Event Impact Analysis for Time Series

Tests, Measures, and Models

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Technischen Universität Dortmund
an der Fakultät für Informatik

von

Erik Scharwächter

Dortmund
2022

Tag der mündlichen Prüfung: 14.11.2022

Dekan: Prof. Dr.-Ing. Gernot Fink
Gutachter: Prof. Dr. Emmanuel Müller, Prof. Dr. Carsten Jentsch

Abstract

Time series arise in a variety of application domains—whenever data points are recorded over time and stored for subsequent analysis. A critical question is whether the occurrence of events like natural disasters, technical faults, or political interventions leads to changes in a time series, for example, temporary deviations from its typical behavior. The vast majority of existing research on this topic focuses on the *specific* impact of a *single* event on a time series, while methods to *generically* capture the impact of a *recurring* event are scarce. In this thesis, we fill this gap by introducing a novel framework for event impact analysis in the case of *randomly recurring* events. We develop a statistical perspective on the problem and provide a generic notion of event impacts based on a statistical independence relation. The main problem we address is that of establishing the presence of event impacts in stationary time series using statistical independence tests. Tests for event impacts should be generic, powerful, and computationally efficient. We develop two algorithmic test strategies for event impacts that satisfy these properties. The first is based on coincidences between events and peaks in the time series, while the second is based on multiple marginal associations. We also discuss a selection of follow-up questions, including ways to measure, model and visualize event impacts, and the relationship between event impact analysis and anomaly detection in time series. At last, we provide a first method to study event impacts in nonstationary time series. We evaluate our methodological contributions on several real-world datasets and study their performance within large-scale simulation studies.

Publications

This thesis is based on the following publications that contain original contributions by the first author. Emmanuel Müller supported these publications in his role as a supervisor, by asking critical questions and providing guidelines for presenting the material. Jonathan Lennartz helped implementing one of the experiments.

[SM20a] Erik Scharwächter and Emmanuel Müller. "Does Terrorism Trigger Online Hate Speech? On the Association of Events and Time Series." *Annals of Applied Statistics* 14(3), 2020, pp. 1285–1303. DOI: 10.1214/20-AOAS1338 (cit. on p. 27).

[SM20b] Erik Scharwächter and Emmanuel Müller. "Statistical Evaluation of Anomaly Detectors for Sequences." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Mining and Learning from Time Series (KDD MiLeTS)*, 2020 (cit. on p. 101).

[SM20c] Erik Scharwächter and Emmanuel Müller. "Two-Sample Testing for Event Impacts in Time Series." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2020. DOI: 10.1137/1.9781611976236.2 (cit. on pp. 57, 63).

[SLM21] Erik Scharwächter, Jonathan Lennartz, and Emmanuel Müller. "Differentiable Segmentation of Sequences." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021 (cit. on p. 113).

[SM22] Erik Scharwächter and Emmanuel Müller. "Discrete Probabilistic Models for Time Warping." In: *Manuscript in review*, 2022 (cit. on p. 79).

Contents

List of Figures

List of Tables

1 Introduction

Technological advances have enabled the large-scale monitoring of many systems—natural or artificial—over time. Examples include monitoring of electricity demand on power grids, air pollution levels in cities, and share values in financial markets. When this data is recorded for subsequent analyses, it forms a *time series* of observations. Time series capture a specific feature of a system over time, and the behavior of a time series, *i.e.*, its trend, variability, or the existence of anomalous patterns, yields insights into the state of the system at any point in time. Natural disasters, technical faults, political interventions and other events often have a multitude of ecological, societal, or economic side-effects. A solid understanding of the impacts of such events on a given system is important for future risk assessment and informed decision making. Time series are perfectly suited to assess the impacts of events, and are used for that purpose in many application domains.

Since the 1960s, research on event impacts on time series has primarily focused on *singular events* [AG03; BT65; CS63; SMP86; WAG+21], as visualized in Figure 1.1. An illustrative present-day example is the study by Silver et al. [SHA+20] who analyze the impacts of the first COVID-19 lockdown in China on time series that measure air quality. In the past decade, an increasing number of publications set out to generalize beyond case studies of singular events. The stated goal is to capture the impacts of *recurring events* on time series [CRK11; KQP+16; LLL+14; ZYW+09]. Since an event is rarely identical to any other event, the notion of a *recurring event* requires meaningful pooling of *prima facie* singular events to a *family of events* that are similar in some regard. This pooling is always application-dependent, and two events that are similar for one analysis may not be similar for another analysis.

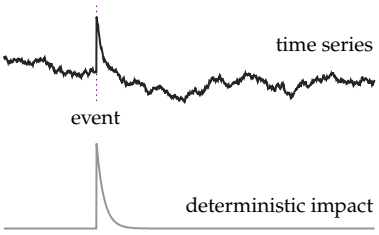

Figure 1.1: Deterministic impacts of a singular event can be estimated, *e.g.*, using intervention analysis [BT75]. Unless otherwise noted, the horizontal axes in our plots denote time, while the vertical axes denote the respective data domains.

An impact analysis with recurring events provides three major advantages over case studies with singular events. On the applied side, the recurring approach justifies *general statements* on the response of the time series to such events. On the technological side, it enables improved *time series forecasting algorithms* that exploit information on recent event occurrences, and it facilitates the development of *event detection algorithms* that operate on the time series. Unfortunately, despite some initial work on the impacts of recurring events on time series, there is still no common understanding of that subject in the scientific literature.

1.1 Contributions

This dissertation provides—to the best of our knowledge—the first framework to systematically analyze the impacts of *recurring events* on time series. We translate the information needs expressed in well-established approaches for the analysis of singular event impacts into a novel, *probabilistic definition* of event impacts for the recurring case. The most important step in this direction is to regard event occurrences as realizations of a stochastic process that, from now on, will be referred to as an *event series*. In our framework, events appear *randomly* at any point in time according to some probability distribution. The stochastic perspective on events is standard procedure in many fields [DV03; DHL+16; Wei18] but has not found its way into studies of event impacts on time series. We also view the time series as a stochastic process, which is in line with the existing literature [BJR+16]. In a nutshell, we propose to study the impacts of recurring events by analyzing the statistical associations between the event series and the time series.

We formalize our novel framework for event impact analysis in Section 1.3, on the level of elementary probability theory. Inspired by the concept of Granger causality [Gra69], we define event impacts very generically via the absence of a specific independence relation across the two stochastic processes. We assume that events occur *rarely*, so that we can distinguish the *typical behavior* of the time series from its (potentially) *deviant behavior* in the presence of an event, using probability distributions only. We further assume that there are no long-range dependencies within the time series. These assumptions considerably simplify our statistical arguments, since they allow taking observations from the time series after two distinct event occurrences as approximately independent. We provide a detailed survey of related work on singular event impacts and on statistical associations between stochastic processes in Section 1.4, and highlight the differences to our framework.

While events *may* be the root cause for the deviant behavior of the time series, the methods we develop in this work are not meant for causal inference. *Statistical association does not imply causation.*

In Chapter 2, we develop statistical tests for event impacts based on *coincidences* between events and *peaks* in the time series. We implement these tests by combining concepts from event coincidence analysis [DSS+16] with results from extreme value theory [Col01]. We use the *trigger coincidence rate* as a measure for the association of events and peaks, and analytically derive its null distribution for a large class of stationary time series. We extend this methodology to peaks over multiple thresholds, and introduce *quantile-trigger rate plots* as a novel visualization of the strength of association.

A focus on peaks is already quite generic, since many patterns of interest within a time series can be *reduced* to peaks when a suitable feature transformation function is applied before the analysis. In

Chapter 3, we propose an alternative test procedure that is not based on peaks. It is directly applicable even on multivariate time series and on time series of other objects like strings or graphs, without preprocessing with a feature transformation function. The procedure is based on a simplified criterion for event impacts that considers only *marginal* associations. This criterion can be tested effectively with multiple two-sample testing, using an algorithm that is computationally efficient even for very long time series.

