which is important for discussion of results in Chapter 3 is defined in Subsection 2.3.5.

2.3.1 Basic notions

2.3.1.1 Ordinal patterns

Let us first recall the definition of ordinal pattern. For $d \in \mathbb{N}$ denote the set of permutations of $\{0, 1, \ldots, d\}$ by Π_d.

Definition 2.12. We say that a finite sequence of real numbers $(x_0, x_1, \ldots, x_d)$ has an *ordinal pattern* $\pi = (r_0, r_1, \ldots, r_d) \in \Pi_d$ of order $d \in \mathbb{N}$ if

$$x_{r_0} \geq x_{r_1} \geq \ldots \geq x_{r_d}$$

and

$$r_{l-1} > r_l \text{ for } x_{r_{l-1}} = x_{r_l}.$$

According to this definition, there are $(d+1)!$ different ordinal patterns of order d.

Remark. The representation of ordinal patterns used in Definition 2.12 (*permutation representation*) is convenient for theoretical considerations, but cumbersome and disadvantageous from the computational viewpoint [KSE07]. Throughout this thesis we use a *number representation* of ordinal patterns when it is convenient: an ordinal pattern of order d is given by a unique number from the set $\{0, 1, \ldots, (d+1)! - 1\}$ (see [UK13] for details, note that various enumerations are possible and here it is not important which one to use).

Consider a sequence $\big(\mathbf{x}(0), \mathbf{x}(1), \ldots, \mathbf{x}(d)\big)$ of vectors $\mathbf{x}(k) = (x_1(k), x_2(k), \ldots, x_N(k))$, $\mathbf{x}(k) \in \mathbb{R}^N$ for $k = 0, 1, \ldots, d$. We say that this sequence has an *N-dimensional ordinal pattern* $(\pi_1, \pi_2, \ldots, \pi_N) \in \Pi_d^N$, where π_j is the ordinal pattern of the sequence $\big(x_j(0), x_j(1), \ldots, x_j(d)\big)$ for $j \in 1, 2, \ldots, N$. There are $\big((d+1)!\big)^N$ possible N-dimensional ordinal patterns, thus one can associate with every $\boldsymbol{\pi} \in \Pi_d^N$ in some fixed way a number

$\mathbf{i} \in \{0, 1, \ldots, \big((d+1)!\big)^N - 1\}$. We call $\mathbf{i}$ an ordinal pattern, as well, and use either this number representation or the representation $(\pi_1, \pi_2, \ldots, \pi_N)$, whichever is more convenient at the moment.

2.3.1.2 Ordinal partitions

The idea of ordinal partition can be considered as a special case of symbolic dynamics, where the set of ordinal patterns is used as an alphabet. We define first an ordinal partition of a probability space $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ for a stochastic process.

Definition 2.13. An *ordinal partition of order* $d \in \mathbb{N}$ of the space $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ induced by a stochastic process $\mathbf{Y} = \big(\mathbf{Y}(t)\big)_{t\in\mathbb{T}}$ with $\mathbf{Y}(t) = \big(Y_1(t), Y_2(t), \ldots, Y_N(t)\big)$ and $N \in \mathbb{N}$ is defined by

$$\mathcal{P}^{\mathbf{Y}}(d) = \{P_{(\pi_1,\pi_2,\ldots,\pi_N)} \mid \pi_j \in \Pi_d \text{ for } j = 1, 2, \ldots, N\}$$

with

$$P_{(\pi_1,\pi_2,\ldots,\pi_N)} = \Big\{\omega \in \Omega \mid \Big(\big(Y_j(d)(\omega)\big), \big(Y_j(d-1)(\omega)\big), \ldots, \big(Y_j(1)(\omega)\big), \big(Y_j(0)(\omega)\big)\Big) \\ \text{has an ordinal pattern } \pi_j \text{ for } j = 1, 2, \ldots, N\Big\}.$$

We use further probabilities of ordinal patterns defined as follows.

Definition 2.14. The *probability of an ordinal pattern* $(\pi_1, \pi_2, \ldots, \pi_N) \in \Pi_d^N$ of order $d \in \mathbb{N}$ is defined as

$$p_{(\pi_1,\pi_2,\ldots,\pi_N)} = \mu(P_{(\pi_1,\pi_2,\ldots,\pi_N)}).$$

Correspondingly, the probability vector $\mathbf{p} = (p_{\boldsymbol{\pi}})_{\boldsymbol{\pi}\in\Pi_d^N}$ is said to be a *distribution of ordinal patterns.*

Remark. Distributions of ordinal patterns are known only for some special cases of stochastic processes [BS07, SK11]. In general one can estimate probabilities of ordinal patterns by their empirical probabilities, see Subsection 2.3.1.4.

Ordinal partitions for measure-preserving dynamical systems are defined in equivalent way. Given a real-valued "observable" $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, ordinal partition for a measure-preserving dynamical system $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ is defined as follows:

$$\mathcal{P}^{\mathbf{X}}(d) := \mathcal{P}^{(\mathbf{X}\circ T^t)}(d)$$

(see Section 2.2 for the discussion of the relationship between stochastic processes and dynamical systems). We consider ordinal partitions for dynamical systems to state theoretical results, while ordinal partitions for stochastic processes are more convenient for discussing applications.

Using of observables allows to define ordinal partition for a set Ω different from $\mathbb{R}^N$. To exemplify, let us discuss an observable, which is intensively used further.

Example 2.6. Consider a Markov shift over an alphabet $A = \{0, 1, \ldots, l\}$. For the sequences $r = (r_0, r_1, \ldots)$, $s = (s_0, s_1, \ldots) \in A^{\mathbb{N}}$ the natural order relation is the lexicographic order: the inequality $r \prec s$ holds if and only if either $r_0 < s_0$ or there exists some $k \in \mathbb{N}$ with $r_i = s_i$ for $i = 0, 1, \ldots, k-1$ and $r_k < s_k$.

Definition 2.15. Let us say that the observable $X : A^{\mathbb{N}} \to \mathbb{R}$ for a Markov shift $\Big(A^{\mathbb{N}}, \mathbb{B}_{\Pi}\big(A^{\mathbb{N}}\big), m, \sigma\Big)$ is *lexicographic-like* if

- for almost all $s \in A^{\mathbb{N}}$ it is injective;
- for all $s \in A^{\mathbb{N}}$ and $k, n \in \mathbb{N}_0$ the inequality $X(\sigma^k s) \leq X(\sigma^n s)$ holds if and only if $\sigma^k s \preceq \sigma^n s$.

In other words, a lexicographic-like observable X induces a natural order relation on $A^{\mathbb{N}}$. A simple example of a lexicographic-like X is provided by considering $s \in A^{\mathbb{N}}$ as $(l+1)$-expansions of a number in $[0, 1]$:

$$X_{\exp}\big((s_0, s_1, \ldots)\big) = \sum_{j=0}^{\infty} \left(\frac{1}{l+1}\right)^{j+1} s_j. \tag{2.15}$$

2.3.1.3 Examples of ordinal partitions

In order to give the reader a feeling of what the ordinal partitions are, we consider here two examples for dynamical systems that are used in Chapter 3. Those familiar with the concept of ordinal partition may skip this subsection.

Example 2.7. Consider ordinal partitions for the system $\Big([0, 1], \mathbb{B}\big([0, 1]\big), \mu_{gm}, T_{gm}\Big)$ introduced in Example 2.1 with the observable $\mathbf{X} = \mathrm{id}$. According to Definition 2.13, $\mathcal{P}^{\mathrm{id}}(1) = \{P_{(01)}, P_{(10)}\}$, where

$$\begin{aligned} P_{(01)} &= \big\{\omega \in [0, 1] \mid \omega < T_{gm}(\omega)\big\} = \left(0, \frac{1}{\varphi}\right], \\ P_{(10)} &= \big\{\omega \in [0, 1] \mid \omega \geq T_{gm}(\omega)\big\} = \{0\} \cup \left(\frac{1}{\varphi}, 1\right], \end{aligned}$$

since it holds

$$\big(T_{gm}(\omega), \omega\big) = \begin{cases} (\varphi\omega, \omega), & 0 \leq \omega \leq \frac{1}{\varphi}, \\ (\varphi\omega - 1, \omega), & \frac{1}{\varphi} < \omega \leq 1. \end{cases}$$

Ordinal partition $\mathcal{P}^{\mathrm{id}}(2)$ consists of $3! = 6$ sets given by

$$\begin{aligned} P_{(012)} &= \big\{\omega \in [0, 1] \mid \omega < T_{gm}(\omega) < T_{gm}^2(\omega)\big\} = \left(0, \frac{1}{\varphi^2}\right], \\ P_{(021)} &= P_{(102)} = \emptyset, \\ P_{(120)} &= \big\{\omega \in [0, 1] \mid T_{gm}^2(\omega) \leq \omega < T_{gm}(\omega)\big\} = \left(\frac{1}{\varphi^2}, \frac{1}{\varphi}\right], \\ P_{(201)} &= \big\{\omega \in [0, 1] \mid T_{gm}(\omega) < T_{gm}^2(\omega) \leq \omega\big\} = \left(\frac{1}{\varphi}, 1\right), \end{aligned}$$

$$P_{(210)} = \big\{\omega \in [0,1] \mid T_{gm}^2(\omega) \le T_{gm}(\omega) \le \omega\big\} = \{0\} \cup \{1\}.$$

In the same manner, one constructs ordinal partitions for higher orders d, the structure of $\mathcal{P}^{\mathrm{id}}(3)$ is shown on Figure 2.3.

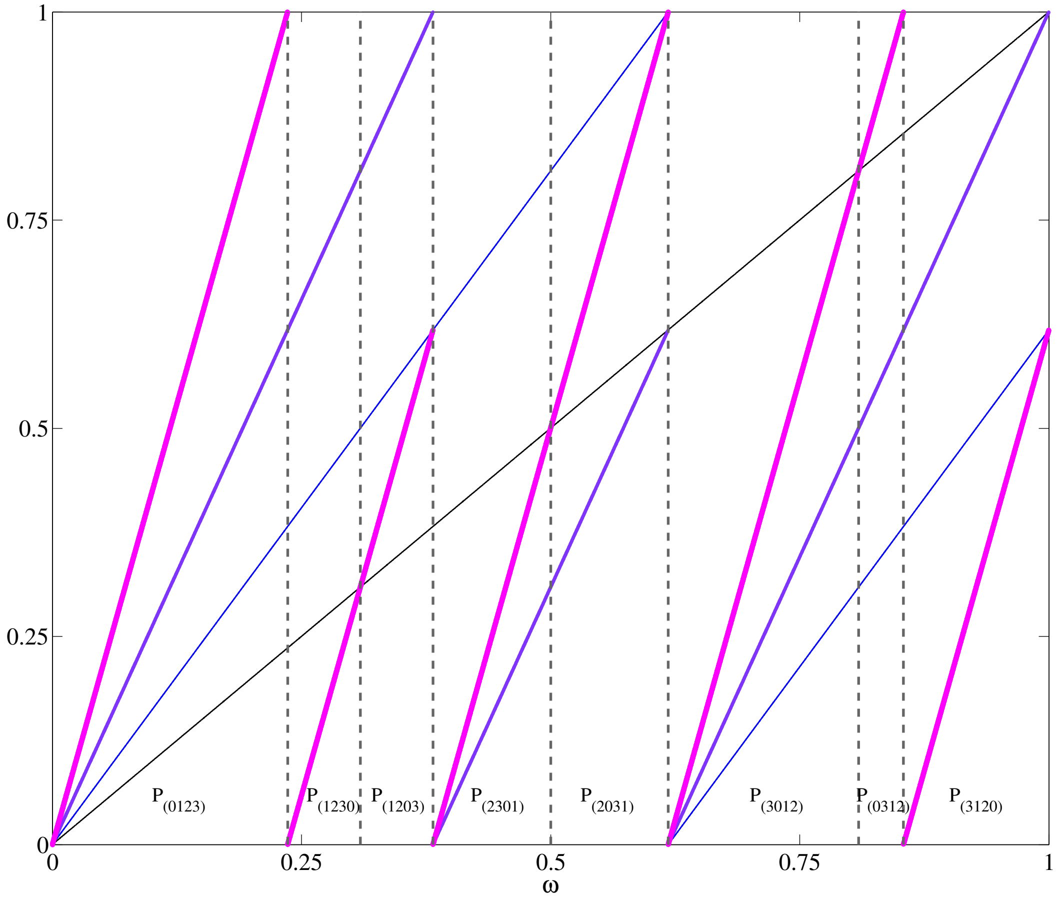


Figure 2.3: Ordinal partition for the golden mean map T_{gm} for $d=3$, the set $P_{(3210)}$ consisting of single points is not shown

Example 2.8. Consider a Markov shift over two symbols $\Big(\{0,1\}^{\mathbb{N}}, \mathbb{B}_{\Pi}\big(\{0,1\}^{\mathbb{N}}\big), m, \sigma\Big)$ and the lexicographic-like observable $X_{\exp}$ given by (2.15), which interprets $s \in \{0,1\}^{\mathbb{N}}$ as a binary expansion of a number from $[0,1]$. Observe, that for $s \in A^{\mathbb{N}}$ given $x = X_{\exp}(s)$, for all $k \in \mathbb{N}_0$ it holds:

$$X_{\exp}(\sigma^k s) = \sum_{j=0}^{\infty} s_{j+k}\left(\frac{1}{2}\right)^{j+1} = 2^k \sum_{j=0}^{\infty} s_j \left(\frac{1}{2}\right)^{j+1} - 2^k \sum_{j=0}^{k-1} s_j \left(\frac{1}{2}\right)^{j+1} = 2^k x - \lfloor 2^k x \rfloor.$$

Let $T_2(x) = 2x - \lfloor 2x \rfloor = (2x) \mod 1$ for $x \in [0,1]$, then for all $d \in \mathbb{N}$ it holds

$$\big(X(\sigma^d s), X(\sigma^{(d-1)} s), \ldots, X(\sigma s), X(s)\big) = \big(T_2^d(x), T_2^{(d-1)}(x), \ldots, T_2(x), x\big)$$

for $x = X_{\exp}(s)$. Now the structure of the ordinal partition for the Markov shift over two symbols becomes clear. For instance, for orders $d = 1, 2$ it holds

$$\begin{aligned}
&\mathcal{P}^{\mathrm{X}_{\exp}}(1) = \{P_{(01)}, P_{(10)}\} \text{ with } P_{(01)} = C_0 \setminus \{\overline{0}\},\ P_{(10)} = C_1 \cup \{\overline{0}\};\\
&\mathcal{P}^{\mathrm{X}_{\exp}}(2) = \{P_{(012)}, P_{(021)}, P_{(102)}, P_{(120)}, P_{(201)}, P_{(210)}\} \text{ with}\\
&\qquad P_{(012)} = C_{00} \setminus \{\overline{0}\},\\
&\qquad P_{(021)} = C_{1011} \cup C_{101011} \cup \ldots \cup C_{\overline{10}\,11},\\
&\qquad P_{(120)} = C_{0100} \cup C_{010100} \cup \ldots \cup C_{\overline{01}\,00},\\
&\qquad P_{(102)} = C_{011} \cup C_{01011} \cup \ldots \cup C_{\overline{01}\,1},\\
&\qquad P_{(201)} = C_{100} \cup C_{10100} \cup \ldots \cup C_{\overline{10}\,0},\\
&\qquad P_{(210)} = C_{11} \cup \{\overline{0}\},
\end{aligned}$$

where $C_{a_0 a_1 \ldots a_n}$ is a cylinder set. Note that the ordinal partition does not depend on the choice of the measure and is the same for all Markov shifts over the given alphabet.

2.3.1.4 Sequence and empirical distribution of ordinal patterns

In applications one does not deal with a stochastic process as is, but only with realizations of it for single points $\omega \in \Omega$. Therefore we define an estimate of ordinal pattern probability from a realization of a stochastic process. First we give the following definition.

Definition 2.16. A realization $\big(\mathbf{y}(t)\big)_{t\in\mathbb{T}} = \big(\mathbf{Y}(t)(\omega)\big)_{t\in\mathbb{T}}$ of an $\mathbb{R}^N$-valued stochastic process $\mathbf{Y}$ is said to have the *sequence of ordinal patterns* $\boldsymbol{\pi}(\mathbf{y}) = \big(\boldsymbol{\pi}(t;\mathbf{y})\big)_{t\in\mathbb{T}'}$ of order d for $\mathbb{T}' = \mathbb{T} \setminus \{0, 1, \ldots, d-1\}$ if $\boldsymbol{\pi}(t;\mathbf{y}) = \big(\pi_1(t), \pi_2(t), \ldots, \pi_N(t)\big)$ and the vector $\big(y_j(t), y_j(t-1), \ldots, y_j(t-d)\big)$ has the ordinal pattern $\pi_j(t)$ for all $j = 1, 2, \ldots, N$, $t \in \mathbb{T}'$.

Let us denote $\Pi_d^N = \{(\pi_1, \pi_2, \ldots, \pi_N) \mid \pi_1, \pi_2, \ldots, \pi_N \in \Pi_d\}$. Consider a sequence of ordinal patterns $\boldsymbol{\pi}(\mathbf{y})$. For any $t \in \mathbb{T}'$ the frequency of occurrence of an ordinal pattern $\mathbf{i} \in \Pi_d^N$ among the first $(t - d + 1)$ ordinal patterns of the sequence is given by

$$n_{\mathbf{i}}(t;\mathbf{y}) = \#\{r \in \{d, 1+d, \ldots, t\} \mid \boldsymbol{\pi}(r;\mathbf{y}) = \mathbf{i}\}. \tag{2.16}$$

Definition 2.17. *The empirical probability of an ordinal pattern* $\mathbf{i}$ at the moment of time $t \in \mathbb{T}'$ for the sequence $\boldsymbol{\pi}(\mathbf{y})$ is defined as

$$\widehat{p}_{\mathbf{i}}(t;\mathbf{y}) = \frac{n_{\mathbf{i}}(t;\mathbf{y})}{t - d + 1}$$

for all $\mathbf{i} \in \Pi_d^N$. Correspondingly, a probability vector

$$\widehat{\mathbf{p}}(t;\mathbf{y}) = \big(\widehat{p}_{\mathbf{i}}(t;\mathbf{y})\big)_{\mathbf{i}\in\Pi_d^N}$$

is said to be an *empirical distribution of ordinal patterns.*

Birkhoff's Ergodic Theorem links empirical probabilities of ordinal patterns to the theoretical probabilities, namely we have:

Lemma 2.3. *Let $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ be an ergodic dynamical system, $\mathbf{X}$ be a real-valued observable. Then for all $d \in \mathbb{N}$ and for all $\mathbf{i} \in \Pi_d^N$ there exists a set $\Omega_0 \subset \mathbb{B}(\Omega)$ with $\mu(\Omega_0) = 1$ such that for $P_{\mathbf{i}} \in \mathcal{P}^{\mathbf{X}}(d)$ it holds:*

$$p_{\mathbf{i}} = \mu(P_{\mathbf{i}}) = \lim_{L \to \infty} \widehat{p}_{\mathbf{i}}(L; \mathbf{y})$$

for the realization $\mathbf{y}(t) = \mathbf{X}\big(T^t(\omega)\big)$ of an arbitrary $\omega \in \Omega_0$.

Note that Birkhoff's ergodic theorem does not provide any information about the rate of convergence. This does not allow to specify what length of the sample L is sufficient for accurate estimation of ordinal patterns probabilities. Usually it is supposed that for rather large L it holds

$$p_{\mathbf{i}} \approx \widehat{p}_{\mathbf{i}}(L; \mathbf{y}).$$

2.3.2 Permutation entropy and sorting entropy

In this subsection we discuss two ordinal-patterns-based measures of complexity for dynamical systems defined as follows.

Definition 2.18. For a measure-preserving dynamical system $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ given a real-valued "observable" $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ the *permutation entropy of order d* (with respect to $\mathbf{X}$) and the *sorting entropy of order d* are respectively defined by

$$h_\mu^{\mathbf{X}}(T, d) = \frac{1}{d} H\big(\mathcal{P}^{\mathbf{X}}(d)\big),$$
$$h_{\mu,\triangle}^{\mathbf{X}}(T, d) = H\big(\mathcal{P}^{\mathbf{X}}(d+1)\big) - H\big(\mathcal{P}^{\mathbf{X}}(d)\big).$$

Note that the original definitions in [BP02] were given for the case $\Omega \subseteq \mathbb{R}$ and $\mathbf{X} = \mathrm{id}$. To see the physical meaning of the permutation entropy let us rewrite it in the explicit form:

$$h_\mu^{\mathbf{X}}(T, d) = -\frac{1}{d} \sum_{\mathbf{i} \in \Pi_d^N} \mu(P_{\mathbf{i}}) \ln \mu(P_{\mathbf{i}}). \tag{2.17}$$

That is the permutation entropy characterizes the diversity of N-dimensional ordinal patterns $\mathbf{i} \in \Pi_d^N$. The sorting entropy represents the increase of diversity of ordinal patterns as the order d increases by one.

The interest to the permutation entropy was initiated by the close relationship between this quantity and the KS entropy. Bandt et al. proved in [BKP02] for the case $\Omega \subseteq \mathbb{R}$, T being a piecewise strictly-monotone interval map and $\mathbf{X} = \mathrm{id}$, that it holds:

$$h_\mu(T) = \lim_{d \to \infty} \frac{1}{d} H\big(\mathcal{P}^{\mathrm{id}}(d)\big). \tag{2.18}$$

This result gave rise to the development of ordinal pattern analysis. Later Keller and Sinn [KS09, KS10, Kel12] showed that in many cases the following ordinal-patterns-based representation of KS entropy is possible:

$$h_\mu(T) = \lim_{d \to \infty} h_\mu\big(T, \mathcal{P}^{\mathbf{X}}(d)\big) = \lim_{d \to \infty} \lim_{n \to \infty} \Big(H\big(\mathcal{P}^{\mathbf{X}}(d)_{n+1}\big) - H\big(\mathcal{P}^{\mathbf{X}}(d)_n\big) \Big). \tag{2.19}$$

Note that if (2.19) holds then the permutation entropy and the sorting entropy for d tending to infinity provide upper bounds for the KS entropy [Kel12]:

$$\overline{\lim_{d\to\infty}}\, h_\mu^{\mathbf{X}}(T,d) \geq h_\mu(T),\ \overline{\lim_{d\to\infty}}\, h_{\mu,\triangle}^{\mathbf{X}}(T,d) \geq h_\mu(T).$$

This inequalities remain correct if one replaces upper limits by lower limits.

Remark. Note that Amigo et al. [AKK05, Ami12] have shown equality of KS entropy and permutation entropy for a concept of permutation entropy that is qualitatively different from the originally given one [BP02]. For a discussion of the relationships between both concepts see [Ami12, AK13].

For a stochastic process $\mathbf{Y}$ the permutation entropy $h_\mu(\mathbf{Y},d)$ and the sorting entropy $h_{\mu,\triangle}(\mathbf{Y},d)$ are defined in the same way as for dynamical systems.

2.3.3 Empirical permutation entropy and sorting entropy

To measure systems complexity in applications, one may want to compute the permutation entropy or the sorting entropy from time series. Simple and natural estimators are the empirical permutation entropy and the empirical sorting entropy, based on estimating $\mu(P_{\mathbf{i}})$ by the empirical probabilities of observing N-dimensional ordinal pattern $\mathbf{i} \in \Pi_d^N$ in the time series [BP02, KSE07] (for a review of applications see also [Ami10, AK13]).

Consider a realization of an ergodic stochastic process as a model of time series. Let us fix some $d \in \mathbb{N}$ and consider a finite realization $y = \big(\mathbf{y}(0), \mathbf{y}(1), \ldots, \mathbf{y}(L)\big)$ of the stochastic process $\mathbf{Y}$ for $L \in \mathbb{N}$. Given the absolute frequencies $n_{\mathbf{i}}(L;\mathbf{y})$ of ordinal patterns defined by (2.16), the naive estimator of the entropy of ordinal partition $H\big(\mathcal{P}^{\mathbf{Y}}(d)\big)$ is defined by

$$\begin{aligned}\widehat{H}\big(L;\mathcal{P}^{\mathbf{Y}}(d)\big) &= -\sum_{\mathbf{i}=0}^{((d+1)!)^N-1} \frac{n_{\mathbf{i}}(L;\mathbf{y})}{L-d+1} \ln \frac{n_{\mathbf{i}}(L;\mathbf{y})}{L-d+1} \\ &= \ln(L-d+1) - \frac{1}{L-d+1} \sum_{\mathbf{i}=0}^{((d+1)!)^N-1} n_{\mathbf{i}}(L;\mathbf{y}) \ln n_{\mathbf{i}}(L;\mathbf{y}). \qquad (2.20)\end{aligned}$$

Then the *empirical permutation entropy* and *empirical sorting entropy* are respectively given by

$$\begin{aligned}\widehat{h}_\mu(L;\mathbf{y},d) &= \frac{\widehat{H}\big(L;\mathcal{P}^{\mathbf{Y}}(d)\big)}{d} \\ &= \frac{1}{d}\ln(L-d+1) - \frac{1}{d(L-d+1)} \sum_{\mathbf{i}=0}^{((d+1)!)^N-1} n_{\mathbf{i}}(L;\mathbf{y}) \ln n_{\mathbf{i}}(L;\mathbf{y}), \\ \widehat{h}_{\mu,\triangle}(L;\mathbf{y},d) &= \widehat{H}\big(L;\mathcal{P}^{\mathbf{Y}}(d+1)\big) - \widehat{H}\big(L;\mathcal{P}^{\mathbf{Y}}(d)\big).\end{aligned}$$

Performance of these estimators of permutation and sorting entropies is investigated empirically in Subsection 3.4.1.

Remark. Note that in general the naive estimator is negatively biased and, by this reason, rather unreliable [Gra03]. A possible solution is to use Grassberger's estimator of entropy [Gra88, Gra03] given by

$$\widehat{H}_G\big(L;\mathcal{P}^{\mathbf{Y}}(d)\big) = \ln(L-d+1) - \frac{1}{L-d+1}\sum_{\mathbf{i}=0}^{((d+1)!)^N-1} n_{\mathbf{i}}(L;\mathbf{y})G\big(n_{\mathbf{i}}(L;\mathbf{y})\big),$$

where $G(0)=0$, $G(1)=-\gamma-\ln 2$ with Euler's constant $\gamma=0.577215\ldots$, and

$$G(2k+2)=G(2k+3)=G(2k)+\frac{2}{2k+1} \text{ for } k\in\mathbb{N}_0$$

(see [PR11] for a discussion of using Grassberger's estimators for ordinal-patterns-based quantities). However, empirical permutation entropy is usually not very sensitive to the size of the sample L (see, for instance, [BP02]). This is due to the fact that usually only few ordinal patterns are realized rather frequently; probabilities of these 'typical' patterns can be estimated from the finite orbit more or less reliably, whereas the rare ordinal patterns have low impact on permutation entropy. By this reason, the naive estimator of permutation entropy is traditionally used.

2.3.4 Topological permutation entropy and sorting entropy

In this subsection we discuss the number of ordinal patterns occurring in dynamics. In particular, we define topological permutation and sorting entropies. A detailed discussion of these quantities is beyond the scope of this thesis and they are used only in this subsection. However, we provide here a couple of own results related to these quantities that could be interesting by themselves.

As we have mentioned in Subsection 2.3.1, there are $(d+1)!$ different ordinal patterns of order d. Therefore, the ordinal partition $\mathcal{P}^{\mathbf{X}}(d)$ for an observable $\mathbf{X}=(X_1,X_2,\ldots,X_N)$ contains at most $\big((d+1)!\big)^N$ elements. However, in most cases, some of ordinal patterns from Π_d^N do not occur in the dynamics and by this reason some of the elements of $\mathcal{P}^{\mathbf{X}}(d)$ has zero measure. These ordinal patterns $\mathbf{i}\in\Pi_d^N$ are said to be *forbidden* [Ami10], while the occurring ordinal patterns are called *allowed*. The *numbers of forbidden* and *allowed ordinal patterns* of order d for the dynamical system $\big(\Omega,\mathbb{B}(\Omega),\mu,T\big)$ (with respect to $\mathbf{X}$) are defined by, respectively

$$\begin{aligned}\mathrm{Forb}_\mu^{\mathbf{X}}(T,d) &= \#\{\mathbf{i}\in\Pi_d^N \mid \mu(P_{\mathbf{i}})=0\},\\ \mathrm{Allow}_\mu^{\mathbf{X}}(T,d) &= \#\{\mathbf{i}\in\Pi_d^N \mid \mu(P_{\mathbf{i}})>0\}.\end{aligned} \tag{2.21}$$

Clearly,

$$\mathrm{Allow}_\mu^{\mathbf{X}}(T,d)+\mathrm{Forb}_\mu^{\mathbf{X}}(T,d)=\big((d+1)!\big)^N.$$

Remark. Originally the number of forbidden ordinal patterns was defined in a bit different way [AEK08, AK08]:

$$\mathrm{Forb}_0^{\mathbf{X}}(T,d) = \#\{\mathbf{i} \in \Pi_d^N \mid P_{\mathbf{i}} \setminus F_d(T) = \emptyset\},$$

where $F_d(T) = \big\{\omega \in \Omega \mid T^i(\omega) = T^d(\omega) \text{ for some } i \in \{0,1,\dots,d\}\big\}$, that is elements of ordinal partition containing only periodic and eventually periodic points of the map T are not taken into account. Our definition can be considered as an adaptation of the traditional one to measure-preserving dynamical systems; in most cases it holds $\mathrm{Forb}_0^{\mathrm{id}}(T,d) = \mathrm{Forb}_\mu^{\mathrm{id}}(T,d)$. For instance, the ordinal partition $\mathcal{P}^{\mathrm{id}}(2)$ in Example 2.7 contains four non-empty elements, however one of them, $P_{(210)}$, consists of only two periodic points and has zero measure, therefore $\mathrm{Forb}_0^{\mathrm{id}}(T_{gm},2) = \mathrm{Forb}_{\mu_{gm}}^{\mathrm{id}}(T_{gm},2) = 3$.

The numbers of forbidden/allowed ordinal patterns have many interesting properties (see [Ami10, Chapters 3-5]), but determining these numbers turns out to be a difficult combinatorial problem [Eli11]. To our knowledge, the only dynamical systems for which the number of allowed patterns is known are the one-sided Bernoulli shifts. For the system $\Big(\{0,1,\dots,l\}^{\mathbb{N}}, \mathbb{B}_\Pi\big(\{0,1,\dots,l\}^{\mathbb{N}}\big), m_B, \sigma\Big)$ and for lexicographic-like random variable X it holds [Eli09]:

$$\mathrm{Allow}_{m_B}^X(\sigma,d) = \sum_{k=2}^{l+1} a_{d,k}, \tag{2.22}$$

where $a_{d,k}$ is given by

$$a_{d,k} = \sum_{i=0}^{k-2} (-1)^i \binom{d+1}{i} \left((k-i-2)(k-i)^{d-1} + \sum_{j=1}^{d} (k-i)^{d-j} \psi_{k-i}(j) \right)$$

with

$$\psi_n(j) = \sum_{r|j} \mathrm{Möb}\left(\frac{j}{r}\right) n^r, \tag{2.23}$$

where Möb() is the Möbius function, see Subsection 2.4.2 for details. In the following proposition we suggest a shorter representation of $\mathrm{Allow}_{m_B}^X(\sigma,d)$.

Proposition 2.4. *Given $\big(A^{\mathbb{N}}, \mathbb{B}_\Pi(A^{\mathbb{N}}), m_B, \sigma\big)$ a Bernoulli shift over $A = \{0,1,\dots,l\}$, then for a lexicographic-like random variable X it holds*

$$\mathrm{Allow}_{m_B}^X(\sigma,d) = \sum_{k=2}^{l+1} (-1)^{(l+1-k)} \binom{d}{l+1-k} \left((k-2)k^{d-1} + \sum_{j=1}^{d} k^{d-j} \psi_k(j) \right), \tag{2.24}$$

with $\psi_n(j)$ given by (2.23). Moreover,

$$\lim_{d\to\infty} \frac{\mathrm{Allow}_{m_B}^X(\sigma,d)}{(d+1)(l+1)^d} = 1.$$

The proof is given in Subsection 2.4.2. Note that in contrast to representation (2.22), formula (2.24) does not emphasize the structure of allowed ordinal patterns. The asymptotic behavior of the number of allowed patterns for Bernoulli shifts was first characterized by Elizalde in [Eli09] without providing a proof.

The *topological permutation entropy of order d* (with respect to $\mathbf{X}$) is defined by

$$h_0^{\mathbf{X}}(T,d) = \frac{1}{d}\ln\left(\mathrm{Allow}_{\mu}^{\mathbf{X}}(T,d)\right).$$

The topological permutation entropy was introduced in [BKP02] for the case $\Omega \subseteq \mathbb{R}$ and $\mathbf{X} = \mathrm{id}$ (for further development see [Mis03]). Similar to the permutation entropy, $h_0^{\mathbf{X}}(T,d)$ characterizes diversity of ordinal patterns, but it does not take into account the measures of elements of the ordinal partition. For all $d \in \mathbb{N}$ it holds

$$h_{\mu}^{\mathbf{X}}(T,d) \leq h_0^{\mathbf{X}}(T,d),$$

where equality holds if and only if all sets $P_{\mathbf{i}} \in \mathcal{P}^{\mathbf{X}}(d)$ with $P_{\mathbf{i}} \neq \emptyset$ have equal measures.

A topological counterpart of the sorting entropy, the *topological sorting entropy of order d* is defined by:

$$h_{0,\triangle}^{\mathbf{X}}(T,d) = \ln\bigl(\mathrm{Allow}_{\mu}^{X}(T,d+1)\bigr) - \ln\bigl(\mathrm{Allow}_{\mu}^{X}(T,d)\bigr) = \ln\frac{\mathrm{Allow}_{\mu}^{X}(T,d+1)}{\mathrm{Allow}_{\mu}^{X}(T,d)}.$$

It is not as useful as the topological permutation entropy; in particular there is no general relationship between sorting and topological sorting entropy. However we mention the following result, which we prove in Subsection 2.4.3.

Proposition 2.5. *Given $\left(A^{\mathbb{N}}, \mathbb{B}_{\Pi}(A^{\mathbb{N}}), m_B, \sigma\right)$ a Bernoulli shift over $A = \{0,1,\ldots,l\}$. Then for a lexicographic-like X it holds*

$$\lim_{d\to\infty} h_{0,\triangle}^{X}(\sigma,d) = \lim_{d\to\infty} h_0^{X}(\sigma,d) = \ln(l+1).$$

Remark. Note that the equality $\lim\limits_{d\to\infty} h_0^{X}(\sigma,d) = \ln(l+1)$ for a Bernoulli shift over $A = \{0,1,\ldots,l\}$ follows from the more general result of C. Bandt, G. Keller and B. Pompe [BKP02, Theorem 1]. However, our proof provided in Subsection 2.4.3 is different from the proof of the general result and therefore may be interesting.

Note that the practical estimation of the topological permutation and sorting entropies from finite orbit is complicated since elements of the ordinal partition of small but finite measure are visited rather rare (see [Ami10, Section 7.7] for possible solutions).

2.3.5 Order isomorphisms

In this subsection we discuss equivalence of dynamical systems, which allows to extend results obtained for a system to all equivalent systems.

Definition 2.19. The system $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ is said to be *isomorphic* to the system $\big(\Upsilon, \mathbb{B}(\Upsilon), \nu, S\big)$, if there are $\Omega_0 \in \mathbb{B}(\Omega)$ and $\Upsilon_0 \in \mathbb{B}(\Upsilon)$ with $\mu(\Omega_0) = 1$, $\nu(\Upsilon_0) = 1$, $T(\Omega_0) \subseteq \Omega_0$, $S(\Upsilon_0) \subseteq \Upsilon_0$, and an invertible measure-preserving map $\phi : \Omega_0 \to \Upsilon_0$ such that $\phi \circ T(\omega) = S \circ \phi(\omega)$ for all $\omega \in \Omega_0$. Then the map ϕ is called an *isomorphism.*

Moreover, isomorphic dynamical systems $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ and $\big(\Upsilon, \mathbb{B}(\Upsilon), \nu, S\big)$ are said to be *order isomorphic* with respect to the observables $\mathbf{X} : \Omega \to \mathbb{R}^N$ and $\mathbf{Y} : \Upsilon \to \mathbb{R}^N$, if for all $\omega_1, \omega_2 \in \Omega'$ and $j = 1, 2, \ldots, N$ it holds

$$X_j(\omega_1) \leq X_j(\omega_2) \Leftrightarrow Y_j \circ \phi(\omega_1) \leq Y_j \circ \phi(\omega_2).$$

Our definition of order isomorphism is equivalent to the original one in [Ami10], but we use different notation. Note that the KS entropy is an invariant of isomorphism, i.e. isomorphic dynamical systems have the same KS entropy [ELW11] (this accounts for the special significance of the KS entropy). Analogously, ordinal patterns and, in particular, the entropy of the ordinal partition $H\big(\mathcal{P}^{\mathbf{X}}(d)\big)$, the permutation and the sorting entropies, as well as their topological counterparts, are invariants of order isomorphism. In particular, this fact allows to extend the result of Bandt, Keller and Pompe [BKP02] linking permutation entropy to the KS entropy, to the class of Markov shifts over a finite alphabet.

Example 2.9. The golden mean dynamical system $\Big([0,1], \mathbb{B}\big([0,1]\big), \mu_{gm}, T_{gm}\Big)$ and the golden mean shift $\Big(\{0,1\}^{\mathbb{N}}, \mathbb{B}_{\Pi}\big(\{0,1\}^{\mathbb{N}}\big), m_{gm}, \sigma\Big)$, (see Examples 2.1 and 2.2) are order isomorphic with respect to the observables id and X given by

$$X\big((s_0, s_1, \ldots)\big) = \sum_{j=0}^{\infty} \left(\frac{1}{2}\right)^{j+1} s_j.$$

Indeed, consider a map $\phi_{\mathcal{M}_{gm}} : [0,1) \to \{0,1\}^{\mathbb{N}}$ providing coding via the partition $\mathcal{M}_{gm}$ defined by (2.4):

$$\phi_{\mathcal{M}_{gm}}(\omega) = (s_0, s_1, \ldots) \text{ with } T^i_{gm}(\omega) \in M_{s_i}.$$

One can rewrite $\phi_{\mathcal{M}_{gm}}$ in an explicit form:

$$\phi_{\mathcal{M}_{gm}}(\omega) = (s_0, s_1, \ldots), \text{ where } s_i = \varphi T^i_{gm}(\omega) - T^{(i+1)}_{gm}(\omega) \text{ for all } i \in \mathbb{N}_0.$$

The map $\phi_{\mathcal{M}_{gm}}$ is invertible, the inverse map is given by

$$\phi^{-1}_{\mathcal{M}_{gm}}((s_0, s_1, \ldots)) = \sum_{j=0}^{\infty} \left(\frac{1}{\varphi}\right)^{j+1} s_j.$$

It is easy to see that $\phi_{\mathcal{M}_{gm}}$ is not only an isomorphism, but also an order isomorphism.

2.4 Proofs

2.4.1 Proof of Theorem 2.1

Proof. Since the partition $\mathcal{G} = \{G_0, G_1, \ldots, G_l\}$ is generating and by (2.8), it holds

$$h_\mu(T) = h_\mu(T, \mathcal{G}) = \lim_{n\to\infty} \big(H(\mathcal{G}_{n+1}) - H(\mathcal{G}_n)\big).$$

It remains to show that $\big(H(\mathcal{G}_{n+1}) - H(\mathcal{G}_n)\big) = \big(H(\mathcal{G}_2) - H(\mathcal{G})\big)$ for all $n \in \mathbb{N}$.

Let $g(a_0, a_1, \ldots, a_n) = \mu\big(G_{a_0} \cap T^{-1}(G_{a_1}) \cap \ldots \cap T^{-n}(G_{a_n})\big)$ to simplify the notation for any $a_0, a_1, \ldots, a_{n-1} \in \{0, 1, \ldots, l\}$ and $n \in \mathbb{N}$. Note that for all $a \in \{0, 1, \ldots, l\}$ it holds $g(a) = \mu(G_a) > 0$. Since $\mathcal{G}$ has the Markov property, it follows that

$$\begin{aligned}
H(\mathcal{G}_{n+1}) &= \sum_{a_0,a_1,\ldots,a_n} g(a_0, a_1, \ldots, a_n) \ln g(a_0, a_1, \ldots, a_n) \\
&= \sum_{a_0,a_1,\ldots,a_n} g(a_0, a_1, \ldots, a_{n-1}) \frac{g(a_{n-1}, a_n)}{g(a_{n-1})} \ln\bigg(g(a_0, a_1, \ldots, a_{n-1}) \frac{g(a_{n-1}, a_n)}{g(a_{n-1})}\bigg) \\
&= \sum_{a_0,a_1,\ldots,a_{n-1}} g(a_0, a_1, \ldots, a_{n-1}) \ln \big(g(a_0, a_1, \ldots, a_{n-1})\big) \sum_{a_n} \frac{g(a_{n-1}, a_n)}{g(a_{n-1})} \\
&+ \sum_{a_{n-1},a_n} \frac{g(a_{n-1}, a_n)}{g(a_{n-1})} \ln\bigg(\frac{g(a_{n-1}, a_n)}{g(a_{n-1})}\bigg) \sum_{a_0,a_1,\ldots,a_{n-2}} g(a_0, a_1, \ldots, a_{n-1}). \quad (2.25)
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{a_n} \frac{g(a_{n-1}, a_n)}{g(a_{n-1})} &= \frac{1}{\mu(G_{a_{n-1}})} \sum_{a_n} \mu\big(G_{a_{n-1}} \cap T^{-1}(G_{a_n})\big) = 1, \\
\sum_{a_0,a_1,\ldots,a_{n-1}} g(a_0, a_1, \ldots, a_{n-1}) &\ln \big(g(a_0, a_1, \ldots, a_{n-1})\big) = H(\mathcal{G}_n),
\end{aligned}$$

and

$$\begin{aligned}
\sum_{a_0,a_1,\ldots,a_{n-2}} g(a_0, a_1, \ldots, a_{n-1}) &= \sum_{a_0,a_1,\ldots,a_{n-2}} \mu\Big(G_{a_0} \cap T^{-1}(G_{a_1}) \cap \ldots \cap T^{-(n-1)}(G_{a_{n-1}})\Big) \\
&= \mu(G_{a_{n-1}}),
\end{aligned}$$

we can rewrite (2.25) as

$$H(\mathcal{G}_{n+1}) = H(\mathcal{G}_n) + \sum_{a_{n-1},a_n} \mu\big(G_{a_{n-1}} \cap T^{-1}(G_{a_n})\big) \ln \frac{\mu\big(G_{a_{n-1}} \cap T^{-1}(G_{a_n})\big)}{\mu(G_{a_{n-1}})}.$$

Hence for all $n \in \mathbb{N}$ it holds

$$H(\mathcal{G}_{n+1}) - H(\mathcal{G}_n) = \sum_{i=0}^{l} \sum_{j=0}^{l} \mu\big(G_i \cap T^{-1}(G_j)\big) \ln \frac{\mu\big(G_i \cap T^{-1}(G_j)\big)}{\mu(G_i)} = H(\mathcal{G}_2) - H(\mathcal{G}),$$

and we are done. □

2.4.2 Proof of Proposition 2.4

Here we prove the formula for the number of allowed ordinal patterns for Bernoulli shifts and we start from an auxiliary statement.