Once the existence of event impacts has been established by one of the tests from above, a natural follow-up question is how these impacts can be characterized. For this purpose, in Chapter 4, we propose a probabilistic model to capture the *deviant behavior* of univariate time series in a meaningful way. We assume that there is a *prototypical pattern* that appears in the time series whenever an event occurs. However, this pattern is not replicated exactly, but appears in a *temporally distorted* form. Our model captures the data-generating process for such patterns by employing a *probabilistic time warping* mechanism based on multivariate statistics. We study the performance of this model with different choices for the distribution of the probabilistic warping component.

In Chapter 5, we briefly revisit the association measures from Chapter 2, and draw some connections to *evaluation measures* commonly used in machine learning to assess the performance of *anomaly detectors* for time series, in particular, precision and recall. We demonstrate that these measures potentially overestimate the performance of a detection algorithm when employed with temporal tolerance, and argue that they are *uninformative* unless their null distributions are provided.

While all of the methods sketched so far assume stationarity of the time series, Chapter 6 addresses a specific *nonstationary* case. We assume that the time series follows a *segmented model*, where the data-generating process changes its dynamics at discrete change points, and remains stationary within each segment. We develop a continuous relaxation of the standard segmented model formulation that allows estimating the change points and all other model parameters with standard algorithms for gradient descent. Event impact analysis can then be performed individually within each segment, or by correlating the occurrence of events with the occurrence of change points.

We conclude this dissertation in Chapter 7 by summarizing our key findings and pointing out interesting directions for future work.

1.2 Terminology

We briefly describe elementary probabilistic concepts following the textbooks of Shao [Sha03] and Wasserman [Was04], using our own notation. Further results and intuitions are taken from Rosenthal [Ros06], Anderson [And03], Gallager [Gal13] Park [Par18] and Koller and Friedman [KF09]. The purpose of this section is to introduce our notation to the reader, along with some important results that we use. For generality, we describe most concepts in this work in terms of *cumulative distribution functions*.

Random variables

probability space

Let $(\Omega, \mathcal{A}, \Pr)$ be a **probability space** with sample space Ω, σ-field $\mathcal{A}$, and probability measure Pr. The sample space Ω is the set of possible outcomes of the phenomenon that we study, the σ-field $\mathcal{A}$ contains subsets of outcomes $A \subseteq \Omega$ for which we would like to make probability statements, and the probability measure $\Pr : \mathcal{A} \longrightarrow \mathbb{R}$ is the function that consistently assigns these probabilities. The **axioms of probability** enforce that $\Pr(A) \geq 0$ for all $A \in \mathcal{A}$, $\Pr(\Omega) = 1$, and

axioms of probability

$$\Pr\Big(\bigcup_i A_i\Big) = \sum_i \Pr(A_i) \tag{1.1}$$

random variable

measurable function

for disjoint sets $A_i \in \mathcal{A}$. A **random variable** for the probability space $(\Omega, \mathcal{A}, \Pr)$ captures features of the phenomenon that can be expressed numerically. Formally, a random variable is a **measurable function** $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$. All functions on countable sample spaces Ω are measurable, and so are continuous functions over real-valued sample spaces, indicator functions, and many other functions encountered in practice. Importantly, concatenations of measurable functions are also measurable functions. A vector-valued measurable function $\mathbf{x} : \Omega \longrightarrow \mathbb{R}^D$ is called a **random vector**. We write $\mathrm{x}(\omega) = x$ for outcomes of random variables, and $\mathbf{x}(\omega) = \boldsymbol{x}$ for outcomes of random vectors. Vectors of random variables are random vectors, and vice-versa. Sometimes, we also consider sets of random variables as random vectors, where the dimensions are ordered lexicographically. A measurable function to an arbitrary target set $\xi : \Omega \longrightarrow \Omega'$ is called a **random element**.

random vector

random element

distribution

The probability measure Pr induces a **distribution** $\Pr_{\mathrm{x}}$ for every random variable x, where

$$\Pr_{\mathrm{x}}(A) := \Pr(\mathrm{x}^{-1}(A)) = \Pr(\{\omega \in \Omega \mid \mathrm{x}(\omega) \in A\}) \tag{1.2}$$

denotes the probability that the random variable x has an outcome from the set $A \subseteq \mathbb{R}$. Formally, the distribution $\Pr_{\mathrm{x}}$ is a new

probability measure over a standardized probability space (the Borel space). Throughout this work, we use the common simplified notation with logical statements such as $\Pr(\mathrm{x} \in A) := \Pr_{\mathrm{x}}(A)$, $\Pr(\mathrm{x} \leq x) := \Pr_{\mathrm{x}}((-\infty, x])$, or $\Pr(a < \mathrm{x} \leq b) := \Pr_{\mathrm{x}}((a, b])$, *etc.*, to refer to probabilities under the distribution of x. In the same way, the probability measure Pr induces a (joint) distribution $\Pr_{\mathbf{x}} = \Pr_{\mathrm{x}_1,\ldots,\mathrm{x}_D}$ for the random vector $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_D)$. We use the same logical notation to refer to probabilities under the joint distribution, such that, in the bivariate case, $\Pr(\mathrm{x} \in A, \mathrm{y} \in B) := \Pr(\mathrm{x} \in A \wedge \mathrm{y} \in B) := \Pr_{\mathrm{xy}}(\{(x, y) \mid x \in A \wedge y \in B\})$.

Characterizing distributions

The distribution of a random variable x is fully specified by the **cumulative distribution function** (cdf)

cumulative distribution function

$$F_{\mathrm{x}} : \mathbb{R} \longrightarrow [0, 1], x \mapsto \Pr(\mathrm{x} \leq x), \tag{1.3}$$

and we have that $\Pr(a < \mathrm{x} \leq b) = F_{\mathrm{x}}(b) - F_{\mathrm{x}}(a)$. Similarly, the distribution of a D-dimensional random vector $\mathbf{x}$ is fully specified by the (joint) cdf

$$F_{\mathbf{x}} : \mathbb{R}^D \longrightarrow [0, 1], (x_1, \ldots, x_D) \mapsto \Pr(\mathrm{x}_1 \leq x_1, \ldots, \mathrm{x}_D \leq x_D). \tag{1.4}$$

We omit the subscript from the cdf F, if it is clear from the context, and write $\mathrm{x} \sim F$ or $\mathbf{x} \sim F$ to denote that the random variable (random vector) has a distribution with cdf F. Any joint cdf $F(x_1, \ldots, x_D)$ for the random vector $\mathbf{x}$ yields D marginal cdfs $F_d(x_d)$ for its component random variables x_d via

$$F_d(x_d) := \lim_{x_{d'} \to \infty,\ d' \neq d} F(x_1, \ldots, x_D) \tag{1.5}$$

The marginal cdf of x_d derived from the joint cdf F of $\mathbf{x}$ is identical to the cdf of x_d, and we use the terms interchangeably. Usually, we refer to the distribution of x_d as the **marginal distribution** of x_d if it was inferred from a joint distribution. Marginal cdfs and distributions can be defined analogously for any subset of random variables from a random vector.

marginal distribution

We distinguish two types of random variables according to the functional forms of their cdfs:

Definition 1.2.1 *A random variable* x *is* ***discrete****, if there is a sequence of real numbers* $a_1 < a_2 < \ldots$*, and a sequence of non-negative numbers* $\pi_1, \pi_2, \ldots$ *with* $\sum_k \pi_k = 1$*, such that*

discrete

$$F_{\mathrm{x}}(x) = \begin{cases} \sum_{i=1}^{k} \pi_i, & \text{if } a_k \leq x < a_{k+1}, k = 1, 2, \ldots \\ 0, & \text{if } x < a_1. \end{cases}$$

probability mass function

The function $p_{\mathrm{x}} : \mathbb{R} \longrightarrow \mathbb{R}$ *with* $p_{\mathrm{x}}(a_k) = \pi_k$ *for* $k = 1, 2, \ldots$, *and* $p_{\mathrm{x}}(x) = 0$ *everywhere else is called the* ***probability mass function*** *(pmf) of* x, *and* $\mathrm{x} \sim p_{\mathrm{x}}$ *denotes that* x *is discrete with mass function* p_{x}.

continuous

Definition 1.2.2 *A random variable* x *is* ***continuous****, if there is a non-negative function* f_{x} *with* $\int_{-\infty}^{\infty} f_{\mathrm{x}}(u)\, du = 1$ *such that*

$$F_{\mathrm{x}}(x) = \int_{-\infty}^{x} f_{\mathrm{x}}(u)\, du.$$

probability density function

The function f_{x} *is the* ***probability density function*** *(pdf) of* x, *and* $\mathrm{x} \sim f_{\mathrm{x}}$ *denotes that* x *is continuous with density function* f_{x}.

In the discrete case, the random variable takes a *countable* number of outcomes a_k, and we have that $\Pr(\mathrm{x} \in A) = \sum_{a_k \in A} p_{\mathrm{x}}(a_k)$, which entails $\Pr(\mathrm{x} = x) = p_{\mathrm{x}}(x)$. In the continuous case, the random variable takes an *uncountable* number of outcomes, and we have that $\Pr(a < \mathrm{x} < b) = \Pr(a < \mathrm{x} \leq b) = \Pr(a \leq \mathrm{x} < b) = \Pr(a \leq \mathrm{x} \leq b) = F_{\mathrm{x}}(b) - F_{\mathrm{x}}(a)$, which entails $\Pr(\mathrm{x} = x) = 0$.

In the same way, we say that a random vector is discrete (continuous) if its joint cdf can be expressed via a joint pmf (pdf). Marginal pmfs (pdfs) can be obtained from the joint pmf (pdf) by summing (integrating) over the unwanted dimensions. They completely specify the marginal cdfs of the marginal distributions. A random vector is **mixed** if it can be partitioned into a set of discrete random variables and a set of continuous random variables.

mixed

We also make *conditional* probability statements for a random variable x given that another random variable y takes the value y. The **conditional distribution** $\Pr_{\mathrm{x}|\mathrm{y}=y}$ is completely specified by the conditional cdf $F_{\mathrm{x}|\mathrm{y}=y}$, and we write

conditional distribution

$$\mathrm{x} \mid \mathrm{y} = y \sim F_{\mathrm{x}|\mathrm{y}=y} \tag{1.6}$$

to define the conditional distribution of x given $\mathrm{y} = y$ via the conditional cdf. If x and y are discrete random variables with joint pmf p_{xy} and marginal pmf p_{y}, the conditional cdf can be expressed by the conditional pmf

$$p_{\mathrm{x}|\mathrm{y}=y}(x) := \frac{p_{\mathrm{xy}}(x, y)}{p_{\mathrm{y}}(y)} \tag{1.7}$$

for $p_{\mathrm{y}}(y) > 0$. This entails

$$p_{\mathrm{x}}(x) = \sum_{y} p_{\mathrm{xy}}(x, y) = \sum_{y} p_{\mathrm{x}|\mathrm{y}=y}(x) \cdot p_{\mathrm{y}}(y). \tag{1.8}$$

The case with two continuous random variables is analogue, with pdfs and integrals instead of pmfs and sums, and the generalizations for discrete and continuous random vectors are straightfor-

ward. We can express the cdf of mixed random variables, where x is continuous and y is discrete with outcomes $a_1, a_2, \ldots$, via the conditional cdf of x given $\mathrm{y} = a_k$ and the marginal pmf of y:

$$F_{\mathrm{xy}}(x, y) = \sum_{a_k \leq y} F_{\mathrm{x}|\mathrm{y}=a_k}(x) \cdot p_{\mathrm{y}}(a_k) \tag{1.9}$$

In this case, the marginal distribution of x,

$$F_{\mathrm{x}}(x) = \sum_{a_k} F_{\mathrm{x}|\mathrm{y}=a_k}(x) \cdot p_{\mathrm{y}}(a_k). \tag{1.10}$$

is called a **mixture distribution**. mixture distribution

In the following, we use the common simplified notation with logical statements $\Pr(\mathrm{x} \in A \mid \mathrm{y} = y) := \Pr_{\mathrm{x}|\mathrm{y}=y}(A)$ to refer to probabilities under the conditional distribution of x given $\mathrm{y} = y$. Conditional distributions $\Pr_{\mathrm{x}|\mathrm{y}\in B}$ for the random variable x given that y takes any value from the set B are defined analogously.

An important property of a random variable is its **expected value**. If x is discrete with outcomes $a_1, a_2, \ldots$, its expected value is expected value

$$\mathrm{E}[\mathrm{x}] := \sum_k a_k p_{\mathrm{x}}(a_k). \tag{1.11}$$

If x is continuous, its expected value is defined as

$$\mathrm{E}[\mathrm{x}] := \int x f_{\mathrm{x}}(x)\, dx. \tag{1.12}$$

If the expected value is infinity, it is said to not exist. Sometimes, we add a subscript $\mathrm{E}_{\mathrm{x}}[\mathrm{x}]$ to distinguish the expected value from the conditional expectation defined below. The expected value of a random vector is the vector of expected values of its component random variables; conversely, we refer to the components of the expected value of a random vector as the **marginal expectations**. marginal expectations
A key result is that the expected value of the continuous random variable y with $\mathrm{y} = g(\mathrm{x})$ can be expressed as

$$\mathrm{E}_{\mathrm{y}}[\mathrm{y}] = \int y f_{\mathrm{y}}(y)\, dy = \int g(x) f_{\mathrm{x}}(x)\, dx = \mathrm{E}_{\mathrm{x}}[g(\mathrm{x})], \tag{1.13}$$

and analogously for the discrete case. The **variance** of the random variable x is the expected value $\mathrm{Var}[\mathrm{x}] := \mathrm{E}[(\mathrm{x} - \mathrm{E}[\mathrm{x}])^2]$. The **covariance** of two random variables x and y is the expected value $\mathrm{Cov}[\mathrm{x}, \mathrm{y}] := \mathrm{E}_{\mathrm{xy}}[(\mathrm{x} - \mathrm{E}_{\mathrm{x}}[\mathrm{x}])(\mathrm{y} - \mathrm{E}_{\mathrm{y}}[\mathrm{y}])]$ under the joint distribution $\Pr_{\mathrm{xy}}$. The **conditional expectation** $\mathrm{E}_{\mathrm{x}|\mathrm{y}=y}[\mathrm{x}]$ is the expected value of x under the conditional distribution $\Pr_{\mathrm{x}|\mathrm{y}=y}$. It is computed with the conditional pmf or conditional pdf, respectively.

variance

covariance

conditional expectation

Independence

A key concept for the present work is *independence*. The most fundamental definition of independence in the probability space $(\Omega, \mathcal{A}, \Pr)$ is a statement on the probability of the *intersection* of two sets of outcomes from the σ-field:

independent

Definition 1.2.3 *Let $(\Omega, \mathcal{A}, \Pr)$ be a probability space. The sets $A \in \mathcal{A}$ and $B \in \mathcal{A}$ are **independent** if $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$.*

This definition directly yields a notion of independence for random variables defined on the same probability space. The formal notion of independence for random variables is rather technical, but can be characterized more conveniently with the help of cdfs:

Theorem 1.2.1 *Let $\{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$ be a set of random variables. Furthermore, let F be the joint cdf of the random vector $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_N)$, and F_n be the marginal cdf of x_n for $n = 1, \ldots, N$.*

The random variables are independent if and only if

$$F(x_1, \ldots, x_N) = \prod_n F_n(x_n).$$

For independent random variables x and y, we have that

$$\Pr(\mathrm{x} \in A, \mathrm{y} \in B) = \Pr(\mathrm{x} \in A) \cdot \Pr(\mathrm{y} \in B) \tag{1.14}$$

and

$$\Pr(\mathrm{x} \in A \mid \mathrm{y} \in B) = \Pr(\mathrm{x} \in A) \tag{1.15}$$

for all sets of outcomes A and B. This entails that, if x and y are independent, the conditional cdf of x given $\mathrm{y} \in B$ is equal to its marginal cdf, $F_{\mathrm{x}|\mathrm{y}\in B} = F_{\mathrm{x}}$. If the random variables $\mathrm{x}_1, \ldots, \mathrm{x}_N$ are independent and have the same marginal cdf F, we say that they are **iid** (independent and identically distributed) and write $\mathrm{x}_n \overset{\text{iid}}{\sim} F$.

iid

In the case of discrete random variables, independence holds if and only if the joint pmf factorizes over the marginal pmfs

$$p_{\mathrm{x}_1,\ldots,\mathrm{x}_N}(x_1, \ldots, x_N) = \prod_n p_{\mathrm{x}_n}(x_n). \tag{1.16}$$

As a consequence, two discrete random variables x and y are independent if and only if the conditional pmf $p_{\mathrm{x}|\mathrm{y}=y}(x) = p_{\mathrm{x}}(x)$ equals the marginal pmf. Again, the continuous case is analogue, with pdfs instead of pmfs.