Lemma 2.6. *For $\psi_n(j)$ defined by (2.23) for $n, j \in \mathbb{N}$ with $n > 1$, it holds*

$$\lim_{d\to\infty} \frac{1}{d+1} \sum_{j=1}^{d} \frac{\psi_n(j)}{n^j} = 1.$$

Proof. Recall that the Möbius function is defined for $i \in \mathbb{N}$ as follows [NB99, Appendix 2a]

$$\text{Möb}(i) = \begin{cases} 1, & i = 1, \\ 0, & i \text{ is not } \textit{square-free} \text{ (is divisible by a square of an integer)}, \\ 1, & i \text{ is square-free and has an even number of prime factors}, \\ -1, & i \text{ is square-free and has an odd number of prime factors}. \end{cases}$$

It is easily seen that for all $n, j \in \mathbb{N}$ with $n > 1$ it holds

$$n^j \geq \psi_n(j) = \sum_{r|j} \text{Möb}\left(\frac{j}{r}\right) n^r \geq n^j - \sum_{r|j, r<j} (-1) n^r \geq n^j - \sum_{r=1}^{\lfloor j/2 \rfloor} n^r.$$

This implies that

$$1 \geq \frac{\psi_n(j)}{n^j} \geq 1 - \sum_{r=\lfloor (j+1)/2 \rfloor}^{j-1} \frac{1}{n^r},$$

and hence for all $d \in \mathbb{N}$ it holds

$$\frac{d}{d+1} \geq \frac{1}{d+1} \sum_{j=1}^{d} \frac{\psi_n(j)}{n^j} \geq \frac{d}{d+1} - \frac{1}{d+1} \sum_{j=1}^{d} \sum_{r=\lfloor (j+1)/2 \rfloor}^{j-1} \frac{1}{n^r}.$$

Finally, from the obvious inequality

$$\sum_{j=1}^{d} \sum_{r=\lfloor (j+1)/2 \rfloor}^{j-1} \frac{1}{n^r} \leq \sum_{j=1}^{d} \frac{j}{n^j}$$

and from the fact that for all $n > 1$ it holds

$$\lim_{d\to\infty} \frac{1}{d+1} \sum_{j=1}^{d} \frac{j}{n^j} = 0$$

we conclude that

$$1 = \lim_{d\to\infty} \frac{d}{d+1} \geq \lim_{d\to\infty} \frac{1}{d+1} \sum_{j=1}^{d} \frac{\psi_n(j)}{n^j} \geq \lim_{d\to\infty} \frac{d}{d+1} + 0 = 1.$$

This completes the proof. □

Now we come to the proof of Proposition 2.4.

Proof. Let us first prove that representation (2.24) of the number of allowed patterns holds. Under assumptions of the present proposition holds equality (2.22):

$$\mathrm{Allow}_{m_B}^X(\sigma, d) = \sum_{k=2}^{l+1}\sum_{i=0}^{k-2}(-1)^i\binom{d+1}{i}\left((k-i-2)(k-i)^{d-1} + \sum_{j=1}^{d}(k-i)^{d-j}\psi_{k-i}(j)\right).$$

For fixed d, given $b_i = (-1)^i\binom{d+1}{i}$ and $c_n = \left((n-2)n^{d-1} + \sum_{j=1}^{d} n^{d-j}\psi_n(j)\right)$, we have

$$\begin{aligned}
\mathrm{Allow}_{m_B}^X(\sigma, d) &= \sum_{k=2}^{l+1}\sum_{i=0}^{k-2} b_i c_{k-i} \\
&= (b_0c_2) + (b_0c_3 + b_1c_2) + \ldots + (b_0c_{l+1} + b_1c_l + \ldots + b_{l-1}c_2) \\
&= c_2(b_0 + b_1 + \ldots + b_{l-1}) + c_3(b_0 + b_1 + \ldots + b_{l-2}) + \ldots + c_{l+1}b_0 \\
&= \sum_{k=2}^{l+1} c_k \sum_{i=0}^{l+1-k} b_i \\
&= \sum_{k=2}^{l+1}\left((k-2)k^{d-1} + \sum_{j=1}^{d} k^{d-j}\psi_k(j)\right)\sum_{i=0}^{l+1-k}(-1)^i\binom{d+1}{i}.
\end{aligned}$$

One can easily show by induction that for all $n \in \mathbb{N}$ it holds

$$\sum_{i=0}^{n}(-1)^i\binom{d+1}{i} = (-1)^n\binom{d}{n}.$$

Therefore we obtain

$$\mathrm{Allow}_{m_B}^X(\sigma, d) = \sum_{k=2}^{l+1}(-1)^{l+1-k}\binom{d}{l+1-k}\left((k-2)k^{d-1} + \sum_{j=1}^{d} k^{d-j}\psi_k(j)\right).$$

This finishes the first part of the proof. It remains to show that for all $l \in \mathbb{N}$ it holds

$$\lim_{d\to\infty}\frac{\mathrm{Allow}_{m_B}^X(\sigma, d)}{(d+1)(l+1)^d} = 1.$$

It follows immediately that

$$\begin{aligned}
\frac{\mathrm{Allow}_{m_B}^X(\sigma, d)}{(d+1)(l+1)^d} &= \sum_{k=2}^{l+1}(-1)^{l+1-k}\binom{d}{l+1-k}\frac{(k-2)k^{d-1}}{(d+1)(l+1)^d} \\
&\quad + \sum_{k=2}^{l+1}(-1)^{l+1-k}\binom{d}{l+1-k}\sum_{j=1}^{d}\frac{k^{d-j}\psi_k(j)}{(d+1)(l+1)^d}.
\end{aligned}$$

Now observe that the following limits exists.

1. For all $k, l \in \mathbb{N}$ with $k \le l$ it holds

$$\lim_{d\to\infty}(-1)^{l+1-k}\binom{d}{l+1-k}\frac{(k-2)k^{d-1}}{(d+1)(l+1)^d} = \lim_{d\to\infty}(-1)^{l+1-k}d^{l-k}\left(\frac{k}{l+1}\right)^d = 0.$$

2. For all $l \in \mathbb{N}$ with $k = l+1$ it holds
$$\lim_{d\to\infty} (-1)^{l+1-k} \binom{d}{l+1-k} \frac{(k-2)k^{d-1}}{(d+1)(l+1)^d} = \lim_{d\to\infty} \left(\frac{1}{d+1}\right) = 0.$$

3. As one can easily see for all $k, j \in \mathbb{N}$ it holds $|\psi_k(j)| \leq k^j$, therefore for all $k, l \in \mathbb{N}$ with $k \leq l$ it holds
$$\begin{aligned}\lim_{d\to\infty} \binom{d}{l+1-k} \sum_{j=1}^{d} \frac{k^{d-j}\psi_k(j)}{(d+1)(l+1)^d} &\leq \lim_{d\to\infty} \frac{d^{l+1-k}}{(d+1)(l+1)^d} \sum_{j=1}^{d} k^d \\ &= \lim_{d\to\infty} d^{l+1-k} \left(\frac{k}{l+1}\right)^d = 0.\end{aligned}$$

4. For all $l \in \mathbb{N}$ and $k = l+1$ by applying Lemma 2.6 we obtain
$$\lim_{d\to\infty} (-1)^{l+1-k} \binom{d}{l+1-k} \sum_{j=1}^{d} \frac{k^{d-j}\psi_k(j)}{(d+1)(l+1)^d} = \lim_{d\to\infty} \frac{1}{d+1} \sum_{j=1}^{d} \frac{\psi_{l+1}(j)}{(l+1)^j} = 1.$$

Summarizing these four statements we obtain
$$\begin{aligned}1 &= \sum_{k=2}^{l+1} \lim_{d\to\infty} (-1)^{l+1-k} \binom{d}{l+1-k} \frac{(k-2)k^{d-1}}{(d+1)(l+1)^d} \\ &\quad + \sum_{k=2}^{l+1} \lim_{d\to\infty} (-1)^{l+1-k} \binom{d}{l+1-k} \sum_{j=1}^{d} \frac{k^{d-j}\psi_k(j)}{(d+1)(l+1)^d} \\ &= \lim_{d\to\infty} \sum_{k=2}^{l+1} (-1)^{l+1-k} \binom{d}{l+1-k} \left(\frac{(k-2)k^{d-1}}{(d+1)(l+1)^d} + \sum_{j=1}^{d} \frac{k^{d-j}\psi_k(j)}{(d+1)(l+1)^d} \right) \\ &= \lim_{d\to\infty} \frac{\mathrm{Allow}_{m_B}^X(\sigma, d)}{(d+1)(l+1)^d},\end{aligned}$$
which finishes the proof. □

2.4.3 Proof of Proposition 2.5

Here we prove that for a lexicographic-like X the topological permutation entropy $\lim_{d\to\infty} h_0^X(\sigma, d)$ as well as the topological sorting entropy $\lim_{d\to\infty} h_{0,\triangle}^X(\sigma, d)$ of Bernoulli shifts over $(l+1)$ symbols are given by $\ln(l+1)$ for all $l \in \mathbb{N}$.

Proof. Let us first show the equality for the topological permutation entropy. It holds
$$\begin{aligned}\lim_{d\to\infty} h_0^X(\sigma, d) &= \lim_{d\to\infty} \frac{1}{d} \ln\big(\mathrm{Allow}_{m_B}^X(\sigma, d)\big) = \lim_{d\to\infty} \frac{1}{d} \ln\left(\frac{\mathrm{Allow}_{m_B}^X(\sigma, d)}{(d+1)(l+1)^d} (d+1)(l+1)^d \right) \\ &= \lim_{d\to\infty} \frac{1}{d} \left(\ln \frac{\mathrm{Allow}_{m_B}^X(\sigma, d)}{(d+1)(l+1)^d} + \ln(d+1) + d\ln(l+1) \right) \\ &= \lim_{d\to\infty} \frac{1}{d} \ln \frac{\mathrm{Allow}_{m_B}^X(\sigma, d)}{(d+1)(l+1)^d} + \lim_{d\to\infty} \frac{\ln(d+1)}{d} + \lim_{d\to\infty} \ln(l+1) = \ln(l+1).\end{aligned}$$

The latter equality is a consequence of Lemma 2.4: recall that for a lexicographic-like X it holds

$$\lim_{d\to\infty} \frac{\mathrm{Allow}^X_{m_B}(\sigma, d)}{(d+1)(l+1)^d} = 1. \tag{2.26}$$

It remains to show that for lexicographic-like X the topological sorting entropy is given by $\ln(l+1)$. Indeed, by the definition of topological sorting entropy and (2.26) it holds

$$\begin{aligned}
\lim_{d\to\infty} h^X_{0,\triangle}(\sigma, d) &= \lim_{d\to\infty} \ln \frac{\mathrm{Allow}^X_{m_B}(\sigma, d+1)}{\mathrm{Allow}^X_{m_B}(\sigma, d)} \\
&= \lim_{d\to\infty} \ln\left((l+1) \frac{(d+1)(l+1)^d}{(d+2)(l+1)^{d+1}} \frac{\mathrm{Allow}^X_{m_B}(\sigma, d+1)}{\mathrm{Allow}^X_{m_B}(\sigma, d)} \right) \\
&= \ln(l+1) + \lim_{d\to\infty} \left(\ln \frac{\mathrm{Allow}^X_{m_B}(\sigma, d+1)}{(d+2)(l+1)^{d+1}} - \ln \frac{\mathrm{Allow}^X_{m_B}(\sigma, d)}{(d+1)(l+1)^d} \right) \\
&= \ln(l+1).
\end{aligned}$$

□

Chapter 3

Conditional entropy of ordinal patterns

In this chapter we introduce the conditional entropy of ordinal patterns.

Definition 3.1. Let $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ be a measure-preserving dynamical system. For a real-valued observable $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ we define the *conditional entropy of ordinal patterns of order* $d \in \mathbb{N}$ as

$$h^{\mathbf{X}}_{\mu,\mathrm{cond}}(T, d) = H(\mathcal{P}^{\mathbf{X}}(d)_2) - H(\mathcal{P}^{\mathbf{X}}(d)). \tag{3.1}$$

Conditional entropy of ordinal patterns can be used as a measure of time series complexity. It has some nice properties and, as we demonstrate, it often provides a much better practical estimation of the KS entropy than the permutation entropy. For brevity, we refer to $h^{\mathbf{X}}_{\mu,\mathrm{cond}}(T, d)$ as the "conditional entropy of order d" when no confusion can arise.

This chapter is organized as follows. We start from clarifying the physical meaning of conditional entropy and provide an example motivating the discussion of the conditional entropy in Section 3.1. In Section 3.2 we consider the interrelation between the conditional entropy of ordinal patterns, the permutation entropy and the sorting entropy. Section 3.3 is devoted to the relationship between the conditional entropy of ordinal patterns and the KS entropy. In particular, we show that in some cases conditional entropy approaches the KS entropy faster than the permutation entropy as order d tends to infinity (Theorem 3.4); that the conditional entropy for finite d coincides with the KS entropy for systems with periodic dynamics (Theorem 3.6) and for Markov shifts over two symbols (Theorem 3.10). In Section 3.4 we discuss properties of the empirical conditional entropy, namely robustness with respect to noise and sensitivity to the size of the sample. In Section 3.5 we summarize the results and discuss some open questions. Finally, in Section 3.6 we provide those proofs that are mainly technical.

Some of results presented in this chapter were published in article [UK14].

3.1 Physical meaning of conditional entropy and a motivating example: measuring complexity of logistic maps

Conditional entropy of ordinal patterns represents the first element of the sequence

$$\Big(\big(H(\mathcal{P}^{\mathbf{X}}(d)_{n+1}) - H(\mathcal{P}^{\mathbf{X}}(d)_n)\big)\Big)_{n\in\mathbb{N}},$$

which provides the ordinal representation (2.19) of the KS entropy as both n and d tend to infinity. To see the physical meaning of the conditional entropy recall (see Subsection 2.3.2) that the entropies of the ordinal partitions $\mathcal{P}^{\mathbf{X}}(d)$ and $\mathcal{P}^{\mathbf{X}}(d)_2$ are given by

$$H(\mathcal{P}^{\mathbf{X}}(d)) = -\sum_{\boldsymbol{\pi}\in\Pi_d^N} \mu(P_{\boldsymbol{\pi}})\ln\mu(P_{\boldsymbol{\pi}}),$$

$$H(\mathcal{P}^{\mathbf{X}}(d)_2) = -\sum_{\boldsymbol{\pi}\in\Pi_d^N}\sum_{\boldsymbol{\xi}\in\Pi_d^N} \mu\big(P_{\boldsymbol{\pi}}\cap T^{-1}(P_{\boldsymbol{\xi}})\big)\ln\mu\big(P_{\boldsymbol{\pi}}\cap T^{-1}(P_{\boldsymbol{\xi}})\big),$$

where $\Pi_d^N = \{\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_N) \mid \pi_1,\pi_2,\ldots,\pi_N \in \Pi_d\}$ is the set of all N-dimensional ordinal patterns. Then we can rewrite the conditional entropy (3.1) as

$$h^{\mathbf{X}}_{\mu,\mathrm{cond}}(T,d) = -\sum_{\boldsymbol{\pi}\in\Pi_d^N}\sum_{\boldsymbol{\xi}\in\Pi_d^N} \mu\big(P_{\boldsymbol{\pi}}\cap T^{-1}(P_{\boldsymbol{\xi}})\big)\ln\frac{\mu\big(P_{\boldsymbol{\pi}}\cap T^{-1}(P_{\boldsymbol{\xi}})\big)}{\mu(P_{\boldsymbol{\pi}})} \tag{3.2}$$

(with $0/0 := 0$ and $0\ln 0 := 0$), cf. with representation (2.17) of the permutation entropy. Let ω be a point from $P_{\boldsymbol{\pi}}\cap T^{-1}(P_{\boldsymbol{\xi}})$ for some $P_{\boldsymbol{\pi}}, P_{\boldsymbol{\xi}} \in \mathcal{P}^{\mathbf{X}}(d)$ with $\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_N)$ and $\boldsymbol{\xi} = (\xi_1,\xi_2,\ldots,\xi_N)$. Then we say that in ω the ordinal pattern $\boldsymbol{\xi}$ is a *successor* of the ordinal pattern $\boldsymbol{\pi}$. The conditional entropy characterizes the diversity of successors of given ordinal patterns $\boldsymbol{\pi}$, whereas the permutation entropy characterizes the diversity of ordinal patterns $\boldsymbol{\pi}$ themselves (see Subsection 2.3.2).

Example 3.1. Consider the family of logistic maps $f_r : [0,1] \hookleftarrow$ defined by $f_r(\omega) = r\omega(1-\omega)$ for $r \in [3.5, 4]$ (we have already mentioned this family in Example 1.1). The complexity of the dynamical system $\big([0,1], \mathbb{B}\big([0,1]\big), f_r, \mu_r\big)$, where μ_r is the SRB measure, is measured by the KS entropy $h_{\mu_r}(f_r)$. For almost all $r \in [3.5, 4]$ by Pesin's formula it holds $h_{\mu_r}(f_r) = \max\{\mathrm{LE}(f_r), 0\}$ (see Subsection 2.1.4.3 and Remark after the example for details). Since the Lyapunov exponent for the logistic maps can be estimated rather accurately [Spr03], this family provides a convenient object for comparing practical measures of complexity with the KS entropy.

For the logistic maps the permutation entropy of order d converges to the KS entropy as d tends to infinity [BKP02]. However, Figure 3.1 shows that the permutation entropy of order $d = 9$ is relatively far from the Lyapunov exponent in comparison with the conditional entropy of the same order (values of both entropies are numerically estimated from orbits of length $L = 4 \cdot 10^6$ of a "random point" in $[0, 1]$). For other values of d the behavior of entropies is rather similar.

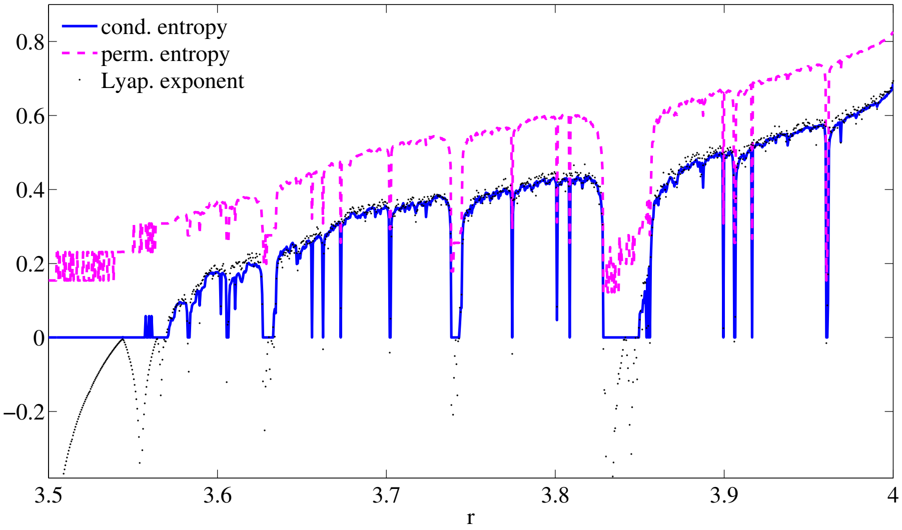


Figure 3.1: Empirical conditional entropy and permutation entropy in comparison with the Lyapunov exponent for logistic maps

Throughout this chapter we will return to the Example 3.1. Some general theoretical underpinnings for the fact that the conditional entropy estimates the KS entropy better than the permutation entropy are provided in Subsection 3.3.1.

Remark. For the family of logistic maps it has been shown that for almost all $r \in [1, 4]$ there exists the SRB measure μ_r, and the map f_r belongs to one of following types [Thu01, Lyu12] (see also [MN00, Lyu02] for original results):

(i) *Regular maps* providing relatively simple dynamics: Lebesgue-almost all orbits are attracted by a periodic cycle of Lebesgue measure zero. The invariant SRB measure for f_r is supported on this periodic cycle [MN00]. The Lyapunov exponent $\mathrm{LE}(f_r)$ is non-positive for regular maps and the KS entropy is equal to 0, thus it holds $h_{\mu_r}(f_r) = 0 \geq \mathrm{LE}(f_r)$.

(ii) *Stochastic maps* with more complex behavior. The invariant SRB measure for f_r is supported on a set of positive Lebesgue measure. In this case the Lyapunov exponent is positive, then by Theorem 2.2 the KS entropy coincides with the Lyapunov exponent: $h_{\mu_r}(f_r) = \mathrm{LE}(f_r) > 0$.

In the Example 3.1 we consider only $r \in [3.5, 4]$ since for almost all $r \in [1, 3.5]$ the map f_r is regular and represents rather simple behavior [Spr03], which is out of interest.

3.2 Interrelationship between conditional entropy of ordinal patterns, permutation and sorting entropy

In this section we consider the relationship between the conditional entropy of order d, the permutation entropy $h_{\mu}^{\mathbf{X}}(T,d)$ and the sorting entropy $h_{\mu,\triangle}^{\mathbf{X}}(T,d)$.

Lemma 3.1. *Let $(\Omega, \mathbb{B}(\Omega), \mu, T)$ be a measure-preserving dynamical system. Then for all $d \in \mathbb{N}$ it holds*

$$h_{\mu,cond}^{\mathbf{X}}(T,d) \leq h_{\mu,\triangle}^{\mathbf{X}}(T,d). \tag{3.3}$$

Moreover, if for some $d_0 \in \mathbb{N}$ it holds $h_{\mu}^{\mathbf{X}}(T,d_0+1) \leq h_{\mu}^{\mathbf{X}}(T,d_0)$, then we get

$$h_{\mu,cond}^{\mathbf{X}}(T,d_0) \leq h_{\mu,\triangle}^{\mathbf{X}}(T,d_0) \leq h_{\mu}^{\mathbf{X}}(T,d_0). \tag{3.4}$$

Proof. It can easily be shown (for details see [KS10]) that for all $d \in \mathbb{N}$ it holds

$$H(\mathcal{P}^{\mathbf{X}}(d)_2) \leq H(\mathcal{P}^{\mathbf{X}}(d+1)),$$

which implies

$$h_{\mu,\mathrm{cond}}^{\mathbf{X}}(T,d) = H(\mathcal{P}^{\mathbf{X}}(d)_2) - H(\mathcal{P}^{\mathbf{X}}(d)) \leq H(\mathcal{P}^{\mathbf{X}}(d+1)) - H(\mathcal{P}^{\mathbf{X}}(d)) = h_{\mu,\triangle}^{\mathbf{X}}(T,d),$$

and the proof of (3.3) is complete.

If $h_{\mu}^{\mathbf{X}}(T,d_0+1) \leq h_{\mu}^{\mathbf{X}}(T,d_0)$ for some $d_0 \in \mathbb{N}$ then we have

$$d_0 H(\mathcal{P}^{\mathbf{X}}(d_0+1)) \leq (d_0+1) H(\mathcal{P}^{\mathbf{X}}(d_0)).$$

Consequently, it holds

$$d_0\big(H(\mathcal{P}^{\mathbf{X}}(d_0+1)) - H(\mathcal{P}^{\mathbf{X}}(d_0))\big) \leq H(\mathcal{P}^{\mathbf{X}}(d_0)),$$

which establishes (3.4). □

By Lemma 3.1 we have that the conditional entropy under certain assumption is not greater than the permutation entropy and that in the general case the conditional entropy is not greater than the sorting entropy. Moreover, one can show that in the strong-mixing case the conditional entropy and the sorting entropy asymptotically approach each other. According to [UUK13], if Ω is an interval in $\mathbb{R}$ and T is strong-mixing then it holds

$$\lim_{d\to\infty}\Big(H(\mathcal{P}^{\mathrm{id}}(d+1)) - H(\mathcal{P}^{\mathrm{id}}(d)_2)\Big) = 0.$$

Together with Lemma 3.1 this implies the following statement.

Corollary 3.2. *Let $(\Omega, \mathbb{B}(\Omega), \mu, T)$ be a measure-preserving dynamical system, where Ω is an interval in $\mathbb{R}$ and T is strong-mixing. Then*

$$\lim_{d\to\infty}\Big(h_{\mu,\triangle}^{\mathrm{id}}(T,d) - h_{\mu,cond}^{\mathrm{id}}(T,d)\Big) = 0.$$

3.3 Conditional entropy of ordinal patterns and Kolmogorov-Sinai entropy

In this section we discuss the relationship between the conditional entropy of ordinal patterns and the KS entropy. We demonstrate that under certain assumptions the conditional entropy of ordinal patterns estimates the KS entropy better than the permutation entropy (Subsection 3.3.1). Besides, we prove that for some dynamical systems the conditional entropy of ordinal patterns for a finite order d coincides with the KS entropy (Subsections 3.3.2 and 3.3.3), while the permutation entropy only asymptotically approaches the KS entropy. Finally we summarize some results related to the properties of ordinal partitions for unimodal maps (Subsection 3.3.4).

3.3.1 Relationship in the general case

We start from a direct consequence of representation (2.19).

Proposition 3.3. *Let $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ be a measure-preserving dynamical system, $\mathbf{X}$ be a random vector on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ such that (2.19) is satisfied. Then the equality*

$$h_\mu(T) = \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,cond}(T,d)$$

holds if and only if for every $n \in \mathbb{N}$ it holds

$$\overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,cond}(T,d) = \overline{\lim_{d\to\infty}}\, H(\mathcal{P}^{\mathbf{X}}(d)_{n+1}) - H(\mathcal{P}^{\mathbf{X}}(d)_n).$$

Though being intriguing, this condition does not seem to provide a deep insight into the problem. Below we formulate another result linking conditional entropy and KS entropy, which appears to be more useful. Statements *(i)* and *(ii)* of the following theorem imply that under the given assumptions the conditional entropy of ordinal patterns bounds the KS entropy better than the sorting entropy and the permutation entropy, respectively.

Theorem 3.4. *Let $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ be a measure-preserving dynamical system, $\mathbf{X}$ be a random vector on $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ such that (2.19) is satisfied. Then it holds*

(i) $h_\mu(T) \leq \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,cond}(T,d) \leq \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,\triangle}(T,d)$.

(ii) *Moreover, if for some $d_0 \in \mathbb{N}$ it holds*

$$h^{\mathbf{X}}_{\mu}(T,d) \geq h^{\mathbf{X}}_{\mu}(T,d+1) \text{ for all } d \geq d_0, \tag{3.5}$$

or the limit of the sorting entropy

$$\lim_{d\to\infty} h^{\mathbf{X}}_{\mu,\triangle}(T,d) \text{ exists,} \tag{3.6}$$

then it holds:

$$h_\mu(T) \leq \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,cond}(T,d) \leq \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,\triangle}(T,d) \leq \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu}(T,d). \tag{3.7}$$

Proof. For any given partition $\mathcal{P}$, the difference $H(\mathcal{P}_{n+1}) - H(\mathcal{P}_n)$ decreases monotonically with increasing n [CT06, Section 4.2]. In particular, for the ordinal partition $\mathcal{P}^{\mathbf{X}}(d)$ it holds

$$\begin{aligned} h_\mu(T, \mathcal{P}^{\mathbf{X}}(d)) = \lim_{n\to\infty} \big(H(\mathcal{P}^{\mathbf{X}}(d)_{n+1}) - H(\mathcal{P}^{\mathbf{X}}(d)_n)\big) &\le H(\mathcal{P}^{\mathbf{X}}(d)_2) - H(\mathcal{P}^{\mathbf{X}}(d)) \\ &= h^{\mathbf{X}}_{\mu,\mathrm{cond}}(T,d) \end{aligned}$$

and consequently

$$\lim_{d\to\infty} h_\mu(T, \mathcal{P}^{\mathbf{X}}(d)) \le \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,\mathrm{cond}}(T,d).$$

The last inequality together with (3.3) implies *(i)*.

Statement *(ii)* will be proved once we prove the inequality below:

$$\overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,\triangle}(T,d) \le \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu}(T,d). \tag{3.8}$$

If (3.5) is satisfied, we get (3.8) immediately from Lemma 3.1. If (3.6) is satisfied, by Cesaro's mean theorem [CT06, Theorem 4.2.3] it follows that

$$\lim_{d\to\infty} \frac{1}{d} H\big(\mathcal{P}^{\mathbf{X}}(d)\big) = \lim_{d\to\infty} \Big(H\big(\mathcal{P}^{\mathbf{X}}(d+1)\big) - H\big(\mathcal{P}^{\mathbf{X}}(d)\big)\Big),$$

which is a particular case of (3.8), and we are done. □

Note that both statements of Theorem 3.4 remain correct if one replaces the upper limits by the lower limits.

As a consequence of Theorem 3.4 we get the following result.

Corollary 3.5. *If assumption* (2.19) *and either of assumptions* (3.5) *or* (3.6) *are satisfied, then* $h_\mu(T) = \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu}(T,d)$ *yields*

$$h_\mu(T) = \overline{\lim_{d\to\infty}}\, h^{\mathbf{X}}_{\mu,cond}(T,d).$$

This sheds some light on the behavior of the conditional entropy of the logistic maps, described in Section 3.1 (recall that due to [BKP02] the permutation entropy of the logistic maps converges to the KS entropy for d tending to infinity). Nevertheless, it is not clear whether the statements (3.5) or (3.6) hold, neither in the general case nor for the logistic maps. Note that a sufficient condition for (3.6) is the monotone decrease of the sorting entropy $h^{\mathbf{X}}_{\mu,\triangle}(T,d)$ with increasing d. However, the sorting entropy and the permutation entropy do not necessarily decrease for all d, which is illustrated by the following example.

Example 3.2. Consider the golden mean dynamical system $\Big([0,1], \mathbb{B}\big([0,1]\big), \mu_{gm}, T_{gm}\Big)$. The values of permutation, sorting and conditional entropies for this dynamical system[1]

[1] As we have seen in Example 2.7, p. 35, for the golden mean dynamical system the ordinal partition $\mathcal{P}^{\mathrm{id}}(d)$ has a quite simple structure. This allows to compute permutation, sorting and conditional entropies theoretically, in contrast to Example 3.1, where we have to estimate them.

are shown in Figure 3.2. Note that neither sorting nor permutation entropy is monotonically decreasing with increasing d. (The interesting fact that for all $d = 1, 2, \ldots, 9$ the conditional entropy coincides with the KS entropy is explained in Subsection 3.3.3.)

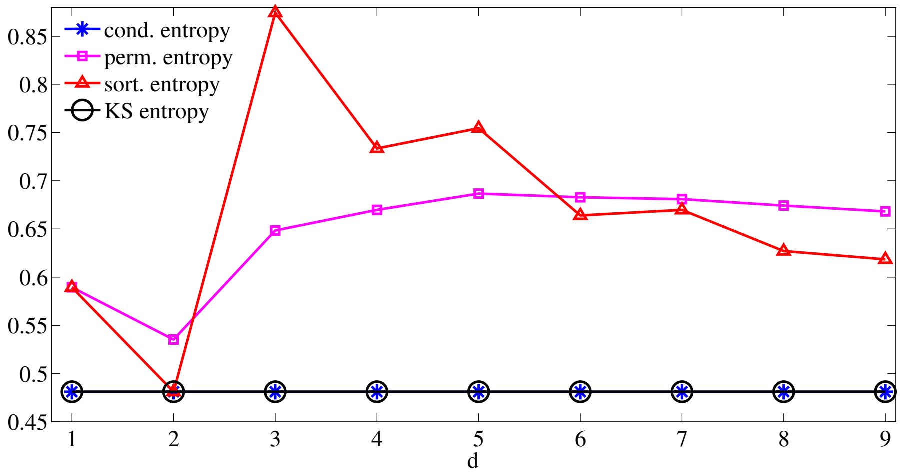


Figure 3.2: Conditional entropy, permutation entropy and sorting entropy in comparison with the KS entropy of the golden mean map

The questions when $h^{\mathbf{X}}_{\mu,\triangle}(T,d)$ or $h^{\mathbf{X}}_{\mu}(T,d)$ decrease starting from some $d_0 \in \mathbb{N}$ is still open. For instance, for the logistic map with $r = 4$ estimated values of permutation entropy and sorting entropy decrease starting from $d = 7$ and $d = 4$, respectively (see Figure 3.3). However, at this point we do not have theoretical results in this direction, see Subsection 3.3.4 for some discussion.

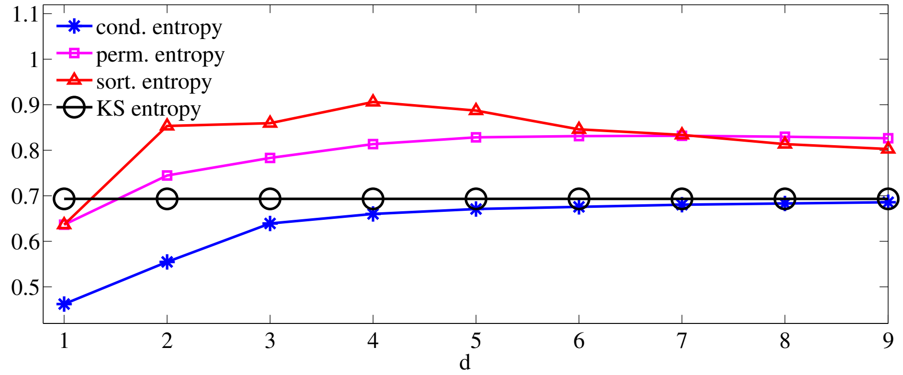


Figure 3.3: Empirical conditional entropy, permutation entropy and sorting entropy in comparison with the KS entropy of the logistic map with $r = 4$

3.3.2 Periodic case

Here we relate the conditional entropy to the KS entropy in the case of periodic dynamics. By a periodic dynamical system we mean a system such that the set of periodic points has measure 1. Though it is well-known that the KS entropy of a periodic dynamical system is equal to zero [KAH+06], the permutation entropy of order d can be arbitrarily large in this case (see [UK14, Section 3.5] for a discussion and examples) and thus does not provide a reliable estimate for the KS entropy. In Theorem 3.6 we show that the conditional entropy of a periodic dynamical system is equal to the KS entropy starting from some finite order d, which advantages the conditional entropy over the permutation entropy.

Theorem 3.6. *Let $\big(\Omega, \mathbb{B}(\Omega), \mu, T\big)$ be a measure-preserving dynamical system. Suppose that the set of periodic points of Ω with period not exceeding $k \in \mathbb{N}$ has measure 1, then for all $d \in \mathbb{N}$ with $d \geq k$ it holds*

$$h^{\mathbf{X}}_{\mu,cond}(T,d) = 0 = h_\mu(T). \tag{3.9}$$

Proof. One has to prove only the equality on the left-hand side of (3.9); for this it is sufficient to show that $h^{X_j}_{\mu,\mathrm{cond}}(T,d) = 0$ for every component X_j of the random vector $\mathbf{X}$, so consider $X = X_j$ for $j \in \{1, 2, \ldots, N\}$. By assumption there exists a set $\Omega_0 \subset \mathbb{B}(\Omega)$ such that $\mu(\Omega_0) = 1$ and for all $\omega \in \Omega_0$ it holds $T^l(\omega) = \omega$ for some $l \in \{1, 2, \ldots, k\}$. Let us fix an order $d \geq k$ and take ordinal patterns $\pi, \xi \in \Pi_d$ such that $\mu(P_\pi \cap T^{-1}(P_\xi)) > 0$. We show now that it holds

$$\mu(P_\pi \cap T^{-1}(P_\xi)) = \mu(P_\pi),$$

which together with (3.2) provides (3.9). Consider some $\omega_1 \in \Omega_0 \cap P_\pi \cap T^{-1}(P_\xi)$, which is periodic with a (minimal) period $l \in \{1, 2, \ldots, k\}$, that is $X(\omega_1) = X(T^l(\omega_1))$. By Definition 2.12 of an ordinal pattern, $\pi = (\ldots, d, d-l, \ldots)$. All points $\omega \in \Omega_0 \cap P_\pi$ have the same period; indeed, for any point with period $l_2 \leq k$ such that $l_2 \neq l$, the ordinal pattern is $(\ldots, d, d-l_2, \ldots) \neq \pi$. Therefore, for all $\omega \in \Omega_0 \cap P_\pi$ it holds $X(T^{(d+1)}(\omega)) = X(T^{(d+1-l)}(\omega))$ and the ordinal pattern for $T(\omega)$ is obtained from the ordinal pattern π in a well-defined way [KSE07]: by deleting the entry d, adding 1 to all remaining entries and inserting the entry 0 to the left of the entry l. Since $T(\omega_1) \in P_\xi$, for every other $\omega \in \Omega_0 \cap P_\pi$ it also holds $T(\omega) \in P_\xi$. Hence for all $\pi, \xi \in \Pi_d$ with $\mu(P_\pi \cap T^{-1}(P_\xi)) > 0$ it holds

$$\mu(P_\pi \cap T^{-1}(P_\xi)) = \mu(\Omega_0 \cap P_\pi \cap T^{-1}(P_\xi)) = \mu(\Omega_0 \cap P_\pi) = \mu(P_\pi),$$

which yields (3.9) and we are done. □

Example 3.3. In order to illustrate the behavior of permutation and conditional entropies of periodic dynamical systems, consider the rotation maps $g_\alpha(\omega) = (\omega + \alpha) \mod 1$ on the interval $[0,1]$ with the Lebesgue measure λ. Let α be rational, then $g_\alpha(\omega)$ provides a periodic behavior and it holds $h^{\mathrm{id}}_\lambda(g_\alpha) = 0$. Figure 3.4 illustrates conditional and permutation entropy of the rotation maps for $d = 4$ and $d = 8$ for α varying with step 0.001. For both values of d the conditional entropy is more close to zero than the permutation entropy since periodic orbits provide various ordinal patterns, but most of these patterns have only one successor. Note that for those values of α forcing periods shorter than d (for instance for $\alpha = 0.25$ all $\omega \in [0, 1]$ have period 4) it holds $h^{\text{id}}_{\lambda,\text{cond}}(g_\alpha, d) = 0 = h_\lambda(g_\alpha)$ as provided by Theorem 3.6.

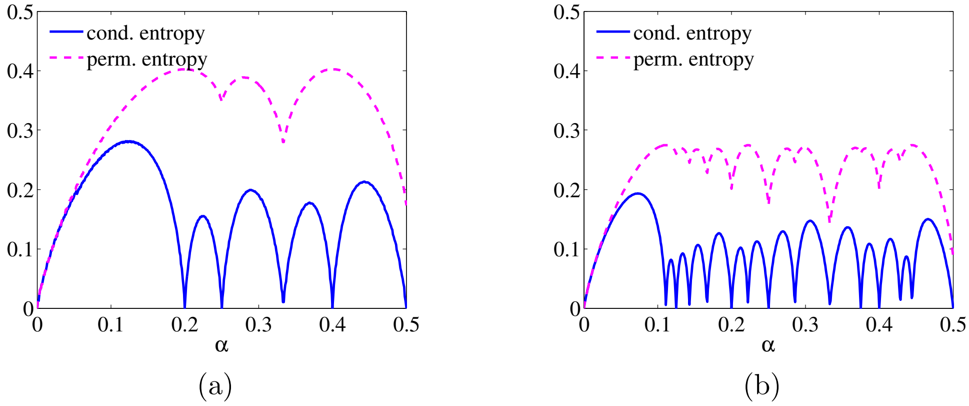


Figure 3.4: Conditional and permutation entropy of rotation maps for $d = 4$ (a) and $d = 8$ (b)

Since for a logistic map f_r with regular behavior the f_r-invariant SRB measure μ_r is supported on a finite periodic cycle (see Section 3.1), we have immediately the following result, giving a partial explanation for the behavior of the conditional entropy in Figure 3.1.

Corollary 3.7. *For a measure-preserving dynamical system* $\Big([0, 1], \mathbb{B}\big([0, 1]\big), \mu_r, f_r\Big)$, *where* $f_r = r\omega(1 - \omega)$ *is a regular logistic map and* μ_r *is the* f_r*-invariant SRB measure, let* k *be length of the attractive periodic cycle supporting the measure* μ_r*. Then for all* $d \in \mathbb{N}$ *with* $d \geq k$ *it holds*

$$h^{\text{id}}_{\mu_r,cond}(f_r, d) = 0 = h_{\mu_r}(f_r).$$

3.3.3 Markov property of ordinal partition

Recall (see Subsection 2.1.4) that for the entropy rate of the partition $\mathcal{M}$ with the Markov property it holds

$$h_\mu(T, \mathcal{M}) = H(\mathcal{M}_2) - H(\mathcal{M}).$$

Moreover by Theorem 2.1 if a partition $\mathcal{M}$ is both generating and has the Markov property, then it holds

$$h_\mu(T) = H(\mathcal{M}_2) - H(\mathcal{M}).$$

From the last two statements follows a sufficient condition for the coincidence between the conditional entropy and the KS entropy.

Lemma 3.8. *Let* $(\Omega, \mathbb{B}(\Omega), \mu, T)$ *be a measure-preserving dynamical system,* $\mathbf{X}$ *be an* $\mathbb{R}$*-valued random vector on* $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$ *such that* (2.19) *is satisfied. Then the following two statements hold:*

(i) *If* $\mathcal{P}^{\mathbf{X}}(d)$ *has the Markov property for all* $d \geq d_0$*, then*

$$h_\mu(T) = \lim_{d \to \infty} h^{\mathbf{X}}_{\mu, cond}(T, d).$$

(ii) *If* $\mathcal{P}^{\mathbf{X}}(d)$ *is generating and has the Markov property for some* $d \in \mathbb{N}$ *then*

$$h_\mu(T) = h^{\mathbf{X}}_{\mu, cond}(T, d). \tag{3.10}$$

In general, it is complicated to verify whether ordinal partitions are generating or have the Markov property; however below this is done for Markov shifts over two symbols.

In Subsection 3.6.2 we prove the following statement.

Lemma 3.9. *Let* $\big(\{0,1\}^{\mathbb{N}}, \mathbb{B}_\Pi(\{0,1\}^{\mathbb{N}}), m, \sigma\big)$ *be an ergodic Markov shift over two symbols. If the random variable* X *is lexicographic-like (see Definition 2.15, p. 35), then the ordinal partition* $\mathcal{P}^X(d)$ *is generating and has the Markov property for all* $d \in \mathbb{N}$.

Note that for any alphabet A a lexicographic-like X is injective for almost all $s \in A^{\mathbb{N}}$, thus X provides the ordinal representation (2.19) of the KS entropy (see [Kel12]). Hence as a direct consequence of Lemmas 3.8 and 3.9 we obtain the following.

Theorem 3.10. *Under the assumptions of Lemma 3.9 for all* $d \in \mathbb{N}$ *it holds*

$$h_m(\sigma) = h^X_{m, cond}(\sigma, d).$$

Example 3.4. Figure 3.5 illustrates Theorem 3.10 for the Bernoulli shift over two symbols with $m_B(C_0) = 0.663$, $m_B(C_1) = 0.337$. For all $d = 1, 2, \ldots, 9$ the empirical conditional entropy computed from an orbit of length $L = 3.6 \cdot 10^6$ nearly coincides with the theoretical KS entropy $h_{m_B}(\sigma)$. Meanwhile, the empirical permutation entropy differs from the KS entropy significantly.

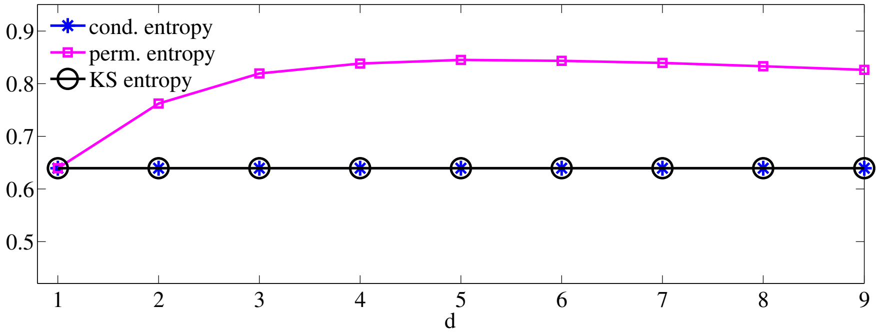


Figure 3.5: Empirical conditional entropy and permutation entropy in comparison with the KS entropy of the Bernoulli shift over two symbols

The result established in Theorem 3.10 naturally extends to the class of maps that are order-isomorphic to an ergodic Markov shift over two symbols (see Subsection 2.3.5 for details). For instance, the golden mean Markov shift and the golden mean dynamical system are order-isomorphic (see Example 2.9), which explains the coincidence of the conditional entropy and the KS entropy in Figure 3.2. Note that the logistic map for $r = 4$ is isomorphic, but not order-isomorphic to an ergodic Markov shift over two symbols [Ami10, Subsection 3.4.1]. Therefore in the case of the logistic map with $r = 4$ the conditional entropy for finite d does not coincide with the KS entropy (see Figure 3.3).

Remark. We consider in this thesis one-sided (i.e. non-invertible) Markov shifts, however using the same reasoning, one can show that Corollary 3.10 also holds for two-sided Markov shifts over two symbols (with the state space $\{0,1\}^{\mathbb{Z}}$ instead of $\{0,1\}^{\mathbb{N}}$). In particular, this provides that given $\Omega = [0,1] \times [0,1]$ with the two-dimensional Lebesgue measure λ_2 and the baker's map T_{baker} defined by

$$T_{baker}(\omega, \upsilon) = \begin{cases} (2\omega, \frac{\upsilon}{2}), & 0 \le \omega \le \frac{1}{2}, \\ (2\omega - 1, \frac{\upsilon+1}{2}), & \frac{1}{2} < \omega \le 1, \end{cases}$$

the KS entropy and the conditional entropy of $\big(\Omega, \mathbb{B}(\Omega), \lambda_2, T_{baker}\big)$ coincide.

Remark. Note that Theorem 3.10 can be reformulated in a statistical way: the order statistic from a Markov chain with two states forms a Markov chain. This generalizes the result for binary IID process obtained in [Nag82].

Theorem 3.10 cannot be extended to Markov shifts over a general alphabet since for them the conditions of Lemma 3.8 are not satisfied. (This is similar to the fact that the order statistic does not form a Markov chain for IID processes with more than two states [Nag82].)

Example 3.5. Figure 3.6 represents estimated values of the empirical conditional and permutation entropies of various orders d for the Bernoulli shifts over alphabets consisting of three and four symbols. Although these shifts have the same KS entropy as the shift in Figure 3.5, their conditional entropies differ significantly.

In order to explain the result in Example 3.5, we show in Proposition 3.11 that for the Bernoulli shifts over more than two symbols the ordinal partition $\mathcal{P}^X(d)$ does not refine the known generating partition $\mathcal{C} = \{C_0, C_1, \ldots, C_l\}$, consisting of all cylinders C_a. Moreover, for the shifts over more than two symbols, $\mathcal{P}^X(d)$ does not necessarily have the Markov property (Proposition 3.12).

Proposition 3.11. *Let $\big(A^{\mathbb{N}}, \mathbb{B}_{\Pi}(A^{\mathbb{N}}), m_B, \sigma\big)$ be a Bernoulli shift over $A = \{0, 1, \ldots, l\}$ with $l \ge 2$ and $m_B(C_a) > 0$ for all $a \in A$. Then the partition $\mathcal{P}^X(d)$ does not refine $\mathcal{C} = \{C_0, C_1, \ldots, C_l\}$.*

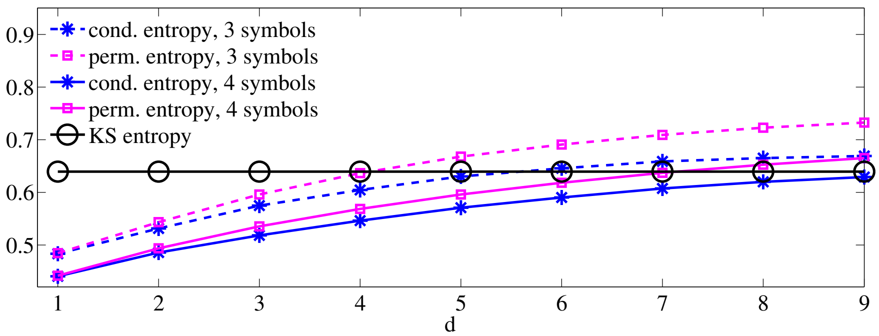


Figure 3.6: Empirical conditional entropy and permutation entropy in comparison with the KS entropy of Bernoulli shifts over three and four symbols

Proof. Consider the element of the ordinal partition corresponding to the "increasing" ordinal pattern:

$$P_{(0,1,\ldots,d)} = \left\{s \in A^{\mathbb{N}} \mid X(s) < X(\sigma s) < \ldots < X(\sigma^d s)\right\}.$$

For all $a_0, a_1, \ldots, a_{d+1} \in A$ with $a_0 \leq a_1 \leq a_2 \leq \ldots \leq a_d < a_{d+1}$ it holds

$$C_{a_0 a_1 \ldots a_{d+1}} \subset P_{(0,1,\ldots,d)}.$$

Since the Bernoulli measure of the set $C_{a_0 a_1 \ldots a_{d+1}}$ is strictly positive, for all $a_0 \in A \setminus \{l\}$ it holds

$$m_B(C_{a_0} \cap P_{(0,1,\ldots,d)}) \geq m_B(C_{a_0 a_1 \ldots a_{d+1}} \cap P_{(0,1,\ldots,d)}) > 0.$$

Therefore, for $l \geq 2$ for all $d \in \mathbb{N}$, the set $P_{(0,1,\ldots,d)} \in \mathcal{P}^X(d)$ is not a subset of any cylinder set. □

Proposition 3.12. *Given $\left(A^{\mathbb{N}}, \mathbb{B}_\Pi(A^{\mathbb{N}}), m_B, \sigma\right)$ a Bernoulli shift over $A = \{0, 1, \ldots, l\}$ with $l \in \mathbb{N}$ and $m_B(C_i) = \frac{1}{l+1}$ for all $i \in A$. Then the ordinal partition $\mathcal{P}^X(d)$ with $d \in \mathbb{N}$ and X being lexicographic-like, has the Markov property if and only if $l = 1$, that is for the two-symbol alphabet.*

The proof is provided in Subsection 3.6.2.

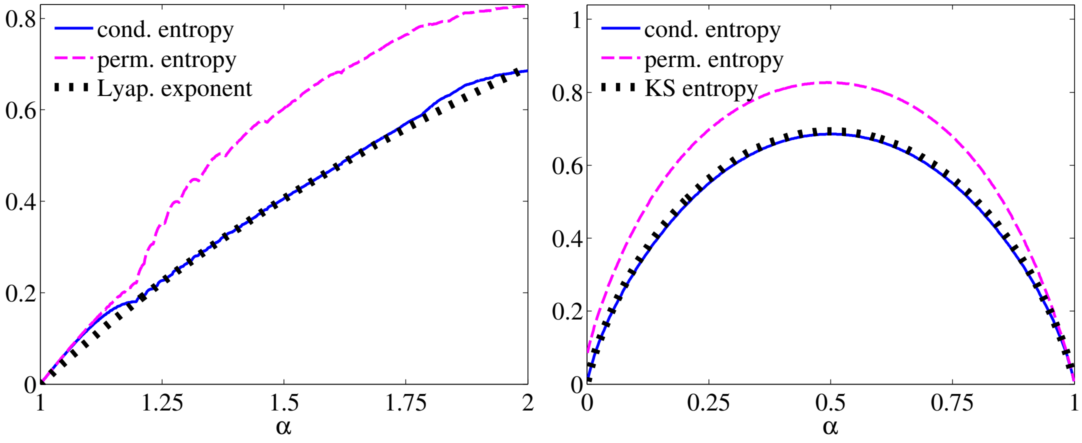
3.3.4 Ordinal partitions for unimodal maps

In this subsection we return to Example 3.1 and discuss the striking similarity between the conditional entropy of finite order d and the KS entropy in case of logistic maps. We would like to emphasize that this is not a single instance: consider the families of tent maps $\widehat{T}_\alpha$ and of skew tent maps $\grave{T}_\alpha$ defined on the interval $[0, 1]$ as

$$\widehat{T}_\alpha(\omega) = \begin{cases} \alpha\omega, & 0 \leq \omega \leq \frac{1}{2}, \\ \alpha(1-\omega), & \frac{1}{2} < \omega \leq 1, \end{cases} \quad \text{for } \alpha \in (1, 2],$$

$$\grave{T}_\alpha(\omega) = \begin{cases} \frac{\omega}{\alpha}, & 0 \leq \omega \leq \alpha, \\ \frac{1-\omega}{1-\alpha}, & \alpha < \omega \leq 1, \end{cases} \quad \text{for } \alpha \in (0, 1).$$

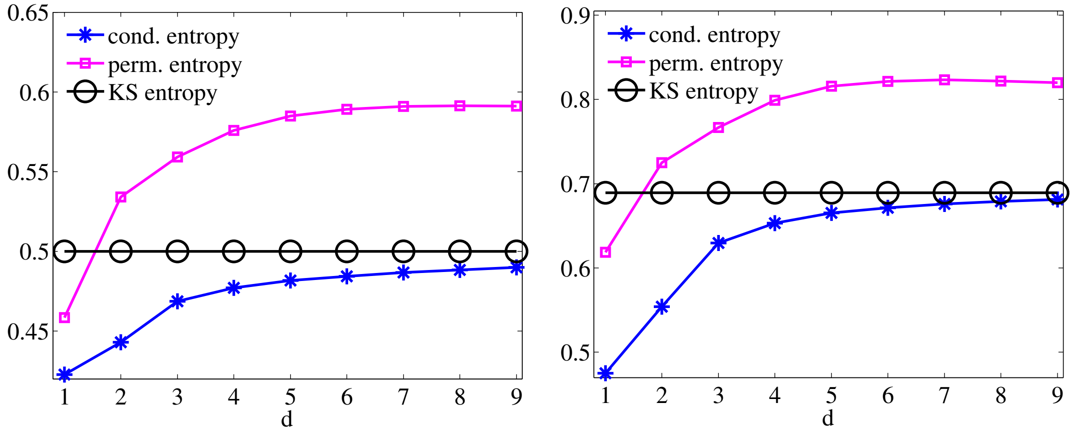
In these cases the estimated conditional entropy of ordinal patterns is also rather similar to the KS entropy (see Figure 3.7).



(a) (b)

Figure 3.7: Empirical conditional entropy and permutation entropy for $d = 9$ in comparison with the Lyapunov exponent of tent maps (a) and with the KS entropy of skew tent maps (b)

The families of logistic, tent and skew tent maps belong to a broader class of *unimodal maps*. They are defined on an interval I and map it into itself; they have a single critical point $c \in I$, are monotone increasing on the left of c and decreasing on the right of c. Figure 3.8 shows that the conditional entropy seemingly converges to the KS entropy with increasing order d for two more examples of unimodal maps.



(a) (b)

Figure 3.8: Empirical conditional entropy and permutation entropy for various order d in comparison with the KS entropy of the cusp map $T(\omega) = 1 - 2\sqrt{\omega}$ (a) and of the sine map $T(\omega) = \sin(\pi\omega)$ (b)

At this point we cannot provide a theoretical explanation for this behavior of the conditional entropy neither for logistic maps nor for the whole class of unimodal maps. By Lemma 3.8, the conditional entropy converges to the KS entropy if the ordinal partition $\mathcal{P}^{\mathrm{id}}(d)$ has the Markov property for all $d \geq d_0$ for some $d_0 \in \mathbb{N}$. Below we check whether this is the case for logistic maps and find out that $\mathcal{P}^{\mathrm{id}}(d)$ does not have the Markov property, at least for relatively small d.

Consider the dynamical system $\Big([0,1], \mathbb{B}\big([0,1]\big), \mu_r, f_r\Big)$, where $f_r(\omega) = r\omega(1-\omega)$ is a logistic map and μ_r is the f_r-invariant SRB measure. By Definition 2.5 if $\mathcal{P}^{\mathrm{id}}(d)$ has the Markov property then for $P_{i_0}, P_{i_1}, \ldots, P_{i_n} \in \mathcal{P}^{\mathrm{id}}(d)$ all sets in the form

$$P_{i_0 i_1 \ldots i_n} = P_{i_0} \cap f_r^{-1}(P_{i_1}) \cap \ldots \cap f_r^{-n}(P_{i_n})$$

satisfy the equality:

$$\mu_r(P_{i_0 i_1 \ldots i_n}) = \mu_r(P_{i_0}) \prod_{k=0}^{n-1} \frac{\mu_r(P_{i_k i_{k+1}})}{\mu_r(P_{i_k})}. \tag{3.11}$$

Let us say that a set $P_{i_0 i_1 \ldots i_n}$ is *Markov* if it satisfies (3.11). To check the Markov property for the partition $\mathcal{P}^{\mathrm{id}}(d)$, one has to verify whether all sets $P_{i_0 i_1 \ldots i_n}$ are Markov for all $n \in \mathbb{N}$. Since the analytical representation for the measure μ_r is in general unknown, we carry out the following numerical experiment to provide at least an approximate solution to this problem.

Experiment 3.1: For what $r \in [3.5, 4]$ all sets $P_{i_0 i_1 \ldots i_n}$ are Markov at least for small $n \in \mathbb{N}$?

Objects: orbits with length $L = 10^6$ of logistic maps f_r for randomly chosen points $\omega_0 \in [0,1]$, $r \in \{3.5000, 3.5005, \ldots, 4.0000\}$.

Technique. Recall (see Subsection 2.1.1) that the SRB measure of a set B can be estimated as a relative frequency of visiting B by orbits of Lebesgue-almost every ω_0:

$$\widetilde{\mu}_r(B, L) = \frac{1}{L} \#\big\{i \in \{0, 1, \ldots, L-1\} \mid f_r^i(\omega_0) \in B\big\},$$

with $\lim\limits_{L \to \infty} \widetilde{\mu}_r(B, L) = \mu_r(B)$. We deduce from (3.11) that for every Markov set $P_{i_0 i_1 \ldots i_n}$ and for every $\varepsilon > 0$ there exists $L \in \mathbb{N}$ such that it holds

$$\left| \widetilde{\mu}_r(P_{i_0 i_1 \ldots i_n}, L) - \widetilde{\mu}_r(P_{i_0}, L) \prod_{k=0}^{n-1} \frac{\widetilde{\mu}_r(P_{i_k i_{k+1}}, L)}{\widetilde{\mu}_r(P_{i_k}, L)} \right| < \varepsilon\, \widetilde{\mu}_r(P_{i_0 i_1 \ldots i_n}, L). \tag{3.12}$$

There is no method for choosing such L, so we can only take it rather large and say that if a set $P_{i_0 i_1 \ldots i_n}$ satisfies (3.12) then it is *nearly Markov*, and otherwise – that it is *nearly non-Markov (nnM)*. Existence of nnM sets indicates that the ordinal partition $\mathcal{P}^{\mathrm{id}}(d)$ does not have the Markov property for f_r.

Here we check whether equality (3.12) holds for all $P_{i_0 i_1 \ldots i_n} \in \mathcal{P}^{\mathrm{id}}(d)_n$ for given r, d and n. To this end we estimate measures in (3.12) for an orbit of a randomly chosen initial point, and calculate the total measure $S(r, d, n)$ of nnM sets for $\varepsilon = 0.01$:

$$S(r, d, n) = \sum_{\text{nnM } P_{i_0 i_1 \ldots i_n} \in \mathcal{P}^{\mathrm{id}}(d)_n} \widetilde{\mu}_r(P_{i_0 i_1 \ldots i_n}, L).$$

If $S(r,d,n) > 0$ then the partition $\mathcal{P}^{\mathrm{id}}(d)$ most likely does not have the Markov property for f_r. Note that $S(r,d,n) = 0$ for a fixed n does not imply that $\mathcal{P}^{\mathrm{id}}(d)$ has the Markov property since one needs to check all $n \geq 2$.

Results for various r, $n = 3$ and $d = 1, 2$ are presented in Figure 3.9 (we do not consider larger n and d due to the high computational cost). Additionally we present the results for $r = 3.58$, 4, $n = 2$ and $d = 1, 2, \ldots, 6$ in Figure 3.10.

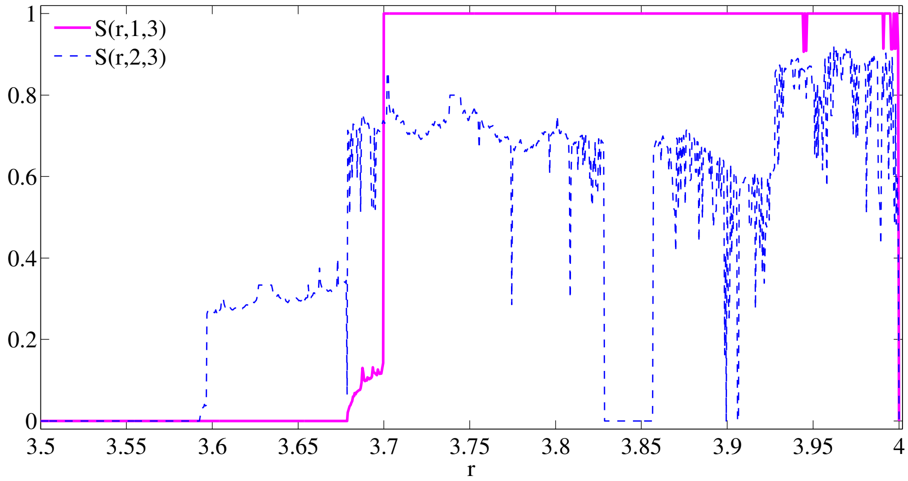


Figure 3.9: Total measure $S(r,d,n)$ of nnM sets for $n = 3$, $d = 1$, 2

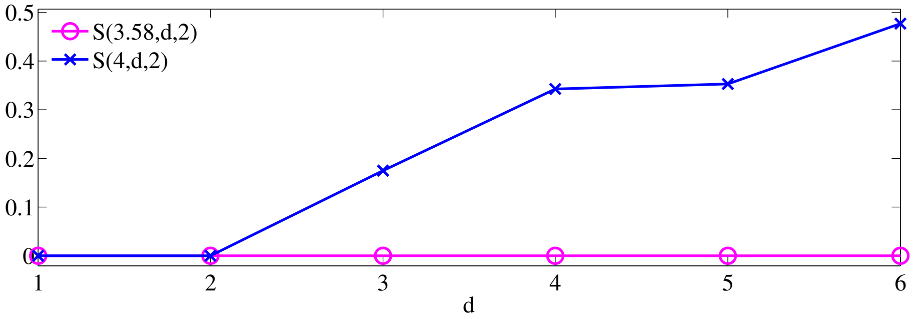


Figure 3.10: Total measure $S(r,d,n)$ of nnM sets for $n = 2$, $r = 3.58$ and $r = 4$

Discussion and conclusions. Figure 3.9 shows that $S(r,d,n) = 0$ holds for regular logistic maps (see Section 3.1 for the description of this type of maps), which is out of interest since this case is described by Corollary 3.7. On the contrary, $S(r,d,n) > 0$ holds for most of the values of $r \in [3.5, 4]$ corresponding to the stochastic behavior of logistic map with the following two exceptions.

For $r \in [r_0, r_1)$, where $r_0 \approx 3.57$ is the so-called accumulation point [Spr03], and $r_1 \approx 3.59$, all sets of $\mathcal{P}^{\mathrm{id}}(d)$ are nearly Markov for $d = 1$, 2. One can rigorously show that the ordinal partition $\mathcal{P}^{\mathrm{id}}(d)$ has the Markov property for $d = 1$, but it is not clear, whether this remains true for higher d. Some hope in this direction gives the fact that for $r = 3.58$ all sets of $\mathcal{P}^{\mathrm{id}}(d)$ are nearly Markov for $d \leq 6$ and $n = 2$ (Figure 3.10).

The second exception is the point $r = 4$. In this case the invariant measure μ_4 can be obtained analytically; though the ordinal partition has the Markov property for $d = 1$, 2, already for $d = 3$ this is not the case. As d increases the ordinal partition "loses" the Markov property (Figure 3.10).

Therefore the experiment demonstrates that the ordinal partition $\mathcal{P}^{\mathrm{id}}(d)$ in general does not have the Markov property for logistic maps, at least for small d. The same holds for tent and skew tent maps, thus the study of the conditional entropy of unimodal maps seems to be an interesting subject of further investigation.

3.4 Empirical conditional entropy of ordinal patterns

In this section we focus on practical aspects of the conditional entropy of ordinal patterns. It can be used for measuring complexity of time series modeled by realization of a stationary stochastic process. Therefore we define first the conditional entropy of ordinal patterns for a stochastic process $\mathbf{Y}$ on a probability space $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$:

$$h_{\mu,\mathrm{cond}}(\mathbf{Y}, d) = H(\mathcal{P}^{\mathbf{Y}}(d)_2) - H(\mathcal{P}^{\mathbf{Y}}(d)). \tag{3.13}$$

Consider a realization of a stochastic process $\big(\mathbf{y}(t)\big)_{t\in\mathbb{T}} = \Big(\big(\mathbf{Y}(t)\big)(\omega)\Big)_{t\in\mathbb{T}}$ and the corresponding sequence of ordinal patterns $\boldsymbol{\pi}(t;\mathbf{y})_{t\in\mathbb{T}'}$ for $\mathbb{T}' = \mathbb{T} \setminus \{0, 1, \ldots, d-1\}$. Recall that the frequency of occurrence of an ordinal pattern $\mathbf{i} \in \Pi_d^N$ of order d among the first $(t-d+1)$ ordinal patterns of the sequence $\boldsymbol{\pi}(t;\mathbf{y})$ is given by (see Subsection 2.3.1.4):

$$n_{\mathbf{i}}(t;\mathbf{y}) = \#\big\{r \in \{d, 1+d, \ldots, t\} \mid \boldsymbol{\pi}(r;\mathbf{y}) = \mathbf{i}\big\}.$$

for $t \in \mathbb{T}'$. In the same manner we determine the frequency of an ordinal pattern pair $\mathbf{i}, \mathbf{j}$ for $\mathbf{i}, \mathbf{j} \in \Pi_d^N$:

$$n_{\mathbf{i},\mathbf{j}}(t;\mathbf{y}) = \#\big\{r \in \{d, 1+d, \ldots, t\} \mid \boldsymbol{\pi}(r;\mathbf{y}) = \mathbf{i}, \boldsymbol{\pi}(r+1;\mathbf{y}) = \mathbf{j}\big\}$$

for $t \in \mathbb{T}'$. The *empirical conditional entropy of ordinal patterns* of order $d \in \mathbb{N}$ for $\mathbf{y}$ is given by

$$\begin{aligned}\widehat{h}_{\mathrm{cond}}(t;\mathbf{y},d) &= -\frac{1}{t-d}\sum_{\mathbf{i}=0}^{((d+1)!)^N-1}\ \sum_{\mathbf{j}=0}^{((d+1)!)^N-1} n_{\mathbf{i},\mathbf{j}}(t-1;\mathbf{y}) \ln \frac{n_{\mathbf{i},\mathbf{j}}(t-1;\mathbf{y})}{n_{\mathbf{i}}(t-1;\mathbf{y})} \qquad (3.14)\\ &= \frac{1}{t-d}\sum_{\mathbf{i}=0}^{((d+1)!)^N-1} n_{\mathbf{i}}(t-1;\mathbf{y}) \ln n_{\mathbf{i}}(t-1;\mathbf{y})\\ &\quad -\frac{1}{t-d}\sum_{\mathbf{i}=0}^{((d+1)!)^N-1}\ \sum_{\mathbf{j}=0}^{((d+1)!)^N-1} n_{\mathbf{i},\mathbf{j}}(t-1;\mathbf{y}) \ln n_{\mathbf{i},\mathbf{j}}(t-1;\mathbf{y}).\end{aligned}$$

(cf. with the naive estimator of entropy of ordinal partition (2.20), p. 39).

As a direct consequence of Birkhoff's Ergodic Theorem we have that under certain assumptions the empirical conditional entropy approaches to the conditional entropy. Namely it holds the following.

Theorem 3.13. *Given a realization* $\mathbf{y}$ *of an ergodic stationary stochastic process* $\mathbf{Y} = \big(\mathbf{Y}(t)\big)_{t\in\mathbb{T}}$ *with* $\mathbb{T} = \{0, 1, \ldots, L\}$ *on a probability space* $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$*, for any* $d \in \mathbb{N}$ *it holds almost sure that*

$$\lim_{L\to\infty} \widehat{h}_{cond}(L; \mathbf{y}, d) = h_{\mu,cond}(\mathbf{Y}, d). \tag{3.15}$$

In particular, given an ergodic measure-preserving map T *on* $\big(\Omega, \mathbb{B}(\Omega), \mu\big)$*, for any real-valued observable* $\mathbf{X} : \Omega \to \mathbb{R}^N$ *and for almost all* $\omega \in \Omega$ *it holds*

$$\lim_{L\to\infty} \widehat{h}_{cond}\Big(L; \Big(\mathbf{X}(\omega), \mathbf{X}\big(T(\omega)\big), \ldots, \mathbf{X}\big(T^L(\omega)\big)\Big), d\Big) = h^{\mathbf{X}}_{\mu,cond}(T, d). \tag{3.16}$$

Proof. Statement (3.16) follows from Lemma 2.3, and equality (3.15) follows from the fact that every ergodic stationary stochastic process can be represented in form (2.14) on the basis of an observable of some ergodic dynamical system (see Section 2.2, p. 31). □

In practice it is problematic to apply the empirical conditional entropy (as well as the empirical permutation entropy) to data with dimension higher than one (for more discussion see Subsection 3.4.1). Therefore we consider below only one-dimensional processes $\mathbf{Y} = Y$. In this case we have

$$\widehat{h}_{\mathrm{cond}}(t+1; y, d) = \frac{1}{t-d+1} \sum_{i=0}^{(d+1)!-1} \Bigg(n_i(t; y) \ln n_i(t; y) - \sum_{j=0}^{(d+1)!-1} n_{i,j}(t; y) \ln n_{i,j}(t; y) \Bigg).$$

3.4.1 Sensitivity to the size of a sample

Since in applications a complete orbit or an infinite realization of a stochastic process is not accessible, we estimate the conditional entropy by the empirical conditional entropy $\widehat{h}_{\mathrm{cond}}(L; y, d)$ for a finite sample length L.

For estimation of the permutation entropy using the naive estimator, Amigó et al. recommend to take $L \geq 5(d+1)!$ [AZS08]. Empirical conditional entropy is more sensitive to the size of a sample than empirical permutation entropy since rare ordinal patterns being successors of frequent ones have considerable impact on the value of $\widehat{h}_{\mathrm{cond}}(t; y, d)$. Moreover, to calculate the conditional entropy one takes into account pairs of successive ordinal patterns. As one can easily check, there are $(d+1)(d+1)!$ possible pairs, thus reliable estimation of the conditional entropy requires larger sample than estimation of the permutation entropy. For this reason we consider Grassberger's estimator of conditional entropy:

$$\begin{aligned}\widehat{h}^{G}_{\mathrm{cond}}(t+1; y, d) &= \widehat{H}_G\big(\mathcal{P}^Y(d)_2\big) - \widehat{H}_G\big(\mathcal{P}^Y(d)\big)\\ &= \frac{1}{t-d+1} \sum_{i=0}^{(d+1)!-1} \Bigg(n_i(t; y) G\big(n_i(t; y)\big) - \sum_{j=0}^{(d+1)!-1} n_{i,j}(t; y) G\big(n_{i,j}(t; y)\big) \Bigg).\end{aligned}$$

Grassberger's estimator has very small systematic errors unless the number of possible outcomes $(d+1)(d+1)!$ is much larger than L [Gra03]. Therefore, to get a reliable estimation of the empirical conditional entropy, we propose to take $L \sim (d+1)(d+1)!$. The following example shows how sensitive estimators of empirical conditional entropy are to the size of a sample, in comparison with estimators of permutation and sorting entropies.

Example 3.6. Consider an orbit generated by the 3-adic sawtooth map[2] $T_3 = (3\omega)$ mod 1. Figure 3.11 illustrates, how the estimated values of empirical conditional, permutation and sorting entropies of order $d = 8$ depend on the size of a sample (in fractions of $(d+1)(d+1)!$). The empirical permutation entropy converges to the true value rather quickly, whereas the empirical sorting entropy and empirical conditional entropy are far from the true values for relatively short orbits. Implementation of Grassberger's estimator for the conditional entropy provides a significant improvement, however, the estimation of the conditional entropy remains unreliable for a sample smaller than $\frac{1}{16}(d+1)(d+1)!$.

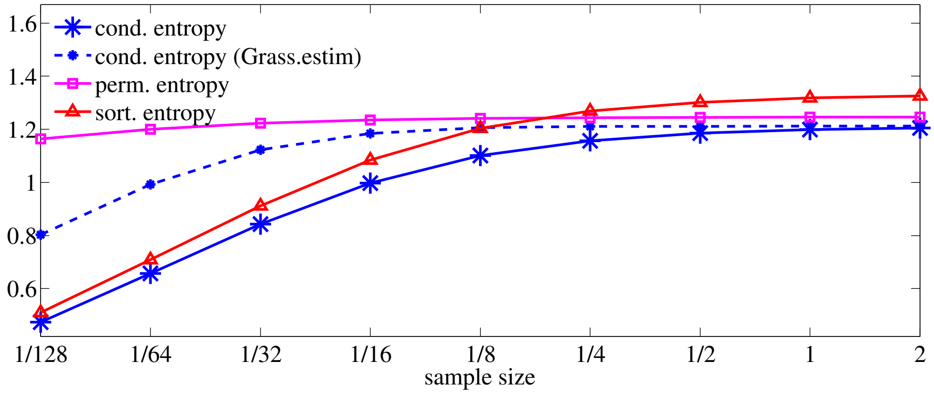


Figure 3.11: Empirical conditional, permutation and sorting entropies of order $d = 8$ for the 3-adic sawtooth map $T_3 = (3\omega)$ mod 1 for various sizes of the sample (in fractions of $9 \cdot 9!$)

Remark. Example 3.6 shows that for a reliable computation of the empirical conditional entropy for a univariate realization y, one has to take a sample of size $L \sim (d+1)(d+1)!$, which corresponds to the number of all possible pairs of (one-dimensional) ordinal patterns. Since there are $\big((d+1)(d+1)!\big)^N$ N-dimensional ordinal patterns (for a description of N-dimensional ordinal patterns see Subsubsection 2.3.1.1, p. 33), it is necessary to take $L \sim \big((d+1)(d+1)!\big)^N$ to compute the empirical conditional entropy of order d for an N-dimensional stochastic process. (This provides $L \sim 5832$ already for $d = 2$, $N = 3$ and $L \sim 9604$ for $d = 3$, $N = 2$.) Therefore we consider the empirical conditional entropy of ordinal patterns only for univariate stochastic processes.

[2] We choose this map since it has higher entropy than the golden mean map and logistic maps ($h_\lambda(T_3) = \ln 3$ for the Lebesgue measure λ), which makes estimation of entropies for T_3 more challenging.

The estimation of the permutation and the sorting entropy from a finite orbit is especially problematic when the entropy is large. This fact can be easily seen for the permutation entropy: indeed, the larger the entropy, the more uniform the distribution of ordinal patterns is. In turn, this means that a larger number of ordinal patterns influences the entropy and the size of the sample should be sufficient to estimate frequencies of all these ordinal patterns. The same reasonings explain problems with the estimation of the conditional entropy of ordinal patterns. We provide Example 3.7 to illustrate the problems with measuring large complexity using the empirical conditional entropy.

Example 3.7. Consider a sawtooth map $T_\beta(\omega) = \beta\omega \mod 1$ on the unit interval $[0, 1]$ for $\beta = 3, 5, \ldots, 15$. This is a particular case of the beta-transformation, thus it holds $h_\lambda(T_\beta) = \ln\beta$ for the Lebesgue measure λ [Par60]. Figure 3.12 presents the empirical conditional entropy of order $d = 8$ computed from orbits of these maps for different lengths L. One can see that the length $L = (d + 1)(d + 1)! = 9 \cdot 9!$ is insufficient for estimating conditional entropy of T_β with $\beta \geq 5$. The result for $L = 5(d + 1) \cdot (d + 1)! = 45 \cdot 9!$ is much better; note that taking larger length does not provide further improvement. As we have already mentioned in Example 3.6, Grassberger's estimator provides a much better estimation of conditional entropy.

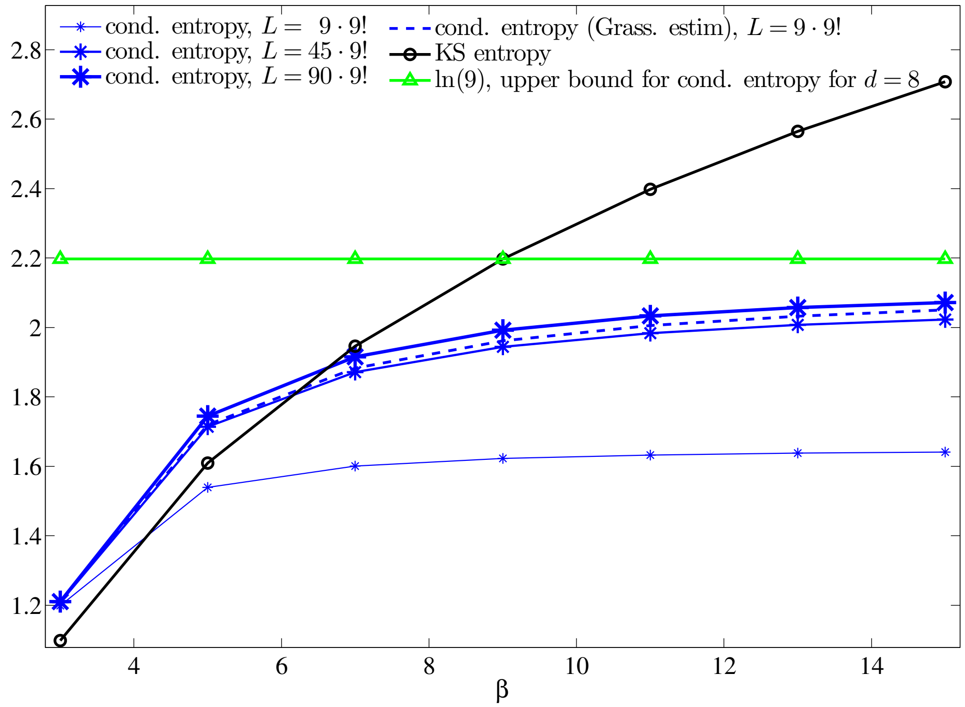


Figure 3.12: Empirical conditional entropies of order $d = 8$ computed for various lengths of an orbit of the sawtooth map $T_\beta(\omega) = \beta\omega \mod 1$ with various values of β

Remark. From (3.13) it follows that the conditional entropy of order d is bounded from above:

$$h_{\mu,\mathrm{cond}}(Y,d) \le \ln(d+1) \tag{3.17}$$

for all $d \in \mathbb{N}$ and for every stationary stochastic process Y. Due to the upper bound (3.17), the conditional entropy of order $d = 8$ does not provide a somehow reliable estimate of the KS entropy of T_β for $\beta > 9$. Note that the permutation entropy of order d is also bounded from above and the bound is even lower than for the conditional entropy, see [KUU14] for details.

3.4.2 Robustness with respect to noise

Real-world data are usually corrupted with some noise. Generally speaking, all ordinal-patterns-based quantities are rather robust to observational noise since it distorts ordinal structure less then values. However, the extent of this robustness may differ. Relatively small observational noise creates some new ordinal patterns, that are not observed in the noiseless dynamics [AZS08, Ami10]. Since these new patterns are relatively rare, the empirical permutation entropy is rather robust to noise. By contrast, the empirical conditional entropy of ordinal patterns is quite sensitive to noise since even rare new patterns can significantly change the "transition probabilities" $\frac{n_{\mathbf{i},\mathbf{j}}(t-1;\mathbf{y})}{n_{\mathbf{i}}(t-1;\mathbf{y})}$ (see (3.14)) of ordinal patterns. Let us consider an example to illustrate the effect of noise on the permutation, sorting and conditional entropies.

Example 3.8. Consider the noisy logistic stochastic process $\mathrm{NL}(t;r,\sigma)$ for $t \in \mathbb{T}$ given by

$$\mathrm{NL}(t;r,\sigma) = f_r^t + \sigma\epsilon(t),$$

where $f_r : [0,1] \hookleftarrow$ is the logistic map with control parameter $r \in [1,4]$, ϵ is the standard additive white Gaussian noise (see p. 32) with the level $\sigma > 0$. Figure 3.13 illustrates the increase of empirical entropies computed from a realization of $\mathrm{NL}(t;4,\sigma)$ for various σ relative to values of entropies for $\mathrm{NL}(t;4,0)$. The size of sample $L = 4 \cdot 10^5$ is taken.

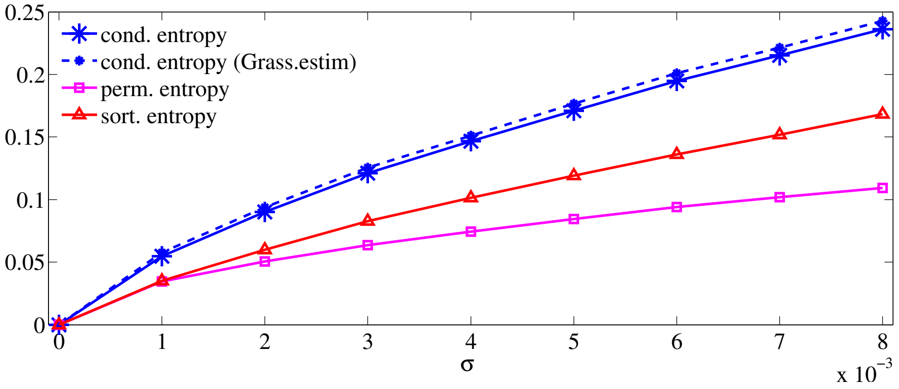


Figure 3.13: Relative increase of empirical conditional, permutation and sorting entropies of order $d = 6$ for the level σ of observational noise for a noisy logistic stochastic process

One can see that the empirical permutation entropy is much more robust to noise than the empirical conditional entropy of ordinal patterns. Note that the estimated value of the conditional entropy has almost the same sensitivity to noise for both estimators.

So when applying the empirical conditional entropy one has to take in mind that it is less robust to noise than the permutation entropy. A method for correcting the overestimation of the conditional entropy caused by noise is of interest.

3.5 Conclusions

Let us briefly summarize the results of this chapter. As we have discussed, the conditional entropy of ordinal patterns has rather good properties, namely:

1. The conditional entropy for finite order d coincides with the KS entropy for systems with periodic dynamics (Theorem 3.6) and for Markov shifts over two symbols (Theorem 3.10). The latter result can be extended to dynamical systems that are order-isomorphic to Markov shifts over two symbols, such as the golden mean dynamical system, the dyadic map on the unit interval and the baker's map on the unit square (all with the Lebesgue measure).

2. The conditional entropy converges to the KS entropy as d tends to infinity in the following cases:

   (a) If the ordinal partition has the Markov property for all $d \geq d_0$ (statement (i) of Lemma 3.8).

   (b) If the permutation entropy $h_\mu^{\mathbf{X}}(T, d)$ converges to the KS entropy as d tends to infinity and it holds $h_\mu^{\mathbf{X}}(T, d) \geq h_\mu^{\mathbf{X}}(T, d+1)$ for all $d \geq d_0$ (Corollary 3.5).

   (c) If the permutation entropy converges to the KS entropy as d tends to infinity and there exits a limit of the sorting entropy as d tends to infinity (Corollary 3.5).

3. According to experiments, the conditional entropy provides a good practical estimation of the KS entropy (see Theorem 3.4 for some theoretical underpinnings), for instance, for unimodal maps (Subsection 3.3.4) and for Markov shifts (see Figure 3.6).

The empirical conditional entropy of ordinal patterns can be used as a practical measure of complexity for time series. Moreover, in Chapter 4 we successfully apply this quantity to the problem of time series segmentation. In this regard it is important to note that the conditional entropy is computationally simple: it has the same computational complexity as the permutation entropy (The algorithm for fast computing the conditional entropy of ordinal patterns, on the basis of the ideas suggested in [UK13], is presented in [Una15]).

Meanwhile, some questions concerning the conditional entropy of ordinal patterns remain open. Possible directions of a future work are.

1. Find dynamical systems different from the considered here for that the conditional entropy coincides with the KS entropy starting from some finite d. In this regard statement (ii) of Lemma 3.8 may be helpful.

2. Show for some concrete systems that the conditional entropy converges to the KS entropy as d tends to infinity (the first candidates for this are unimodal maps).

3. Find an improved estimator of conditional entropy, being more robust with respect to noise than naive and Grassberger's estimators.

3.6 Proofs

In Subsection 3.6.2 we prove Lemma 3.9, to do this we use an auxiliary result established in Subsection 3.6.1. In Subsection 3.6.3 we provide a proof of Proposition 3.12.

3.6.1 Markov property of a partition for Markov shifts

Hereafter we call the partition $\mathcal{C} = \{C_0, C_1, \dots, C_l\}$, where $C_0, C_1, \dots, C_l$ are cylinders, a *cylinder partition.* By the definition of Markov shifts, the cylinder partition is generating and has the Markov property.

To prove Lemma 3.9 we need the following result.

Lemma 3.14. *Let $(A^{\mathbb{N}}, \mathbb{B}_{\Pi}(A^{\mathbb{N}}), m, \sigma)$ be a Markov shift over $A = \{0, 1, \dots, l\}$. Suppose that $\mathcal{P} = \{P_0, P_1, \dots, P_n\}$ is a partition of $A^{\mathbb{N}}$ with the following properties:*

(i) any set P_i is either a subset of some cylinder C_a ($P_i \subseteq C_a$) or an invariant set of zero measure ($\sigma^{-1}(P_i) = P_i$ and $m(P_i) = 0$);

(ii) for any $P_i, P_j \in \mathcal{P}$ it holds either $P_i \cap \sigma^{-1}(P_j) = C_a \cap \sigma^{-1}(P_j)$ for some $a \in A$ or $m\left(P_i \cap \sigma^{-1}(P_j)\right) = 0$.

Then the partition $\mathcal{P}$ is generating and has the Markov property.

Let us first prove the following lemma.

Lemma 3.15. *Suppose that the assumptions of Lemma 3.14 hold and that P_i, P_j are sets from $\mathcal{P}$ with $m(P_i \cap \sigma^{-1}(P_j)) \neq 0$, $P_i \subseteq C_{a_0}$, and $P_j \subseteq C_{a_1}$, where C_{a_0} and C_{a_1} are cylinders. Then for any $P \subseteq P_j$, $P \in \mathbb{B}_{\Pi}(A^{\mathbb{N}})$ it holds*

$$m(P_i \cap \sigma^{-1}(P)) = \frac{m(C_{a_0} \cap \sigma^{-1}(C_{a_1}))}{m(C_{a_1})} m(P). \tag{3.18}$$

Proof. Given $m(P) = 0$, equality (3.18) holds automatically; thus assume that $m(P) > 0$. As a consequence of assumption (*ii*) of Lemma 3.14, for any $P \subseteq P_j$ we have

$$\begin{aligned} P_i \cap \sigma^{-1}(P) = P_i \cap \left(\sigma^{-1}(P_j) \cap \sigma^{-1}(P)\right) &= \left(C_{a_0} \cap \sigma^{-1}(P_j)\right) \cap \sigma^{-1}(P) \\ &= C_{a_0} \cap \sigma^{-1}(P). \end{aligned}$$

Since $P \subseteq P_j \subseteq C_{a_1}$, for all $s \in P$ the first symbol is fixed. One can also decompose the set P into a union of sets with two fixed elements:

$$P = \bigcup_{a_2 \in A_0} \left(C_{a_1} \cap \sigma^{-1}(B_{a_2})\right),$$

where it holds $B_{a_2} \subseteq C_{a_2}$ for all $a_2 \in A_0 \subseteq A$. Then it follows

$$m(P_i \cap \sigma^{-1}(P)) = m(C_{a_0} \cap \sigma^{-1}(P)) = m(C_{a_0} \cap \sigma^{-1}(C_{a_1}) \cap \sigma^{-2}(\bigcup_{a_2 \in A_0} B_{a_2})).$$

Finally, since m is a Markov measure, we get

$$\begin{aligned} m(C_{a_0} \cap \sigma^{-1}(C_{a_1}) \cap \sigma^{-2}(\bigcup_{a_2 \in A_0} B_{a_2})) &= \frac{m(C_{a_0} \cap \sigma^{-1}(C_{a_1}))}{m(C_{a_1})} m(C_{a_1} \cap \sigma^{-1}(\bigcup_{a_2 \in I} B_{a_2})) \\ &= \frac{m(C_{a_0} \cap \sigma^{-1}(C_{a_1}))}{m(C_{a_1})} m(P). \end{aligned}$$

This completes the proof. □

Now we come to the proof of Lemma 3.14.