The concept of independence can be generalized to multiple sets of random variables, in the following way:

Theorem 1.2.2 *Let* $\{\mathrm{x}_1, ..., \mathrm{x}_N\}$ *and* $\{\mathrm{y}_1, ..., \mathrm{y}_M\}$ *be two sets of random variables, and let* $\mathbf{x} = (\mathrm{x}_1, ..., \mathrm{x}_N)$ *and* $\mathbf{y} = (\mathrm{y}_1, ..., \mathrm{y}_M)$ *be the corresponding random vectors with joint cdfs* $F_{\mathbf{x}}$ *and* $F_{\mathbf{y}}$*. Furthermore, let* $F_{\mathbf{xy}}$ *be the joint cdf of the random vector* $(\mathrm{x}_1, ..., \mathrm{x}_N, \mathrm{y}_1, ..., \mathrm{y}_M)$*.*

The two sets of random variables are independent if and only if

$$F_{\mathbf{xy}}(x_1, ..., x_N, y_1, ..., y_M) = F_{\mathbf{x}}(x_1, ..., x_N) \cdot F_{\mathbf{y}}(y_1, ..., y_M).$$

We write $\mathrm{x}_1, ..., \mathrm{x}_N \perp\!\!\!\perp \mathrm{y}_1, ..., \mathrm{y}_M$ *to express this independence.*

The notion of event impacts defined in Section 1.3 below is, in fact, a statement on the independence of two sets of random variables. At last, the random variables x and y can also be **conditionally independent** given the random variable z. Conditional independence is characterized by a factorization of the conditional joint cdf into the conditional marginal cdfs,

conditionally independent

$$F_{\mathrm{xy}|\mathrm{z}=z}(x, y) = F_{\mathrm{x}|\mathrm{z}=z}(x) \cdot F_{\mathrm{y}|\mathrm{z}=z}(y), \tag{1.17}$$

for all values of z, and expressed notationally by $\mathrm{x} \perp\!\!\!\perp \mathrm{y} \mid \mathrm{z}$. If x and y are conditionally independent given z, we have that

$$\begin{aligned} &\Pr(\mathrm{x} \in A, \mathrm{y} \in B \mid \mathrm{z} \in C) \\ &\qquad = \Pr(\mathrm{x} \in A \mid \mathrm{z} \in C) \cdot \Pr(\mathrm{y} \in B \mid \mathrm{z} \in C) \end{aligned} \tag{1.18}$$

and

$$\Pr(\mathrm{x} \in A \mid \mathrm{y} \in B, \mathrm{z} \in C) = \Pr(\mathrm{x} \in A \mid \mathrm{z} \in C). \tag{1.19}$$

for all sets of outcomes A, B and C. Same as the unconditional notion of independence, conditional independence can be characterized by a factorization of the conditional pmfs in the case of discrete random variables, and by a factorization of the conditional pdfs in the case of continuous random variables. Conditional independence for random vectors (or sets of random variables) is characterized and expressed in the straightforward way.

1.3 Event impacts

We are now in the position to formalize our framework for event impact analysis. The two primal concepts are time series and event series, which are special **stochastic processes**:

stochastic processes

Definition 1.3.1 *A* ***time series*** $\mathbf{X} = (\mathrm{x}_t)_{t \in \mathbb{Z}}$ *is a sequence of continuous random variables over a common probability space* $(\Omega, \mathcal{A}, \Pr)$*.*

time series

We interpret the index t as the time of observation. All time series encountered in practice are finite, and we restrict our attention to indices $t \in \{1, ..., T\}$ in the main part of this work, where T

length

is the **length** of the time series. Strictly speaking, a finite time series is a T-dimensional random vector. In some chapters, we encounter more general time series that are sequences of random vectors or random elements. Therefore, we retain separate terms and notations for these concepts. Sometimes, we refer to a time series in the sense of Definition 1.3.1 as a **univariate** time series to distinguish it from a **multivariate** time series of random vectors.

univariate

multivariate

We develop event impact analysis to study the effect of *randomly recurring* events on the behavior of a time series. We assume that these events also follow a stochastic process:

event series

Definition 1.3.2 *An **event series** $\mathbf{E} = (\mathrm{e}_t)_{t\in\mathbb{Z}}$ is a sequence of discrete random variables over a common probability space $(\Omega, \mathcal{A}, \Pr)$, with only two possible outcomes $\mathrm{e}_t(\Omega) = \{0, 1\}$ for all t.*

Semantically, the outcome 0 means that no event has occurred at time t, while the outcome 1 indicates an event occurrence at time t. As before, we restrict our attention to the finite index set $t \in \{1, ..., T\}$. Our notion of an event series is a discrete-time analogue of point processes [DV03], and a special case of a discrete-valued time series [DHL+16; JR19; Wei18]. Viewing the event series as a stochastic process is a key aspect that distinguishes our work from previous approaches to study event impacts.

Stationarity and sparsity

An important property of stochastic processes that we need for our exposition is *stationarity*. Stationarity means that the statistical properties of the process do not change over time [BJR+16]:

stationary

Definition 1.3.3 *A stochastic process $\mathbf{Z} = (\mathrm{z}_t)_{t\in\mathbb{Z}}$ with random variables over a common probability space $(\Omega, \mathcal{A}, \Pr)$ is **stationary** if, for all finite index sets $\mathcal{T} = \{t_1, ..., t_N\} \subset \mathbb{Z}$ with $N = 1, 2, ...$, we have*

$$F_{\mathrm{z}_{t_1},...,\mathrm{z}_{t_N}} = F_{\mathrm{z}_{t_1+\delta},...,\mathrm{z}_{t_N+\delta}}$$

for every $\delta \in \mathbb{Z}$, i.e., all finite cdfs are shift-invariant.

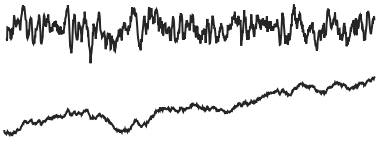

Figure 1.2: A stationary (top) and a nonstationary time series (bottom).

An immediate consequence of stationarity is that the marginal distribution of z_t is identical for every t. If a stochastic process is stationary and all finite selections of random variables are independent, we call it an iid process. Furthermore, we say that two processes $\mathbf{Z}$ and $\mathbf{Z}'$ are **jointly stationary** if all finite *joint* cdfs $F_{\mathrm{z}_{t_1},...,\mathrm{z}_{t_N},\mathrm{z}'_{t_1},...,\mathrm{z}'_{t_N}}$ are shift-invariant.

jointly stationary

Throughout this work, we assume that the event series $\mathbf{E}$ is stationary. Furthermore, we assume that the event series is **sparse**, *i.e.*, $\Pr(\mathrm{e}_t = 1) := \epsilon$ for a very small $\epsilon > 0$. Sparsity formally reflects

sparse

our focus on rare and potentially extreme events. In a sparse event series, on average, we have a lag of $L = \frac{1-\epsilon}{\epsilon} \gg 0$ time steps between two event occurrences. If we restrict our attention to finite event series of length T, sparsity can be expressed by assuming that the number of event occurrences is much smaller than the length of the time series, *i.e.*, $\sum_{t=1}^{T} \mathrm{e}_t := N_{\mathbf{E}} \ll T$ with $\frac{T}{N_{\mathbf{E}}} \approx L$.