Proof. By assumption (*i*), the partition $\mathcal{P}$ is finer than the generating partition $\mathcal{C}$ except for an invariant set of measure zero; hence $\mathcal{P}$ is generating as well. To show that $\mathcal{P}$ has the Markov property let us fix some $n \in \mathbb{N}$ and consider $P_{i_0}, P_{i_1}, \ldots, P_{i_n} \in \mathcal{P}$ with

$$m\left(P_{i_0} \cap \sigma^{-1}(P_{i_1}) \cap \ldots \cap \sigma^{-n}(P_{i_{n-1}})\right) > 0.$$

We need to show that the following equality holds:

$$\frac{m\left(P_{i_0} \cap \sigma^{-1}(P_{i_1}) \cap \ldots \cap \sigma^{-n}(P_{i_n})\right)}{m\left(P_{i_0} \cap \sigma^{-1}(P_{i_1}) \cap \ldots \cap \sigma^{-(n-1)}(P_{i_{n-1}})\right)} = \frac{m\left(P_{i_{n-1}} \cap \sigma^{-1}(P_{i_n})\right)}{m(P_{i_{n-1}})}.$$

According to assumption (*i*), there exist $a_0, a_1, \ldots, a_n \in A$ with $P_{i_k} \subset C_{a_k}$ for all $k = 0, 1, \ldots, n$. Therefore by successive application of (3.18) we have:

$$\begin{aligned} & m\left(P_{i_0} \cap \sigma^{-1}(P_{i_1}) \cap \ldots \cap \sigma^{-n}(P_{i_n})\right) = \\ & \quad = m\left(P_{i_1} \cap \sigma^{-1}(P_{i_2}) \cap \ldots \cap \sigma^{-(n-1)}(P_{i_n})\right) \frac{m(C_{a_0} \cap \sigma^{-1}(C_{a_1}))}{m(C_{a_1})} = \ldots \\ & \quad = m(P_{i_{n-1}} \cap \sigma^{-1}(P_{i_n})) \prod_{k=0}^{n-2} \frac{m\left(C_{a_k} \cap \sigma^{-1}(C_{a_{k+1}})\right)}{m(C_{a_{k+1}})}. \end{aligned}$$

Analogously,

$$\begin{aligned}
&m\left(P_{i_0}\cap\sigma^{-1}(P_{i_1})\cap\ldots\cap\sigma^{-(n-1)}(P_{i_{n-1}})\right)=\\
&\quad=m\left(P_{i_1}\cap\sigma^{-1}(P_{i_2})\cap\ldots\cap\sigma^{-(n-2)}(P_{i_{n-1}})\right)\frac{m(C_{a_0}\cap\sigma^{-1}(C_{a_1}))}{m(C_{a_1})}=\ldots\\
&\quad=m(P_{i_{n-2}}\cap\sigma^{-1}(P_{i_{n-1}}))\prod_{k=0}^{n-3}\frac{m\left(C_{a_k}\cap\sigma^{-1}(C_{a_{k+1}})\right)}{m(C_{a_{k+1}})}\\
&\quad=m(P_{i_{n-1}})\prod_{k=0}^{n-2}\frac{m\left(C_{a_k}\cap\sigma^{-1}(C_{a_{k+1}})\right)}{m(C_{a_{k+1}})},
\end{aligned}$$

and we are done. □

Corollary 3.16. *Let $\mathcal{P}=\{P_0,P_1,\ldots,P_n\}$ and $\widetilde{\mathcal{P}}=\{P\setminus O\mid P\in\mathcal{P}\}\cup\{O\}$, where $m(O)=0$ and $\sigma^{-1}(O)=O$, be partitions of $A^{\mathbb{N}}$. If $\widetilde{\mathcal{P}}$ satisfies the assumptions of Lemma 3.14, then $\mathcal{P}$ is generating and has the Markov property.*

3.6.2 Proof of Lemma 3.9

Now we show that for ergodic Markov shifts over two symbols the ordinal partitions are generating and have the Markov property. The idea of the proof is to construct for an ordinal partition $\mathcal{P}^X(d)$ a partition $\widetilde{\mathcal{P}}^X(d)$ as in Corollary 3.16 and to show that $\widetilde{\mathcal{P}}^X(d)$ satisfies the assumptions of Lemma 3.14. Then the partition $\mathcal{P}^X(d)$ is generating and has the Markov property by Corollary 3.16.

The proof is divided into a sequence of three lemmas. First, Lemma 3.17 relates the partition $\mathcal{P}^X(1)$ with the cylinder partition $\mathcal{C}$. Then we construct the partition $\widetilde{\mathcal{P}}^X(d)$ and show in Lemma 3.18 that it satisfies assumption (i) of Lemma 3.14. Finally, in Lemma 3.19 we prove that $\widetilde{\mathcal{P}}^X(d)$ satisfies assumption (ii) of Lemma 3.14.

Given $\overline{0}=(0,0,\ldots,0,\ldots),\overline{1}=(1,1,\ldots,1,\ldots)$, the following holds.

Lemma 3.17. *Let $P_{(0,1)},P_{(1,0)}\in\mathcal{P}^X(1)$ be elements of the ordinal partition corresponding to the increasing and decreasing ordinal pattern of order $d=1$, respectively:*

$$P_{(0,1)}=\{s\in\{0,1\}^{\mathbb{N}}\mid X(s)<X(\sigma s)\},\ P_{(1,0)}=\{s\in\{0,1\}^{\mathbb{N}}\mid X(s)\geq X(\sigma s)\},$$

where X is lexicographic-like. Then it holds

$$P_{(0,1)}=C_0\setminus\{\overline{0}\}\ \text{ and }\ P_{(1,0)}=C_1\cup\{\overline{0}\}.$$

Proof. We show first that for all $s\in C_0\setminus\{\overline{0}\}$ it holds $X(s)<X(\sigma s)$. Indeed, assume $s=(s_0,s_1,\ldots)\in C_0\setminus\{\overline{0}\}$. Then for the smallest $k\in\mathbb{N}$ with $s_k=1$ it holds $s_j=(\sigma s)_j$ for $j=0,\ldots,k-1$ and $s_{k-1}<(\sigma s)_{k-1}=s_k$, that is $s\prec\sigma s$. Since X is lexicographic-like, this implies $X(s)<X(\sigma s)$.

By the same reason, for all $s\in C_1\setminus\{\overline{1}\}$ it holds $X(s)>X(\sigma s)$. Finally, as one can easily see, $s\in\{\overline{0}\}\cup\{\overline{1}\}$ implies $X(s)=X(\sigma s)$. According to Definition 2.12 of an ordinal pattern, in this case $s\in P_{(1,0)}$ and we are done. □

In order to apply Corollary 3.16, consider the set

$$O = \bigcup_{n=0}^{\infty} \sigma^{-n}(\{\overline{0}\}).$$

By Definition 2.4 of a Markov shift $m(C_0), m(C_1) > 0$, hence no fixed point has full measure. Together with the assumption of ergodicity of the shift, this implies that the measure of a fixed point is zero, thus $m(O) = 0$. As is easy to check, $\sigma^{-1}(O) = O$. Therefore, to prove that the partition $\mathcal{P}^X(d)$ is generating and has the Markov property, it is sufficient to show that the partition

$$\widetilde{\mathcal{P}}^X(d) = \{P \setminus O \mid P \in \mathcal{P}^X(d)\} \cup \{O\} \tag{3.19}$$

satisfies the assumptions of Lemma 3.14.

Lemma 3.18. *Let $d \in \mathbb{N}$ and $\widetilde{\mathcal{P}}^X(d)$ be the partition defined by (3.19) for an ergodic Markov shift over two symbols. For every $P \in \widetilde{\mathcal{P}}^X(d) \setminus \{O\}$ it holds*

$$P \subset C_{a_0 a_1 \ldots a_{d-1}},$$

where $C_{a_0 a_1 \ldots a_{d-1}}$ is a cylinder set.

Proof. Consider the partition consisting of the cylinder sets:

$$\mathcal{C}_d = \{C_{a_0 a_1 \ldots a_{d-1}} \mid a_0, a_1, \ldots, a_{d-1} \in \{0, 1\}\},$$

for $d \in \mathbb{N}$. According to Lemma 3.17, $\mathcal{P}^X(1)$ coincides with the partition $\mathcal{C} = \{C_0, C_1\}$ except for the only point $\overline{0}$, consequently for all $d \in \mathbb{N}$, partition $\mathcal{P}^X(1)_d$ coincides with $\mathcal{C}_d$ except for the points from the set $\sigma^{-(d-1)}(\{\overline{0}\}) \subset O$. Since $\widetilde{\mathcal{P}}^X(d) \setminus \{O\}$ is finer than $\mathcal{P}^X(1)_d$, we are done. □

Lemma 3.19. *Let $d \in \mathbb{N}$ and $\widetilde{\mathcal{P}}^X(d)$ be the partition defined by (3.19) for an ergodic Markov shift over two symbols. Given $P_i, P_j \in \widetilde{\mathcal{P}}^X(d)$ with $P_i \subset C_{a_0}$ for $a_0 \in \{0, 1\}$, it holds either*

$$P_i \cap \sigma^{-1}(P_j) = C_{a_0} \cap \sigma^{-1}(P_j)$$

or

$$m(P_i \cap \sigma^{-1}(P_j)) = 0.$$

Proof. Fix some $d \in \mathbb{N}$ and let $P_i, P_j \in \widetilde{\mathcal{P}}^X(d)$. If $P_i = O$ or $P_j = O$, then it follows immediately that $m(P_i \cap \sigma^{-1}(P_j)) = 0$; thus we put $P_i \neq O$, $P_j \neq O$. Further, let us define the set P as follows:

$$P = C_{a_0} \cap \sigma^{-1}(P_j) = \{s = (s_0, s_1, \ldots) \mid s_0 = a_0, (s_1, s_2, \ldots) \in P_j\}.$$

It is sufficient to prove that it holds either $P \subset P_i$ or $P \cap P_i = \emptyset$. To do this we show that the ordering of $\big(X(s), X(\sigma s), \ldots, X(\sigma^d s)\big)$ is the same for all $s \in P$. Since

$(s_1, s_2, \ldots) \in P_j$, the ordering of $\big(X(\sigma s), X(\sigma^2 s), \ldots, X(\sigma^d s)\big)$ is the same for all $s \in P$. It remains to show that the relation between $X(s)$ and $X(\sigma^{(k)} s)$ for $k = 1, 2, \ldots, d$ is the same for all $s \in P$.

Note that the order relations between $X(\sigma s)$ and $X(\sigma^{(k+1)} s)$ for $k = 1, 2, \ldots, d$ is given by the fact that $\sigma s = (s_1, s_2, \ldots) \in P_j$. Next, by Lemma 3.18 for every P_j there exists a cylinder set $C_{a_1 a_2 \ldots a_d}$, such that if $(s_1, s_2, \ldots) \in P_j$ then $s_k = a_k$ for $k = 1, 2, \ldots, d$. Now it remains to consider two cases:

First case: assume that $s_0 = a_0 = 0$ and consider $k = 1, 2, \ldots, d$. If $s_k = 1$ then $X(s) < X(\sigma^k s)$. Further, if $s_k = 0$, then $X(\sigma s) < X(\sigma^{(k+1)} s)$ implies $X(s) < X(\sigma^k s)$, and $X(\sigma s) \geq X(\sigma^{(k+1)} s)$ implies $X(s) \geq X(\sigma^k s)$.

Second case: analogously, assume that $s_0 = a_0 = 1$ and consider $k = 1, 2, \ldots, d$. If $s_k = 0$ then $X(s) \geq X(\sigma^k s)$. If $s_k = 1$, then $X(\sigma s) < X(\sigma^{(k+1)} s)$ implies $X(s) < X(\sigma^k s)$, and $X(\sigma s) \geq X(\sigma^{(k+1)} s)$ implies $X(s) \geq X(\sigma^k s)$.

Therefore all $s \in P$ are in the same set of the ordinal partition, and consequently it holds either $P \subset P_i$ or $P \cap P_i = \emptyset$. This finishes the proof. □

3.6.3 Proof of Proposition 3.12

The proof comprises two lemmas.

Lemma 3.20. *Given $\big(A^{\mathbb{N}}, \mathbb{B}_{\Pi}(A^{\mathbb{N}}), m_B, \sigma\big)$ a Bernoulli shift over $A = \{0, 1, \ldots, l\}$ with $l \geq 2$ and $m_B(C_i) = p_i = \frac{1}{l+1}$ for all $i \in A$. Then for all $d \in \mathbb{N}$ it holds*

$$m_B(P_{(0,1,\ldots,d)}) = \frac{1}{l(l+1)^d} \sum_{i=0}^{l-1} (l-i) \binom{i+d-1}{d-1}, \tag{3.20}$$

where $P_{(0,1,\ldots,d)} \in \mathcal{P}^X(d)$ with lexicographic-like X.

Proof. Let us discuss the structure of the set $P_{(0,1,\ldots,d)}$. Deduce that $s = (s_0, s_1, \ldots) \in P_{(0,1,\ldots,d)}$ holds if and only if $X(\sigma^{(d-1)} s) < X(\sigma^d s)$ and $s_0 \leq s_1 \leq \ldots \leq s_{d-1}$. Consider first the former condition: the inequality $X(\sigma^{(d-1)} s) < X(\sigma^d s)$ holds if

(i) either $s_{d-1} < s_d$;

(ii) or there exists some $k \in \mathbb{N}$ such that $s_{d-1} = s_d = \ldots = s_{d+k-1} < s_{d+k}$.

Case (i) can be written in form (ii) for $k = 0$. By taking $s_{d-1} = i$, $s_{d+k} = j$ we have

$$\begin{aligned} m_B\Big(\big\{s \mid X(\sigma^{(d-1)} s) < X(\sigma^d s)\big\}\Big) &= \sum_{i=0}^{l-1} p_i \sum_{j=i+1}^{l} p_j + \sum_{i=0}^{l-1} p_i \sum_{k=1}^{\infty} {p_i}^k \sum_{j=i+1}^{l} p_j \\ &= \sum_{i=0}^{l-1} \left(\sum_{k=1}^{\infty} {p_i}^k \right) \sum_{j=i+1}^{l} p_j = \sum_{i=0}^{l-1} \frac{p_i}{1-p_i} \sum_{j=i+1}^{l} p_j. \end{aligned}$$

Consider now the latter condition for $(s_0, s_1, \ldots) \in P_{(0,1,\ldots,d)}$, namely $s_0 \leq s_1 \leq \ldots \leq s_{d-1} = i$. It holds

$$m_B\Big(\{s \mid s_0 \leq s_1 \leq \ldots \leq s_{d-1} = i\}\Big) = \sum_{s_{d-2}=0}^{i} p_{s_{d-2}} \sum_{s_{d-3}=0}^{s_{d-2}} p_{s_{d-3}} \ldots \sum_{s_0=0}^{s_1} p_{s_0}.$$

Combining both conditions yields:

$$m_B(P_{(0,1,\ldots,d)}) = \sum_{i=0}^{l-1} \frac{p_i}{1-p_i} \sum_{j=i+1}^{l} p_j \left(\sum_{s_{d-2}=0}^{i} p_{s_{d-2}} \sum_{s_{d-3}=0}^{s_{d-2}} p_{s_{d-3}} \ldots \sum_{s_0=0}^{s_1} p_{s_0} \right). \qquad (3.21)$$

By assumption $p_0 = p_1 = \ldots = p_l = \frac{1}{l+1}$, and (3.21) can be rewritten as

$$m_B(P_{(0,1,\ldots,d)}) = \sum_{i=0}^{l-1} \frac{1}{l} \frac{(l-i)}{(l+1)} \frac{1}{(l+1)^{d-1}} \left(\sum_{s_{d-2}=0}^{i} \sum_{s_{d-3}=0}^{s_{d-2}} \ldots \sum_{s_0=0}^{s_1} 1 \right). \qquad (3.22)$$

In (3.22) the expression in brackets is nothing else but the number of combinations of $(d-1)$ numbers from the set $\{0, 1, \ldots, i\}$ with repetitions, thus

$$m_B(P_{(0,1,\ldots,d)}) = \frac{1}{l(l+1)^d} \sum_{i=0}^{l-1} (l-i) \binom{i+d-1}{d-1},$$

which completes the proof. $\square$

Note that for $l = 1$ Lemma 3.20 remains true, though equality (3.20) degenerates to

$$m_B(P_{(0,1,\ldots,d)}) = \frac{1}{2^d}.$$

Lemma 3.21. *Given* $\big(A^{\mathbb{N}}, \mathbb{B}_{\Pi}(A^{\mathbb{N}}), m_B, \sigma\big)$ *and* $P_{(0,1,\ldots,d)}$ *as in Lemma 3.20, the sequence*

$$\frac{m_B(P_{(0,1,2)})}{m_B(P_{(0,1)})}, \frac{m_B(P_{(0,1,2,3)})}{m_B(P_{(0,1,2)})}, \ldots, \frac{m_B(P_{(0,1,\ldots,d+1)})}{m_B(P_{(0,1,\ldots,d)})}, \ldots$$

is decreasing for the alphabet $A = \{0, 1, \ldots, l\}$ *with* $l \geq 2$.

Proof. According to Lemma 3.20, for all $d \in \mathbb{N}$ it holds

$$\begin{aligned}\frac{m_B(P_{(0,1,\ldots,d+1)})}{m_B(P_{(0,1,\ldots,d)})} &= \frac{\frac{1}{l(l+1)^{d+1}} \sum_{i=0}^{l-1} (l-i) \binom{i+d}{d}}{\frac{1}{l(l+1)^d} \sum_{j=0}^{l-1} (l-j) \binom{j+d-1}{d-1}} = \frac{\sum_{i=0}^{l-1} (l-i) \frac{(i+d)!}{i!\, d!}}{(l+1) \sum_{j=0}^{l-1} (l-j) \frac{(j+d-1)!}{j!\, (d-1)!}} \\ &= \frac{\sum_{i=0}^{l-1} (l-i) \frac{(i+d-1)!}{i!} (i+d)}{(l+1)d \sum_{j=0}^{l-1} (l-j) \frac{(j+d-1)!}{j!}} = \frac{1}{l+1} \left(1 + \frac{\sum_{i=0}^{l-1} i \frac{(l-i)}{i!} (i+d-1)!}{d \sum_{j=0}^{l-1} \frac{(l-j)}{j!} (j+d-1)!} \right).\end{aligned}$$

Therefore, it remains to show that for all $d \in \mathbb{N}$

$$\frac{\sum\limits_{i=0}^{l-1} i \frac{(l-i)}{i!}(i+d-1)!}{d \sum\limits_{j=0}^{l-1} \frac{(l-j)}{j!}(j+d-1)!} > \frac{\sum\limits_{i=0}^{l-1} i \frac{(l-i)}{i!}(i+d)!}{(d+1) \sum\limits_{j=0}^{l-1} \frac{(l-j)}{j!}(j+d)!}. \tag{3.23}$$

Consider an obvious inequality

$$(d+1) \sum_{i=0}^{l-1} \sum_{j=0}^{l-1} i b_i b_j c_i c_{j+1} > d \sum_{i=0}^{l-1} \sum_{j=0}^{l-1} i b_i b_j c_{i+1} c_j, \tag{3.24}$$

where $b_i, c_i > 0$ for all $i = 0, 1, \ldots, l-1$. Inequality (3.24) is equivalent to

$$\frac{\sum\limits_{i=0}^{l-1} i b_i c_i}{d \sum\limits_{j=0}^{l-1} b_j c_j} > \frac{\sum\limits_{i=0}^{l-1} i b_i c_{i+1}}{(d+1) \sum\limits_{j=0}^{l-1} b_j c_{j+1}}.$$

By taking $b_i = \frac{(l-i)}{i!}$, $c_i = (i+d-1)!$ one obtains inequality (3.23). □

Finally we come to the proof of Proposition 3.12.

Proof. For $l = 1$ the considered Bernoulli shift is a particular case of a Markov shift over two symbols. Therefore, according to Lemma 3.9, the partition $\mathcal{P}^X(d)$ has the Markov property for all $d \in \mathbb{N}$. It remains to show that for $l \geq 2$, for all $d \in \mathbb{N}$, $n \geq 2$ there exist $P_{a_0}, P_{a_1}, \ldots, P_{a_n} \in \mathcal{P}^X(d)$ such that

$$\frac{m\big(P_{a_0} \cap \sigma^{-1}(P_{a_1}) \cap \ldots \cap \sigma^{-n}(P_{a_n})\big)}{m\big(P_{a_0} \cap \sigma^{-1}(P_{a_1}) \cap \ldots \cap \sigma^{-(n-1)}(P_{a_{n-1}})\big)} \neq \frac{m\big(P_{a_{n-1}} \cap \sigma^{-1}(P_{a_n})\big)}{m(P_{a_{n-1}})}. \tag{3.25}$$

Let us take $P_{a_0} = P_{a_1} = \ldots = P_{a_n} = P_{(0,1,\ldots,d)}$. For all $d, n \in \mathbb{N}$

$$m_B\big(P_{(0,1,\ldots,d)} \cap \sigma^{-1}(P_{(0,1,\ldots,d)}) \cap \ldots \cap \sigma^{-n}(P_{(0,1,\ldots,d)})\big) = m_B(P_{(0,1,\ldots,d+n)});$$

thus by Lemma 3.20,

$$m_B\big(P_{(0,1,\ldots,d)} \cap \sigma^{-1}(P_{(0,1,\ldots,d)}) \cap \ldots \cap \sigma^{-n}(P_{(0,1,\ldots,d)})\big) > 0$$

and we can rewrite (3.25) as follows:

$$\frac{m_B\big(P_{(0,1,\ldots,d+n)}\big)}{m_B\big(P_{(0,1,\ldots,d+n-1)}\big)} \neq \frac{m_B\big(P_{(0,1,\ldots,d+1)}\big)}{m_B\big(P_{(0,1,\ldots,d)}\big)}.$$

Indeed, by Lemma 3.21, for $d \in \mathbb{N}$, $n \geq 2$ it holds

$$\frac{m_B\big(P_{(0,1,\ldots,d+n)}\big)}{m_B\big(P_{(0,1,\ldots,d+n-1)}\big)} < \frac{m_B\big(P_{(0,1,\ldots,d+1)}\big)}{m_B\big(P_{(0,1,\ldots,d)}\big)},$$

and we are done. □

Chapter 4

Ordinal change-point detection

In Chapter 3 we have considered models for stationary time series; however, most of real-world time series are non-stationary, that is some of their characteristics change over time [BN93]. Moments of time when a characteristic of a time series changes are called *change-points*; methods for change-point detection are intensively developed [CMS94, BD00, LT07, BHM$^+$13] and have many applications [HS95, KRDF01, LT07, KMPS09]. In particular, detection of change-points provides a segmentation of time series into *pseudo-stationary segments*, parts of time series between the detected change-points. Such segmentation is of interest since many practical complexity measures require stationarity of time series [GR84, BP02, BD09] and may be unreliable when stationarity condition fails.

This chapter is devoted to the detection of change-points in a time series using a novel approach, *ordinal change-point detection*, first considered in [SGK12]. We suggest here new methods for detecting change-points that can be especially helpful as a preprocessing for ordinal-patterns-based methods since most of them are well-defined only for stationary time series [BP02].

We provide a general description of change-points detection problems in Section 4.1. In Section 4.2 we consider the first ordinal-patterns-based method introduced in [SGK12] and suggest several new ordinal-patterns-based methods:

- two methods on the basis of the well-known likelihood ratio and χ^2 statistics;
- a method on the basis of a new statistic CEofOP, which is strongly related to the conditional entropy of ordinal patterns (see Chapter 3).

In Section 4.3 we compare the considered methods for change-points detection by performing experiments on artificially generated time series. Since the results show that the method on the basis of the CEofOP statistic is more effective than other ordinal-patterns-based methods, we investigate its theoretical properties in Section 4.4. In Section 4.5 we provide technical details that are important for the implementation of the ordinal-patterns-based methods for change-point detection.

4.1 General framework and problem statement

For detecting change-points, time series are usually considered as realizations of stochastic processes, which allows to utilize powerful statistical methods [BN93, BD93]. In this chapter we consider stochastic processes with the set of times given by $\mathbb{T} = \{0, 1, \ldots, L\}$ for $L \in \mathbb{N}$ and put $L = \infty$, when $\mathbb{T}$ is supposed to be countable. We consider only one-dimensional stochastic processes to keep notation simple, though there are no principal restrictions on the dimension of a process. A convenient model for studying changes in time series is provided by the following class of processes.

Definition 4.1. A stochastic process $\big(Y(t)\big)_{t\in\mathbb{T}}$ is said to be *piecewise stationary* if there exist $t_0^*, t_1^*, \ldots, t_{N_{\mathrm{st}}}^* \in \mathbb{T}$ with $0 = t_0^* < t_1^* < \ldots < t_{N_{\mathrm{st}}}^* = L$, $N_{\mathrm{st}} \in \mathbb{N}$ such that for every $i \in \{0, 1, \ldots, N_{\mathrm{st}} - 1\}$ the sub-process $\big(Y(t_i^* + 1), Y(t_i^* + 2), \ldots, Y(t_{i+1}^*)\big)$ is stationary (and a prolongation of this sub-process is not stationary any more). The boundaries t_i^* of stationary segments for $i \in \{1, 2, \ldots, N_{\mathrm{st}} - 1\}$ are called *change-points.*

(See [Sto12, Section 3.1] for an alternative, though equivalent, definition of piecewise stationarity.) Simply speaking, a piecewise stationary stochastic process is obtained by gluing N_{st} stationary stochastic processes. Definition 4.1 is important for the entire chapter so we illustrate it by the following example.

Example 4.1. A simple piecewise stationary stochastic process is given by

$$\begin{aligned} Y_{\mathrm{norm}}(t) &= Y_{\mathrm{norm}}\big(t; (m_1, m_2, \ldots, m_{N_{\mathrm{st}}}), (t_1^*, t_2^*, \ldots, t_{N_{\mathrm{st}}-1}^*)\big) \\ &= \begin{cases} \epsilon(t) + m_1, & t \in \{0, 1, \ldots, t_1^*\} \\ \epsilon(t) + m_2, & t \in \{t_1^* + 1, t_1^* + 2, \ldots, t_2^*\} \\ \ldots \\ \epsilon(t) + m_{N_{\mathrm{st}}}, & t \in \{t_{N_{\mathrm{st}}-1}^* + 1, t_{N_{\mathrm{st}}-1}^* + 2, \ldots, L\}, \end{cases} \end{aligned}$$

where $m_1, m_2, \ldots, m_{N_{\mathrm{st}}} \in \mathbb{R}$ are the expected values of the process in corresponding intervals, $t_1^*, t_2^*, \ldots, t_{N_{\mathrm{st}}-1}^* \in \mathbb{T}$ are change-points, ϵ is the standard additive white Gaussian noise (see p. 32). Figure 4.1 shows a realization of these process with a single change-point at $t^* = 150$, note the change in mean of the realization at t^*.

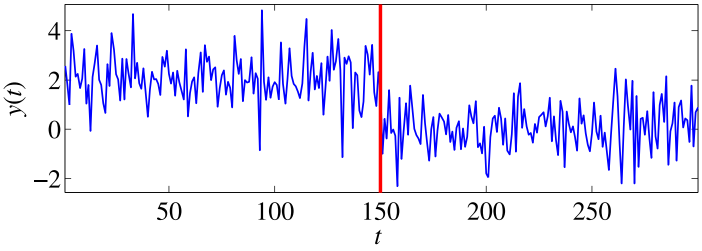


Figure 4.1: A realization of a stochastic process $Y_{\mathrm{norm}}\big(t; (2, 0), 150\big)$, the change-point is marked by a vertical line

Detection of changes in mean is a simple task, it can be done by various methods (see [BN93, Chapter 2]), but in the general case, change-point detection is much more complicated. If a time series can be modeled by a certain stochastic process, one uses *parametric* methods of change-point detection. As appears from its name, a parametric method detects the changes of certain parameters of the stochastic process. When too little is known about the time series to consider it as a realization of a particular stochastic process, one still may expect that certain characteristics (mean, standard deviation, etc.) of the time series reflect the change. In this case one applies *non-parametric* methods for change-point detection (see [BD00, Chapter 2] for details). Non-parametric methods are generally preferred [BD00, Section 7.3] since they require less a priori information than parametric methods. For an overview of methods for change-point detection we refer to [BN93, BD93, CMS94, BD00].

In Subsection 4.1.1 we consider three basic problems of change-point detection and illustrate them by simple examples. In Subsection 4.1.2 we define a special type of change-points (structural change-points), for detection of those it is reasonable to apply ordinal-patterns-based methods. In Subsection 4.1.3 we explain the idea of ordinal change-point detection and suggest ordinal-patterns-based solutions for the three problems of change-point detection. In Subsection 4.1.4 we introduce notations that will be used throughout the chapter.

4.1.1 Three problems of change-point detection

We are interested in detection of change-points in a realization y of a piecewise stationary stochastic process Y. To fix the framework of change-point detection, we consider three problems on the basis of different assumptions about the number of change-points in Y.

Problem 1. Given a single change-point $t^* \in \mathbb{T}$ in the process Y, one needs to find an estimate $\widehat{t^*} \in \mathbb{T}$ of the change-point by studying the realization y.

Problem 2. Given at most one change-point t^* in the process Y, one needs to find an estimate $\widehat{t^*} \in \mathbb{T}$ from y (solve Problem 1) or conclude that no change has occurred.

Problem 3. Given $N_{\mathrm{st}} \in \mathbb{N}$ stationary segments bounded by the change-points $t_1^*, t_2^*, \ldots, t_{N_{\mathrm{st}}-1}^* \in \mathbb{T}$, one calculates an estimate $\widehat{N}_{\mathrm{st}} \in \mathbb{N}$ of N_{st} and estimates $\widehat{t_1^*}, \widehat{t_2^*}, \ldots, \widehat{t^*}_{\widehat{N}_{\mathrm{st}}-1} \in \mathbb{T}$ of the change-points.

These problems are nested in the sense that to solve Problem 3 one needs to solve Problem 2, and to solve Problem 2 one has to solve Problem 1. Problem 3 represents the most general setting of the change-point detection problem and is known as a "multiple change-points detection" [Lav99, LT07]. Problem 2 coincides with the classical "at most one change" problem and is rather often addressed either as a separate task or as a part of the multiple change-points detection [CMS94, BHM$^+$13]. Problem 1 is seldom considered (see [BN93, Subsection 1.1.2.3]); however, we distinguish this problem for technical reasons.

The solution of **Problem 1** is provided by a *statistic*[3] $S(t;y)$ for $t \in \mathbb{T}$ that tends to reach its maximum in $t = t^*$. Then an estimate of the change-point t^* is given by

$$\widehat{t^*}(y) = \arg\max_{t\in\mathbb{T}} S(t;y).$$

To provide the reader an impression of how a statistic detects a change-point, we consider in Example 4.2 a classical statistic that will be used throughout this chapter.

Example 4.2. One of the classical non-parametric methods for detecting changes in mean of a realization y was suggested by Brodsky and Darkhovsky in [BD93] on the basis of the statistic given by

$$\mathrm{BD}(t;y,\delta) = \left(\frac{t(L-t)}{L^2}\right)^{\delta} \left| \frac{1}{t}\sum_{l=1}^{t} y(l) - \frac{1}{L-t}\sum_{l=t+1}^{L} y(l) \right|, \tag{4.1}$$

where $t \in \mathbb{T} \setminus \{L\} = \{0, 1, \ldots, L-1\}$ for $L \in \mathbb{N}$, and the parameter $\delta \in [0,1]$ regulates properties of the statistic (see [BDKS99] for details). Figure 4.2 shows $\mathrm{BD}(t;y,0)$ for a realization of the process $Y_{\text{norm}}\big(t;(2,0),150\big)$ from Example 4.1, note that the maximum of the statistic indicates the change-point.

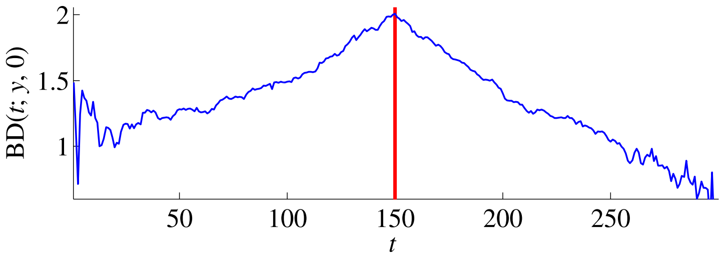


Figure 4.2: Statistic $\mathrm{BD}(t;y,0)$ for a realization y of the process $Y_{\text{norm}}\big(t;(2,0),150\big)$; the change-point is marked by a vertical line

Remark. Let us briefly explain why the maximum of the statistic (4.1) allows to estimate changes in mean. Given y a realization of the process $Y_{\text{norm}}\big(t;(m_1,m_2),t^*\big)$ for $m_1, m_2 \in \mathbb{R}$, $t^* \in \mathbb{T}$, as L tends to infinity for $\theta \in (0,1)$ it holds

$$\mathrm{BD}(\lfloor\theta L\rfloor;y,0) = \begin{cases} |m_1 - m_2|\frac{L-t^*}{L-\lfloor\theta L\rfloor}, & \lfloor\theta L\rfloor \le t^*, \\ |m_1 - m_2|\frac{t^*}{\lfloor\theta L\rfloor}, & \lfloor\theta L\rfloor > t^*. \end{cases}$$

It follows that

$$\mathrm{BD}(t^*;y,0) = \lim_{L\to\infty} \max_{\theta\in(0,1)} \mathrm{BD}(\lfloor\theta L\rfloor;y,0) = |m_1 - m_2|.$$

[3] Here we follow the terminology and notation of [BD00].

If a stationary segment between two change-points is too short, estimation of its boundaries can be difficult. Besides, $S(t; y)$ can be rather high for $t \in \mathbb{T} = \{0, 1, \ldots, L\}$ near to 0 or to L even if it is not a change-point (see Figure 4.2). To overcome these difficulties we introduce for a statistic S the *minimal length $\tau_{min}(S)$ of a (detectable) stationary segment* and estimate change-points by

$$\widehat{t^*}(y) = \arg\max_{t \in \mathbb{T}_0} S(t; y)$$

$$\text{with } \mathbb{T}_0 = \left\{ \frac{\tau_{\min}(S)}{2}, \frac{\tau_{\min}(S)}{2} + 1, \ldots, L - \frac{\tau_{\min}(S)}{2} \right\}. \tag{4.2}$$

For instance, looking on Figure 4.2 one may suggest to take $\tau_{\min}(\mathrm{BD}) \geq 50$. The choice of $\tau_{\min}(S)$ for the statistics used in this chapter will be discussed in Section 4.5.

To solve **Problem 2** one finds an estimate $\widehat{t^*}$ of a change-point (solves Problem 1) and then checks, whether the parameters of the stochastic process Y before and after $\widehat{t^*}$ are the same, by testing between the two following hypotheses [BN93, Subsection 1.1.2.2]:

H_0**:** parts $y(1), y(2), \ldots, y(\widehat{t^*})$ and $y(\widehat{t^*} + 1), \ldots, y(L)$ of a realization come from the same distribution;

H_A**:** parts $y(1), y(2), \ldots, y(\widehat{t^*})$ and $y(\widehat{t^*}+1), \ldots, y(L)$ of a realization come from different distributions.

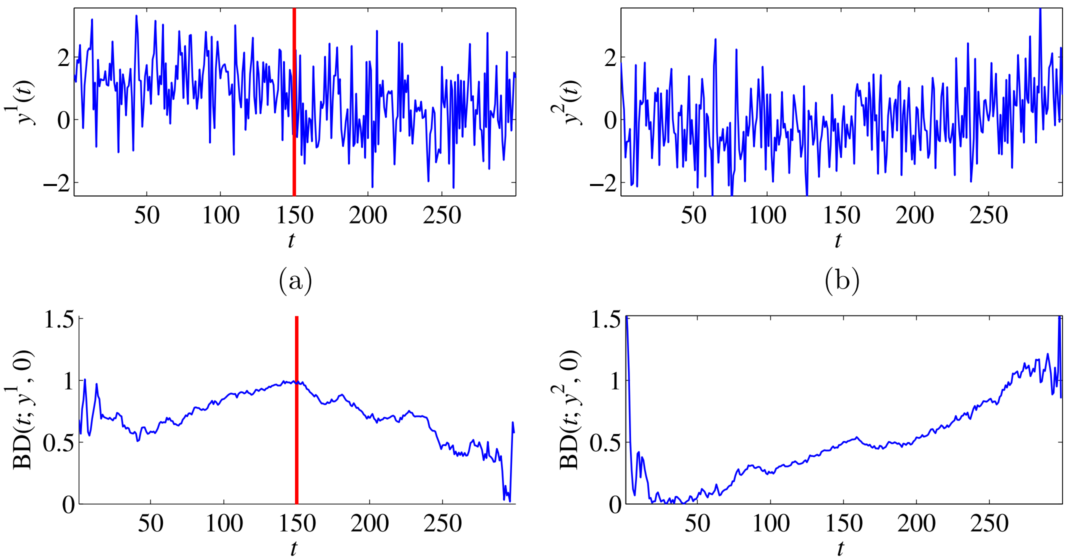
To perform the test one equips the statistic S with a threshold th_S such that if $S(\widehat{t^*}; y) \geq th_S$ then one accepts H_A, otherwise H_0 is chosen. The choice of the threshold is ambiguous: the empirical distributions of y before and after $\widehat{t^*}$ usually do not coincide even if y is stationary, and they can still differ not that much when $\widehat{t^*}$ is a change-point. The following example illustrates the ambiguity of the threshold selection for the statistic $\mathrm{BD}(t; y, 0)$ from Example 4.2.

Example 4.3. Consider a realization $y^1 = \big(y^1(t)\big)_{t \in \mathbb{T}}$ of the process $Y_{\mathrm{norm}}\big(t; (1.5, 0), 150\big)$ with change in mean (Figure 4.3a) and a realization $y^2 = \big(y^2(t)\big)_{t \in \mathbb{T}}$ of a stationary stochastic process $Y(t) = \epsilon(t)$ (Figure 4.3b). Values of the BD statistic given by (4.1) for y^1 and y^2 are shown in Figure 4.3c and 4.3d, respectively. In this case it holds

$$\max_{t \in \mathbb{T}_0} \mathrm{BD}(t; y^1, 0) < \max_{t \in \mathbb{T}_0} \mathrm{BD}(t; y^2, 0)$$

for $\mathbb{T}_0$ given by (4.2) with $\tau_{\min}(\mathrm{BD}) = 50$. That is any choice of a threshold leads either to the detection of a false change-point in the realization y^2 or to the non-detection of a change-point in the realization y^1. Even though changes in mean are generally easy to detect, one cannot choose a “perfect” threshold to distinguish realizations of processes with change-points and of stationary processes.

The higher th_S, the higher the possibility of false acceptance of the hypothesis H_0 (*false positive error*, [Faw06]) is; on the contrary, the lower th_S, the higher the



(a) (b)

(c) (d)

Figure 4.3: Realizations y^1 of a process $Y_{\text{norm}}\big(t; (1.5, 0), 150\big)$ with change-point marked by a vertical line (a) and y^2 of the standard additive white Gaussian noise $\epsilon(t)$ (b); statistics $\mathrm{BD}(t; y^1, 0)$ (c) and $\mathrm{BD}(t; y^2, 0)$ (d)

possibility of false acceptance of the hypothesis H_A. As it is usually done, consider the threshold th_S as a function of the desired probability α of false positive errors. Given N realizations $\upsilon^j = \big(\upsilon^j(1), \upsilon^j(2), \ldots, \upsilon^j(L)\big)$ of a stationary stochastic process for $j = 1, 2, \ldots, N$, $th_{\mathrm{S}}(\alpha)$ is determined by

$$\#\big\{j = 1, 2, \ldots, N \mid \max_{t \in \mathbb{T}_0} S(t; \upsilon^j) \geq th_{\mathrm{S}}(\alpha)\big\} = \lfloor \alpha N \rfloor \tag{4.3}$$

for $\mathbb{T}_0$ given by (4.2) and for sufficiently large N. A commonly-used approach for computing the threshold $th_S(\alpha)$ is bootstrapping (see [DH97] for an overview). The most simple bootstrapping technique is *resampling without replacement* (see [ST01] for a theoretical discussion of this technique and [Pol07, KMPS09] for its applications with detailed and clear explanations). To solve Problem 2 for the realization y one generates pseudo-stationary sequences $\upsilon^1, \upsilon^2, \ldots, \upsilon^N$ by shuffling elements of y, then one can compute $th_S(\alpha)$ directly by (4.3):

$$th_S(\alpha) = c_k \text{ for } k : \#\{j = 1, 2, \ldots, N \mid c_j \geq c_k\} = \lfloor \alpha N \rfloor, \tag{4.4}$$

where $c_j = \max\limits_{t \in \mathbb{T}_0} S(t; \upsilon^j)$ for all $j = 1, 2, \ldots, N$.

A simple and effective technique for solving **Problem 3** is the binary segmentation procedure introduced in [Vos81] (see [Lav99, LT07] for an alternative approach, which has more difficult implementation and is not considered here). The idea of binary segmentation is simple: one applies a single change-point detection procedure to the realization y; if a change-point is detected then it splits y into two segments. This procedure is repeated iteratively for the obtained segments until all of them either do not contain change-points or are shorter than $\tau_{\min}(S)$.

4.1.2 Structural change-points

The methods suggested in this chapter require less information than most of non-parametric methods. Roughly speaking, we restrict ourselves to the time series for those the future values depend on the past values, and changes occur in the evolution rule that links the past of the time series with its future. Such changes affect ordinal patterns, therefore we consider further the change-points described by the following definition.

Definition 4.2. Let $\big(Y(t)\big)_{t\in\mathbb{T}}$ be a piecewise stationary stochastic process with a change-point $t^* \in \mathbb{T}$. We say that t^* is a *structural change-point* if for the sub-processes $\big(Y(t)\big)_{t\in\{0,1,\dots,t^*\}}$ and $\big(Y(t)\big)_{t\in\{t^*,t^*+1,\dots,L\}}$ the distributions of ordinal patterns do not coincide.

This assumption is realistic for many time series, though not all change-points are structural. For instance, a change-point, where only the mean of a time series changes (Example 4.1), is not structural since mean is irrelevant for the distribution of ordinal patterns [Ami10, Subsection 3.4.3]. We illustrate Definition 4.2 by Examples 4.4 and 4.5; the processes, introduced there, are used throughout the chapter for empirical investigation of change-point detection methods.

Example 4.4. A *piecewise stationary noisy logistic* (NL) *process* for a given number of stationary segments $N_{\text{st}} \in \mathbb{N}$ and for change-points $t_1^*, t_2^*, \dots, t_{N_{\text{st}}-1}^* \in \mathbb{T}$ is defined by

$$\begin{aligned}\mathrm{NL}(t) &:= \mathrm{NL}\big(t;(r_1,r_2,\dots,r_{N_{\text{st}}}),(\sigma_1,\sigma_2,\dots,\sigma_{N_{\text{st}}}),(t_1^*,t_2^*,\dots,t_{N_{\text{st}}-1}^*)\big)\\ &= \begin{cases} f_{r_1}^t + \sigma_1\epsilon(t), & t\in\{0,1,\dots,t_1^*\}\\ f_{r_2}^{t-t_1^*}\circ f_{r_1}^{t_1^*} + \sigma_2\epsilon(t), & t\in\{t_1^*+1,\dots,t_2^*\}\\ \dots \\ f_{r_{N_{\text{st}}}}^{(t-t_{N_{\text{st}}-1}^*)}\circ f_{r_{N_{\text{st}}-1}}^{(t_{N_{\text{st}}-1}^*-t_{N_{\text{st}}-2}^*)}\circ\dots\circ f_{r_1}^{t_1^*} + \sigma_{N_{\text{st}}}\epsilon(t), & t\in\{t_{N_{\text{st}}-1}^*+1,\dots,L\}\end{cases}\end{aligned}$$

where $f_r : [0,1] \hookleftarrow$ is a logistic map, $r_1, r_2, \dots, r_{N_{\text{st}}} \in [1,4]$ are the values of control parameter, $\sigma_1, \sigma_2, \dots, \sigma_{N_{\text{st}}} > 0$ are the levels of noise. A realization of NL is, in fact, an orbit generated by a logistic map with a piecewise-constant control parameter and observed with a noise of piecewise-constant level.