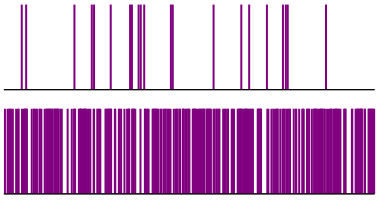

Figure 1.3: Sparse (top) and dense event series (bottom).

Event impacts via statistical association

The assumption that the time series and the event series are *both* stochastic processes comes with two major benefits. First and foremost, we can now study event impacts by focusing on *statistical associations* between the two processes; in particular, we can develop statistical independence tests for this purpose. Second, we can use *probability theory* to characterize the nature of the event impacts, *e.g.*, by estimating probability distributions or complex probabilistic models for the impact. Formally, the event series $\mathbf{E}$ has no association with the time series $\mathbf{X}$ if the two are independent processes in the following sense:

independent processes

Definition 1.3.4 *The time series* $\mathbf{X}$ *and the event series* $\mathbf{E}$ *are* ***independent processes*** *if for all finite index sets* $\mathcal{T} = \{t_1, \ldots, t_N\} \subset \mathbb{Z}$ *with* $N = 1, 2, \ldots$ *we have that* $\mathrm{x}_{t_1}, \ldots, \mathrm{x}_{t_N} \perp\!\!\!\perp \mathrm{e}_{t_1}, \ldots, \mathrm{e}_{t_N}$. *We write* $\mathbf{X} \perp\!\!\!\perp \mathbf{E}$ *to denote that* $\mathbf{X}$ *and* $\mathbf{E}$ *are independent processes.*

If $\mathbf{X}$ and $\mathbf{E}$ are independent processes, the behavior of one of the two processes does not yield any information about the behavior of the other process. However, if we find that the two processes are *not* independent, we do not know how they are associated: there is an infinite number of independence relations that could be violated, and countless ways to describe the corresponding association. Therefore, most existing independence tests for stochastic processes do not test for independent processes, but for a subset of independence relations that has a specific semantic interpretation and can be described in a meaningful way—see the related work in Section 1.4.2 (Statistical associations).

The notion of event impacts that we develop here focuses on a novel set of independence relations between an event series and a time series: Informally, event impact analysis studies the statistical association of the value of the event series $\mathbf{E}$ *at* time t, and the values of the time series $\mathbf{X}$ within a *window around* time t. A focus on this association is compatible with previous works on impact analysis for singular events—see the related work in Section 1.4.1 (Singular events). Formally, we distinguish two cases that we treat separately, depending on whether $\mathbf{X}$ is stationary or not.

1.3.1 Stationary case

In the first case, we assume that the time series $\mathbf{X}$ and the event series $\mathbf{E}$ are jointly stationary, which entails that both processes are also marginally stationary. The methods developed in Chapter 2 to Chapter 5 are designed for the stationary case, and only Chapter 6 discusses a specific nonstationary case.

temporary

In the stationary case, event impacts can only be **temporary**. There is a specific notion of *typical behavior* of the time series—a term that will be specified more precisely soon. An event occurrence may lead to (or coincide with) a temporary *deviation* from this typical behavior, and if enough time passes after the event occurrence, the time series will eventually return to its *typical behavior*. However, we cannot expect that the individual occurrences of a *recurring* event will always lead to *exactly* the same behavior in the time series. Randomness is central in our definition of event impacts.

Random impacts of random events

Similar challenges were discussed earlier by Chi et al. [CSH+16].

There are three main challenges directly associated with the randomness that we allow. First, there is an unknown and possibly non-deterministic *temporal lag* between each event occurrence and an observable event impact in the time series. Second, the *duration* of the event impact, *i.e.*, the number of time steps until the time series returns to its typical behavior, is unknown and possibly non-deterministic. Third, the *observable impact* itself is unknown and possibly non-deterministic. We circumvent the first two challenges by limiting our attention to event impacts within a certain **window of interest**, parametrized by a maximum lag Δ that must be chosen prior to the analysis. The parameter Δ must be selected such that an event occurrence at time t is expected to affect the time series somewhere between time steps t and $t + \Delta$. To address the third challenge, we define event impacts in the most generic probabilistic way possible, via the absence of a specific independence relation:

window of interest

The window of interest may also be defined symmetrically around t or with an explicit time lag.

Definition 1.3.5 *Let* $\mathbf{X}$ *be a time series and* $\mathbf{E}$ *be an event series. We say that* $\mathbf{X}$ *has* ***event impacts*** *if* $x_t, \ldots, x_{t+\Delta} \not\perp\!\!\!\perp e_t$ *for any* $t \in \mathbb{Z}$.

event impacts

If we assume that $\mathbf{X}$ and $\mathbf{E}$ are jointly stationary, the independence relations that do or do not hold at any point in time $t \in \mathbb{Z}$ also hold at any other point in time. In contrast, in the nonstationary setting, the independence relations potentially change over time. The key advantage of this formalization of event impacts is that it does not impose any restrictions as to where in the window of interest the impact materializes, how many time steps it encompasses, or how it looks like. Moreover, it captures scenarios where not every single event occurrence has an observable impact on the time series.

Formal characterizations

Event impacts in the sense of Definition 1.3.5 can be characterized very generically with the help of cdfs, either by lack of the factorization of the joint cdf into two marginal cdfs,

$$\begin{aligned} &F_{\mathrm{e}_t,\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}}(e_t, x_t, \ldots, x_{t+\Delta}) \\ &\qquad \neq F_{\mathrm{e}_t}(e_t) \cdot F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}}(x_t, \ldots, x_{t+\Delta}), \end{aligned} \tag{1.20}$$

or by diverging conditional and marginal cdfs:

$$F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=1} \neq F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}} \tag{1.21}$$

$$F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=0} \neq F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}} \tag{1.22}$$

$$F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=1} \neq F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=0} \tag{1.23}$$

The latter characterization allows us to clarify the notion of *typical behavior* of the time series mentioned earlier, and how events lead to a *deviation* from this typical behavior. The marginal cdf $F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}}$ can be expressed as a mixture of the conditional cdfs,

$$F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}} = \epsilon \cdot F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=1} + (1-\epsilon) \cdot F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=0} \tag{1.24}$$

where $\epsilon = \Pr(\mathrm{e}_t = 1)$ is the probability to observe an event at time t. Since we assume a sparse event series with a very small $\epsilon > 0$, this marginal cdf is dominated by the conditional cdf given $\mathrm{e}_t = 0$,

$$F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}} \approx F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=0}. \tag{1.25}$$

The sparser the event series, the more similar the marginal cdf to the conditional cdf given $\mathrm{e}_t = 0$, with the extreme case $F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}} = F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=0}$ in the limit $\epsilon \longrightarrow 0$. For this reason, we say that, when the event series is sparse, the marginal cdf $F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}}$ represents the **typical behavior** in a window of interest starting at time t, while the conditional cdf $F_{\mathrm{x}_t,\ldots,\mathrm{x}_{t+\Delta}|\mathrm{e}_t=1}$ represents (potentially) **deviant behavior** in that window induced by an event at time t.

typical behavior
deviant behavior

These characterizations hold regardless of whether the time series and the event series are jointly stationary or not. However, in the stationary case, the typical behavior and the deviant behavior do not depend on the time step t—they are shift-invariant. Therefore, we can say that, in the stationary case, the two cdfs represent the typical behavior and the deviant behavior *of the time series*.