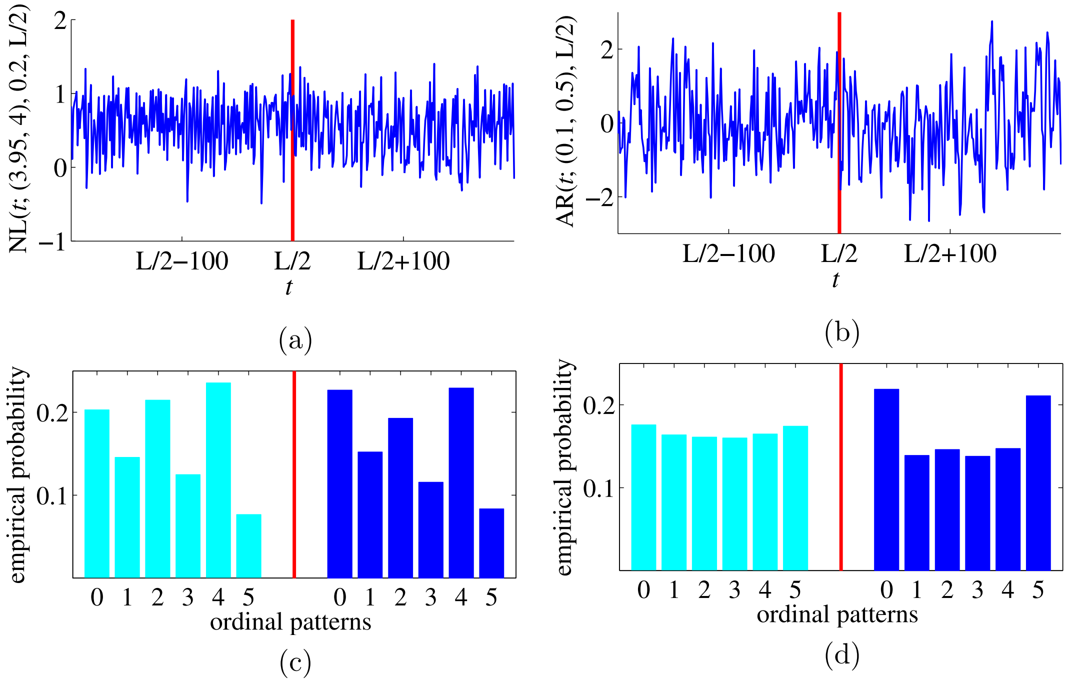
Figure 4.4a shows a realization of an NL process; as one can see in Figure 4.4c, the empirical distributions of ordinal patterns of order $d = 2$ before the change-point and after the change-point do not coincide. In general, as one can check, change-points in NL processes are reflected by distributions of ordinal patterns of order $d \geq 1$.

Example 4.5. The first order *piecewise stationary autoregressive* (AR) *process* is given by:

$$\mathrm{AR}(t) := \mathrm{AR}\big(t; (\phi_1, \phi_2, \ldots, \phi_{N_{\mathrm{st}}}), (t_1^*, t_2^*, \ldots, t_{N_{\mathrm{st}}-1}^*)\big),$$

$$\mathrm{AR}(t) = \begin{cases} \phi_1 \mathrm{AR}(t-1) + \epsilon(t), & t \in \{1, 2, \ldots, t_1^*\} \\ \phi_2 \mathrm{AR}(t-1) + \epsilon(t), & t \in \{t_1^* + 1, t_1^* + 2, \ldots, t_2^*\} \\ \ldots \\ \phi_{N_{\mathrm{st}}} \mathrm{AR}(t-1) + \epsilon(t), & t \in \{t_{N_{\mathrm{st}}-1}^* + 1, t_{N_{\mathrm{st}}-1}^* + 2, \ldots, L\}, \end{cases}$$

where $\phi_1, \phi_2, \ldots, \phi_{N_{\mathrm{st}}} \in [0, 1)$ are parameters of an autoregressive model, $\mathrm{AR}(0) = \epsilon(0)$. Figure 4.4b illustrates a realization of an AR process. Note that a change-point in an AR process is not change in mean since the expected value of an AR process is always zero (the simplest characteristic that reflects changes in AR process is the correlation function $\mathrm{corr}(\mathrm{AR}(t), \mathrm{AR}(t+1))$). From the results in [BS07, Section 5] it follows that the distributions of ordinal patterns of order $d \geq 2$ reflect change-points for the AR processes. Figure 4.4d illustrates this for the realization from Figure 4.4b: empirical distributions of ordinal patterns of order $d = 2$ before and after the change-point differ significantly.



(a)

(b)

(c)

(d)

Figure 4.4: Upper row: parts of realizations of an NL (a) and of an AR (b) process with change-points marked by vertical lines. Lower row: empirical distributions of ordinal patterns of order $d = 2$ before and after the change-point for the realizations of NL (c) and AR (d) process

The NL and AR processes have rather different ordinal patterns distributions. For this reason we use these processes for empirical investigation of methods for ordinal change-point detection in Sections 4.2, 4.3. When the positions of the change-points are given, we use a shorten notation: "AR, $\phi_1 \to \phi_2 \to \ldots \to \phi_{N_{\mathrm{st}}}$" for an AR process and "NL, $r_1 \to r_2 \to \ldots \to r_{N_{\mathrm{st}}}$, $\sigma = \sigma_1$" for an NL process when the level σ of noise is constant.

4.1.3 Ordinal change-point detection

The idea of ordinal change-point detection is to find structural change-points in a realization of a stochastic process $\big(Y(t)\big)_{t\in\mathbb{T}}$ by detecting changes in the sequence $\pi = \big(\pi(t)\big)_{t\in\mathbb{T}'}$ of ordinal patterns with $\mathbb{T}' = \{d, 1+d, \ldots, L\}$. We suggest ordinal-patterns-based solutions for the three problems of change-point detection formulated in Subsection 4.1.1.

To solve **Problem 1** we need to define an ordinal-patterns-based statistic $S(t;\pi)$ for $t \in \mathbb{T}'$ of the sequence π of ordinal patterns. If t^* is a change-point for Y, then

- $\pi(d), \pi(1+d), \ldots, \pi(t^*)$ characterize the process before the change;
- $\pi(t^*+1), \pi(t^*+2), \ldots, \pi(t^*+d-1)$ correspond to the transitional state;
- $\pi(t^*+d), \pi(t^*+1+d), \ldots, \pi(L)$ characterize the process after the change.

Therefore, $S(t;\pi)$ should measure dissimilarity between the distributions of ordinal patterns for $\pi(d), \pi(1+d), \ldots, \pi(t)$ and for $\pi(t+d), \pi(t+d), \ldots, \pi(L)$. Then an estimate of the change-point is given by

$$\widehat{t^*} = \arg\max_{t\in\mathbb{T}'_0} S(t;\pi) \tag{4.5}$$

with $\mathbb{T}'_0 = \left\{\frac{\tau_{\min}(S)}{2} + d, \frac{\tau_{\min}(S)}{2} + d + 1, \ldots, L - \frac{\tau_{\min}(S)}{2}\right\}$, where $\tau_{\min}(S)$ is the minimal length of a stationary segment for the statistic S. Ordinal-patterns-based statistics are considered in Section 4.2, the choice of $\tau_{\min}(S)$ is discussed in Section 4.5.

To solve **Problem 2** we estimate $\widehat{t^*}$ by (4.5) and then test between the hypotheses:

H_0**:** parts $\pi(d), \pi(1+d), \ldots, \pi(\widehat{t^*})$ and $\pi(\widehat{t^*}+d), \ldots, \pi(L)$ of sequence π come from the same distribution;

H_A**:** parts $\pi(d), \pi(1+d), \ldots, \pi(\widehat{t^*})$ and $\pi(\widehat{t^*}+d), \ldots, \pi(L)$ of sequence π come from different distributions.

As well as in the general case (Subsection 4.1.1), we perform this test by comparing $S\big(\widehat{t^*};\pi\big)$ with a threshold th_S. We formulate the solution of Problem 2 in Algorithm 1; there we use bootstrapping from the sequence of ordinal patterns π for computing the threshold th_S (details are provided in Section 4.5).

We solve **Problem 3** by using the binary segmentation procedure [Vos81], our algorithm for solving Problem 3 consists of two steps:

Algorithm 1 Solution of Problem 2

Input: sequence $\pi = \big(\pi(t_{\text{start}}), \ldots, \pi(t_{\text{end}})\big)$ of ordinal patterns, nominal probability α of false positive errors, statistic S, minimal length $\tau_{\min}(S)$ of a stationary segment

```
function PROBLEM2($\pi$, $\alpha$, $S$, $\tau_{\min}(S)$)
    if $t_{\text{end}} - t_{\text{start}} < \tau_{\min}(S)$ then
        return 0;        ▷ sequence is too short, no change-point can be detected
    end if
    $\widehat{t^*} \leftarrow \underset{t \in \{t_{\text{start}} + \frac{1}{2}\tau_{\min}(S), \ldots, t_{\text{end}} - \frac{1}{2}\tau_{\min}(S)\}}{\arg\max} S(t; \pi)$;        ▷ solve Problem 1 for $\pi$
    $th_S \leftarrow$ Bootstrapping($\alpha, \pi$);
    if $S(\widehat{t^*}; \pi) < th_S$ then
        return 0;
    else
        return $\widehat{t^*}$;
    end if
end function
```

Step 1: preliminary estimation of boundaries of the stationary segments with doubled nominal probability of false positive errors (that is with a higher risk of detecting false change-points).

Step 2: verification of the boundaries and rejection of false change-points: Problem 2 is solved for every two adjacent intervals.

Details of these two steps are displayed in Algorithm 2. Note that Step 1 is the usual binary segmentation procedure as suggested in [Vos81], while Step 2 is added to improve the obtained solution (the idea of such an improvement was suggested in [BDKS99]).

Definition 4.3. We call the segments of y *pseudo-stationary*[4] if they are bounded by the estimates $\widehat{t_0^*}, \widehat{t_1^*}, \ldots, \widehat{t^*_{\widehat{N}_{\text{st}}}}$ of change-points obtained by Algorithm 2.

4.1.4 Notation for change-point detection

Throughout this chapter we use the following notation.

- Y is a piecewise stationary stochastic process; all change-points of Y are supposed to be structural.
- $y := \big(y(t)\big)_{t \in \mathbb{T}}$ is a realization of Y for $\mathbb{T} = \{0, 1, \ldots, L\}$ with $L \in \mathbb{N}$.
- $\pi(y) := \big(\pi(t; y)\big)_{t \in \mathbb{T}'}$ for $\mathbb{T}' = \{d, 1 + d, \ldots, L\}$ is the sequence of ordinal patterns of order $d \in \mathbb{N}$ corresponding to y. For brevity, we denote the sequence simply by π when the origin of the sequence π of ordinal patterns is clear or unimportant.

[4] At least some of these segments may be non-stationary since there is no guaranty that Algorithm 2 provides detection of all change-points.

Algorithm 2 Solution of Problem 3

Input: sequence $\pi = \big(\pi(d), \pi(d+1), \ldots, \pi(L)\big)$ of ordinal patterns of order d, nominal probability α of false positive errors, statistic S, minimal length $\tau_{\min}(S)$ of a stationary segment.

```
function Problem3($\pi$, $\alpha$, $S$, $\tau_{\min}(S)$)
    $\widehat{N}_{\text{st}} \leftarrow 1$; $\widehat{t}^*_0 \leftarrow 0$; $\widehat{t}^*_1 \leftarrow L$; $i \leftarrow 0$          ▷ Step 1
    repeat
        $\widehat{t}^* \leftarrow$ Problem2$\Big(\big(\pi(\widehat{t}^*_i + d), \pi(\widehat{t}^*_i + d + 1), \ldots, \pi(\widehat{t}^*_{i+1})\big)$, $2\alpha$, $S$, $\tau_{\min}(S)\Big)$;
        if $\widehat{t}^* > 0$ then
            Insert $\widehat{t}^*$ to the list of change-points after $\widehat{t}^*_i$;
            $\widehat{N}_{\text{st}} \leftarrow \widehat{N}_{\text{st}} + 1$;
        else
            $i \leftarrow i + 1$;
        end if
    until $i < \widehat{N}_{\text{st}}$;
    $i \leftarrow 0$;          ▷ Step 2
    repeat
        $\widehat{t}^* \leftarrow$ Problem2$\Big(\big(\pi(\widehat{t}^*_i + d), \pi(\widehat{t}^*_i + d + 1), \ldots, \pi(\widehat{t}^*_{i+2})\big)$, $\alpha$, $S$, $\tau_{\min}(S)\Big)$;
        if $\widehat{t}^* > 0$ then
            $\widehat{t}^*_{i+1} \leftarrow \widehat{t}^*$;
            $i \leftarrow i + 1$;
        else
            Delete $\widehat{t}^*_{i+1}$ from the change-points list;
            $\widehat{N}_{\text{st}} \leftarrow \widehat{N}_{\text{st}} - 1$;
        end if
    until $i < \widehat{N}_{\text{st}} - 1$;
    return $\widehat{N}_{\text{st}}, \big(\widehat{t}^*_0, \widehat{t}^*_1, \ldots, \widehat{t}^*_{\widehat{N}_{\text{st}}}\big)$;
end function
```

- $S(t; y)$ stands for the statistic of a realization y as a function of time t. Ordinal-patterns-based statistics may be also denoted by $S(t; \pi)$.

- $t^* \in \mathbb{T}'$ stands for the change-point and $\widehat{t}^* \in \mathbb{T}'$ – for its estimate. In particular, $\widehat{t}^*(S; y)$ is an estimate of the change-point by the statistic S from the realization y. We omit S when it is clear from the context what statistic we use.

- $N_{\text{st}} \in \mathbb{N}$ stands for the number of stationary segments in a process Y, that it the number of change-points is $N_{\text{st}} - 1$. $\widehat{N}_{\text{st}}(S; y) \in \mathbb{N}$ is an estimate of N_{st} by the statistic S from the realization y.

- $\tau_{\min}(S)$ is the minimal length of a stationary segment for the statistic S.

- $th_S(\alpha, y)$ is the threshold for the statistic S computed by bootstrapping from y for the given probability α of false positive errors. We omit y when it is clear from the context or unimportant.

- $\epsilon = \big(\epsilon(t)\big)_{t \in \mathbb{T}}$ is the standard additive white Gaussian noise (see p. 32).

Some additional notation is required for the detection of changes in distributions of ordinal patterns (Subsection 4.2.1).

- $w_0, w_1, \ldots, w_M \in \mathbb{N}$ are adjacent boundaries of non-overlapping windows for computing distributions of ordinal patterns, M is the number of windows.
- W stands for the length of the sliding windows (if constant).
- $z_i(m) := z_i(m; y)$ is the relative frequency of an ordinal pattern i in m-th window for the realization y.
- $\mathbf{z}(m) := \big(z_0(m), z_1(m), \ldots, z_{(d+1)!-1}(m)\big)$.
- $\mathbf{z} := \big(\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(M)\big)$.
- $Z_i(m) := Z_i(m; y)$ is the absolute frequency of an ordinal pattern i in the m-th window for the realization y.
- $\mathbf{Z}(m) := \big(Z_0(m), Z_1(m), \ldots, Z_{(d+1)!-1}(m)\big)$.
- $\mathbf{Z} := \big(\mathbf{Z}(1), \mathbf{Z}(2), \ldots, \mathbf{Z}(M)\big)$.

4.2 Methods for ordinal change-point detection

In this section we consider an existing ordinal-patterns-based method for change-point detection and introduce three new methods. We discuss the solution of Problem 1, since for solving Problems 2 and 3 one uses Algorithms 1 and 2 described in Subsection 4.1.3. In Subsection 4.2.1 we consider statistics for detecting changes in the distributions of ordinal patterns. In particular, we consider the corrected maximum mean discrepancy introduced in [SKC13], and we suggest to use for ordinal change-point detection two classical statistics, the likelihood ratio and the χ^2 statistics. In Subsection 4.2.2 we introduce a new statistic CEofOP for ordinal change-point detection.

4.2.1 Detection of changes in distributions of ordinal patterns

In this subsection we consider the approach to ordinal change-point detection suggested in [SGK12]. The idea of this approach is to split the sequence $\pi(y)$ of ordinal patterns of order d into $M \in \mathbb{N}$ adjacent non-overlapping windows with the boundaries $w_0, w_1, \ldots, w_M \in \mathbb{N}$, such that $d = w_0 < w_1 < \ldots < w_M = L + 1$:

$$m\text{-th window}: \ \big(\pi(w_{m-1}; y), \pi(w_{m-1} + 1; y), \ldots, \pi(w_m - 1; y)\big), \tag{4.6}$$

and to consider the relative frequency of each ordinal pattern $i \in \{0, 1, \ldots, (d+1)! - 1\}$ in the m-th window:

$$z_i(m) = z_i(m; y) = \frac{\#\bigl\{t \in \{w_{m-1}, w_{m-1} + 1, \ldots, w_m - 1\} \mid \pi(t; y) = i\bigr\}}{w_m - w_{m-1}} \tag{4.7}$$

for $m \in \{1, 2, \ldots, M\}$. When the lengths of windows are sufficiently large, frequencies of ordinal patterns $z_i(m; y)$ for stationary y vary only slightly. Meanwhile, $z_i(m; y)$ usually changes drastically once there is a structural change in y (see Example 4.6). Sinn et al. [SGK12] suggested to use this fact for detecting structural[5] change-points, namely – for the estimation of the number of the window where the change-point occurs.

Example 4.6. Given a realization $\mathrm{nl}(t)$ of the process $\mathrm{NL}\bigl(t; (3.95, 4), (0.2, 0.2), 2 \cdot 10^4\bigr)$ (Figure 4.5a), consider the relative frequencies $z_i(m; \mathrm{nl})$ of ordinal patterns of order $d = 3$ in windows with the boundaries $w_m = d + 256m$ for $m = 1, 2, \ldots, M$. To visualize the frequencies we draw in Figure 4.5c curves $\eta_j(m) = \sum\limits_{i=0}^{j} z_i(m; \mathrm{nl})$ for $j = 1, 2, \ldots, (d+1)! - 1$. For any window number m the space between the bottom line and the first curve represents the relative frequency $z_0(m; \mathrm{nl})$ of 0-th ordinal pattern, the space between the first and the second curve represents $z_1(m; \mathrm{nl})$, and so on, the space between the upper curve and the top line represents $z_{(d+1)!-1}(m; \mathrm{nl})$. In the same manner, Figure 4.5d illustrates the frequencies of ordinal patterns ($d = 3$, $W = 256$) for a realization (Figure 4.5b) of $\mathrm{AR}\bigl(t; (0.1, 0.5), t^*\bigr)$. For both realizations, the frequencies before and after the change-point differ significantly, while they vary only slightly on stationary segments.

4.2.1.1 Ordinal change-point detection via maximum mean discrepancy

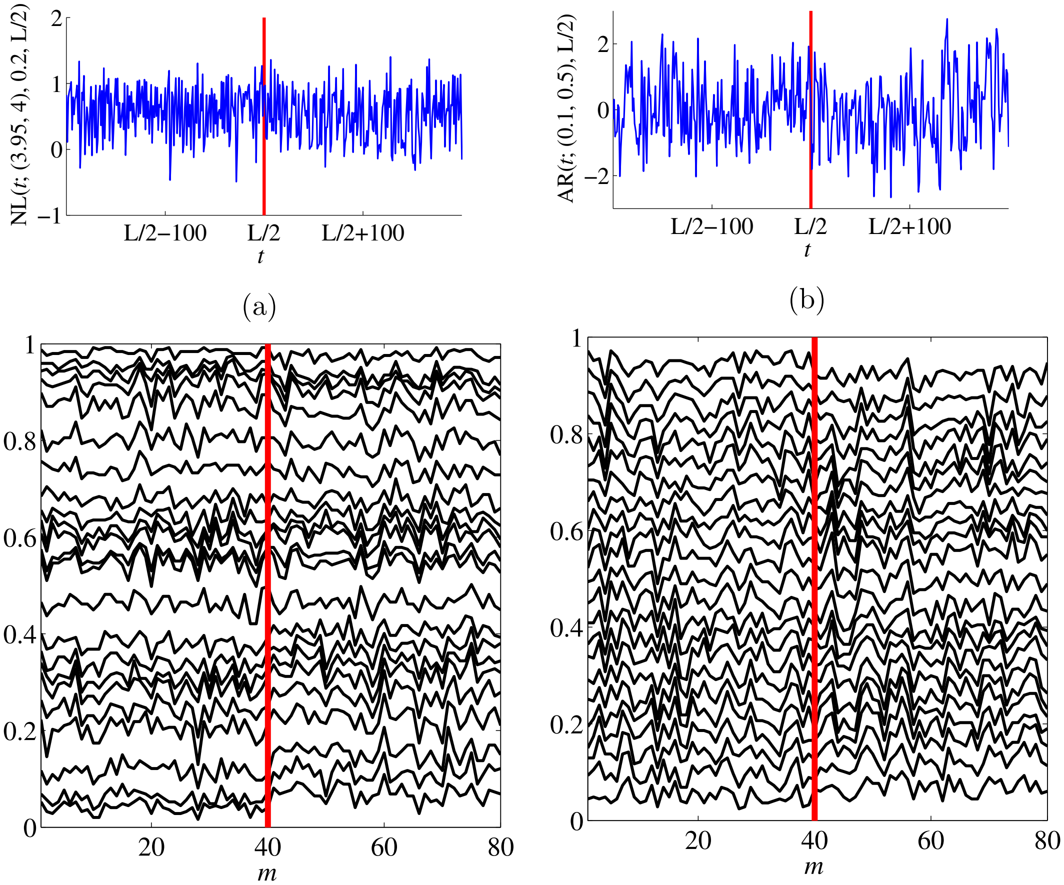
Here we provide a detailed description of the method for ordinal change-point detection introduced in [SGK12, SKC13] in order to simplify its comparison with the methods for change-point detection suggested in this chapter; in Subsection 4.5.1 we propose some improvements of this method.

Let a structural change-point occur inside the m^*-th window for $m^* \in \{1, 2, \ldots, M\}$. Then for $m_1, m_2 \in \{1, 2, \ldots, M\}$ the distributions of ordinal patterns $\mathbf{z}(m_1)$ and $\mathbf{z}(m_2)$ are similar if either $m_1 < m_2 < m^*$ or $m^* < m_1 < m_2$, and differ if $m_1 < m^* < m_2$. Estimation of m^* on the basis of this property is provided by a statistic called *maximum mean discrepancy* (MMD) introduced in [GBR$^+$07] (for theoretical details see also [GBR$^+$12]):

$$\mathrm{MMD}(m; \mathbf{z}) = \left(\frac{K_1(m; \mathbf{z})}{m^2} - \frac{2K_2(m; \mathbf{z})}{m(M-m)} + \frac{K_3(m; \mathbf{z})}{(M-m)^2} \right)^{\frac{1}{2}},$$

for $m = 1, 2, \ldots, M - 1$, where $K_1(m; \mathbf{z})$ and $K_3(m; \mathbf{z})$ characterize dissimilarities within the sets of vectors $\bigl\{\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(m)\bigr\}$ and $\bigl\{\mathbf{z}(m+1), \mathbf{z}(m+2), \ldots, \mathbf{z}(M)\bigr\}$,

[5] In fact, in [SGK12] it is assumed that change-points are structural, though a different terminology is used.



(a)

(b)

(c)

(d)

Figure 4.5: Upper row: parts of realizations of NL process (a) and AR process (b) with change-points marked by vertical lines. Lower row: relative frequencies of ordinal patterns for the realization of NL process (c) and of AR process (d)

respectively, while $K_2(m;\mathbf{z})$ represents dissimilarity between these two sets [SKC13]. These dissimilarities for $m = 1, 2, \ldots, M-1$ are given by

$$K_1(m;\mathbf{z}) = \sum_{m_1=1}^{m} \sum_{m_2=1}^{m} k\big(\mathbf{z}(m_1), \mathbf{z}(m_2)\big),$$

$$K_2(m;\mathbf{z}) = \sum_{m_1=1}^{m} \sum_{m_2=m+1}^{M} k\big(\mathbf{z}(m_1), \mathbf{z}(m_2)\big),$$

$$K_3(m;\mathbf{z}) = \sum_{m_1=m+1}^{M} \sum_{m_2=m+1}^{M} k\big(\mathbf{z}(m_1), \mathbf{z}(m_2)\big),$$

where $k\big(\mathbf{z}(m_1), \mathbf{z}(m_2)\big)$ is a measure of dissimilarity between $\mathbf{z}(m_1)$ and $\mathbf{z}(m_2)$ (for details see [GBR$^+$12]). Various choices of k are possible [GFHS09], in [SGK12] the Radial Basis Function kernel [VTS04] is used:

$$k\big(\mathbf{z}(m_1), \mathbf{z}(m_2)\big) = \exp\left(- \sum_{i=0}^{(d+1)!-1} \big(z_i(m_1) - z_i(m_2)\big)^2 \right).$$

The MMD statistic is undefined for $m = M$, thus for technical reasons we set $\mathrm{MMD}(M; \mathbf{z}) = \min\limits_{m \in \{1,\dots,M-1\}} \mathrm{MMD}(m; \mathbf{z})$.

The number m^* of the window containing the structural change-point is estimated by

$$\widehat{m}^*(\mathbf{z}) = \operatorname*{arg\,max}_{m \in \{1,2,\dots,M\}} \mathrm{MMD}(m; \mathbf{z}).$$

However, $\mathrm{MMD}(m; \mathbf{z})$ is overestimated for the values of m near to 1 and to $M - 1$ [SGK12]. In order to overcome this difficulty, Sinn et al. introduced the *corrected maximum mean discrepancy* CMMD [SKC13]:

$$\mathrm{CMMD}(m; \mathbf{z}) = \mathrm{MMD}(m; \mathbf{z}) - \left(\frac{M-1}{m(M-m)} \max_{j=1,2,\dots,M} \mathrm{MMD}(j; \mathbf{z}) \right)^{\frac{1}{2}}. \tag{4.8}$$

Remark. Note that a modified version of the CMMD statistic (mCMMD) is introduced in [SGK12]:

$$\mathrm{mCMMD}(m; \mathbf{z}) = \mathrm{MMD}(m; \mathbf{z}) - \left(\frac{M-1}{m(M-m)} \max_{j=1,2,\dots,M} \mathrm{MMD}(j; \mathbf{z}) \right). \tag{4.9}$$

This statistic has no clear theoretical justification, so we do not discuss it here. However, in certain cases mCMMD provides a better estimation of the window containing a structural change-point than the CMMD statistic[6]; therefore we present some empirical results related to mCMMD in Subsection 4.5.1.1.

Figure 4.6 shows the CMMD statistic for NL and AR processes, order $d = 3$ of ordinal patterns is used. Windows are defined by $w_m = d + 256m$ for $m = 0, 1, \dots, M$; change-points are marked by vertical lines. One can see that the CMMD statistic has two noticeable drawbacks:

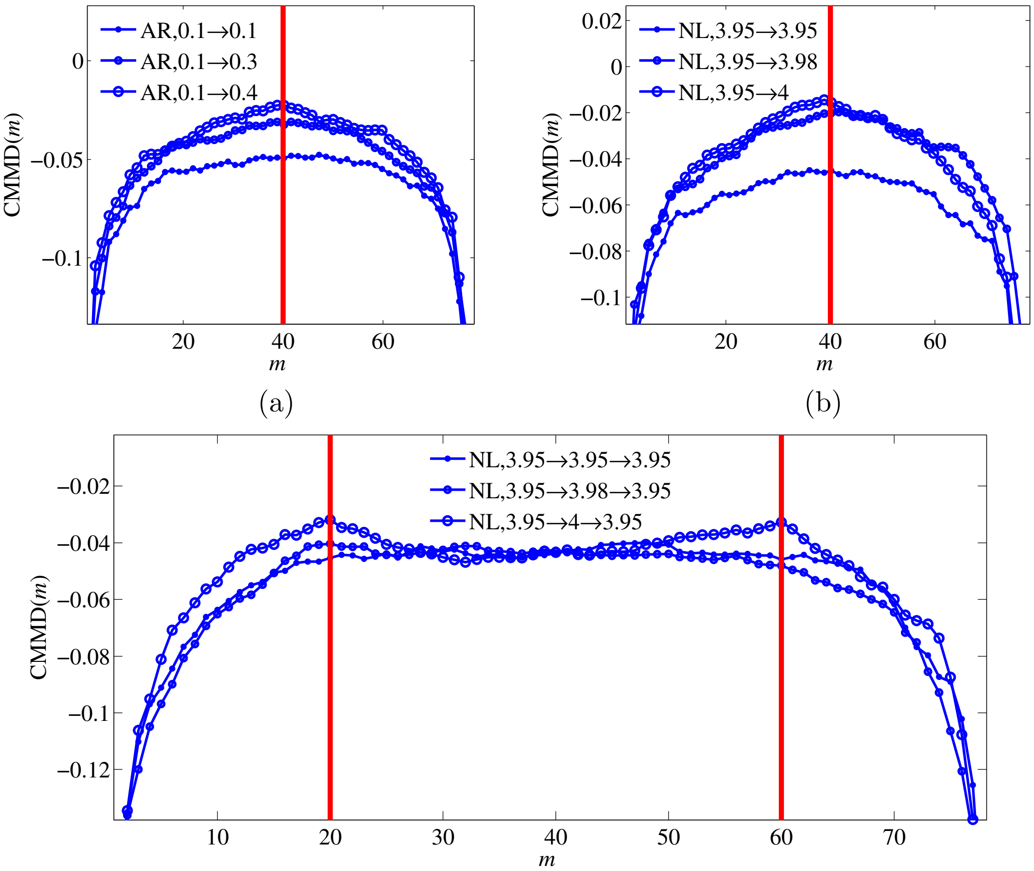
- maximums of CMMD do not always coincide with locations of change-points (see, for instance, Figure 4.6b), which is disadvantageous in view of Problem 1.
- the CMMD statistic does not provide a clear distinction between the processes with several changes and without changes (see Figure 4.6c), which is disadvantageous in view of Problem 3.

In the original papers [SGK12, SKC13] authors do not estimate a change-point, but only a number of the window containing it:

$$\widehat{m}^* = \widehat{m}^*(\mathbf{z}) = \operatorname*{arg\,max}_{m \in \{1,2,\dots,M\}} \mathrm{CMMD}(m; \mathbf{z}).$$

Estimating the number of the window containing the change-point is only a part of Problem 1. In order to have a complete method for the change-points detection via CMMD for a comparison with other methods, we provide solutions of Problems 1–2 using MMD and CMMD statistics in Subsection 4.5.1.

[6] In particular, the results of the experiments described in [SGK12, Section 4.1] as the empirical justification of the CMMD statistic, are obtained for the mCMMD statistic. In those conditions the CMMD statistic provides much worse results, see Subsection 4.5.1.1 for details.



(a)

(b)

(c)

Figure 4.6: CMMD for AR processes with one change-point (a), for NL processes with one change-point (b), and with two change-points (c) for different values of control parameters. For NL processes the constant level of noise $\sigma = 0.2$ is used

4.2.1.2 Ordinal change-point detection via likelihood ratio statistic and χ^2-statistic

Here we suggest to use two classical statistics for detecting changes in distributions of ordinal patterns for the non-overlapping windows (4.6) with the boundaries $w_0, w_1, \ldots, w_M \in \mathbb{N}$. Both statistics are computed from the absolute frequencies of ordinal patterns in the m-th window given by

$$Z_i(m) = \#\big\{t \in \{w_{m-1}, w_{m-1}+1, \ldots, w_m - 1\} \mid \pi(t) = i\big\}$$

for $m \in \{1, 2, \ldots, M\}$. Denote the vector of absolute frequencies in the m-th window by $\mathbf{Z}(m) = \big(Z_i(m)\big)_{i=0}^{(d+1)!-1}$ and the sequence of these vectors by $\mathbf{Z} = \big(\mathbf{Z}(1), \mathbf{Z}(2), \ldots, \mathbf{Z}(M)\big)$. Detection of a single change-point in $\mathbf{Z}$ can be considered as testing between the following hypotheses:

H_0**:** vectors $\mathbf{Z}(1), \mathbf{Z}(2), \ldots, \mathbf{Z}(M)$ come from the same distribution;

H_m**:** vectors $\mathbf{Z}(1), \mathbf{Z}(2), \ldots, \mathbf{Z}(m)$ and $\mathbf{Z}(m+1), \ldots, \mathbf{Z}(M)$ come from different distributions.

A basic statistic for testing between these hypotheses is the *likelihood ratio* statistic [BN93, Subsection 2.2.3]:

$$\mathrm{LR}_0(m;\mathbf{Z}) = -2\ln\frac{\mathrm{Lkl}\big(H_0 \mid \mathbf{Z}\big)}{\mathrm{Lkl}\big(H_m \mid \mathbf{Z}\big)} = -2\ln\mathrm{Lkl}\big(H_0 \mid \mathbf{Z}\big) + 2\ln\mathrm{Lkl}\big(H_m \mid \mathbf{Z}\big),$$

where $\mathrm{Lkl}\big(H \mid \mathbf{Z}\big)$ is the likelihood of the hypothesis H given a sequence $\mathbf{Z}$. Assume that the absolute frequencies of ordinal patterns in windows are multinomial independent random variables (we have no theoretical justification for this). Then $\mathrm{LR}_0(m;\mathbf{Z})$ is given by (see [HS95])

$$\mathrm{LR}_0(m;\mathbf{Z}) = 2\sum_{i=0}^{(d+1)!-1}\left(P_i(m)\ln\frac{P_i(m)}{w_m} + Q_i(m)\ln\frac{Q_i(m)}{v_m} - P_i(M)\ln\frac{P_i(M)}{w_M}\right),$$

where $v_m = w_M - w_m$, $P_i(m) = \sum\limits_{j=1}^{m} Z_i(j)$ represents frequency of the ordinal pattern i before the m-th window and $Q_i(m) = \sum\limits_{j=m+1}^{M} Z_i(j)$ – after it.

Another statistic for testing between H_0 and H_m is the χ^2*-statistic*. For a sequence of multinomial independent random variables it was introduced in [HS95] (for theoretical details see also [BHM$^+$13]):

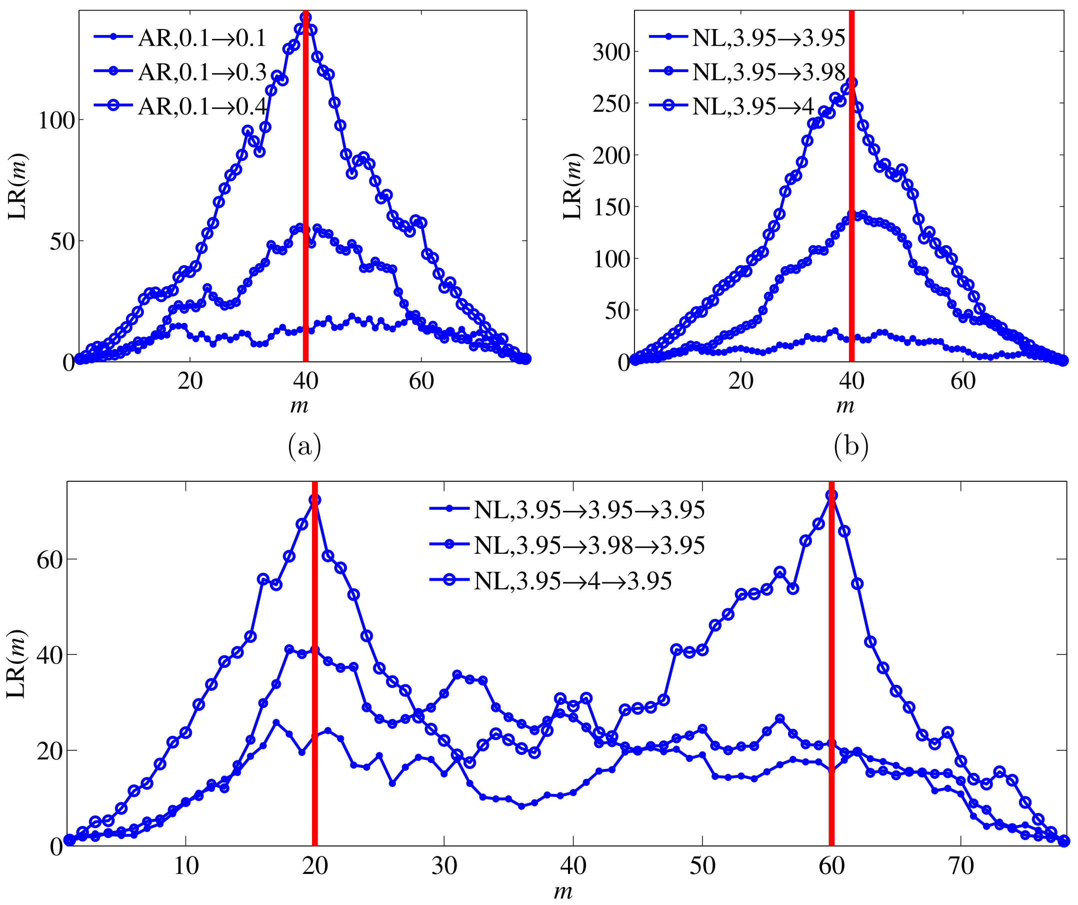
$$\begin{aligned}\mathrm{Chi}_0(m;\mathbf{Z}) &= w_m \sum_{i=0}^{(d+1)!-1}\frac{\left(\frac{P_i(M)}{w_M} - \frac{P_i(m)}{w_m}\right)^2}{\frac{P_i(m)}{w_m}} + v_m\sum_{i=0}^{(d+1)!-1}\frac{\left(\frac{P_i(M)}{w_M} - \frac{Q_i(m)}{v_m}\right)^2}{\frac{Q_i(m)}{v_m}}\\ &= \sum_{i=0}^{(d+1)!-1}\left(\frac{P_i(M)}{w_M}\right)^2\left(\frac{(w_m)^2}{P_i(m)} + \frac{(v_m)^2}{Q_i(m)}\right) - w_M\\ &= \sum_{i=0}^{(d+1)!-1}\frac{\big(P_i(m)w_M - P_i(M)w_m\big)^2}{P_i(M)w_m v_m}.\end{aligned}$$

On the basis of the results in [BHM$^+$13], we suggest to use for detecting structural change-points the following modified versions of the likelihood ratio (LR) and χ^2 (Chi) statistics:

$$\begin{aligned}\mathrm{LR}(m;\mathbf{Z}) &= \frac{w_m v_m}{w_M^2}\mathrm{LR}_0(m;\mathbf{Z})\\ &= 2\frac{w_m v_m}{w_M^2}\sum_{i=0}^{(d+1)!-1}\left(P_i(m)\ln\frac{P_i(m)}{w_m} + Q_i(m)\ln\frac{Q_i(m)}{v_m} - P_i(M)\ln\frac{P_i(M)}{w_M}\right),\end{aligned}$$

$$\mathrm{Chi}(m;\mathbf{Z}) = \frac{w_m v_m}{w_M^2}\mathrm{Chi}_0(m;\mathbf{Z}) = \sum_{i=0}^{(d+1)!-1}\frac{(P_i(m)w_M - P_i(M)w_m)^2}{P_i(M)w_M^2}.$$

Figure 4.7 shows the values of $\mathrm{LR}(m;\mathbf{Z})$ for various processes (the behavior of $\mathrm{Chi}(m;\mathbf{Z})$ is similar). Note that change-points are indicated more clearly than by the CMMD statistic (cf. Figure 4.6). Details related to the solution of Problems 1 and 2 for LR and Chi statistics are addressed in Subsection 4.5.1.



(a)

(b)

(c)

Figure 4.7: LR statistic for AR processes with one change-point (a), for NL processes with one change-point (b), and with two change-points (c) for different values of control parameters. For NL processes the constant level of noise $\sigma = 0.2$ is used

4.2.2 Change-point detection via the CEofOP statistic

Here we introduce a new method for ordinal change-point detection on the basis of the empirical conditional entropy of ordinal patterns defined in Section 3.4. For a given order $d \in \mathbb{N}$, let us consider the empirical conditional entropy as a function of the sequence π of ordinal patterns:

$$\widehat{h}_{\text{cond}}\big(\pi(d), \ldots, \pi(t)\big) = \frac{1}{t-d} \sum_{i=0}^{(d+1)!-1} \sum_{j=0}^{(d+1)!-1} n_{i,j}(t)\big(\ln n_i(t) - \ln n_{i,j}(t)\big),$$

$$\text{where } n_i(t) = \#\{r \in \{d, 1+d, \ldots, t-1\} \mid \pi(r) = i\},$$

$$n_{i,j}(t) = \#\{r \in \{d, 1+d, \ldots, t-1\} \mid \pi(r) = i, \pi(r+1) = j\}.$$

We suggest to use the following statistic for detecting structural change-points

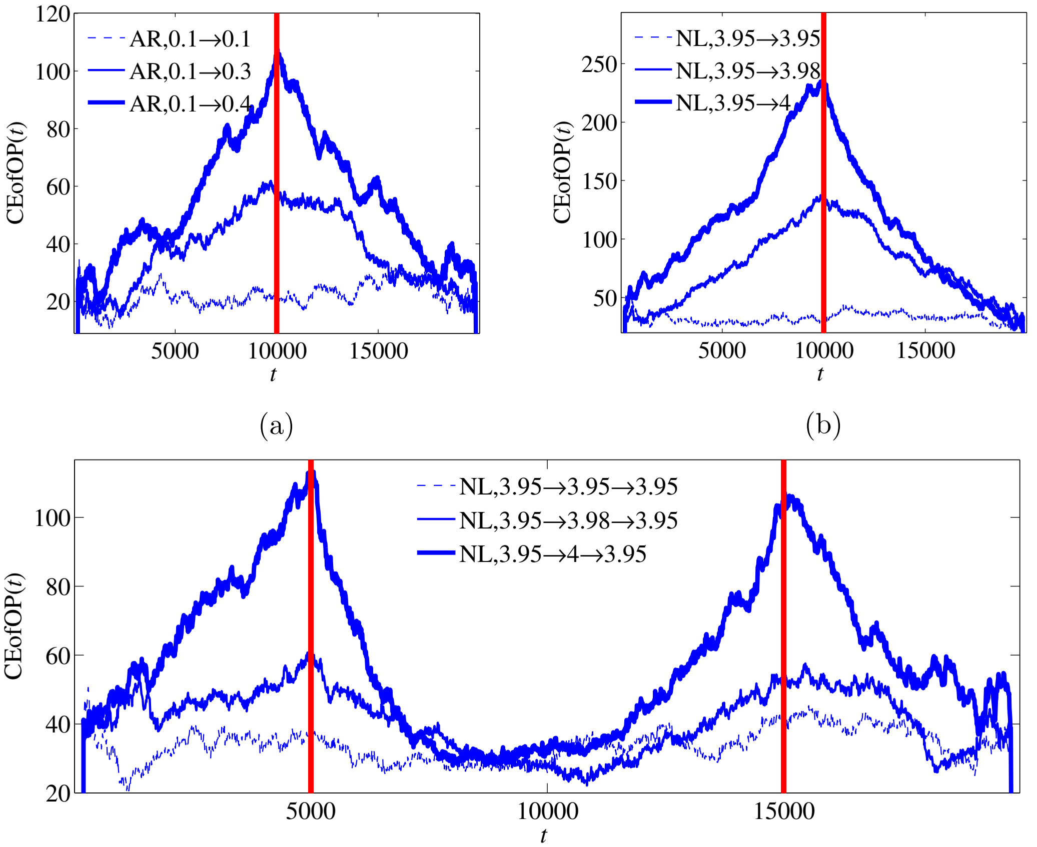
$$\begin{aligned}\text{CEofOP}(t;\pi) = (L-2d)\,\widehat{h}_{\text{cond}}\big(\pi(d), \ldots, \pi(L)\big) - (t-d)\,\widehat{h}_{\text{cond}}\big(\pi(d), \ldots, \pi(t)\big) \\ - \big(L-(t+d)\big)\,\widehat{h}_{\text{cond}}\big(\pi(t+d), \ldots, \pi(L)\big).\end{aligned} \tag{4.10}$$

Indeed, the conditional entropy is concave (not only for ordinal patterns but in general, see [HK02, Subsection 2.1.3]), thus for all $t \in \mathbb{T}'$ it holds

$$\begin{aligned} \widehat{h}_{\text{cond}}\big(\pi(d), \pi(1+d), \ldots, \pi(L)\big) \geq {} & \frac{t-d}{L-2d}\, \widehat{h}_{\text{cond}}\big(\pi(d), \pi(1+d), \ldots, \pi(t)\big) \\ & + \frac{L-(t+d)}{L-2d}\, \widehat{h}_{\text{cond}}\big(\pi(t+d), \ldots, \pi(L)\big). \end{aligned} \tag{4.11}$$

Therefore if probabilities of ordinal patterns change at some point t^*, then $\text{CEofOP}\big(t;\pi\big)$ tends to attain its maximum at $t = t^*$. If the sequence π is stationary, then for L being sufficiently large, (4.11) holds with equality (see p. 106, Theorem 4.3).

In contrast to the statistics considered in Subsection 4.2.1, CEofOP does not estimate the window, where the change-point is located, but provides an immediate estimate of the change-point t^*. Figure 4.8 shows that for the NL processes CEofOP indicates change-points as clear as the LR statistic, but provides a better distinction between the cases with change and without change (cf. Figures 4.7c and 4.8c).



(a)

(b)

(c)

Figure 4.8: CEofOP for AR processes with one change-point (a), for NL processes with one change-point (b), and with two change-points (c) for different values of control parameters. For NL processes the constant level of noise $\sigma = 0.2$ is used

Properties of the CEofOP statistic are discussed in Section 4.4, details related to the application of CEofOP for solving Problems 1 and 2 are provided in Subsection 4.5.2.