Figure 1.4 illustrates random event impacts in a univariate time series. The general shape of the impacts is the same as in the deterministic example from Figure 1.1, but with a random delay before the onset of the impact, random magnitude, and random decay rate. We can gain insights into the nature of these event impacts by studying the statistical properties of the deviant behavior

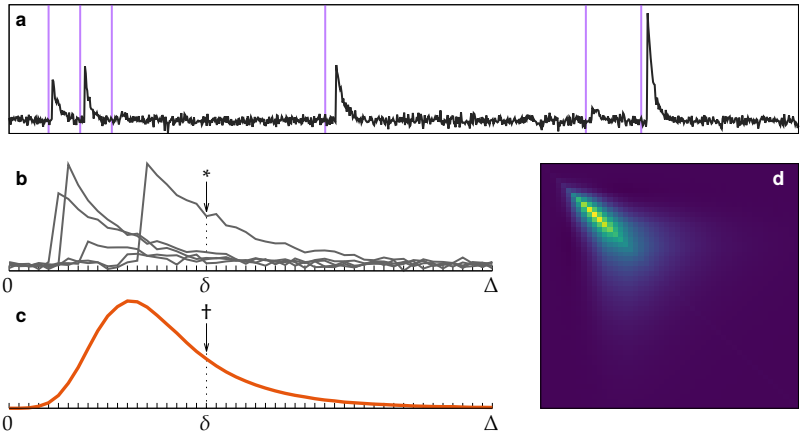

Figure 1.4: Random event impacts. **a**: Excerpts from the time series and the event series. **b**: Samples from the deviant behavior $F_{x_t,\dots,x_{t+\Delta}|e_t=1}$; (∗) samples from the marginal cdf $F_{x_{t+\delta}|e_t=1}$ at lag δ. **c**: Expected value of the deviant behavior; (†) marginal expectation at lag δ. **d**: Covariance matrix with entries $\mathrm{Cov}[x_{t+\delta}, x_{t+\delta'}]$ for all δ, δ' in the window of interest, computed from $F_{x_{t+\delta},x_{t+\delta'}|e_t=1}$.

$F_{x_t,\dots,x_{t+\Delta}|e_t=1}$, *e.g.*, its expected value or its covariance structure. For comparison, the expected value of the *typical behavior* is always a straight horizontal line at $\mathrm{E}[x_{t+\delta}] = \mu$ for all δ, while the covariance matrix is always a Toeplitz matrix, *i.e.*, a matrix with constant diagonals [BJR+16]. This allows simple visual assessment of event impacts in the first two moments of the distribution. However, this visual approach cannot easily be extended to event impacts in higher-order moments of the distribution, or to scenarios with multivariate or more general time series.

Statistical independence tests for event impacts

The characterizations of event impacts via diverging conditional and marginal cdfs from Equation 1.21 to Equation 1.23 provide a means to implement statistical tests for event impacts based on observed data: we have to find statistical evidence in the data that the respective cdfs are, in fact, diverging. Due to Equation 1.25, the most effective approach is to seek evidence that the *deviant behavior* diverges from the *typical behavior*—similar to the visual approach from above. However, in a *finite* sparse event series of length T, there are only $N_{\mathbf{E}} \ll T$ event occurrences, so that we have access to only few examples of the deviant behavior. This renders statistical inference on the multivariate function $F_{x_t,\dots,x_{t+\Delta}|e_t=1}$ less reliable, especially for larger values of Δ. For example, for reliable estimation of the covariance matrix in Figure 1.4, we need $N_{\mathbf{E}} \geq \Delta^2$ event occurrences. Therefore, instead of working directly with the multivariate cdf, we simplify the statistical procedures involved in our tests by exploiting the following result:

Lemma 1.3.1 *Let* $g(x_t, \dots, x_{t+\Delta})$ *be a measurable function.*

If $x_t, \dots, x_{t+\Delta} \perp\!\!\!\perp e_t$, *then* $g(x_t, \dots, x_{t+\Delta}) \perp\!\!\!\perp e_t$.

An immediate consequence is that if we find a measurable function g such that $g(x_t, \dots, x_{t+\Delta}) \not\perp\!\!\!\perp e_t$, we know that $x_t, \dots, x_{t+\Delta} \not\perp\!\!\!\perp e_t$,

i.e., we know that $\mathbf{X}$ has event impacts. We can translate the independence relation into an equivalent statement on the cdfs, and say that $\mathbf{X}$ has event impacts if

$$F_{g(x_t,\ldots,x_{t+\Delta})|e_t=1} \neq F_{g(x_t,\ldots,x_{t+\Delta})}. \tag{1.26}$$

The advantage of this perspective over the perspective from Equation 1.21 is that—at least for scalar-valued functions g—the involved cdfs are univariate and thus more easily accessible for statistical inference. Moreover, the function g can be viewed as a statistic that captures properties of the time series that we expect to be associated with event occurrences. Importantly, the converse of Lemma 1.3.1 does not hold: There exist functions g such that $g(x_t, \ldots, x_{t+\Delta}) \perp\!\!\!\perp e_t$, but not $x_t, \ldots, x_{t+\Delta} \perp\!\!\!\perp e_t$. If we cannot find event impacts with a specific choice of function g, we cannot conclude that there are no event impacts in the general sense of Definition 1.3.5.

The key question in the stationary case is how to define the function g such that we obtain a powerful statistical test procedure for the kind of event impacts that we expect. The procedures that we propose in Chapter 2 and Chapter 3 are based on different choices for the function g in Lemma 1.3.1. In Chapter 4, we devise a generic probabilistic model for the deviant behavior $F_{x_t,\ldots,x_{t+\Delta}|e_t=1}$.

1.3.2 Nonstationary case

In the nonstationary case, we do not assume that the time series $\mathbf{X}$ and the event series $\mathbf{E}$ are jointly stationary. If we have access to only a single paired realization from the event series and the time series, we are rather limited in the statistical methodology that we can apply in this case, unless other structural assumptions are made. In this work, we discuss methods for event impact analysis for two types of nonstationarity that may be present *within the time series*. For both types, we can eliminate the nonstationarity by preprocessing, and reduce the problem to the stationary case.

In the first type that we treat below, we assume that the time series follows an *integrated* process. This type of nonstationarity results in stochastic trends in the time series, and can be eliminated by the *differencing* operation. In the second case that we treat in Chapter 6, we assume that the time series is *segmented*, *i.e.*, its data-generating process changes its dynamics at specific points in time, and is stationary within each segment. The two types can be combined, so that we can first eliminate stochastic trends by differencing, and capture nonstationarity in other statistical properties by partitioning the time series into stationary segments.

In contrast to the stationary case, in a nonstationary time series, event impacts can be temporary *or* permanent. However, it is not trivial to formalize the concepts of temporary and permanent event impacts in a nonstationary time series concisely and comprehensively, *i.e.*, in a way that reflects all reasonable meanings of the terms. We defer this typological discussion to future work.

Integrated time series

The notion of an integrated time series that we use here is motivated from the autoregressive integrated moving average (ARIMA) model in time series analysis. It is meant to capture what Box et al. [BJR+16] call a form of *homogeneous* nonstationarity. Let $\mathbf{X}$ be a time series. Formally, we use the classical **difference operator** ∇ to express differences between consecutive observations within the time series, *i.e.*, $\nabla x_t := x_t - x_{t-1}$. We write $\nabla^q x_t$ to apply the difference operator q times, so that the random variable

difference operator

$$z_t := \nabla^q x_t = z(x_{t-q}, \ldots, x_t) \tag{1.27}$$

is a function of the past q values of the time series up to time step t. We say that $\mathbf{X}$ is an **integrated time series** of order q if q is the smallest value such that the time series $\mathbf{Z} = (z_t)_{t \in \mathbb{Z}}$ is stationary. In the literature on time series analysis, integrated time series are also called *difference-stationary processes* or *unit root processes* [Ham94].

integrated time series

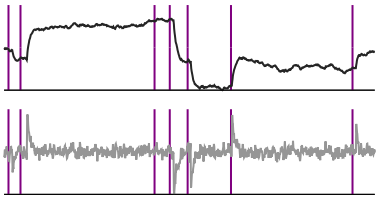

Figure 1.5: Integrated time series (top) can be studied with our methods when differencing is applied as a preprocessing step (bottom).

Figure 1.5 shows an example of an integrated time series with event impacts that appear to be permanent in an informal sense. We observe that, after one differencing operation, the resulting time series is stationary with temporary event impacts.

If there are reasons to believe that $\mathbf{X}$ is an integrated time series of unknown order q, we have to perform iterative differencing on the observed data as a preprocessing step, until the resulting time series is stationary. We can then apply the methods for event impact analysis in the stationary case developed in the main part of this work. Strictly speaking, we must assume that the time series after differencing $\mathbf{Z}$ and the event series $\mathbf{E}$ are *jointly* stationary for our methods to be applicable.

There is an additional subtlety that needs to be considered: Event impacts in the sense of Definition 1.3.5 are defined via the lack of independence $e_t \not\perp\!\!\!\perp x_t, \ldots, x_{t+\Delta}$ for a specific window of interest parametrized by Δ. If we apply the methods for event impact analysis in the stationary case *unchanged* on the time series $\mathbf{Z}$, strictly speaking, we do not test for event impacts in $\mathbf{X}$, but for event impacts in $\mathbf{Z}$. More precisely, if we find that $e_t \not\perp\!\!\!\perp z_t, \ldots, z_{t+\Delta}$,

we cannot conclude that $e_t \not\perp\!\!\!\perp x_t, \ldots, x_{t+\Delta}$. The reason is that due to the differencing operation,

$$(z_t, \ldots, z_{t+\Delta}) = g(x_{t-q}, \ldots, x_{t+\Delta}) \tag{1.28}$$

contains information from the time series that lies *outside* of the original window of interest—information that temporally *precedes* the event. Therefore, the window of interest within $\mathbf{Z}$ must be shrunk to the random variables $z_{t+q}, \ldots, z_{t+\Delta}$ that only contain information from the original window of interest in the time series $\mathbf{X}$. Moreover, the nature of the event impacts can only be understood by studying the statistical properties of the deviant behavior of the time series $\mathbf{Z}$—which is shift-invariant—, in contrast to the deviant behavior of the time series $\mathbf{X}$—which changes over time. Apart from that, no changes are necessary. This simple observation makes the methods developed in the main part of this work applicable to a large class of problems with nonstationary time series.

Segmented time series

The other type of nonstationarity that we consider in this work is the case of a segmented time series with stationary segments. An introduction to event impact analysis in this case is provided in Chapter 6. The idea is to first detect the *change points*, *i.e.*, the points in time at which the data-generating process changes its dynamics, and perform event impact analysis separately for each segment. Alternatively, we can correlate the occurrence of change points itself with the occurrence of events in the event series.

1.4 Related work

Our research is motivated by a continuous stream of application-oriented papers that discuss the impacts of *recurring* events on time series. For example, Zhang, Lai, and Wang [ZLW08] and Zhang et al. [ZYW+09] estimate the impacts of extreme events on crude oil prices, while Chesney, Reshetar, and Karaman [CRK11] analyze the behavior of stock markets after terrorist attacks. Luo et al. [LLL+14] seek methods to correlate performance metrics of a computing system with critical system incidents. Chi, Han, and Wang [CHW16] and Chi et al. [CSH+16] identify more application domains that may benefit from such methodology, including traffic control, customer behavior analytics and finance. Kalyanam et al. [KQP+16] study the activity bursts in social media time series after news events. More recently, Madaan et al. [MSK+19] studied the impact of weather events and hoarding events on prices of agricultural commodities in India, while Dortmont, Elzen, and

Wijk [DEW19] discussed event impacts from the perspective of visual analytics, with applications in electronic health and IT security. The set of tools employed in these works is highly diverse, and to date, there seems to be no common understanding of how to capture the impacts of recurring events in a meaningful way.

In this thesis, we make the case for a *statistical perspective* on the impacts of recurring events. Some of the works mentioned above contain initial ideas to study such impacts with statistical methods, in particular, with two-sample tests. Luo et al. [LLL+14] propose to compare all subsequences within a fixed window of interest before and after event occurrences with a set of subsequences sampled randomly from the time series, using two-sample tests. Similarly, Chi, Han, and Wang [CHW16] and Chi et al. [CSH+16] directly compare the subsequences after event occurrences with the subsequences before event occurrences, with a two-sample test. A shortcoming of these works is that—although they make use of statistical methodology—they do not discuss the statistical assumptions that allow the use of such methodology. Many of the ideas from these works are compatible with our notion of event impacts from Definition 1.3.5, either directly or with some minor adaptations. In fact, these works have inspired the multiple two-sample testing procedure that we develop in Chapter 3. In the following sections, we give an overview of methodologically related problems and approaches.

1.4.1 Singular events

There is a large body of research that studies the impact of a *singular event* on the behavior of a time series; we summarize the most popular approaches below. The occurrence of the event is often controlled by an experimenter, and the event is referred to as the "intervention." The ultimate goal of many of these studies is to assess whether the intervention *caused* a change in the behavior of the time series, but, strictly speaking, they can only assess whether it *coincides* with a change in the time series. These methods have been developed and matured concurrently in many disciplines, from econometrics to ecology and medicine, often using similar terms for different ideas or different terms for similar ideas.

As pointed out earlier, this limitation applies in the same way to the methods that we develop in this work.

In all of these approaches, the time series can be split into a *pre-intervention phase* and a *post-intervention phase*. If the time series follows an iid process in both phases, a standard two-sample test would be sufficient to assess whether there is a change from one phase to the other [BT65], potentially caused by the event. Research in this field focuses primarily on the development of methods for various types of non-iid time series, and on the discussion of techniques that allow for more compelling causal inferences.

Intervention analysis

A well-established methodology to study singular event impacts is *intervention analysis* [BT65; BT75; Gla72], sometimes referred to as *interrupted time series analysis* [MMM+80]. Classical intervention analysis seeks to estimate the deterministic *linear impact* of a single event on the behavior of a time series that follows an *autoregressive model* (ARIMA), as illustrated in Figure 1.1. It is a special case of a *linear transfer function model*, where the input series is either a pulse function or a step function at the time point of the event. The impact of this event has one of several parametric forms defined by a linear transfer function [BJR+16]. The representation of the event as a pulse or step function and the choice of the transfer function are application-dependent and usually hypothesis-driven. Originally formulated only for univariate time series, it has been extended to multivariate time series [Abr80; ES93] and to models other than ARIMA [FF10; FF12; FLE+14; LKF+14; SL21; WO20]. Intervention analysis can also be applied in settings with multiple events, either by assuming that every event has *the same* deterministic impact, or *its own* deterministic impact, unrelated with the other impacts. A potential downside of intervention analysis is that it assumes parametric models both for the event impacts and for the other components of the time series.

Regression-based approaches

Similar in spirit, but quite different methodologically, the *segmented regression approach of interrupted time series* [CS63; GMS81] captures changes in the *linear trend* and *level* of a time series that coincide with an intervention. The key idea is to estimate a segmented linear regression model for the pre- and post-intervention phase, using time as the covariate. The approach can be extended to capture time series with seasonal patterns and known outliers [WSZ+02], and to generalized linear models with additional covariates that account for any other potential confounding factors [KDS+15; LCL+21; LCG17]. Guidelines for correct model specification are given by Huitema and McKean [HM00] and, more recently, Lopez Bernal, Soumerai, and Gasparrini [LSG18].

The segmented regression approach can also use *control time series* that are unaffected by the intervention. This allows more reliable causal inferences [BSI19; Lin15; LA11; LCG18; Sim77]. Estimation of regression models with control time series and additional covariates is also known as the *difference-in-differences method* [Aba05; AC85; LCL+21]. However, the classical difference-in-difference model captures changes in the *level* of the time series only.

in contrast to *observed* confounders that can be included as covariates

The *synthetic control method* [ADH11; ADH+10; AG03] is a *factor model* that extends the difference-in-differences approach to settings with *unobserved* confounders that potentially vary with time. The key idea is to learn a weighted combination of all available control time series that best approximates the time series of interest within the *pre*-intervention phase. The weighted combination of the control time series in the *post*-intervention phase then yields a counterfactual "synthetic control" for the time series of interest, had the event not happend. Any departures of the time series of interest from the synthetic control within the post-intervention phase are attributed to the intervention. The Bayesian counterfactual approach of Brodersen et al. [BGK+15] is a variation of this scheme based on a *structural state-space model* for time series.