4.3 Change-point detection for the noisy logistic and autoregressive processes: simulation studies

In this section we investigate effectiveness of the methods for change-point detection that were considered in Section 4.2:

- three ordinal-patterns-based methods for detecting changes in distributions of ordinal patterns (via MMD[7], LR and Chi statistics);
- the ordinal-patterns-based method for change-point detection via the CEofOP statistic;
- two versions of the classical Brodsky-Darkhovsky method [BDKS99] on the basis of statistic (4.1). The first version is intended to detect changes in mean and uses statistic (4.1) directly (see Example 4.3):

  $$\mathrm{BD}^{\mathrm{exp}}(t; y, \delta) = \mathrm{BD}(t; y, \delta).$$

  The second version of the Brodsky-Darkhovsky method detects changes in the correlation function $\mathrm{corr}(y(t), y(t+1))$:

  $$\mathrm{BD}^{\mathrm{corr}}(t; y, \delta) = \mathrm{BD}(t; \upsilon, \delta) \text{ with } \upsilon(t) = y(t)y(t+1) \text{ for all } t \in \mathbb{T} \setminus \{L\}.$$

  The mean is just the basic characteristic, while the correlation function reflects relations between the future and the past of a time series and seems to be a natural choice for detecting structural change-points.

We carry out experiments on realizations of the NL and AR piecewise stationary stochastic processes (see Subsection 4.1.3) for order $d = 3$ and window size $W = 256$[8]. Methods for ordinal change-point detection are implemented according to the general algorithms presented in Subsection 4.1.3, technical details are discussed in Section 4.5. The Brodsky-Darkhovsky method is implemented according to [BDKS99] with the only exception: to compute a threshold th_{BD} we use bootstrapping, which in our case provided better results than the technique described in [BDKS99].

4.3.1 Performance for Problem 1

In this subsection we study how well the methods for change-point detection estimate the position of a single change-point (Problem 1). First we consider Problem 1 for realizations of the processes with fixed length (Experiment 4.1). Second, since we expect that the performance of ordinal-patterns-based methods for change-point detection may strongly depend on the length of realization, we check this in Experiment 4.2.

[7] In fact we use the MMD statistic in combination with CMMD, see Subsection 4.5.1 for details

[8] We use $W = 256$ just because it is convenient from the computational viewpoint; this length of the window is also sufficient for estimating frequencies of ordinal patterns of order $d = 3$ inside the windows, since $256 > 120 = 5(d+1)!$ [Ami10, Section 9.3]. Results of the experiments remain almost the same for $200 \leq W \leq 1000$.

Experiment 4.1: comparing performance of methods for change-point detection with respect to Problem 1.

Objects: $N_T = 10000$ realizations $y^j = \big(y^j(t)\big)_{t\in\{0,1,\dots,L\}}$ for $j = 1, 2, \dots, N_T$ of processes listed in Table 4.1. A single change occurs at a random time t^* uniformly distributed in $\big\{\frac{L}{4} - W, \frac{L}{4} - W + 1, \dots, \frac{L}{4} + W\big\}$. For all processes, length $L = 80\,W$ is taken.

Short name	Complete designation
NL, $3.95 \to 3.98$, $\sigma = 0.2$	$\mathrm{NL}\big(t; (3.95, 3.98), (0.2, 0.2), t^*\big)$
NL, $3.95 \to 3.80$, $\sigma = 0.3$	$\mathrm{NL}\big(t; (3.95, 3.80), (0.3, 0.3), t^*\big)$
NL, $3.95 \to 4.00$, $\sigma = 0.2$	$\mathrm{NL}\big(t; (3.95, 4.00), (0.2, 0.2), t^*\big)$
AR, $0.1 \to 0.3$	$\mathrm{AR}\big(t; (0.1, 0.3), t^*\big)$
AR, $0.1 \to 0.4$	$\mathrm{AR}\big(t; (0.1, 0.4), t^*\big)$
AR, $0.1 \to 0.5$	$\mathrm{AR}\big(t; (0.1, 0.5), t^*\big)$

Table 4.1: Processes used in Experiment 4.1

Technique. Problem 1 consists in estimation of the position of a change-point t^*, so performance of a change-point detection for Problem 1 is characterized by the accuracy of this estimation. The error of the change-point estimation provided by method S for the j-th realization of process Y is given by

$$\mathrm{err}^j(S, Y) = \Big(\widehat{t^*}(S; y^j) - t^*\Big),$$

where t^* is the actual position of the change-point and $\widehat{t^*}(S; y^j)$ is its estimate by S. To measure the overall accuracy of a method for change-point detection via statistic S as applied to the process Y we use three quantities:

- the *fraction of satisfactory estimated change-points* sE:

  $$\mathrm{sE}(S, Y) = \frac{\#\big\{j \in \{1, 2, \dots, N_T\} : |\mathrm{err}^j(S, Y)| \le \mathrm{MaxErr}\big\}}{N_T},$$

  where MaxErr is the maximal satisfactory error, we take $\mathrm{MaxErr} = W = 256$;
- the *bias* $\mathrm{B}(S, Y) = \dfrac{1}{N_T} \sum\limits_{j=1}^{N_T} \mathrm{err}^j(S, Y)$.
- the *root mean squared error* $\mathrm{RMSE}(S, Y) = \sqrt{\dfrac{1}{N_T} \sum\limits_{j=1}^{N_T} \big(\mathrm{err}^j(S, Y)\big)^2}$.

The larger sE is and the more near to zero the bias and the RMSE are, the better the solution of Problem 1.

Results are presented in Tables 4.2, 4.3 for NL and AR processes, respectively. For every process the values of performance measures that are the best among the ordinal-patterns-based statistics are shown in **bold**. If the value obtained for a version of the Brodsky-Darkhovsky method is better, it is also shown in **bold**.

Statistic	NL, 3.95 → 3.98 $\sigma = 0.2$			NL, 3.95 → 3.80 $\sigma = 0.3$			NL, 3.95 → 4.00 $\sigma = 0.2$		
	sE	B	RMSE	sE	B	RMSE	sE	B	RMSE
MMD	0.34	698	1653	0.50	-51	306	0.68	**-13**	206
LR	0.40	468	1262	0.62	53	353	0.77	61	238
Chi	0.41	391	1174	0.62	52	351	0.78	22	179
CEofOP	**0.61**	**53**	**397**	**0.65**	**1**	**256**	**0.88**	20	**99**
$\mathrm{BD}^{\mathrm{exp}}$	**0.62**	78	**351**	**0.78**	-6	**145**	**0.89**	43	**96**
$\mathrm{BD}^{\mathrm{corr}}$	0.44	85	656	0.71	13	202	0.77	43	189

Table 4.2: Performance of different statistics for Problem 1, NL processes

Statistic	AR, 0.1 → 0.3			AR, 0.1 → 0.4			AR, 0.1 → 0.5		
	sE	B	RMSE	sE	B	RMSE	sE	B	RMSE
MMD	0.32	616	1626	0.54	-14	368	0.68	-48	184
LR	0.37	447	**1342**	0.62	62	389	0.78	17	166
Chi	0.37	448	1354	0.62	56	386	0.78	12	164
CEofOP	**0.39**	**126**	1838	**0.68**	**0**	**234**	**0.86**	**0**	**110**
$\mathrm{BD}^{\mathrm{exp}}$	0.00	8942	11757	0.00	10535	12372	0.00	11686	12791
$\mathrm{BD}^{\mathrm{corr}}$	**0.79**	**31**	**151**	**0.92**	21	**73**	**0.97**	21	**50**

Table 4.3: Performance of different statistics for Problem 1, AR processes

Discussion and conclusions:

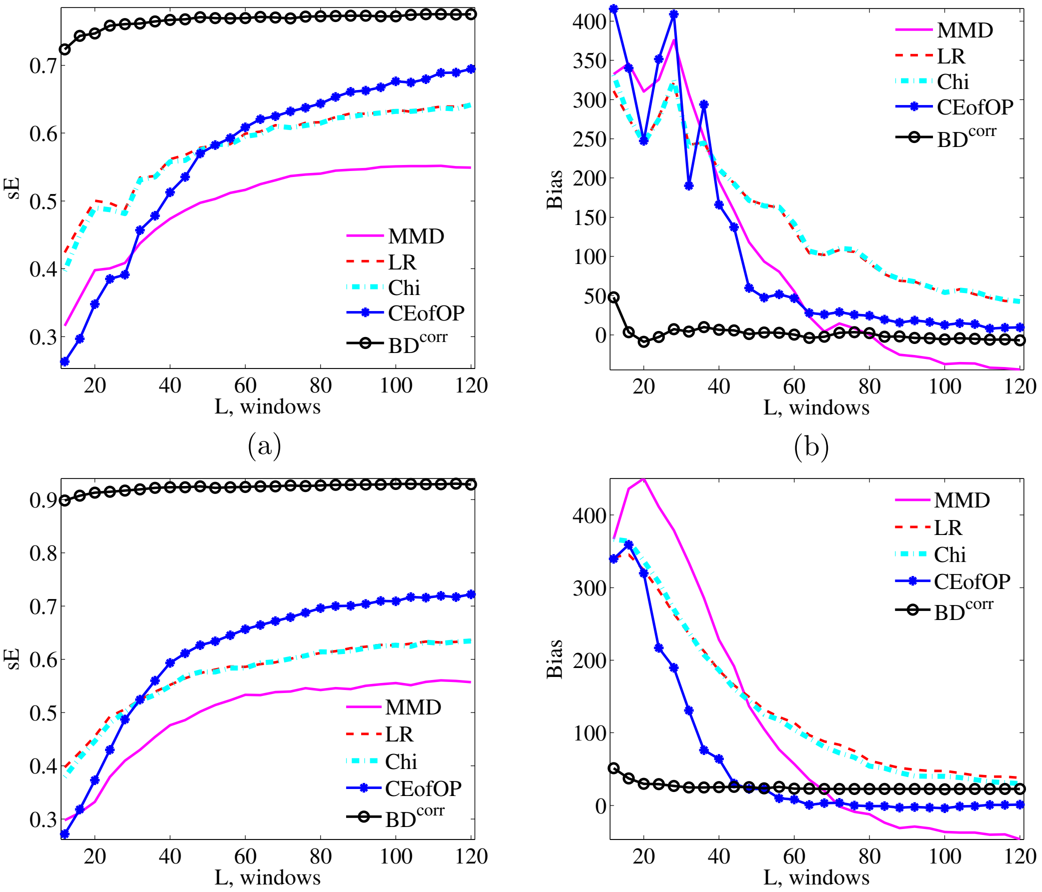
1. For the considered processes the change-point detection via the CEofOP statistic shows better performance than other ordinal-patterns-based methods. For the NL processes the CEofOP statistic has almost the same performance as the classical Brodsky-Darkhovsky method; for the AR processes performance of the classical method is better, though CEofOP has lower bias.

2. In contrast to the ordinal-patterns-based methods, the Brodsky-Darkhovsky method is unreliable when there is lack of a priori information about the time series. For instance, changes in NL processes only slightly influence the correlation function and $\mathrm{BD}^{\mathrm{corr}}$ indicated the changes not very well (cf. performance of $\mathrm{BD}^{\mathrm{corr}}$ and CEofOP in Table 4.2). Meanwhile, changes in the AR processes do not influence the expected value (see Example 4.5), which does not allow to detect them using $\mathrm{BD}^{\mathrm{exp}}$ (see Table 4.3). Therefore we do not consider the $\mathrm{BD}^{\mathrm{exp}}$ statistic in further experiments.

Experiment 4.2: studying the dependence of the change-point detection performance with respect to Problem 1 on the length L of a time series.

Objects: $N_T = 10000$ realizations of NL, 3.95 → 3.80, $\sigma = 0.3$ and AR, 0.1 → 0.4 for realization lengths $L = 12\,W, 16\,W, \ldots, 120\,W$. Again, we consider a single change at a random time $t^* \in \left\{\frac{L}{4} - W, \frac{L}{4} - W + 1, \ldots, \frac{L}{4} + W\right\}$.

Results are presented in Figure 4.9.



(a) (b)

(c) (d)

Figure 4.9: Measures of change-point detection performance for NL (a, b) and AR (c, d) processes with different lengths

Discussion and conclusions:

1. Performance of the CEofOP statistic strongly depends on the length of time series. For sufficiently long stationary segments, CEofOP has better performance than other ordinal-patterns-based methods. In comparison with the classical method, CEofOP has almost the same performance for NL processes (see Figures 4.9a,b), and lower bias for AR processes (see Figure 4.9d).
2. For short stationary segments, LR and Chi statistics have better performance than CEofOP, but much worse than the Brodsky-Darkhovsky method.

4.3.2 Performance for Problem 2

Recall (Algorithm 1) that the solution of Problem 2 via statistic S is provided by performing the following test

$$S\big(\widehat{t}^*(S;y);y\big) \geq th_S(\alpha,y), \tag{4.12}$$

where α is the (nominal) probability of false positive errors, $th_S(\alpha)$ is the threshold for the statistic S. To assess performance of such tests the receiver operating characteristic (ROC) curve analysis is often used [KMPS09], [ZLB$^+$11, Section 1.4]. In our case the ROC curve represents the probability to detect a change-point in a process with a single change (*true positive rate*, TPR), plotted as a function of the frequency of false positive errors (*false positive rate*, FPR).

Recall (see Section 4.1) that we use bootstrapping to compute $th_S(\alpha, y)$ (details are provided in Section 4.5). Bootstrapping does not guarantee that FPR coincides with the nominal probability α, so we compare them empirically in Experiment 4.3. Then we construct the ROC curves for NL and AR processes in Experiment 4.4.

Experiment 4.3: comparing values of FPR for various statistics with the nominal probability α of false positive errors.
Objects: $N_T = 10000$ realizations y^j for $j = 1, 2, \ldots, N_T$ of the stationary noisy logistic stochastic process with $r = 3.98$ and $\sigma = 0.2$. (Results for stationary autoregressive processes are similar, so we omit them).

Technique. As soon as (4.12) is satisfied for y^j, one gets the false positive error. Hence, the empirical value of the FPR for a statistic S and given probability α is computed by

$$\mathrm{FPR}(S, \alpha) = \frac{\#\Big\{ j \in \{1, 2, \ldots, N_T\} \mid S\Big(\widehat{t^*}(S; y^j); y^j\Big) \geq th_S(\alpha, y^j) \Big\}}{N_T}.$$

Results: the obtained values of FPR are presented in Table 4.4.

S \ α	0.001	0.005	0.015	0.025	0.050	0.075	0.125	0.20	0.40	0.75
MMD	0.066	0.168	0.326	0.454	0.669	0.788	0.890	0.97	1.00	1.00
LR	0.004	0.022	0.058	0.092	0.166	0.236	0.378	0.53	0.77	0.97
Chi	0.006	0.026	0.076	0.108	0.184	0.220	0.328	0.47	0.75	0.97
CEofOP	0.002	0.006	**0.012**	**0.020**	**0.042**	**0.067**	**0.113**	**0.19**	**0.37**	**0.71**
BD$^{\mathrm{corr}}$	0.002	**0.005**	**0.017**	**0.020**	0.036	0.052	0.096	**0.16**	**0.35**	**0.70**

Table 4.4: Empirically estimated values of frequency $\mathrm{FPR}(S, \alpha)$ of false positive errors. Values satisfying $\mathrm{FPR}(S, \alpha) \in [0.8\alpha, 1.1\alpha]$ are shown in **bold**

Discussion and conclusions: for the CEofOP and BD$^{\mathrm{corr}}$ statistics the values of FPR are close to the values of α. For other statistics the empirical values of FPR are notably higher than the nominal values of α, hence one has to set α much lower than the desired FPR when applying these statistics. For instance, in order to get false positive errors with probability not exceeding 0.1, one has to take $\alpha \approx 0.025$ for the methods on the basis of LR and Chi statistics, and $\alpha \approx 0.002$ for the method on the basis of MMD statistic.

Experiment 4.4: comparing ROC curves for various statistics in order to evaluate performance for Problem 2.

Objects: $N_T = 10000$ realizations of the processes NL, $3.95 \to 3.98$, $\sigma = 0.2$ and AR, $0.1 \to 0.3$ from Experiment 4.1. For every statistic S we have chosen only realizations y^j, for those Problem 1 was solved satisfactory, that is it holds $\mathrm{err}^j(S, Y) \leq \mathrm{MaxErr} = W = 256$.

Results. For the given process Y, statistic S and probability α, TPR is estimated by

$$\mathrm{TPR}(S, Y, \alpha) = \frac{\#\Big\{j \in \{1, 2, \dots, N_T\} \mid S\Big(\widehat{t^*}(S; y^j); y^j\Big) \geq th_S(\alpha, y^j)\Big\}}{N_T}.$$

Figure 4.10 shows the values of $\mathrm{TPR}(S, Y, \alpha)$ plotted against values of $\mathrm{FPR}(S, \alpha)$.

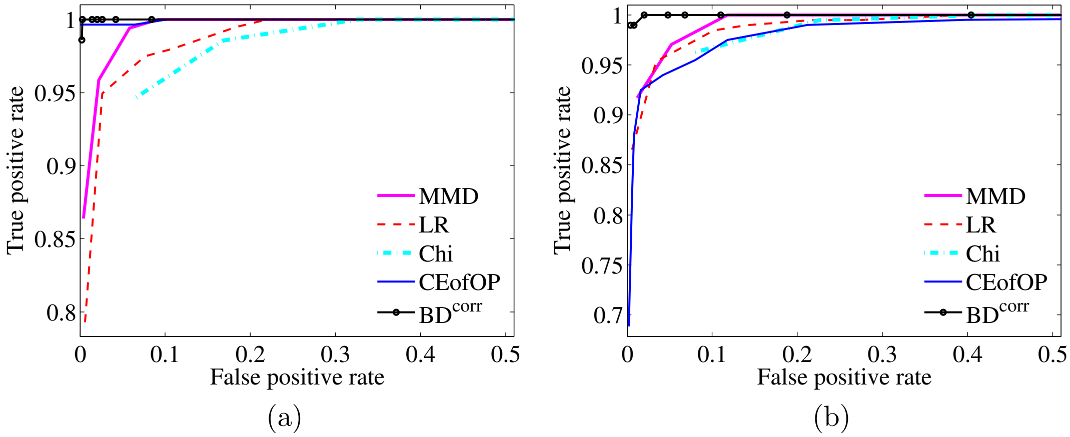


Figure 4.10: ROC curves for detection of a single change-point in realizations of stochastic processes: $\mathrm{NL}\big(t; (3.95, 3.98), (0.2, 0.2), t^*\big)$ (a) and $\mathrm{AR}\big(t; (0.1, 0.3), t^*\big)$ (b)

Discussion and conclusions: for the considered processes all methods for change-points detection provide almost perfect solutions of Problem 2. Indeed, already for small values of FPR (that is for low risk of detecting a change-point in case of no-change), the values of TPR are near to 1 (that is if there is a change-point in a process and the position of this change-point was correctly estimated, then this change-point will be detected with probability near to 1).

4.3.3 Performance for Problem 3

Experiment 4.5: comparing performance of methods for change-point detection with respect to Problem 3 (multiple change-points detection).

Objects: $N_T = 10000$ realizations y^j of the processes $\mathrm{AR}\big(t; (0.3, 0.5, 0.1, 0.4), (t_1^*, t_2^*, t_3^*)\big)$ and $\mathrm{NL}\big(t; (3.98, 4, 3.95, 3.8), (0.2, 0.2, 0.2, 0.3), (t_1^*, t_2^*, t_3^*)\big)$ with three independent change-points t_k^* uniformly distributed in $\Big\{\overline{t_k^*} - W, \overline{t_k^*} - W + 1, \dots, \overline{t_k^*} + W\Big\}$ for $k = 1, 2, 3$ with $\overline{t_1^*} = 0.3L$, $\overline{t_2^*} = 0.7L$, $\overline{t_3^*} = 0.9L$, $N_{\mathrm{st}} = 4$ and $L = 100\,W$ (we consider unequal lengths of stationary intervals to study methods for change-point detection in realistic conditions).

Technique. As we apply Algorithm 2 with a statistic S to y^j, we obtain estimates of number $\widehat{N}_{\mathrm{st}}(S; y^j)$ of stationary segments and of change-points positions $\widehat{t}_l^*(S; y^j)$ for $l = 1, 2, \ldots, \widehat{N}_{\mathrm{st}}(S; y^j) - 1$. Since the number of estimated change-points may be different from the actual number of changes, we suppose that the estimate for t_k^* is provided by the nearest $\widehat{t}_l^*(S; y^j)$. Therefore the error of estimation of the k-th change-point provided by S is given by

$$\mathrm{err}_k^j(S, Y) = \min_{l \in \{1,2,\ldots,\widehat{N}_{\mathrm{st}}(S;y^j)-1\}} \left| \widehat{t}_l^*(S; y^j) - t_k^* \right|.$$

To assess the overall accuracy of change-point detection, we compute two following quantities

- the fraction sE_k of satisfactory estimates of a change-point t_k^*, $k = 1, 2, 3$:

  $$\mathrm{sE}_k(S, Y) = \frac{\#\big\{j \in \{1, 2, \ldots, N_T\} \mid \mathrm{err}_k^j(S, Y) \leq \mathrm{MaxErr}\big\}}{N_T},$$

  where MaxErr is the maximal satisfactory error; we take $\mathrm{MaxErr} = W = 256$.

- average number of false change-points:

  $$\mathrm{fCP}(S, Y) = \frac{\sum_{j=1}^{N_T} \Big(\widehat{N}_{\mathrm{st}}(S; y^j) - 1 - \#\big\{k \in \{1, 2, 3\} \mid \mathrm{err}_k^j(S, Y) \leq \mathrm{MaxErr}\big\} \Big)}{N_T}.$$

  Note that we count as false change-points both false alarms (that is detecting a change-point in a stationary segment) and inaccurately estimated change-points.

Values of probability α of false positive errors have been taken as minimal values providing TPR near to 1 according to the results of Experiments 4.3, 4.4 (see Tables 4.5 and 4.6).

Results are presented in Tables 4.5 and 4.6. Best overall results and best results among ordinal-patterns-based methods are shown in **bold**.

Statistic	α	fCP	Fraction sE_k of satisfactory estimates			
			1st change	2nd change	3rd change	average
MMD	0.001	1.17	0.465	0.642	0.747	0.618
LR	0.001	0.98	0.470	0.749	0.850	0.690
Chi	0.001	1.70	0.470	0.740	0.771	0.660
CEofOP	0.05	**0.62**	**0.753**	**0.882**	**0.930**	**0.855**
$\mathrm{BD}^{\mathrm{corr}}$	0.05	1.34	0.296	0.737	0.751	0.595

Table 4.5: Problem 3: performance of change-point detection methods for NL process with three change-points

Statistic	α	fCP	Fraction sE_k of satisfactory estimates			
			1st change	2nd change	3rd change	average
MMD	0.001	1.17	0.340	0.640	0.334	0.438
LR	0.001	**0.98**	0.345	0.764	0.350	0.486
Chi	0.001	1.72	0.354	0.748	0.466	0.523
CEofOP	0.05	1.12	**0.368**	**0.834**	**0.517**	**0.573**
$\mathrm{BD}^{\mathrm{corr}}$	0.05	**0.53**	**0.783**	**0.970**	**0.931**	**0.895**

Table 4.6: Problem 3: performance of change-point detection methods for AR process with three change-points

Discussion and conclusions:

1. Since distributions of ordinal patterns for NL and AR processes have different properties (see Subsection 4.1.2), results for them differ significantly. For the NL processes performance of the ordinal-patterns-based methods is better, while for the AR processes the Brodsky-Darkhovsky method surpasses them.
2. The CEofOP statistic provides good results for the NL processes. However, for the AR processes performance is much worse: only the most prominent change is detected rather well. Weak results for two other change-points are caused by insufficient lengths of stationary segments: as we have seen in Experiment 4.2, the CEofOP statistic is rather sensitive to these lengths.

Our general conclusion is that the suggested method for ordinal change-point detection via the CEofOP statistic shows better performance than other ordinal-patterns-based methods. It also has comparable performance to the classical Brodsky-Darkhovsky method. Therefore we consider properties of the method for ordinal change-point detection via CEofOP in Section 4.4 and apply this method to real-world time series in Subsection 5.3.2.

4.4 Properties of the CEofOP statistic

We start with a simple example that shows how the CEofOP statistic estimates a change-point (Subsection 4.4.1). Then we provide a theoretical justification of the CEofOP statistic: in Subsection 4.4.2 we show its relation to the likelihood ratio statistic for a piecewise stationary Markov chain [AG57], and in Subsection 4.4.3 we consider the asymptotic properties of CEofOP.

4.4.1 CEofOP statistic: a toy example

We provide the following example to demonstrate convexity of conditional entropy and to illustrate detection of changes via the CEofOP statistic.

Example 4.7. Consider a realization y of a process with a single change-point $t^* \in \mathbb{T}$. Let the sequence $\pi = \big(\pi(t;y)\big)_{t\in\mathbb{T}'}$ of ordinal patterns of order $d = 1$ be a realization of a piecewise stationary Markov chain, such that the probabilities of ordinal patterns are equal to $\frac{1}{2}$ both before and after the change, while the transition probabilities are given by the matrices

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Let $\gamma = \frac{t^*}{L}$ and let $t = \theta L \in \mathbb{T}$ for $\theta \in (0,1)$. Then from the general representation (4.10) of CEofOP it follows that

$$\mathrm{CEofOP}(\theta L;\pi) = \begin{cases} L\big(\frac{\gamma}{2}\ln\frac{2-\gamma}{\gamma} - \ln(2-\gamma) + \frac{2-\theta-\gamma}{2}\ln\frac{2-\theta-\gamma}{1-\theta} + \frac{\gamma-\theta}{2}\ln\frac{\gamma-\theta}{1-\theta}\big), & \theta < \gamma \\ L\big(\frac{\gamma}{2}\ln\frac{2-\gamma}{\theta} - \ln(2-\gamma) + \frac{2\theta-\gamma}{2}\ln\frac{2\theta-\gamma}{\theta} + (1-\theta)\ln 2\big), & \theta \geq \gamma \end{cases}.$$

When one sets to zero the derivative of $\mathrm{CEofOP}(\theta L;\pi)$ with respect to θ, it becomes clear that $\mathrm{CEofOP}(\theta L;\pi)$ has a unique maximum at $\theta = \gamma$, which provides a detection of the change-point. In particular, for $\gamma = \frac{1}{2}$ we obtain:

$$\mathrm{CEofOP}(\theta L;\pi) = \begin{cases} L\Big(\ln 2 - \frac{3}{4}\ln 3 + \frac{3-2\theta}{4}\ln\frac{3-2\theta}{2-2\theta} + \frac{1-2\theta}{4}\ln\frac{1-2\theta}{2-2\theta}\Big), & \theta < \frac{1}{2} \\ L\Big((2-2\theta)\ln 2 - \frac{3}{4}\ln 3 - \theta\ln\theta + \frac{4\theta-1}{4}\ln(4\theta-1)\Big), & \theta \geq \frac{1}{2} \end{cases}.$$

Figure 4.11 illustrates that when $\gamma = \frac{1}{2}$, $\mathrm{CEofOP}(\theta L;\pi)$ attains a distinct maximum at $\theta = \frac{1}{2}$ (shown by the vertical line).

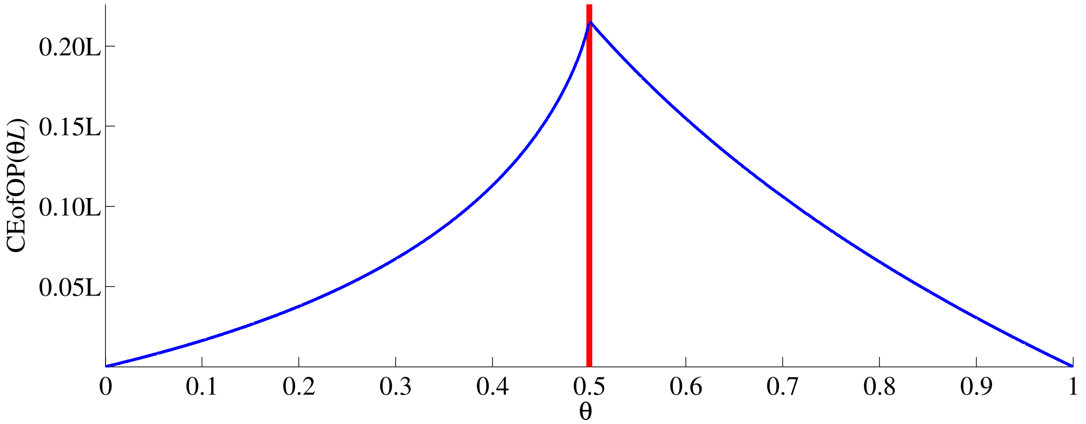


Figure 4.11: $\mathrm{CEofOP}(\theta L;\pi)$ when sequence π of ordinal patterns of order 1 is a Markov chain with a single change-point at $L/2$

4.4.2 Relation between the CEofOP and the likelihood ratio statistic

In this subsection we show that there is a connection between the CEofOP statistic and the classical likelihood ratio statistic. Though taking place only in a particular case, this connection reveals the nature of the CEofOP statistic. Before stating the result, we set up necessary notation.

Consider a realization y of a piecewise stationary process $\big(Y(t)\big)_{t\in\mathbb{T}}$ with at most one structural change-point. The basic statistic for testing whether there is a change in the transition probabilities of ordinal patterns at time t, is the likelihood ratio statistic [BN93, Subsection 2.2.3]:

$$\mathrm{LR}^{\mathrm{M}}\big(t;\pi\big) = -2\ln\frac{\mathrm{Lkl}\big(H_0 \mid \pi\big)}{\mathrm{Lkl}\big(H_t \mid \pi\big)} = -2\ln\mathrm{Lkl}\big(H_0 \mid \pi\big) + 2\ln\mathrm{Lkl}\big(H_t \mid \pi\big), \tag{4.13}$$

where $\mathrm{Lkl}\big(H \mid \pi\big)$ is the likelihood of the hypothesis H given a sequence $\pi = \big(\pi(t;y)\big)_{t\in\mathbb{T}'}$ of ordinal patterns, and the hypotheses are given by

$$\begin{aligned} H_0 : \big(p_{i,j}(t)\big)_{i,j=0}^{(d+1)!-1} &= \big(q_{i,j}(t)\big)_{i,j=0}^{(d+1)!-1}, \\ H_t : \big(p_{i,j}(t)\big)_{i,j=0}^{(d+1)!-1} &\neq \big(q_{i,j}(t)\big)_{i,j=0}^{(d+1)!-1}, \end{aligned}$$

where $p_{i,j}(t)$, $q_{i,j}(t)$ are transition probabilities of ordinal patterns before and after t, respectively.

Proposition 4.1. *If a sequence π of ordinal patterns of order $d\in\mathbb{N}$ is a realization of a Markov chain with at most one change-point, then it holds*

$$\mathrm{LR}^{\mathrm{M}}(t;\pi) = 2\,\mathrm{CEofOP}(t;\pi) + 2d\cdot\widehat{h}_{cond}\Big(\pi(d),\pi(1+d),\ldots,\pi(L)\Big).$$

Proof. First we estimate the probabilities and the transition probabilities before and after the change [AG57, Section 2]:

$$\begin{aligned} \widehat{p}_i(t) &= \frac{n_i(t)}{t-d}, & \widehat{p}_{i,j}(t) &= \frac{n_{i,j}(t)}{n_i(t)}, \\ \widehat{q}_i(t) &= \frac{m_i(t)}{L-(t+d)}, & \widehat{q}_{i,j}(t) &= \frac{m_{i,j}(t)}{m_i(t)}, \end{aligned}$$

$$\begin{aligned} \text{where } n_i(t) &= \#\{r\in\{d,1+d,\ldots,t-1\} \mid \pi(r)=i\}, \\ m_i(t) &= n_i(L) - n_i(t+d), \\ n_{i,j}(t) &= \#\{r\in\{d,1+d,\ldots,t-1\} \mid \pi(r)=i, \pi(r+1)=j\}, \\ m_{i,j}(t) &= n_{i,j}(L) - n_{i,j}(t+d). \end{aligned}$$

Then, as one can see from [AG57, Section 3.2], we have

$$\mathrm{Lkl}(H_0 \mid \pi) = \widehat{p}_{\pi(d)}(L)\prod_{l=d}^{L-1}\widehat{p}_{\pi(l),\pi(l+1)}(L) = \widehat{p}_{\pi(d)}(L)\prod_{i=0}^{(d+1)!-1}\prod_{j=0}^{(d+1)!-1}\big(\widehat{p}_{i,j}(L)\big)^{n_{i,j}(L)},$$

$$\begin{aligned} \mathrm{Lkl}(H_t \mid \pi) &= \widehat{p}_{\pi(d)}(t)\prod_{l=d}^{t}\widehat{p}_{\pi(l),\pi(l+1)}(t)\prod_{l=t+d}^{L-1}\widehat{q}_{\pi(l),\pi(l+1)}(t) \\ &= \widehat{p}_{\pi(d)}(t)\prod_{i=0}^{(d+1)!-1}\prod_{j=0}^{(d+1)!-1}\big(\widehat{p}_{i,j}(t)\big)^{n_{i,j}(t-1)}\prod_{i=0}^{(d+1)!-1}\prod_{j=0}^{(d+1)!-1}\big(\widehat{q}_{i,j}(t)\big)^{m_{i,j}(t-1)}. \end{aligned}$$

Assume that the first ordinal pattern $\pi(d)$ is fixed (non-random) in order to simplify the computations. Then $\widehat{p}_{\pi(d)}(L) = \widehat{p}_{\pi(d)}(t)$ and it holds:

$$\begin{aligned}
\mathrm{LR}^{\mathrm{M}}(t;\pi) = -2 &\sum_{i=0}^{(d+1)!-1} \sum_{j=0}^{(d+1)!-1} n_{i,j}(L)\big(\ln n_{i,j}(L) - \ln n_i(L)\big) \\
+2 &\sum_{i=0}^{(d+1)!-1} \sum_{j=0}^{(d+1)!-1} n_{i,j}(t)\big(\ln n_{i,j}(t) - \ln n_i(t)\big) \\
+2 &\sum_{i=0}^{(d+1)!-1} \sum_{j=0}^{(d+1)!-1} m_{i,j}(t)\big(\ln m_{i,j}(t) - \ln m_i(t)\big).
\end{aligned}$$

Since $\sum\limits_{j=0}^{(d+1)!-1} n_{i,j}(t) = n_i(t)$, one finally obtains:

$$\begin{aligned}
\mathrm{LR}^{\mathrm{M}}(t;\pi) &= 2(L-d)\cdot \widehat{h}_{\mathrm{cond}}\big(\pi(d), \pi(1+d), \ldots, \pi(L)\big) \\
&\quad - 2(t-d)\cdot \widehat{h}_{\mathrm{cond}}\big(\pi(d), \pi(1+d), \ldots, \pi(t)\big) \\
&\quad - 2\big(L-(t+d)\big)\cdot \widehat{h}_{\mathrm{cond}}\big(\pi(t+d), \pi(t+1+d), \ldots, \pi(L)\big) \\
&= 2\,\mathrm{CEofOP}(t;\pi) + 2d\cdot \widehat{h}_{\mathrm{cond}}\Big(\pi(d), \pi(1+d), \ldots, \pi(L)\Big).
\end{aligned}$$

□

Note that given y a realization of a process Y, $\pi(y)$ forms a Markov chain if and only if the ordinal partition for Y has the Markov property (see Subsection 3.3.3). An example of a stochastic process for that π forms a Markov chain is provided by the following consequence of Lemma 3.9 (p. 58).

Corollary 4.2. *Let $\big(Y(t)\big)_{t\in\mathbb{T}}$ be a Markov chain with $Y(t) \in \{0,1\}$ for all $t \in \mathbb{T}$. Then for every realization y of Y and for all $d \in \mathbb{N}$, the sequence $\pi = \big(\pi(t;y)\big)_{t\in\mathbb{T}'}$ of ordinal patterns of order d is a realization of a Markov chain.*

4.4.3 Asymptotic properties of the CEofOP statistic

In this subsection we consider asymptotic properties of the CEofOP statistic. Example 4.7 shows that a rigorous description of CEofOP for the processes with structural change-points is rather complicated, so we provide a theoretical result only for the stationary stochastic processes (without changes).

Theorem 4.3. *Let $\big(\pi(d), \pi(1+d), \ldots, \pi(L)\big)$ be a sequence of ordinal patterns of order $d \in \mathbb{N}$ corresponding to a realization y of an ergodic*[9] *stochastic process $Y = \big(Y(t)\big)_{t\in\mathbb{T}}$ with $\mathbb{T} = \{0,1,\ldots,L\}$. Then for any $\theta \in (0,1)$ it holds almost sure that*

$$\lim_{L\to\infty} \mathrm{CEofOP}\Big(\lfloor \theta L\rfloor; \big(\pi(d), \pi(1+d), \ldots, \pi(L)\big)\Big) = 0. \tag{4.14}$$

[9] See p. 32, Definition 2.10.

Proof. The basis of the proof is provided by Theorem 3.13 (p. 65). Indeed, from (3.16) it follows that for almost all realizations y of Y it holds

$$\begin{aligned}\lim_{L\to\infty}\widehat{h}_{\mathrm{cond}}\Big(\pi(d),\dots,\pi(L)\Big) &= \lim_{L\to\infty}\widehat{h}_{\mathrm{cond}}\Big(\pi(d),\dots,\pi\big(\lfloor\theta L\rfloor\big)\Big)\\ &= \lim_{L\to\infty}\widehat{h}_{\mathrm{cond}}\Big(\pi\big(\lfloor\theta L\rfloor+d\big),\dots,\pi(L)\Big) = h_{\mu,\mathrm{cond}}(Y,d).\end{aligned}$$

Then

$$\begin{aligned}\lim_{L\to\infty}\mathrm{CEofOP}\Big(\lfloor\theta L\rfloor;\big(\pi(d),\pi(1+d),\dots,\pi(L)\big)\Big) &= (L-2d)h_{\mu,\mathrm{cond}}(Y,d)\\ &- (\lfloor\theta L\rfloor-d)h_{\mu,\mathrm{cond}}(Y,d) - (L-\lfloor\theta L\rfloor-d)h_{\mu,\mathrm{cond}}(Y,d) = 0.\end{aligned}$$

□

4.5 Implementation details

In this section we consider technical details of solving Problem 1 by different ordinal-patterns-based methods and describe application of bootstrapping for Problem 2 (for Problem 3, detection of multiple change-points, we use the general scheme exhaustively described by Algorithm 2). These details are not important for understanding the idea of ordinal change-point detection, but are useful for its practical implementation. In Subsection 4.5.1 we suggest an implementation of the methods for detecting changes in distributions of ordinal patterns; we consider the change-point detection via the CMMD, MMD, LR and Chi statistics. In Subsection 4.5.2 we provide an implementation of the CEofOP statistics.

4.5.1 Implementation of detecting changes in distributions of ordinal patterns

Consider a sequence π of ordinal patterns of order d corresponding to a realization of a stochastic process with a single structural change-point $t^* \in \mathbb{T}' = \{d, 1+d, \dots, L\}$. We split π into M non-overlapping windows (4.6) of equal size W (that is we assume that $L-d+1 = MW$), where $M, W \in \mathbb{N}$. Then the left border of the m-th window is given by $w_{m-1} = d + (m-1)W$ for $m = 1, 2, \dots, M$. Let us say that if $w_{m^*-1} \le t^* < w_{m^*}$, then the m^*-th window is a *change-window*.

Recall (Subsection 4.2.1) that the MMD, CMMD, LR and Chi statistics estimate only the change-window, which does not provide a solution of Problem 1 since the exact position of the change-point remains unknown. Experiments show that in the general case the CMMD statistic estimates the change-window with a considerable bias, so we suggest a procedure to correct it in Subsection 4.5.1.1 (the same procedure is applicable to LR and Chi statistics). Then in Subsection 4.5.1.2 we specify the position of a change-point inside the change-window. In Subsection 4.5.1.3 we describe an implementation of the bootstrapping, which is necessary for the solution of Problem 2.

4.5.1.1 Correcting bias of the CMMD statistic

Recall that the distribution $\mathbf{z}(m) = \big(z_0(m), \ldots, z_{(d+1)!-1}(m)\big)$ of relative frequencies of ordinal patterns in the m-th window is given by (4.7), and let $\mathbf{z} = \big(\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(M)\big)$. The estimate of the change-window by the CMMD statistic is computed by [SKC13]:

$$\widehat{m}_0^* = \underset{m \in \{1,2,\ldots,M\}}{\arg\max} \operatorname{CMMD}(m; \mathbf{z}). \tag{4.15}$$

Results of experiments (see Figure 4.12a) demonstrate that if $m^* \neq \frac{M}{2}$ then the estimate $\widehat{m}_0^*$ has a bias towards $\frac{M}{2}$. Here we suggest Algorithm 3 to correct this bias: we estimate the change-window by (4.15) and set the $\widehat{m}_0^*$-th window to the center by omitting several windows. The new sequence $\mathbf{z}_c$ of distributions of ordinal patterns frequencies is called *centered*, from it we compute an improved estimate of the change-window.

Algorithm 3 Estimation of a change-window by CMMD with corrected bias

Input: sequence $\pi = \big(\pi(d), \ldots, \pi(L)\big)$ of ordinal patterns of order d, window size W
Output: centered sequence $\mathbf{z}_c$ of distributions of ordinal patterns frequencies; estimate $\widehat{m}^*$ of the change-window, offset Ofs of this estimate

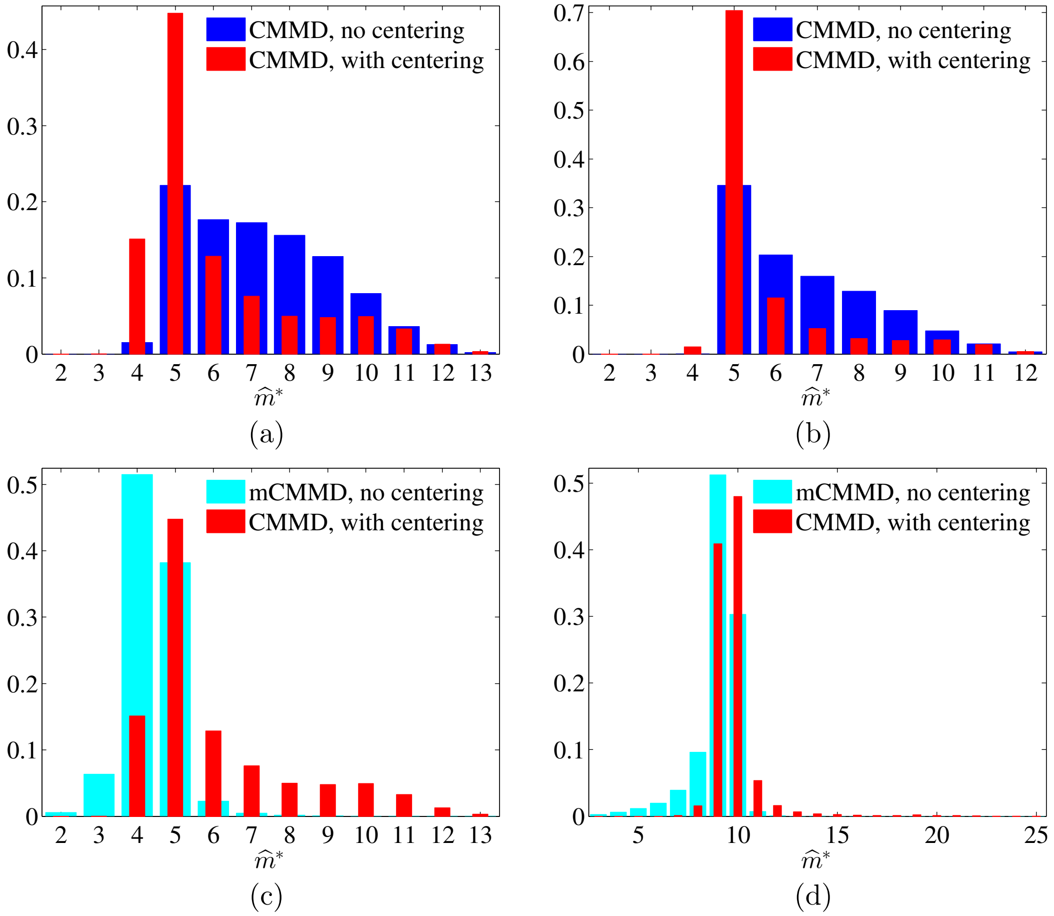
```
function ChangeWindowCMMD($\mathbf{z}$)
    $M \leftarrow (L - d + 1)/W$;
    $w_0 \leftarrow d$;
    for $m = 1, 2, \ldots, M$ do                                    ▷ Compute $\mathbf{z}$
        $w_m \leftarrow d + mW$;
        for $i = 0, 1, \ldots, (d+1)! - 1$ do    ▷ Calculate frequencies of ordinal patterns
            $z_i(m) \leftarrow \dfrac{\#\big\{l \in \{w^j_{m-1}, w^j_{m-1} + 1, \ldots, w^j_m - 1\} \mid \pi(l) = i\big\}}{w^j_m - w^j_{m-1}}$
        end for
    end for
    $\widehat{m}_0^* \leftarrow \underset{m \in \{1,2,\ldots,M\}}{\arg\max} \operatorname{CMMD}(m; \mathbf{z})$;    ▷ Compute a preliminary estimate of $m^*$
    if $\widehat{m}_0^* \leq \frac{M}{2}$ then                                    ▷ Center $m^*$
        Ofs $\leftarrow 0$;
        $M_c \leftarrow 2\widehat{m}_0^*$;
    else
        Ofs $\leftarrow 2\widehat{m}_0^* - M$;
        $M_c \leftarrow 2(M - \widehat{m}_0^*)$;
    end if
    for $m = 1, 2, \ldots, M_c$ do            ▷ Obtain a centered sequence of distributions
        $\mathbf{z}_c(m) \leftarrow \mathbf{z}(m + \mathrm{Ofs})$;
    end for
    $\widehat{m}^* \leftarrow \underset{m \in \{1,2,\ldots,M_c\}}{\arg\max} \operatorname{CMMD}(m; \mathbf{z}_c)$;  ▷ and compute an improved estimate of $m^*$
    return $\mathbf{z}_c, \widehat{m}^*$, Ofs;
end function
```

Experiment 4.6: assessing the effect of correcting bias on the estimate of the change-window by the CMMD statistic.

Objects: $N_T = 20000$ realizations of the process $\mathrm{AR}\big(t; (0.1, 0.4), t^*\big)$ of length $L = 20W + d$ for $W = 500$ and $d = 3$. The change-point is uniformly distributed inside the 5th window: $t^* \in \{4W + d, 4W + 1 + d, \dots, 5W + d - 1\}$.

Technique. For every realization we compute estimates of the change-window: $\widehat{m}_0^*$ (without centering) and $\widehat{m}^*$ (with centering); thus we obtain distributions of two change-window estimates.

Results are presented in Figure 4.12a.



(a)

(b)

(c)

(d)

Figure 4.12: Distributions of estimates of the change-window for the process AR, 0.1 → 0.4 of length $L = 20W + d$: with a random change-point in the 5-th window (a, c), with a fixed change-point $t^* = 5W + d - 1$ (b), and for the process AR, 0.1 → 0.4 of length $L = 40W + d$ with a change-point uniformly distributed in the 10-th window (d)

Discussion and conclusions: Correcting bias improves the detection of the change-window considerably. Note that the share of correct estimates of the change-window without centering obtained in our experiment significantly differs from that obtained in [SGK12, Section 4.1] for the same process: there it reaches 85% [SGK12, Figure 3], while in our experiment it is below 25% (Figure 4.12a). There are two reasons for this.

1. In our experiments the change-point is located randomly, while in [SGK12] it has a fixed position on the boundary between windows, which maximizes the CMMD statistic. The effect of centering for the fixed change-point $t^* = 5W + d - 1$ is even more prominent (see Figure 4.12b).
2. Results in [SGK12, Section 4.1] are obtained not for the CMMD statistic, but for its modified version mCMMD defined by (4.9) on p. 91. This statistic has no clear theoretical justification, but in the given settings it provides results that are better than for the CMMD statistic without centering, and comparable to the results for CMMD with centering (Figure 4.12c). However, this is not the case when the number of windows M is sufficiently large; we illustrate this fact in Figure 4.12d (there we consider $N_T = 20000$ realizations of the same process of length $L = 40W + d$ with the change-point uniformly distributed inside the 10th window: $t^* \in \{9W + d, 9W + 1 + d, \ldots, 10W + d - 1\}$), one can see the CMMD statistic with centering provides better results than mCMMD.

We also use Algorithm 3 for correcting bias of the LR and Chi statistics; the only difference is that the absolute frequencies Z of ordinal patterns are considered instead of the relative frequencies z.

4.5.1.2 Estimation of a change-point

Here we estimate a change-point for the given estimate $\widehat{m}^*$ of the change-window. We suggest Algorithm 4 for the estimation of a change-point via the CMMD statistic, algorithms for the LR and Chi statistics are almost the same.

Let us sketch the idea behind the algorithm. We start from applying Algorithm 3 (see Figure 4.13a) and assume that the estimate $\widehat{m}^*$ of the change-window is correct.

The left boundary of the $\widehat{m}^*$-th window is initially given by $w_{\widehat{m}^*-1} = d + (\widehat{m}^* - 1)W$ (see Figure 4.13b), therefore it holds

$$t^* \in \{d + (\mathrm{Ofs} + \widehat{m}^* - 1)W, d + (\mathrm{Ofs} + \widehat{m}^* - 1)W + 1, \ldots, d + (\mathrm{Ofs} + \widehat{m}^*)W - 1\},$$

where Ofs is the offset of the estimate $\widehat{m}^*$ of the change-window, see Algorithm 3. The dissimilarity between the samples

$$\big(\mathbf{z}_c(1), \mathbf{z}_c(2), \ldots, \mathbf{z}_c(\widehat{m}^* - 1)\big) \text{ and } \big(\mathbf{z}_c(\widehat{m}^*), \mathbf{z}_c(\widehat{m}^* + 1), \ldots, \mathbf{z}_c(M_c)\big). \tag{4.16}$$

should be maximal if $w_{\widehat{m}^*-1} = t^*$ since in this case $\mathbf{z}(\widehat{m}^*)$ characterizes frequencies of ordinal patterns after the change[10]. So we search for the position of the left boundary of the $\widehat{m}^*$-th window that maximizes the MMD statistic[11], see Figure 4.13c.

[10] Note that in the general case $\widehat{m}^*$-th contains ordinal patterns both before and after the change.

[11] We use here MMD statistic instead of CMMD for several reasons. First, the MMD statistic should not have significant bias since samples (4.16) have comparable sizes due to the centering procedure. Second, the MMD statistic is more effective for solving Problem 2 than CMMD (it is problematic to calculate the threshold for the latter statistic). Finally, using MMD is advantageous from the computational viewpoint.

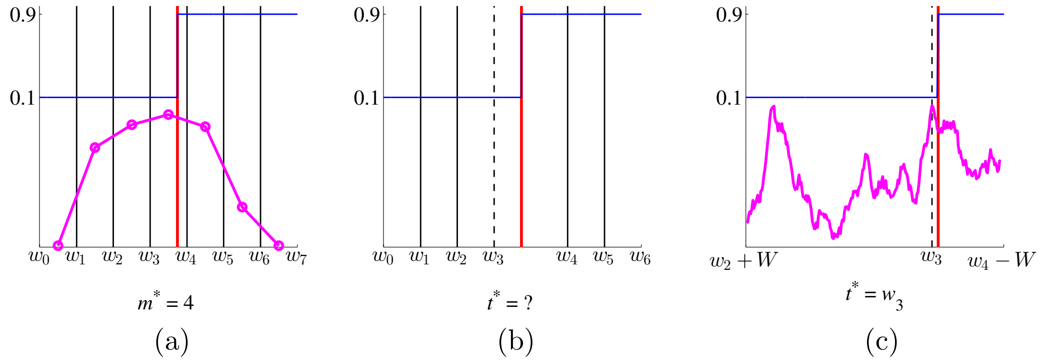


Figure 4.13: Estimation of a change-point in process $\mathrm{AR}\big(t;(0.1,0.9),3.75\,W\big)$ of length $L=7W+d$ by the CMMD and MMD statistics: estimation of a change-window by CMMD (a), reorganizing windows (b) and estimation of a change-point by MMD shifting the left boundary of the $\widehat{m}^*$-th window (c). The value of the control parameter of the AR process is shown by the blue line, the change-point – by the red vertical line, window boundaries - by the black vertical lines, values of CMMD (a) and MMD (c) statistics - by the magenta curves.

The described above procedure determines the choice of the minimal length $\tau_{\min}(S)$ of a stationary segment: in order to detect a change-point using the MMD, LR or Chi statistics it is necessary to have $M \geq 5$ windows, thus one has to take $\tau_{\min}(\mathrm{MMD}) = \tau_{\min}(\mathrm{LR}) = \tau_{\min}(\mathrm{Chi}) = 5W$, where W is window size.

4.5.1.3 Bootstrapping

Here we describe application of a bootstrapping to Problem 2. For the MMD statistic it consists in testing whether it holds

$$\max_{m\in\{1,2,\ldots,M\}} \mathrm{MMD}(m;\mathbf{z}') \geq th_{\mathrm{MMD}}(\alpha)$$

for the probability vectors $\mathbf{z}'(1),\ldots,\mathbf{z}'(M)$ obtained by Algorithm 4 and for the given probability α of false positive errors. The problem is to compute the threshold $th_{\mathrm{MMD}}(\alpha)$, we use for this bootstrapping, namely, a resampling without replacement [ST01, Pol07, KMPS09]. The general idea of resampling without replacement is discussed in Subsection 4.1.1. Here we generate $N \in \mathbb{N}$ sequences $\boldsymbol{\zeta}^j = \big(\boldsymbol{\zeta}^j(m)\big)_{m=1}^{M}$ with $j = 1,2,\ldots,N$, by shuffling (randomly permuting) elements of $\mathbf{z}'$. Though $\mathbf{z}'$ is not necessary stationary, $\boldsymbol{\zeta}^j$ becomes “quasi-independent” and “quasi-stationary” after a proper shuffling. For sufficiently large N it holds

$$\#\big\{j=1,2,\ldots,N \mid c_j \geq th_{\mathrm{MMD}}(\alpha)\big\} = \lfloor \alpha N \rfloor,$$

for $c_j = \max\limits_{m\in\{1,2,\ldots,M\}} \mathrm{MMD}(m;\boldsymbol{\zeta}^j)$ (cf. with (4.3)). Then the threshold can be computed by (4.4) on p.82 for $S = \mathrm{MMD}$.

However, the values of $\mathrm{MMD}(m;\boldsymbol{\zeta})$ and $\mathrm{CMMD}(m;\boldsymbol{\zeta})$ depend strongly on m even

Algorithm 4 Estimation of a change-point by CMMD and MMD

Input: sequence $\pi = \big(\pi(d), \dots, \pi(L)\big)$ of ordinal patterns of order d, window size W

Output: estimate $\widehat{t^*}$ of the change-point, centered sequence $\mathbf{z}'$ of distributions of ordinal patterns frequencies; offset Ofs of this sequence in regard to the original one

```
function ChangePointMMD($\mathbf{z}$)
    ▷ First we estimate the change-window $\widehat{m}^*$ by means of Algorithm 3:
    $\big(\mathbf{z}(1), \dots, \mathbf{z}(M)\big), \widehat{m}^*, \mathrm{Ofs} \leftarrow \mathrm{ChangeWindowCMMD}(\pi)$
    ▷ Next we shift the left boundary of the $\widehat{m}^*$-th window to maximize the MMD.
    $M \leftarrow M - 1$;
    for $j = 0, 1, \dots, W-1$ do
        for $m = 1, \dots, \widehat{m}^* - 2$ do    ▷ windows with $m < \widehat{m}^* - 1$ remain the same
            $w_m^j \leftarrow d + mW$;
            $\mathbf{z}^j(m) \leftarrow \mathbf{z}(m)$;
        end for
        $m \leftarrow \widehat{m}^* - 1$
        $w_m^j \leftarrow d + mW + j$;    ▷ $\widehat{m}^*$-th window: stepwise offset the left boundary
        $w_{m+1}^j \leftarrow d + (m+2)W$;    ▷ merge the $\widehat{m}^*$-th and $(\widehat{m}^*+1)$-th windows
        for $i = 0, 1, \dots, (d+1)! - 1$ do
            $z_i^j(m) \leftarrow \dfrac{\#\big\{l \in \{w_{m-1}^j, w_{m-1}^j + 1, \dots, w_m^j - 1\} \mid \pi(l) = i\big\}}{w_m^j - w_{m-1}^j}$
            $z_i^j(m+1) \leftarrow \dfrac{\#\big\{l \in \{w_m^j, w_m^j + 1, \dots, w_{m+1}^j - 1\} \mid \pi(l) = i\big\}}{w_{m+1}^j - w_m^j}$
        end for
        for $m = \widehat{m}^* + 1, \dots, M$ do    ▷ windows with $m > \widehat{m}^*$ change their numbers
            $w_m^j \leftarrow d + (m+1)W$;
            $\mathbf{z}^j(m) \leftarrow \mathbf{z}(m+1)$;
        end for
    end for
    ▷ Now the change-point is on the boundary of the $(\widehat{m}^*-1)$-th and $\widehat{m}^*$-th windows
    $k \leftarrow \underset{j \in \{0,1,\dots,W-1\}}{\arg\max} \mathrm{MMD}(\widehat{m}^* - 1; \mathbf{z}^j)$;
    $\mathbf{z}' \leftarrow \big(\mathbf{z}^k(1), \mathbf{z}^k(2), \dots, \mathbf{z}^k(M)\big)$;
    $\widehat{t^*} \leftarrow d + (\mathrm{Ofs} + \widehat{m}^* - 1)W + k$;
    return $\widehat{t^*}$, $\mathbf{z}'$, Ofs;
end function
```

for stationary $\boldsymbol{\zeta}$[12]. By this reason in (4.4) we take[13]

$$c_j = \mathrm{MMD}(\widehat{m}^* - 1; \boldsymbol{\zeta}^j), \tag{4.17}$$

where $\widehat{m}^*$ is the estimate of the change-window. This leads to an increased frequency of false positive errors in comparison with the nominal value α (see Experiment 4.3).

For the LR and Chi statistics the threshold is computed by (4.4) for $\boldsymbol{\zeta}^j$ obtained

[12] One can see this in Figure 4.6 for the CMMD statistic: it has maximum near $m = M/2$ even for the stationary processes.

[13] Recall that $\max\limits_{m \in \{1,2,\dots,M\}} \mathrm{MMD}(m; \mathbf{z}') = \mathrm{MMD}(\widehat{m}^* - 1; \mathbf{z}')$.

by shuffling the vectors $\mathbf{Z}(m)$ of absolute frequencies of ordinal patterns. We take here $c_j = \max_{m \in \{1,2,\ldots,M\}} S(m; \boldsymbol{\zeta}^j)$ since $\mathrm{LR}(m; \boldsymbol{\zeta})$ and $\mathrm{Chi}(m; \boldsymbol{\zeta})$ almost do not dependent on m for stationary $\boldsymbol{\zeta}$ (see Figure 4.7 on p. 94).

4.5.2 Implementation of change-point detection via the CEofOP statistic

Recall that calculation of the CEofOP statistic by (4.10) for the given order d requires a reliable estimation of empirical conditional entropy for the ordinal patterns of order d before and after the assumed change-point (see (4.11) on p. 95). For this the length of a time series should be not smaller than $L_{\min} = (d+1)!(d+1)$ (see Subsection 3.4.1, p. 65). Therefore we recommend to consider $\mathrm{CEofOP}(t; \pi)$ only for

$$L > \tau_{\min} = 2(d+1)!(d+1) \tag{4.18}$$

and for $t \in \mathbb{T}_0' = \left\{ \frac{\tau_{\min}(S)}{2} + d, \frac{\tau_{\min}(S)}{2} + d + 1, \ldots, L - \frac{\tau_{\min}(S)}{2} \right\}$, $d \in \mathbb{N}$. Hence we estimate the position of the change-point by

$$\widehat{t^*}(\pi) = \arg\max_{t \in \mathbb{T}_0'} \mathrm{CEofOP}(t; \pi).$$

For the solution of Problem 2 by means of the CEofOP statistic we compute the threshold by

$$th_{\mathrm{CEofOP}}(\alpha) = c_k \text{ for } k : \#\{j = 1, 2, \ldots, N \mid c_j \geq c_k\} = \lfloor \alpha N \rfloor,$$

where $c_j = \max_{t \in \mathbb{T}_0'} \mathrm{CEofOP}(t; \zeta^j)$ and ζ^j is obtained from the sequence π by the block bootstrapping [Lah03, Pol07, KMPS09]: in contrast to the usual bootstrapping, we shuffle not single entries of π, but blocks of certain length since the subsequent ordinal patterns are clearly dependent. To preserve the dependencies between subsequent ordinal patterns, we take the length of block equal to $(d+1)$.

Chapter 5

Ordinal-pattern-distributions clustering of time series

This chapter is devoted to the discrimination of time series segments by ordinal-patterns-based methods (see Chapter 4 for the discussion of time series segmentation). By *discrimination* we understand partitioning the set of time series segments into classes in such a way that segments of one class correspond to the same state of the system underlying the time series, while different classes correspond to different states. This is a typical task with many practical applications [FRMS96, TTF09, GRS05].

First methods for ordinal-patterns-based discrimination were introduced in [Sin10, Bra11, SKC13], where authors split time series into segments of equal length, sufficiently short to assume that the state of the underlying system does not change inside a segment (that is the obtained segments are supposed to be stationary), and then group the segments with similar empirical distributions of ordinal patterns (see Definition 2.17, p. 37) into classes using *ordinal-pattern-distributions* (OPD) *clustering*[14] (see p. 115 for the formal definition).

We suggest to segment a time series by using detection of change-points via the CEofOP statistic (see Chapter 4) and then cluster the obtained segments. Note that these segments are pseudo-stationary in the sense of having no structural change-points (see Definition 4.2, p. 83 and Definition 4.3, p. 86). Our approach is motivated by the criticism of segmenting time series without taking into account their structure (see, for instance, [BD00, Subsection 7.3.1]). Using OPD clustering together with ordinal-patterns-based segmentation seems to be a promising approach to discrimination of time series segments, therefore in this chapter we empirically investigate algorithms of OPD clustering.

The chapter is organized as follows. In Section 5.1 we provide basic information about cluster analysis and explain the main ideas of OPD clustering. In Section 5.2, in order to choose a clustering algorithm providing the best discrimination of OPDs, we

[14] Brandmaier used for this an expression "Permutation distribution clustering" referring to the permutation representation of ordinal patterns (see Remark on p. 33).

apply OPD clustering with several classical clustering algorithms to artificial time series. In Section 5.3 we apply OPD clustering to discrimination between EEG recordings from healthy individuals and from patients with epilepsy (Subsection 5.3.1) and to discrimination of sleep stages on the basis of EEG recordings (Subsection 5.3.2). In Section 5.4 we discuss possible directions of future work.

Remark. We would like to point out that in this chapter we discuss only discrimination, not a *classification* of time series segments. Roughly speaking, discrimination implies partitioning of a set of objects into several meaningful classes, while for classification one has also to associate every new object with one of these classes. Results obtained in this chapter (especially for the real-world data in Subsection 5.3.2) demonstrate the perspectives of using ordinal-patterns-based methods for classification of time series, however the classification itself lies beyond the scope of this thesis.

5.1 Ordinal-pattern-distributions clustering and basic facts from cluster analysis

The general formulation of the clustering problem is to partition a set of objects called a *dataset* into groups called *clusters* so that each object belongs to exactly one cluster[15], and objects from the same cluster are in a certain sense more similar than objects from different clusters. We call the obtained grouping (set of clusters) *a clustering result.* For clustering, objects are usually represented by vectors of certain features; the quantity used to measure dissimilarity between these vectors is called a *dissimilarity measure*, it is common [AF07, Section 1.3] to understand it as a distance since this simplifies the interpretation of clustering results.

Definition 5.1. Consider a dataset consisting of n time series $\big(y^k(0), y^k(1), \dots, y^k(L^k)\big)$ of length $(L^k + 1)$ for $k = 1, 2, \dots, n$. A clustering of this dataset such that every time series is represented by the empirical distribution $\mathbf{u}^k = \big(u_i^k\big)_{i=0}^{(d+1)!-1}$ of ordinal patterns of order $d \in \mathbb{N}$, where

$$u_i^k = \frac{\#\{t = d, 1+d \dots, L^k \mid \big(y^k(t), y^k(t-1), \dots, y^k(t-d)\big) \text{ has ordinal pattern } i\}}{L^k - d + 1}$$

is called *ordinal-pattern-distributions* (OPD) *clustering.*

Given a dataset of n multivariate time series $\big(\mathbf{y}^k(0), \mathbf{y}^k(1), \dots, \mathbf{y}^k(L^k)\big)$ of length $(L^k + 1)$ for $k = 1, 2, \dots, n$ with $\mathbf{y}^k(t) = \big(y_1^k(t), y_2^k(t), \dots, y_N^k(t)\big)$ for dimension $N \in \mathbb{N}$, *OPD clustering* is a clustering of vectors $\mathbf{u}^k = \big(u_i^k\big)_{i=0}^{N(d+1)!-1}$ formed from empirical distributions of ordinal patterns in each component of a time series:

$$u^k_{i+(j-1)(d+1)!} = \frac{\#\{t = d, 1+d, \dots, L^k \mid \big(y_j^k(t), \dots, y_j^k(t-d)\big) \text{ has ordinal pattern } i\}}{L^k - d + 1}. \quad (5.1)$$

[15] We consider here the so-called "hard" clustering; in the "soft" clustering an object can belong to more than one cluster, see [MIH08] for details.

Remark. According to (5.1) we characterize a multivariate time series by banking distributions of ordinal patterns for each time series component one upon the other. We do not consider the full OPD given by Definition 2.17 on p. 37 since a reliable estimation of it requires very long time series, which is often impractical (see Remark on p. 66). To keep notation simple we call $\mathbf{u}^k$ an OPD both for univariate and multivariate time series.

In the rest of the section we recall techniques for evaluating clustering results (Subsection 5.1.1) and describe clustering algorithms (Subsection 5.1.2) that are used further in this chapter. The reader who is familiar with cluster analysis may proceed to Section 5.2. In this section we use the following notation: groupings of a dataset (in particular, clustering results) are denoted by bold capital letters $\mathbf{U}$, $\mathbf{V}$, etc., clusters – by capital letters U, V, etc., and OPDs from the clusters – by bold letters $\mathbf{u}$, $\mathbf{v}$, etc.

5.1.1 Evaluation of clustering results

We describe here several commonly-used techniques for evaluation of clustering results that are employed in this chapter (for a comprehensive overview see [AF07, Section 1.7, Chapter 2] and [ELLS11, Section 9.4]).

Consider two groupings $\mathbf{U} = \{U_1, U_2, \ldots, U_M\}$ and $\mathbf{V} = \{V_1, V_2, \ldots, V_K\}$ of n objects, where $\mathbf{U}$ is a true classification and $\mathbf{V}$ is a clustering result; we assume that the number $K \in \mathbb{N}$ of clusters is greater than or equal to the number $M \in \mathbb{N}$ of true classes. The relation between $\mathbf{U}$ and $\mathbf{V}$ is summarized in a *contingency matrix* $\big(n_{ij}(\mathbf{U}, \mathbf{V})\big)$ [HA85], where an entry $n_{ij}(\mathbf{U}, \mathbf{V})$ denotes the number of common objects for the sets U_i and V_j, and is given by

$$n_{ij}(\mathbf{U}, \mathbf{V}) = |U_i \cap V_j|$$

for $1 \leq i \leq M$ and $1 \leq j \leq K$, where $|U_i \cap V_j|$ is number of elements in $U_i \cap V_j$.

Since a clustering result $\mathbf{V} = \{V_1, V_2, \ldots, V_K\}$ is defined up to the order of clusters, for a given $\mathbf{U}$ and a fixed $\mathbf{V}$ there are $K!$ possible contingency matrices. To avoid this uncertainty we fix the order of clusters in $\mathbf{V}$. Given $k_0, k_1, \ldots, k_M \in \mathbb{N}_0$ such that

$$0 = k_0 < k_1 < \ldots < k_M = K, \;\; (k_0, k_1, \ldots, k_M) = \underset{k_0, k_1, \ldots, k_M}{\arg\max} \sum_{i=1}^{M} \sum_{j=k_{i-1}+1}^{k_i} n_{ij}(\mathbf{U}, \mathbf{V}). \quad (5.2)$$

We order clusters in $\mathbf{V}$ in such a way that clusters $V_1, \ldots, V_{k_1}$ mainly contain elements from the true class U_1, clusters $V_{k_1+1}, \ldots, V_{k_2}$ mainly contain elements from the true class U_2, and so on. Formally, the ordering of clusters in $\mathbf{V}$ should maximize[16]

$$\operatorname{agreem}(\mathbf{U}, \mathbf{V}) = \frac{1}{n} \sum_{i=1}^{M} \sum_{j=k_{i-1}+1}^{k_i} n_{ij}(\mathbf{U}, \mathbf{V}) \qquad (5.3)$$

for $k_0, k_1, \ldots, k_M \in \mathbb{N}_0$ given by (5.2).

[16] If several numberings provide the maximal value in (5.3), we choose one of them by random.

The quantity defined by (5.3) is called *agreement* (sometimes it is also referred as sensitivity). Agreement is often used in biomedical applications, it represents a share of correctly grouped objects and provides a natural measure for evaluation of clustering results. It holds $0 \leq \text{agreem}(\mathbf{U}, \mathbf{V}) \leq 1$; if clustering result $\mathbf{V}$ coincides with the true classification $\mathbf{U}$, then $\text{agreem}(\mathbf{U}, \mathbf{V}) = 1$. When the number of clusters K coincides with the number of true classes M, then agreement is simply given by $\sum\limits_{i=1}^{M} n_{ii}(\mathbf{U}, \mathbf{V})$. We say that the clustering result $\mathbf{V}$ is *good*, if agreement is close to 1.

Another often-used quantity describing a clustering result is the *Fowlkes-Mallows index* (FMI) [HA85], defined by

$$\text{FMI}(\mathbf{U}, \mathbf{V}) = \frac{\sum\limits_{i=1}^{M} \sum\limits_{j=1}^{K} \binom{n_{ij}(\mathbf{U},\mathbf{V})}{2}}{2\sqrt{\sum\limits_{i=1}^{M} \binom{n_{i\cdot}}{2} \sum\limits_{j=1}^{K} \binom{n_{\cdot j}}{2}}},$$

where $n_{i\cdot} = \sum\limits_{j=1}^{K} n_{ij}(\mathbf{U}, \mathbf{V})$, $n_{\cdot j} = \sum\limits_{i=1}^{M} n_{ij}(\mathbf{U}, \mathbf{V})$. It holds $0 \leq \text{FMI}(\mathbf{U}, \mathbf{V}) \leq 1$, with $\text{FMI}(\mathbf{U}, \mathbf{V}) = 1$ for the coincidence of $\mathbf{U}$ and $\mathbf{V}$.

Let us illustrate the meaning of the agreement and FMI by the following example.

Example 5.1. Consider a dataset divided into classes U_1 and U_2 both containing 100 objects. In Table 5.1 we present contingency matrices, values of agreement and FMI for three clustering results obtained for this dataset: $\mathbf{V}^1$ and $\mathbf{V}^2$ are rather good, while $\mathbf{V}^3$ is "bad". Note that though $\text{agreem}(\mathbf{U}, \mathbf{V}^1) = \text{agreem}(\mathbf{U}, \mathbf{V}^2)$, it holds $\text{FMI}(\mathbf{U}, \mathbf{V}^1) > \text{FMI}(\mathbf{U}, \mathbf{V}^2)$ since for $\mathbf{V}^1$ the number of clusters coincides with the number of true classes, whereas in $\mathbf{V}^2$ one cluster corresponds to class U_1 and two clusters – to U_2.

clustering result	$\mathbf{V}^1 = \{V_1^1, V_2^1\}$		$\mathbf{V}^2 = \{V_1^2, V_2^2, V_3^2\}$			$\mathbf{V}^3 = \{V_1^3, V_2^3\}$	
clusters	V_1^1	V_2^1	V_1^2	V_2^2	V_3^3	V_1^3	V_2^3
class U_1	90	10	40	40	20	70	30
class U_2	20	80	0	10	90	50	50
agreem	0.85		0.85			0.60	
FMI	0.74		0.65			0.52	

Table 5.1: The agreement and the Fowlkes-Mallows index for three clustering results

Sometimes a visualization of the obtained clusters also indicates whether the clustering result is good or not. Since visualization of multidimensional vectors is complicated, a common solution is to reduce dimensionality of the vectors using principal component analysis (PCA) [AF07, Subsection 2.1.1]: one depicts only the 2d or 3d vectors of the first principal components[17].

We use FMI and agreement to evaluate clustering results; we use contingency matrices and PCA to visualize clustering results.

5.1.2 Dissimilarity measures and clustering algorithms

Three commonly used dissimilarity measures [ELLS11, Section 3.3] that were previously applied to OPDs are listed in Table 5.2. We consider there dissimilarity between the OPDs $\mathbf{u} = \big(u_i\big)_{i=0}^{N(d+1)!-1}$ and $\mathbf{v} = \big(v_i\big)_{i=0}^{N(d+1)!-1}$ for order $d \in \mathbb{N}$ and dimension of original time series $N \in \mathbb{N}$. To provide the reader an impression of the essential differences of these dissimilarity measures, we show in Table 5.2 the sets of points on a plane equidistant from a central point with respect to the given distances (this two-dimensional case corresponds to the OPD of order $d = 1$, $N = 1$).

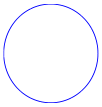
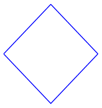
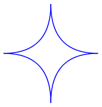
Dissimilarity measure	Dissimilarity between OPDs $\mathbf{u}$ and $\mathbf{v}$	Application to OPDs	Shape of a "circle"
Euclidean distance	$\rho_E(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=0}^{N(d+1)!-1} (u_i - v_i)^2}$	[SKC13]	
city block distance	$\rho_C(\mathbf{u}, \mathbf{v}) = \sum_{i=0}^{N(d+1)!-1} \lvert u_i - v_i \rvert$	[Sin10]	
squared Hellinger distance	$\rho_H(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \sum_{i=0}^{N(d+1)!-1} \big(\sqrt{u_i} - \sqrt{v_i}\big)^2$	[Bra11, PSL13]	

Table 5.2: Dissimilarity measures used for comparing ordinal-pattern-distributions (OPDs)

We consider in this chapter five classical clustering algorithms that are briefly described below (for details we refer to [ELLS11]).

- Three algorithms of *agglomerative clustering*: *complete linkage*, *average linkage* and *single linkage* clustering, having the same scheme: initially every OPD is assigned to a separate cluster, then at every step two nearest clusters are merged. To decide, which clusters are the nearest, a distance $\operatorname{dist}(U, V)$ between clusters U and V is introduced. It is defined in different ways for these three algorithms, namely:

$$\text{complete linkage: } \operatorname{dist}(U, V) = \max_{\mathbf{u} \in U} \max_{\mathbf{v} \in V} \rho(\mathbf{u}, \mathbf{v}),$$

$$\text{average linkage: } \operatorname{dist}(U, V) = \frac{1}{|U||V|} \sum_{\mathbf{u} \in U} \sum_{\mathbf{v} \in V} \rho(\mathbf{u}, \mathbf{v}),$$

$$\text{single linkage: } \operatorname{dist}(U, V) = \min_{\mathbf{u} \in U} \min_{\mathbf{v} \in V} \rho(\mathbf{u}, \mathbf{v}),$$

[17] We follow here [AF07] in using PCA; however, to visualize vectors of relative frequencies one could also use correspondence analysis, see [Gre07] for details of this approach.

where $|U|$ is the number of elements in cluster U and $\rho(\mathbf{u}, \mathbf{v})$ is a dissimilarity measure between OPDs $\mathbf{u}$ and $\mathbf{v}$. Agglomerative clustering can be used with any dissimilarity measure, see Table 5.2. Note that in both previously suggested methods for OPD clustering [Bra11, SKC13], agglomerative clustering (namely, the complete linkage algorithm) was used.

- *Centroid-based clustering* assigns each OPD to the cluster with the nearest center, which is defined depending on the chosen clustering algorithm and dissimilarity measure. We use the most common algorithm of centroid-based clustering called *k-means*. For the given number $K \in \mathbb{N}$ of clusters k-means forms a grouping $\mathbf{U} = \{U_1, U_2, \ldots, U_K\}$ that minimizes the sum for $k = 1, 2, \ldots, K$ of distances between all OPD $\mathbf{u} \in U_k$ and the center m_k of the cluster U_k. The exact function being minimized and the formula for computing the center of a cluster depend on the used dissimilarity measure, see Table 5.3.

Dissimilarity measure	Function being minimized	Center of the cluster U_k
Euclidean distance	$\arg\min\limits_{\mathbf{U}} \sum\limits_{k=1}^{K} \sum\limits_{\mathbf{u}\in U_k} (\rho_E(\mathbf{u}, \mathbf{m}_k))^2$	$\mathbf{m}_k = \frac{1}{\lvert U\rvert} \sum\limits_{\mathbf{u}\in U_k} \mathbf{u}$
city block distance[18]	$\arg\min\limits_{\mathbf{U}} \sum\limits_{k=1}^{K} \sum\limits_{\mathbf{u}\in U_k} (\rho_C(\mathbf{u}, \mathbf{m}_k))$	$\mathbf{m}_k$ is the component-wise median of all $\mathbf{u} \in U_k$
squared Hellinger distance[19]	$\arg\min\limits_{\mathbf{U}} \sum\limits_{k=1}^{K} \sum\limits_{\mathbf{u}\in U_k} (\rho_H(\mathbf{u}, \mathbf{m}_k))$	$\mathbf{m}_k = \left(\frac{1}{\lvert U\rvert} \sum\limits_{\mathbf{u}\in U_k} \sqrt{\mathbf{u}}\right)^2$

Table 5.3: The k-means algorithm for various dissimilarity measures

- *Distribution-based clustering.* Here OPD is considered as a random vector, and similarity of OPD is interpreted as similarity of their distribution laws (that is no external dissimilarity measure is used). Then each cluster contains vectors that are distributed by the same or by a similar law. We use the classical expectation-maximization algorithm (see [ELLS11, Chapter 6] for the description of the method and [Che12] for a MATLAB realization) that considers OPD as a random vector with Gaussian distribution[20].

5.2 Investigation of algorithms for ordinal-pattern-distributions clustering

To our knowledge, a comparison of different clustering algorithms and dissimilarity measures for OPD clustering has been done only in [Bra11], where the Euclidean and the

[18] When the city block distance is used, the clustering algorithm is usually called k-medians [ELLS11].

[19] Using k-means with the squared Hellinger distance is based on the equality $\rho_H(\mathbf{u}, \mathbf{v}) = 2(\rho_E(\sqrt{\mathbf{u}}, \sqrt{\mathbf{v}}))^2$.

[20] We do not really assume that vectors representing OPD have Gaussian distribution; this is just a framework used by expectation-maximization algorithm for measuring dissimilarity between vectors.

squared Hellinger distance were compared for the complete linkage clustering algorithm. In this section we apply five clustering algorithms with three dissimilarity measures defined in Subsection 5.1.2 for OPD clustering of artificial time series and compare the obtained results.

5.2.1 Artificial time series for the experiments

Two types of artificial time series are used for the experiments in this section, namely realizations of (stationary) noisy logistic and autoregressive processes. Recall (Subsection 3.4.2) that a noisy logistic stochastic process NL for $t \in \mathbb{T}$ is given by

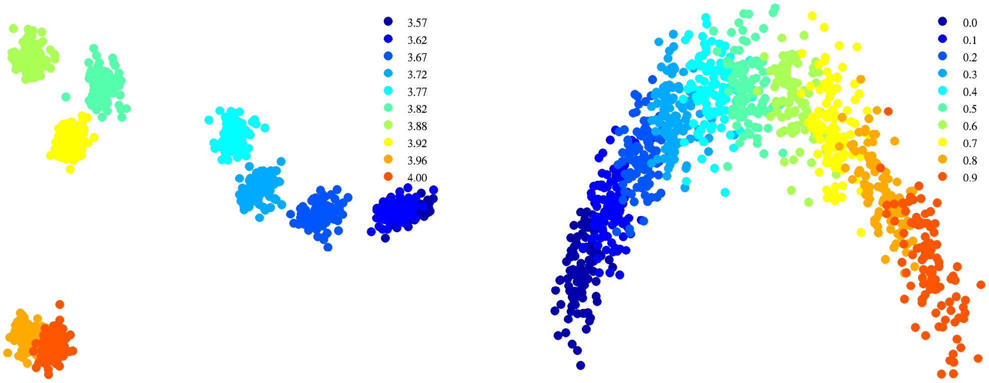
$$\mathrm{NL}(t; r, \sigma) = f_r^t + \sigma\epsilon(t),$$

where $f_r : [0, 1] \hookleftarrow$ is the logistic map with control parameter $r \in [1, 4]$ (as in the previous chapters, $f_r^t = f_r \circ f_r^{t-1}$ for $t > 1$, $f_r^1 = f_r$), ϵ is the standard additive white Gaussian noise (see p. 32), and $\sigma > 0$ is the level of noise. An autoregressive process for $t \in \mathbb{T}$ is defined by

$$\mathrm{AR}(t; \phi) = \phi\mathrm{AR}(t-1; \phi) + \epsilon(t),$$

where $\phi \in [0, 1)$ is a control parameter of the autoregressive model.

In the experiments in Subsection 5.2.2 we cluster a noisy logistic (NL) and an autoregressive (AR) dataset, both consisting of 1000 realizations of the corresponding processes with ten different values of control parameters (100 realizations for each value). In order to make distinguishing of realizations non-trivial, we consider the NL process with $r \in \{3.57, 3.62, 3.67, 3.72, 3.77, 3.82, 3.88, 3.92, 3.96, 4.00\}$, $\sigma = 0.2$, and the AR process with $\phi \in \{0.0, 0.1, \ldots, 0.9\}$. For this choice, OPDs for realizations of the processes differ only slightly. Figure 5.1 demonstrates OPDs of order $d = 4$ for the time series with length $L = 2400$ from the NL and AR datasets.



(a) (b)

Figure 5.1: Distributions of ordinal patterns of order $d = 4$ (two principal components) for realizations of NL (a) and AR (b) stochastic processes; 80% (a) and 67% (b) of total variance between the principal components is explained by the first two components

5.2.2 Ordinal-pattern-distributions clustering of artificial time series

In this subsection we evaluate the results obtained by using clustering algorithms and dissimilarity measures (defined in Subsection 5.1.2) for the NL and AR datasets. OPDs are calculated for order $d = 4$, that is they are characterized by vectors of length $(d+1)! = 120$. We follow the practical recommendations of [Ami10] (see also Subsection 3.4.1) and consider the length of a time series $L \geq L_0 = 5(d+1)! = 600$. However, we expect that even in this case the value of L may strongly influence clustering results, therefore we consider not only fixed L (Experiment 5.1) but also compare clustering results for various values of L (Experiment 5.2).

Experiment 5.1: Which combination of a clustering algorithm and a dissimilarity measure provides the best results of OPD clustering for the NL and AR datasets?

Objects: $n = 1000$ time series with length $L = 2400 = 4L_0$ from the NL and AR datasets (see Subsection 5.2.1 for details).

Technique. We apply to the two datasets four clustering algorithms (complete linkage, single linkage, average linkage, k-means) with three dissimilarity measures listed in Table 5.2, and the expectation-maximization algorithm with its internal dissimilarity measure, altogether 13 combinations of a clustering algorithm with a dissimilarity measure.

Results: the values of FMI are presented in Tables 5.4.

	NL dataset			AR dataset		
Dissimilarity measure Algorithm	ρ_E	ρ_C	ρ_H	ρ_E	ρ_C	ρ_H
complete linkage	0.912	0.885	0.904	0.517	0.522	0.513
single linkage	0.740	0.553	0.556	0.312	0.312	0.312
average linkage	0.912	0.907	0.843	0.587	0.522	0.546
k-means	**0.925**	0.921	0.885	0.716	0.701	**0.722**
expectation-maximization	0.506			0.498		

Table 5.4: Values of FMI characterizing the results of OPD clustering from the NL dataset and from the AR dataset. The best value for each dataset is shown in **bold**

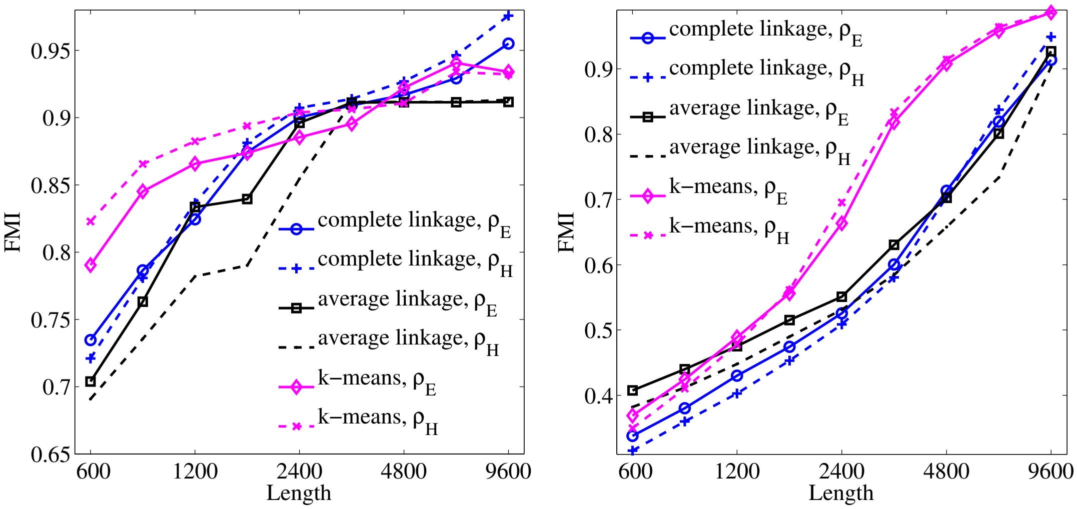
Conclusions: single linkage and expectation-maximization clustering provide much worse results than k-means, complete and average linkage clustering. Based on the results of this experiment we consider further only the three latter algorithms.

Experiment 5.2: How the results of OPD clustering depend on the length L of time series?

Objects: $n = 1000$ time series from the NL dataset and from the AR dataset (see Subsection 5.2.1 for details) with the lengths $L = L_0, \sqrt{2}L_0, 2L_0, \ldots, 16L_0$, where

$L_0 = 5(d+1)! = 600$.

Results are presented in Figure 5.2.



(a) (b)

Figure 5.2: Values of FMI characterizing the results of OPD clustering of time series from the NL dataset (a) and from the AR dataset (b) for various lengths of time series. Results for ρ_E and ρ_C are very similar, so values for ρ_C are not shown

Discussion and conclusions:

1. The k-means clustering for the most values of L either outperforms other clustering algorithms (the AR dataset) or provides comparable results (the NL dataset); the clustering results are a bit better when using k-means with the squared Hellinger distance ρ_H than with ρ_E or ρ_C.

2. Results of OPD clustering strongly depend on the length of time series: the longer the time series is, the better the results of the clustering are. In order to get reliable results of OPD clustering for order $d \leq 5$, we recommend to take

$$L \geq 4L_0 = 20(d+1)!. \tag{5.4}$$

   Note that this is just an empirical finding and we do not have theoretical evidence for this. Condition (5.4) is consistent with the findings in [Bra11], where OPD clustering was efficient for $L > 10(d+1)!$.

Since the best clustering results in Experiments 5.1 and 5.2 are obtained for the k-means algorithm with the squared Hellinger distance ρ_H, we use this combination of an algorithm and a dissimilarity measure for OPD clustering in Section 5.3; however, in Subsection 5.3.1 we also provide a comparison with results for other clustering algorithms.

5.3 Applications to biomedical time series

5.3.1 Applications to epileptic EEG recordings

Detection of epileptic seizures is an important problem in biomedical research, we refer to [LE98, MWWM99, MAEL07] for an overview. We consider in Experiment 5.3 an open access dataset described in [ALM$^+$01] and accessible at [And03], further we refer to it as "Bonn dataset". It consists of 500 single-channel EEG recordings divided into five groups, here we consider three of them, namely:

- Group A: surface EEG of healthy volunteers with eyes open.
- Group D: intracranial EEG during seizure-free intervals measured in the epileptogenic zones.
- Group E: intracranial EEG during seizures selected from recording sites exhibiting ictal activity.

Experiment 5.3: We employ OPD clustering with complete linkage and k-means clustering algorithms, and with several dissimilarity measures to group EEG recordings into three clusters: HEALTHY, SEIZURE-FREE and ICTAL (ideally, they should correspond to the groups A, D and E, respectively) and compare clustering results.
Objects: 300 EEG recordings from the groups A, D and E of the Bonn dataset. Every recording represents 23.6 seconds of artifact-free EEG with a sampling rate 173.61 Hz and has length $L = 4097$.

Technique. We consider two methods of OPD clustering, first of them represents our approach (OPD clustering together with segmentation of time series into pseudo-stationary segments), while the second method is the OPD clustering as it was suggested in [Bra11, SKC13].

Method 1: clustering of pseudo-stationary time series. We take here the entire original time series since they fulfill a weak stationarity criterion formulated in [ALM$^+$01, Section II B2].

Method 2: clustering short equi-sized time series that are supposed to be stationary. We follow [SKC13] in using segments of length $L_{\text{short}} = 528$, which in this case[21] approximately corresponds to 3 s. Each recording is split into 8 segments.

Results are presented in Table 5.5. The best values of FMI are obtained by using Method 1 (clustering of the stationary time series) with the k-means algorithm and the squared Hellinger distance ρ_H for $d = 4$. For this case agreem = 0.87, that is 87% of time series are discriminated correctly. Table 5.6 shows the contingency matrix for this

[21] Note that a rule of thumb is to consider EEG segments of 2 s [BD00, Section 7.3]. In particular, this is done in [SKC13], where the sampling rate is 256 Hz. Here the sampling rate is 173.61 Hz, therefore we take longer segments to obtain reliable estimates of ordinal pattern frequencies.

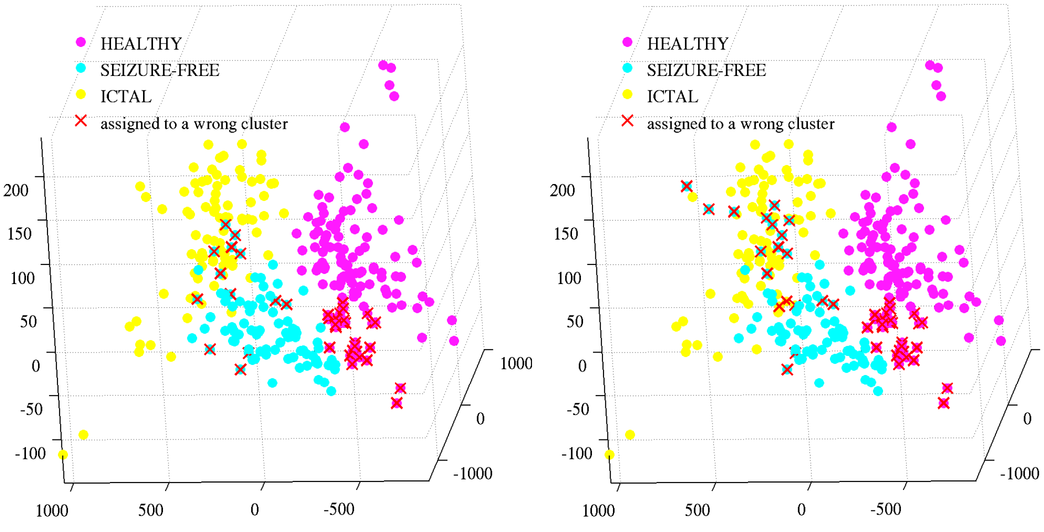
method in comparison with the contingency matrix for k-means with city block distance ρ_C for $d = 4$; Figure 5.3 illustrates the resulting clusters for these two methods.

Method	Method 1				Method 2		
Order	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 2$	$d = 3$	$d = 4$
k-means, ρ_H	0.590	**0.700**	**0.779**	**0.774**	0.584	**0.610**	**0.741**
k-means, ρ_C	**0.593**	0.605	0.746	0.755	0.579	0.577	0.627
k-means, ρ_E	0.592	0.596	0.599	0.591	0.591	0.584	0.582
complete linkage, ρ_H	0.461	0.465	0.656	0.691	0.569	0.503	0.477
complete linkage, ρ_C	0.438	0.598	0.691	0.667	**0.608**	0.583	0.586
complete linkage, ρ_E	0.493	0.536	0.594	0.534	0.550	0.552	0.525

Table 5.5: Values of FMI characterizing the results of OPD clustering of the Bonn dataset, the best values for every d are shown in **bold**

	Method 1, k-means, ρ_H, $d = 4$			Method 1, k-means, ρ_C, $d = 4$		
	HEALTHY	SEIZURE-FREE	ICTAL	HEALTHY	SEIZURE-FREE	ICTAL
Group A	**100**	0	0	**100**	0	0
Group D	25	**75**	0	26	**71**	3
Group E	1	13	**86**	1	16	**83**

Table 5.6: Contingency matrices for the Bonn dataset



(a) (b)

Figure 5.3: Distributions of ordinal patterns of order $d = 4$ (three principal components) for the EEG recordings from groups A, D and E of the Bonn dataset [And03], discriminated using the OPD clustering with the k-means algorithm and the squared Hellinger distance ρ_H (a) and the city block distance ρ_C (b). The first three components explain 99% of total variance between the principal components

Discussion and conclusions:

1. Clustering results provided by the k-means algorithm with ρ_H are considerably better (according to the values of FMI and of agreement) than for other clustering algorithms and dissimilarity measures. This confirms superiority of the k-means clustering with ρ_H for OPD clustering, which was observed in Section 5.2.
2. For most of the considered clustering algorithms, dissimilarity measures and values of d, Method 1 (OPD clustering of pseudo-stationary segments) provides much better results than Method 2 (OPD clustering of short equi-sized segments). We suppose that this is due to the fact that lengths of the segments required to assume their stationarity are too small to provide a reliable estimation of OPD.
3. In general, the higher order d, the better clustering results in this experiment are. The only exception is $d = 5$ for that the results are worse than for $d = 4$, but in this case the empirical condition (5.4) is not satisfied.
4. The obtained results show the potential of OPD clustering, however, some methods provide a better classification for the Bonn dataset (see [TTF09] for the review).

5.3.2 Applications to sleep data

In this section we employ ordinal-patterns-based segmentation and OPD clustering to sleep stage discrimination, which is a relevant problem in biomedical research [SAIB$^+$07, Lib12]. According to the classification in [RK68], there are six sleep stages:

1. the awake state;
2. two stages of light sleep (S1, S2);
3. two stages of deep sleep (S3, S4);
4. rapid eye movement (REM).

A classical approach to the automatic discrimination of sleep stages is based on studying spectral characteristics of EEG [BDKS99, KRDF01], though some empirical studies show that non-linear measures yield better discrimination of sleep stages [FRMS96]. Construction of a generally recognized automatic procedure of sleep stages discrimination is problematic due to the complex nature of EEG signal; thus segmentation and discrimination of sleep EEG are mainly carried out manually by experts [SAIB$^+$07]. The result of manual scoring is called *hypnogram*: it is a sequence of sleep stages computed for 30-s epochs.

Can ordinal-patterns-based methods be helpful for the analysis of sleep EEG? To answer this question we apply to EEG data ordinal-patterns-based discrimination. First we segment each recording of a dataset by using the method for change-point detection via the CEofOP statistic (details of the segmentation of multivariate time series are described in Algorithm 5 in Subsection 5.3.3), then we cluster the obtained segments of all EEG time series with the k-means algorithm and the squared Hellinger distance ρ_H. In Experiments 5.4 and 5.5 we compare the discrimination of sleep stages obtained by using this method with the manual scoring by experts.

Experiment 5.4: investigating efficiency of ordinal-patterns-based segmentation and OPD clustering for discrimination between sleep stages.

Objects: 19 night EEG recordings with sampling rate 500 Hz from a dataset with manually scored hypnograms, kindly provided by Vasil Kolev (Institute of Neurobiology Bulgarian Academy of Sciences)[22]. For each recording we have selected a part that is not shorter than 2 hours and does not contain too many artifacts according to the expert scoring (see section "Discussion and conclusions" of this experiment). For the analysis we use EEG time series from F4 and C4 locations.

Technique. The ordinal-patterns-based discrimination of sleep EEG consists of the following steps:

1. The reference channel (nose) is subtracted from each EEG time series.
2. EEG time series are filtered to the band 1-45 Hz with the Butterworth filter of order 5.
3. The ordinal-patterns-based segmentation procedure described by Algorithm 5 in Subsection 5.3.3 is employed for $d = 4$.
4. OPD clustering for order $d = 4$ using k-means with squared Hellinger distance ρ_H is applied to the segments of all recordings. The number of clusters $K = 6$ is chosen, which corresponds to the number of sleep stages.
5. The obtained clusters were assigned to the following classes:
   - class AWAKE – one cluster;
   - class LIGHT SLEEP – two clusters (one of the clusters may be associated with stage S1, and the other one – with stage S2, but discrimination between these two stages was poor, so we do not distinguish between S1 and S2);
   - class DEEP SLEEP – two clusters (one of the clusters is associated with stage S3, and the other – with stage S4. However, we do not distinguish between S3 and S4 in order to be consistent with the modern classification [SAIB$^+$07].);

[22] The entire dataset consists of 55 EEG recordings, we have rejected 36 recordings after preliminary inspection due to the following reasons:

- too short sleep and too many movement artifacts (recordings 15, 41, 43, 49, 55);
- recording problems, disconnection of relevant electrodes, etc (recordings 04, 11, 27, 45, 56);
- erroneous file of manual scoring (recording 03);
- too many artifacts (recordings 01, 02, 05, 07–10, 12, 13, 16–18, 21, 26, 28, 29, 33, 36, 40, 44, 47, 48, 50, 53, 57)

The selected recordings are listed in Table 5.8.

- class REM – one cluster.

The transition-state segments (see Algorithm 5) are considered as unclassified.

Results: Table 5.7 presents the correspondence between the results of ordinal-patterns-based discrimination vs. manual scoring. Amounts of correctly identified epochs are shown in **bold**. The share of correctly identified epochs for every recording is shown in Table 5.8.

		Results of ordinal-patterns-based discrimination				
		AWAKE	LIGHT SLEEP	DEEP SLEEP	REM	unclassified
Manual score	W	**833**	69	0	347	3
	S1, S2	84	**5550**	398	1222	18
	S3, S4	1	807	**3256**	17	14
	REM	2	896	23	**1340**	2
	unclassified	8	90	19	41	5

Table 5.7: Contingency matrices between the manual score and the results of OPD clustering (for all 19 recordings considered in Experiment 5.4)

Recording	Amount of sleep-related epochs	Agreement
06	243	0.765
14	863	0.796
19	841	0.774
20	1071	0.767
22	700	0.541
23	914	0.786
24	497	0.761
25	728	0.632
30	1067	0.725
31	893	0.779
32	932	0.806
34	1054	0.682
35	822	0.721
38	814	0.849
39	1055	0.694
42	705	0.765
46	886	0.594
52	373	0.756
54	587	0.675
Overall	15045	0.730

Table 5.8: Values of agreement between the manual scoring and the results of OPD clustering for sleep EEG recordings considered in Experiment 5.4

Discussion and conclusions:

1. The overall agreement between the manual scoring and the suggested ordinal-patterns-based method for sleep EEG discrimination is 0.73 (see Table 5.8). This means that the suggested method provides a correct discrimination for 73% epochs, which is a rather good result.
2. Note that the ordinal-patterns-based discrimination is based on completely data-driven procedures of ordinal-patterns-based segmentation and OPD clustering; the only potential problem is the choice of the number K of clusters, which is required for applying the k-means clustering algorithm. In the general case this problem is non-trivial (see [ELLS11, Section 5.5] for a discussion), however here we obtain a discrimination for the natural choice of the number of clusters $K = 6$.
3. We have observed that ordinal-patterns-based segmentation is quite sensitive to EEG artifacts. Here we do not use any artifact removal procedures, but we suppose that this may be useful for further studies.
4. We would like to emphasize that filtering of EEG recordings seems to be important for the suggested method of ordinal-patterns-based discrimination since application of our technique to the unfiltered EEG time series provides much worse results.

Experiment 5.5: investigating efficiency of ordinal-patterns-based segmentation and OPD clustering for discrimination between sleep stages for a publicly available dataset.
Objects: eight EEG recordings with manually scored hypnograms from the dataset described in [KZT$^+$00] and provided by physionet.org [GAG$^+$00]. We refer to the recordings according to their names in the dataset, see Table 5.10. Each recording contains two EEG time series recorded from the Fpz-Cz and Pz-Oz locations and sampled at 100 Hz. Four recordings contain only night EEG and are considered entirely; other recordings contain EEG monitored during 24 hours, for them only the night-related part is investigated.

Technique. We apply the steps 2-4 of the procedure described in the section “Technique” of Experiment 5.4 with the only difference: here we have not obtained a meaningful discrimination of sleep stages for number of clusters $K < 8$ since OPDs for EEG in the waking state and during the light sleep differ significantly for different persons. Therefore, we take $K = 8$, analyze the obtained clusters and group them into larger classes according to the manual scoring for the majority of the epochs in a cluster:

- class AWAKE – three clusters;
- class LIGHT SLEEP – three clusters;
- class DEEP SLEEP – one cluster;
- class REM – one cluster.

The transition-state segments (see Algorithm 5 in Subsection 5.3.3) are considered as unclassified.

Results: Table 5.9 presents the correspondence between the results of ordinal-patterns-based discrimination vs. manual scoring. Amounts of correctly identified epochs are shown in **bold**. The share of correctly identified epochs for every recording is shown in Table 5.10.

		Results of ordinal-patterns-based discrimination				
		AWAKE	LIGHT SLEEP	DEEP SLEEP	REM	unclassified
Manual score	W	**440**	165	0	103	3
	S1, S2	110	**3234**	227	635	19
	S3, S4	0	404	**880**	10	5
	REM	0	244	0	**1365**	0
	unclassified	3	6	1	1	0

Table 5.9: Contingency matrices between the manual score and the results of OPD clustering (for all eight recordings from the dataset [GAG$^+$00])

Recording	Amount of sleep-related epochs	Agreement
sc4002	1050	0.736
sc4012	1150	0.737
sc4102	1050	0.838
sc4112	750	0.792
st7022	944	0.612
st7052	1032	0.813
st7121	1027	0.807
st7132	852	0.689
Overall	7855	0.754

Table 5.10: Agreement between the manual scoring and the results of OPD clustering for sleep EEG recordings from the dataset [GAG$^+$00]

Figure 5.4 illustrates the outcome of the ordinal-patterns-based discrimination of sleep EEG in comparison with the hypnogram.

Discussion and conclusions.

1. The overall agreement between the manual scoring and the suggested ordinal-patterns-based method for sleep EEG discrimination is 0.754 (see Table 5.10), which is comparable with the results for the same dataset reported by researchers that used different discrimination methods. In particular, in studies [BDHS$^+$07] and [RJK$^+$12] authors obtain agreement with the manual scoring equal to 0.745 and 0.815, respectively.

2. We obtain here results, similar to Experiment 5.4, but we use for this $K = 8$ clusters and then manually assign them to classes. We suppose that the bad

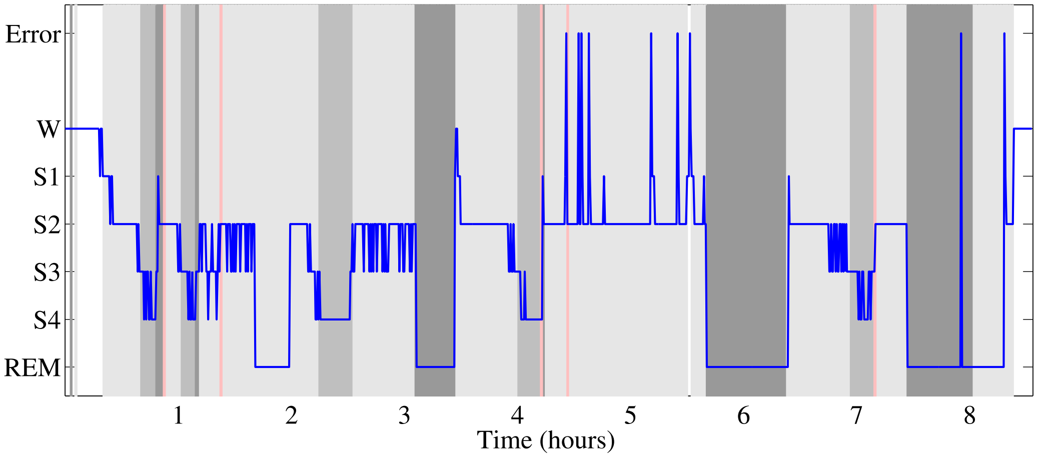


Figure 5.4: Hypnogram (bold curve) and the results of ordinal-patterns-based discrimination of sleep EEG (white color corresponds to class AWAKE, light gray – to LIGHT SLEEP, gray – to DEEP SLEEP, dark gray – to REM, red – to unclassified segments) for recording st7121 from the dataset [GAG+00]

clustering results for $K < 8$ are due to a small number of epochs used for clustering in comparison with Experiment 5.4.

Remark. We have also tried to apply to the EEG recordings from Experiments 5.4 and 5.5 the OPD clustering of 2-s segments as suggested in [SKC13]. We have followed the original paper in using the complete linkage clustering algorithm with the Euclidean distance ρ_E; we have tried $d = 2, 3$ that satisfy condition (5.4). However, OPD clustering of 2-s segments did not provide any sensible discrimination of sleep stages for the most of the considered EEG recordings. In particular, we have encountered the two following problems.

- The method from [SKC13] has very high storage requirements (the discrimination of k recordings requires using a matrix with more than $k^2 \cdot 10^8$ entries), which makes clustering segments of several EEG recordings problematic.
- Considering distributions of ordinal patterns in relatively short 2-s segments leads to a considerable number of outliers. Many clustering algorithms and, especially, complete linkage clustering, are sensitive to outliers [MC02, p.392], [BS12, Section 4.5]; in particular, we observe that when applying OPD clustering to 2-s segments of sleep EEG recordings, clusters are often formed around outliers and discrimination of sleep stages fails.

We obtain in Experiments 5.4 and 5.5 a rather high agreement with the manual scoring by applying the ordinal-patterns-based discrimination to the two entirely different datasets. Indeed, the considered EEG recordings were acquired by a different equipment from different locations of scalp electrodes, and even the sampling rate for the recordings from the two datasets is notably different.

Our results show a potential of ordinal-patterns-based segmentation and OPD clustering for discrimination of sleep stages in EEG data. Though we do not claim that ordinal-patterns-based methods may provide a fully automatic sleep scoring, such methods can be used as a preliminary step for the discrimination of sleep stages by an expert or by more effective and complex automatized methods. Another possibility is to construct a method for discrimination of sleep stages based on combined use of ordinal and non-ordinal features. Finally, one could improve the results of ordinal-patterns-based discrimination by using classification techniques different from clustering, for instance, support vector machines [Bur98].

5.3.3 An algorithm for ordinal-patterns-based segmentation of a multi-channel EEG recording

Here we present Algorithm 5, which is used in Subsection 5.3.2 for discrimination of multi-channel sleep EEG time series. We set there the minimal length τ_{valid} of a valid stationary segment to 3000 time points (that is a valid stationary segment should be at least 30-s., which corresponds to the length of an epoch used in a manual scoring).

5.4 Conclusions

Results of the experiments described in this chapter allow us to draw the following conclusions:

1. Ordinal-pattern-distributions (OPD) clustering is a rather effective technique for discrimination of stationary segments of time series (one can obtain such segments using ordinal-patterns-based segmentation via CEofOP statistic described in Chapter 4).

2. Effectiveness of OPD clustering of time series segments strongly depends on their lengths. According to our findings, the length of a time series segment should satisfy $L > 20(d+1)!$ for $d \leq 5$, however this boundary is completely empirical and needs further investigation.

3. The best candidates for using in OPD clustering are the k-means and complete linkage clustering algorithms with the squared Hellinger, city block or Euclidean distances. Specifically, we obtained the best clustering results when using the k-means clustering algorithms with the squared Hellinger distance.

4. The results obtained by applying ordinal-patterns-based segmentation and OPD clustering to discrimination of sleep stages on the basis of EEG data demonstrate the potential of ordinal-patterns-based methods for analysis of real-world data, though further studies in this direction are required.

Algorithm 5 Ordinal-patterns-based segmentation of a multi-channel EEG recording

Input: N-channel EEG recording $y_j = \big(y_j(0), y_j(1), \ldots, y_j(L)\big)$ of length L for $j = 1, 2, \ldots, N$, order d of ordinal patterns, minimal length τ_{valid} of a valid stationary segment

```
function SegmentEEG($y_1, y_2, \ldots, y_N$, $d$, $\tau_{\text{valid}}$)
    $\alpha \leftarrow 0.08$;                              ▷ corresponds to probability of false alarms 0.07
    $\tau_{\min} \leftarrow 2\big((d+1)^2 d!\big)$;          ▷ minimal time between change-points
                                                           ▷ as it was suggested in Subsection 4.5.2
    $\widehat{N}_{\text{st}} \leftarrow 0$; $\widehat{t}^*_0 \leftarrow 0$;
    for $j = 1, 2, \ldots, N$ do                           ▷ detect change-points for every EEG channel
        $\big(\pi_j(d), \ldots, \pi_j(L)\big) \leftarrow$ sequence of ordinal patterns of order $d$ for $y_j$
        $\widehat{N}^j_{\text{st}}, \big(\widehat{t}^j_0, \widehat{t}^j_1, \ldots, \widehat{t}^j_{\widehat{N}^j_{\text{st}}}\big) \leftarrow$ Problem3$\Big(\big(\pi_j(d), \ldots, \pi_j(L)\big)$, $\alpha$, CEofOP, $\tau_{\text{valid}}\Big)$;
        ▷ insert change-points for each EEG channel into the joint list:
        $\big(\widehat{t}^*_{\widehat{N}_{\text{st}}+1}, \widehat{t}^*_{\widehat{N}_{\text{st}}+2}, \ldots, \widehat{t}^*_{\widehat{N}_{\text{st}}+\widehat{N}^j_{\text{st}}-1}\big) \leftarrow \big(\widehat{t}^j_1, \widehat{t}^j_2, \ldots, \widehat{t}^j_{\widehat{N}^j_{\text{st}}-1}\big)$;
        $\widehat{N}_{\text{st}} \leftarrow \widehat{N}_{\text{st}} + \widehat{N}^j_{\text{st}} - 1$;
    end for
    $\widehat{N}_{\text{st}} \leftarrow \widehat{N}_{\text{st}} + 1$; $\widehat{t}^*_{\widehat{N}_{\text{st}}} \leftarrow L$;
    $\big(\widehat{t}^*_0, \widehat{t}^*_1, \ldots, \widehat{t}^*_{\widehat{N}_{\text{st}}}\big) \leftarrow$ Sort$\Big(\widehat{t}^*_0, \widehat{t}^*_1, \ldots, \widehat{t}^*_{\widehat{N}_{\text{st}}}\Big)$; ▷ sort change-points in increasing order
    $t \leftarrow \widehat{t}^*_1$; $k \leftarrow 2$;
    ▷ If a change-point is too near to the previous, we suppose that these change-points
    ▷ correspond to the same change reflected by different channels not simultaneously.
    ▷ We mark the segment between such change-points as a transition state.
    repeat
        if $\widehat{t}^*_k - t \geq \tau_{\text{valid}}$ then
            $t \leftarrow \widehat{t}^*_k$;
            if TransSegment$_k = -1$ then
                ▷ If previous segment is transitional, then merge it with the current:
                Delete $\widehat{t}^*_{k-1}$ from the change-points list;
                $\widehat{N}_{\text{st}} \leftarrow \widehat{N}_{\text{st}} - 1$;
            else
                TransSegment$_k \leftarrow -1$;        ▷ mark $k$-th segment as a transition state
            end if
        else
            $t \leftarrow \widehat{t}^*_k$;
            TransSegment$_k \leftarrow 0$;
            $k \leftarrow k + 1$;
        end if
    until $k < \widehat{N}_{\text{st}}$;
    return $\widehat{N}_{\text{st}}, \big(\widehat{t}^*_0, \ldots, \widehat{t}^*_{\widehat{N}_{\text{st}}}\big), \big(\text{TransSegment}_0, \ldots, \text{TransSegment}_{\widehat{N}_{\text{st}}-1}\big)$;
end function
```

Bibliography

[Adl98] R.L. Adler. Symbolic dynamics and Markov partitions. *Bulletin of the American Mathematical Society, New Series*, 35(1): 1–56, 1998.

[AEK08] J.M. Amigó, S. Elizalde, M.B. Kennel. Forbidden patterns and shift systems. *Journal of Combinatorial Theory. Series A*, 115(3): 485–504, 2008.

[AF07] J. Abonyi, B. Feil. *Cluster analysis for data mining and system identification.* Springer, 2007.

[AG57] T.W. Anderson, L.A. Goodman. Statistical inference about markov chains. *Annals of Mathematical Statistics*, 28: 89–110, 1957.

[AK08] J.M. Amigó, M.B. Kennel. Forbidden ordinal patterns in higher dimensional dynamics. *Physica D*, 237(22): 2893–2899, 2008.

[AK13] J.M. Amigó, K. Keller. Permutation entropy: One concept, two approaches. *European Physical Journal Special Topics*, 222(2): 263–273, 2013.

[AKK05] J.M. Amigó, M.B. Kennel, L. Kocarev. The permutation entropy rate equals the metric entropy rate for ergodic information sources and ergodic dynamical systems. *Physica D*, 210(1-2): 77–95, 2005.

[AKS92] J. Ashley, B. Kitchens, M. Stafford. Boundaries of Markov partitions. *Transactions of the American Mathematical Society*, 333(1): 177–201, 1992.

[ALM$^+$01] R.G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, C.E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64: 061907, 2001.

[Ami10] J.M. Amigó. *Permutation complexity in dynamical systems. Ordinal patterns, permutation entropy and all that.* Berlin Heidelberg: Springer, 2010.

[Ami12] J.M. Amigó. The equality of Kolmogorov–Sinai entropy and metric permutation entropy generalized. *Physica D*, 241(7): 789–793, 2012.

[And03] R.G. Andrzejak. EEG time series download page. http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html, 2003.

[AZS08] J.M. Amigó, S. Zambrano, M.A.F. Sanjuán. Combinatorial detection of determinism in noisy time series. *Europhysics Letters*, 83: 60005, 2008.

[BD93] B.E. Brodsky, B.S. Darkhovsky. *Nonparametric methods in change-point problems.* Dordrecht: Kluwer Academic Publishers, 1993.

[BD00] B.E. Brodsky, B.S. Darkhovsky. *Non-parametric statistical diagnosis. Problems and methods.* Dordrecht: Kluwer Academic Publishers, 2000.

[BD09] P.J. Brockwell, R.A. Davis. *Time Series: Theory and Methods.* New York: Springer, 2009.

[BDHS$^+$07] C. Berthomier, X. Drouot, M. Herman-Stoïca, P. Berthomier, J. Prado, D. Bokar-Thire, O. Benoit, J. Mattout, M.-P. d'Ortho. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep*, 30(11): 1587–1595, 2007.

[BDKS99] B.E. Brodsky, B.S Darkhovsky, A.Ya. Kaplan, S.L. Shishkin. A nonparametric method for the segmentation of the EEG. *Computer Methods and Programs in Biomedicine*, 60(2): 93–106, 1999.

[BHM$^+$13] A. Batsidis, L. Horvàth, N. Martìn, L. Pardo, K. Zografos. Change-point detection in multinomial data using phi-divergence test statistics. *Journal of Multivariate Analysis*, 118: 53–66, 2013.

[BKP02] C. Bandt, G. Keller, B. Pompe. Entropy of interval maps via permutations. *Nonlinearity*, 15(5): 1595–1602, 2002.

[BN93] M. Basseville, I.V. Nikiforov. *Detection of abrupt changes: theory and application.* Upper Saddle River, NJ: Prentice-Hall, Inc., 1993.

[BP02] C. Bandt, B. Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88(174102), 2002.

[Bra11] A.M. Brandmaier. *Permutation distribution clustering and structural equation model trees.* PhD thesis, University of Saarland, 2011.

[BS02] M. Brin, G. Stuck. *Introduction to dynamical systems.* Cambridge: Cambridge University Press, 2002.

[BS07] C. Bandt, F. Shiha. Order patterns in time series. *Journal of Time Series Analysis*, 28(5): 646–665, 2007.

[BS12] S. Bandyopadhyay, S. Saha. *Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications.* Springer, 2012.

[Bur98] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2): 121–167, 1998.

[Cao97] L. Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D*, 110(1-2): 43–50, 1997.

[CE07] P. Collet, J.-P. Eckmann. *Concepts and Results in Chaotic Dynamics: A Short Course.* Berlin Heidelberg: Springer, 2007.

[CFS82] I.P. Cornfeld, S.V. Fomin, Ya.G. Sinai. *Ergodic theory.* Berlin: Springer, 1982.

[Che12] M. Chen. EM algorithm for Gaussian mixture model. MATLAB Central File Exchange. http://www.mathworks.com/matlabcentral/fileexchange/26184-em-algorithm-for-gaussian-mixture-model, 2012.

[Cho05] G.H. Choe. *Computational ergodic theory.* Berlin: Springer, 2005.

[CMS94] E. Carlstein, H.G. Muller, D. Siegmund. *Change-point problems.* Hayward, CA: Institute of Mathematical Statistics, 1994.

[CT06] T.M. Cover, J.A. Thomas. *Elements of information theory. 2nd ed.* Hoboken: John Wiley & Sons, 2006.

[Den74] M. Denker. Finite generators for ergodic, measure-preserving transformations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 29(1): 45–55, 1974.

[DH97] A.C. Davison, D.V. Hinkley. *Bootstrap Methods and their Applications.* Cambridge: Cambridge University Press, 1997.

[Eli09] S. Elizalde. The number of permutations realized by a shift. *SIAM Journal on Discrete Mathematics*, 23(2): 765–786, 2009.

[Eli11] S. Elizalde. Permutations and β-shifts. *Journal of Combinatorial Theory. Series A*, 118(8): 2474–2497, 2011.

[ELLS11] B.S. Everitt, S. Landau, M. Leese, D Stahl. *Cluster Analysis. 5th ed.* Chichester: Wiley, 2011.

[ELW11] M. Einsiedler, E. Lindenstrauss, T. Ward. Entropy in Dynamics (unpublished). http://maths.dur.ac.uk/∼tpcc68/entropy/welcome.html, 2011.

[ER85] J.-P. Eckmann, D. Ruelle. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57(3): 617–656, 1985.

[ER92] J.-P. Eckmann, D. Ruelle. Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. *Physica D*, 56(2-3): 185–187, 1992.

[Faw06] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27: 861–874, 2006.

[FRMS96] J. Fell, J. Röschke, K. Mann, C. Schäffner. Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures. *Electroencephalography and clinical Neurophysiology*, 98(5): 401–410, 1996.

[GAG$^+$00] A.L Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new Research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000.

[GBR$^+$07] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A.J. Smola. A kernel approach to comparing distributions. *Proceedings of the 22nd national conference on Artificial intelligence*, 2: 1637–1641, 2007.

[GBR$^+$12] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A.J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1): 723–773, 2012.

[GFHS09] A. Gretton, K. Fukumizu, Z. Harchaoui, B.K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[GKK$^+$10] D. Gusak, A. Kukush, A. Kulik, Yu. Mishura, A. Pilipenko. *Theory of Stochastic Processes: With Applications to Financial Mathematics and Risk Theory*. Springer, 2010.

[GR84] U. Grenander, M. Rosenblatt. *Statistical analysis of stationary time series*, volume 320. New York: Chelsea Pub., 1984.

[Gra88] P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7): 369–373, 1988.

[Gra03] P. Grassberger. Entropy estimates from insufficient samplings. *e-print – arXiv preprint physics/0307138*, 2003.

[Gra09] R.M. Gray. *Probability, random processes, and ergodic properties.* Springer, 2009.

[Gre07] M. Greenacre. *Correspondence analysis in practice. 2d ed.* CRC Press, 2007.

[GRS05] S. Gudmundsson, T.P. Runarsson, S. Sigurdsson. Automatic sleep staging using support vector machines with posterior probability estimates. In *Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*, volume 2, pages 366–372. IEEE, 2005.

[HA85] L. Hubert, P. Arabie. Comparing partitions. *Journal of classification*, 2(1): 193–218, 1985.

[HK02] T.S. Han, K. Kobayashi. *Mathematics of information and coding. Transl. from the Japanese by J. Suzuki.* Providence: American Mathematical Society, 2002.

[HN11] T. Haruna, K. Nakajima. Permutation complexity via duality between values and orderings. *Physica D*, 240(17): 1370–1377, 2011.

[HS95] L. Horváth, M. Serbinowska. Testing for changes in multinomial observations: the Lindisfarne Scribes problem. *Scandinavian Journal of Statistics*, 22: 371–384, 1995.

[KAH$^+$06] H. Kantz, E.G. Altmann, S. Hallerberg, D. Holstein, A. Riegert. Dynamical Interpretation of Extreme Events: Predictability and Predictions. In S. Albeverio, V. Jentsch, H. Kantz, editors, *Extreme Events in Nature and Society*, The Frontiers Collection, pages 69–93. Springer Berlin Heidelberg, 2006.

[Kel12] K. Keller. Permutations and the Kolmogorov-Sinai entropy. *Discrete and Continuous Dynamical Systems*, 32(3): 891–900, 2012.

[Kit98] B.P. Kitchens. *Symbolic dynamics. One-sided, two-sided and countable state Markov shifts.* Berlin: Springer, 1998.

[KL03] K. Keller, H. Lauffer. Symbolic analysis of high-dimensional time series. *International Journal of Bifurcation and Chaos*, 13(09): 2657–2668, 2003.

[KMPS09] A.Y. Kim, C. Marzban, D.B. Percival, W. Stuetzle. Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, 89(12): 2529–2536, 2009.

[KRDF01] A. Kaplan, J. Röschke, B. Darkhovsky, J. Fell. Macrostructural EEG characterization based on nonparametric change point segmentation: application to sleep analysis. *Journal of Neuroscience methods*, 106(1): 81–90, 2001.

[Kri70] W. Krieger. On entropy and generators of measure-preserving transformations. *Transactions of the American Mathematical Society*, 149: 453–464, 1970.

[KS09] K. Keller, M. Sinn. A standardized approach to the Kolmogorov-Sinai entropy. *Nonlinearity*, 22(10): 2417–2422, 2009.

[KS10] K. Keller, M. Sinn. Kolmogorov-Sinai entropy from the ordinal viewpoint. *Physica D*, 239(12): 997–1000, 2010.

[KSE07] K. Keller, M. Sinn, J. Emonds. Time series from the ordinal viewpoint. *Stochastics and Dynamics*, 7(2): 247–272, 2007.

[KUU14] K. Keller, A.M. Unakafov, V.A. Unakafova. Ordinal patterns, entropy, and EEG. *Entropy*, 16(12): 6212–6239, 2014.

[KZT$^+$00] B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, J.J.L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9): 1185–1194, 2000.

[Lah03] S.N. Lahiri. *Resampling methods for dependent data.* New York: Springer, 2003.

[Lav99] M. Lavielle. Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1): 79–102, 1999.

[LE98] K. Lehnertz, C.E. Elger. Can Epileptic Seizures be Predicted? Evidence from Nonlinear Time Series Analysis of Brain Electrical Activity. *Physical Review Letters*, 80: 5019–5022, 1998.

[Lee12] K. Lee. Fast Approximate Entropy. MATLAB Central File Exchange. http: //www.mathworks.com/matlabcentral/fileexchange/32427-fast-approximate-entropy, 2012.

[Lib12] M. H. Libenson. *Practical approach to electroencephalography.* Elsevier Health Sciences, 2012.

[LM95] D. Lind, B. Marcus. *An introduction to symbolic dynamics and coding.* Cambridge: Cambridge University Press, 1995.

[LT07] M. Lavielle, G. Teyssière. Adaptive detection of multiple change-points in asset price volatility. In G. Teyssière, A.P. Kirman, editors, *Long Memory in Economics*, pages 129–156. Berlin Heidelberg: Springer, 2007.

[Lyu02] M. Lyubich. Almost every real quadratic map is either regular or stochastic. *Annals of Mathematics. Second Series*, 156(1): 1–78, 2002.

[Lyu12] M. Lyubich. Forty years of unimodal dynamics: on the occasion of Artur Avila winning the Brin prize. *Journal of Modern Dynamics*, 6(2): 183–203, 2012.

[MAAB13] R. Monetti, J.M. Amigó, T. Aschenbrenner, W. Bunk. Permutation complexity of interacting dynamical systems. *The European Physical Journal Special Topics*, 222(2): 421–436, 2013.

[MAEL07] F. Mormann, R.G. Andrzejak, C.E. Elger, K. Lehnertz. Seizure prediction: the long and winding road. *Brain*, 130(2): 314–333, 2007.

[MC02] G.J. Miao, M.A. Clements. *Digital signal processing and statistical classification.* Artech House, 2002.

[MIH08] S. Miyamoto, H. Ichihashi, K. Honda. *Algorithms for fuzzy clustering: methods in c-means clustering with applications.* Springer, 2008.

[Mis03] M. Misiurewicz. Permutations and topological entropy for interval maps. *Nonlinearity*, 16(3): 971, 2003.

[Mis10] M. Misiurewicz. Sensitive dependence on parameters. *Journal of Mathematical Physics*, 51(10): 102704, 2010.

[MN00] M. Martens, T. Nowicki. *Invariant measures for typical quadratic maps.* Paris: Astérisque, 2000.

[MPW$^+$09] C. Merkwirth, U. Parlitz, I. Wedekind, D. Engster, W. Lauterborn. TSTOOL Home Page. http://www.physik3.gwdg.de/tstool/index.html, 2009.

[MWWM99] H.R. Moser, B. Weber, H.G. Wieser, P.F. Meier. Electroencephalograms in epilepsy: analysis and seizure prediction within the framework of Lyapunov theory. *Physica D: Nonlinear Phenomena*, 130(3-4): 291–305, 1999.

[Nag82] H.N. Nagaraja. On the non-Markovian structure of discrete order statistics. *Journal of Statistical Planning and Inference*, 7: 29–33, 1982.

[NB99] H. Nagashima, Yo. Baba. *Introduction to chaos. Physics and mathematics of chaotic phenomena. Transl. from the Japanese by M. Nakahara.* Bristol: Institute of Physics Publishing, 1999.

[Par60] W. Parry. On the β-expansions of real numbers. *Acta Mathematica Academiae Scientiarum Hungaricae*, 11: 401–416, 1960.

[Par98] U. Parlitz. Nonlinear time-series analysis. In J.A.K. Suykens, J. Vandewalle, editors, *Nonlinear Modeling*, pages 209–239. Boston: Kluwer Academic Publishers, 1998.

[Pes97] Ya.B. Pesin. *Dimension theory in dynamical systems: contemporary views and applications.* Chicago: University of Chicago Press, 1997.

[Pin91] S.M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6): 2297–2301, 1991.

[Pol07] A.M. Polansky. Detecting change-points in Markov chains. *Computational Statistics & Data Analysis*, 51(12): 6013–6026, 2007.

[Pom13] B. Pompe. The LE-statistic. *The European Physical Journal Special Topics*, 222(2): 333–351, 2013.

[PR11] B. Pompe, J. Runge. Momentary information transfer as a coupling measure of time series. *Physical Review E*, 83: 051122, 2011.

[PSL13] U. Parlitz, H. Suetani, S. Luther. Identification of equivalent dynamics using ordinal pattern distributions. *The European Physical Journal Special Topics*, 222(2): 553–568, 2013.

[PW77] W. Parry, R.F. Williams. Block coding and a zeta function for finite Markov chains. *Proceedings of the London Mathematical Society, Third Series*, 35: 483–495, 1977.

[RJK$^+$12] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, I. Provazník. Sleep scoring using artificial neural networks. *Sleep Medicine Reviews*, 16(3): 251–263, 2012.

[RK68] A. Rechtschaffen, A. Kales. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects.* Washington: Public Health Service US Government Printing Office, 1968.

[SAIB$^+$07] M.H. Silber, S. Ancoli-Israel, M.H. Bonnet, S. Chokroverty, M.M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S.A. Keenan, M.H. Kryger, T. Penzel, M.R. Pressman, C. Iber. The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, 3(2): 121–131, 2007.

[Sch01] O. Schmitt. *Remarks on the Generator-Problem.* PhD thesis, University of Göttingen, 2001.

[SGK12] M. Sinn, A. Ghodsi, K. Keller. Detecting change-points in time series by maximum mean discrepancy of ordinal pattern distributions. In *Uncertainty in Artificial Intelligence, Proceedings of the 28th Conference*, pages 786–794, 2012.

[Sin10] M. Sinn. *Estimation of Ordinal Pattern Probabilities in Stochastic Processes.* PhD thesis, University of Lübeck, 2010.

[SK11] M. Sinn, K. Keller. Estimation of ordinal pattern probabilities in gaussian processes with stationary increments. *Computational Statistics & Data Analysis*, 55(4): 1781–1790, 2011.

[SKC13] M. Sinn, K. Keller, B. Chen. Segmentation and classification of time series using ordinal pattern distributions. *European Physical Journal Special Topics*, 222(2): 587–598, 2013.

[Spr03] J.C. Sprott. *Chaos and time-series analysis.* Oxford: Oxford University Press, 2003.

[SSH99] Z. Sidák, P.K. Sen, J. Hájek. *Theory of rank tests. 2nd ed.* New York: Academic press, 1999.

[ST01] W.D. Smith, R.L. Taylor. Dependent bootstrap confidence intervals. In I.V. Basawa, C.C. Heyde, R.L. Taylor, editors, *Selected Proceedings of the Symposium on Inference for Stochastic Processes*, pages 91–108. Institute of Mathematical Statistics, 2001.

[Sto12] D.S. Stoffer. Frequency Domain Techniques in the Analysis of DNA Sequences. In T.S. Rao, S.S. Rao, C.R. Rao, editors, *Handbook of Statistics: Time Series Analysis: Methods and Applications*, pages 261–296. Elsevier, 2012.

[Thu01] H. Thunberg. Periodicity versus chaos in one-dimensional dynamics. *SIAM Review*, 43(1): 3–30, 2001.

[TK98] H.M. Taylor, S. Karlin. *An introduction to stochastic modeling. 3rd ed.* Boston: Academic Press, 1998.

[TTF09] A.T. Tzallas, M.G. Tsipouras, D.I. Fotiadis. Epileptic seizure detection in EEGs using time–frequency analysis. *IEEE Transactions on information Technology in Biomedicine*, 13(5): 703–710, 2009.

[UK13] V.A. Unakafova, K. Keller. Efficiently measuring complexity on the basis of real-world data. *Entropy*, 15(10): 4392–4415, 2013.

[UK14] A.M. Unakafov, K. Keller. Conditional entropy of ordinal patterns. *Physica D*, 269: 94–102, 2014.

[Una15] V.A. Unakafova. *Investigating measures of complexity for dynamical systems and for time series.* PhD thesis, University of Lübeck, 2015.

[UUK13] V.A. Unakafova, A.M. Unakafov, K. Keller. An approach to comparing Kolmogorov-Sinai and permutation entropy. *European Physical Journal Special Topics*, 222(2): 353–361, 2013.

[Vos81] L.Yu. Vostrikova. Detecting disorder in multidimensional random processes. *Soviet Mathematics Doklady*, 24(1): 55–59, 1981.

[VTS04] J.-P. Vert, K. Tsuda, B. Schölkopf. A primer on kernel methods. In *Kernel Methods in Computational Biology*, pages 35–70. MIT Press, 2004.

[Wal00] P. Walters. *An introduction to ergodic theory.* New York: Springer, 2000.

[You02] L.-S. Young. What are SRB measures, and which dynamical systems have them? *Journal of Statistical Physics*, 108(5-6): 733–754, 2002.

[You03] L.-S. Young. Entropy in dynamical systems. In A. Greven, G. Keller, G. Warnecke, editors, *Entropy.* Princeton: Princeton University Press, 2003.

[You13] L.-S. Young. Mathematical theory of lyapunov exponents. *Journal of Physics A: Mathematical and Theoretical*, 46(25): 254001, 2013.

[ZLB$^+$11] K.H. Zou, A. Liu, A.I. Bandos, L. Ohno-Machado, H.E. Rockette. *Statistical evaluation of diagnostic performance: topics in ROC analysis.* CRC Press, 2011.