When using any regression-based approach to assess event impacts in time series, it is important to take serial correlations in the error terms into account [BDM04]. Same as intervention analysis, the regression-based approaches to event impact analysis are generally limited by their parametric model assumptions.

BACIP analysis

An alternative methodology that does not require parametric models is known as *BACIP* (Before-After Control-Impact Paired) [Ber83; Smi14; SBO92; SMP86] or *randomized intervention analysis* [CFH+89]. It can be applied to study the impact of a single intervention event on a time series, when a *control time series* is available. The control time series is assumed to follow the same dynamics as the time series of interest, without being affected by the intervention. Event impacts are then assessed by comparing the *mean difference* between the two time series *before* the event with the mean difference *after* the event. If the difference of the mean differences is statistically significant, the event has impact on the time series.

This approach has been extended for multiple control time series [Und92; Und94], for impacts other than level shifts [TKO+17; Und92; Und94; WAG+21], for multivariate time series [TBB+05], and has been augmented with Bayesian methods [CSB+16; CTM96]. Murtaugh [Mur00; Mur02] raised a debate on whether BACIP overstates the statistical evidence for event impacts in the presence of serial correlations [Mur03; Ste03]. Recently, Christie et al. [CAM+19] performed a simulation study that showed that BACIP is better in detecting event impacts than simpler variations thereof. Additional measures to characterize impacts within the BACIP framework have been proposed [CRK19]. However, the main limiting factor of BACIP that it shares with difference-in-differences and synthetic control methods is the availability of control time series.

In the terminology of this branch of research, our approach to capture impacts of *recurring* events can be viewed as an *aggregated* Before-After (BA) analysis *without* control time series: we aggregate event impacts over all individual event occurrences to make statistical inferences on relevant probability distributions. We stress that control time series can—principally—be included in our approach to event impact analysis, by applying our methods on the *difference time series* between the time series of interest and the control time series. The statistical requirement is that the difference time series and the event series are jointly stationary.

This approach should not be confused with the differencing operation described in Section 1.3.2 to handle nonstationarity in integrated time series.

1.4.2 Statistical associations

As noted earlier, we perform event impact analysis by studying the *statistical associations* between a time series and an event series. There is a long history of research on the statistical associations of two or more *time series*, and we have seen an increasing interest in the statistical associations of two or more *event series* (or rather, continuous-time point processes) in the past two decades. However, literature on statistical associations *across* time series and event series is virtually non-existent. A possible explanation is that an event series can be viewed as a special case of a time series, so that there seems to be no reason for a separate treatment of this case.

We believe that the specific properties of event series—in particular, discreteness and sparsity—demand for novel ways to capture their statistical associations with time series in a meaningful way. This is corroborated by the fact that the approaches to study impacts of *singular events* on time series outlined above are remarkably different from the existing measures and tests for statistical associations between time series or between event series.

In the following, we provide an overview of existing measures and tests for statistical associations between two time series $\mathbf{X}$ and $\mathbf{Y}$. Some of these methods can be applied for event impact analysis by simply exchanging one of the time series with an event series $\mathbf{E}$. However, such applications stand on shaky statistical grounds, as, more often than not, these statistical methods are developed under the assumption of *continuous* random variables.

Pairwise associations

The simplest way to study associations across two time series is to consider *pairwise associations*, *i.e.*, associations between a single random variable x_t from $\mathbf{X}$ and a single random variable $y_{t+\delta}$ from $\mathbf{Y}$.

Pairwise associations can be characterized nominally with association measures, or directly tested with statistical independence tests. For example, the *cross-correlation function*

$$\rho_t^{\mathbf{XY}}(\delta) := \frac{\mathrm{Cov}[\mathrm{x}_t, \mathrm{y}_{t+\delta}]}{\sqrt{\mathrm{Var}[\mathrm{x}_t]\,\mathrm{Var}[\mathrm{y}_{t+\delta}]}} \tag{1.29}$$

for $\delta \in \mathbb{Z}$ measures the pairwise *linear associations* between the random variable x_t and every random variable $\mathrm{y}_{t+\delta}$. A value of 0 indicates a lack of linear association at lag δ, and a value of ± 1 indicates a fully deterministic linear relationship at that lag. If $\mathbf{X}$ and $\mathbf{Y}$ are jointly stationary, the cross-correlation function is independent of the time index t and fully determined by the temporal lag between the two time points [BJR+16], so that

$$\rho_t^{\mathbf{XY}}(\delta) = \rho^{\mathbf{XY}}(\delta) = \rho^{\mathbf{YX}}(-\delta) = \rho_t^{\mathbf{YX}}(-\delta) \tag{1.30}$$

for all $t \in \mathbb{Z}$. In principle, any measure designed to capture associations between two random variables can be applied to assess pairwise associations across time series at arbitrary lags. An overview of classical association measures for random variables can be found in Tjøstheim [Tjø96], and more recent developments in Tjøstheim, Otneim, and Støve [TOS18]. In the time series context, the number of *co-movements* [GG61; HP91; YS81; YH86] and *ranks* [Sch89] have been explored as nonparametric alternatives to the cross-correlation function that capture *monotonic associations*. Nonparametric measures for *quantile dependence* across time series have attracted some attention recently [BK19; HLO+16].

A non-zero value of an association measure implies lack of independence of the two random variables x_t and $\mathrm{y}_{t+\delta}$, but not necessarily vice-versa. Only if the association measure reflects the *complete* cdf, or the *complete* pdf (pmf) in the continuous (discrete) case, lack of association implies independence. Such measures are based, *e.g.*, on the *Kolmogorov-Smirnov distance* or the *mutual information* [Tjø96]. An alternative to interpretable association measures is to directly test independence with a suitable test statistic. For example, Fernandes and Néri [FN09] proposed a test for the *instantaneous independence* of two time series $\mathbf{X}$ and $\mathbf{Y}$, *i.e.*, the case $\mathrm{x}_t \perp\!\!\!\perp \mathrm{y}_t$ for all $t \in \mathbb{Z}$, with an entropy-based test statistic, while Chwialkowski and Gretton [CG14] use the Hilbert-Schmidt independence criterion (HSIC) [GBS+05] for the same problem. These tests can easily be adapted for pairwise independence at an arbitrary lag δ.

The boundary between association measures and other test statistics for independence can be fuzzy.

The key challenge when applying association measures or statistical independence tests across time series lies in establishing statistical significance in the presence of serial dependencies. In the context of event impact analysis, pairwise associations of e_t and $\mathrm{x}_{t+\delta}$ for $\delta = 1, \ldots, \Delta$ are a special case of event impacts in the sense

of Definition 1.3.5. They provide a limited perspective on the overall statistical association, but may yield powerful tests in many scenarios. In fact, the method that we develop in Chapter 3 is a test for pairwise independence $e_t \perp\!\!\!\perp x_{t+\delta}$ for $\delta = 1, \ldots, \Delta$.

Tests for pairwise independent processes

A more complex problem that has received much attention in the literature is to establish *pairwise independence* of $\mathbf{X}$ and $\mathbf{Y}$, which is defined via an *infinite number* of pairwise independence relations $x_t \perp\!\!\!\perp y_{t'}$ *for all* $t \neq t'$. However, the vast majority of approaches test whether the time series $\mathbf{X}$ and $\mathbf{Y}$ are *uncorrelated processes*, *i.e.*, whether they have an all-zero cross-correlation function. These works assume either ARMA [DR03; Gew81; Hau76; KY86] or AR(∞) [Hon96; Sha09] representations for the time series. The proposed tests have been generalized to tests for non-correlation of *multivariate time series* with VARMA [ER97; HS05; RF15], VAR(p) [HS07], or VAR(∞) [BR06] representations, to time series with more generic stationarity assumptions [ERD03], and to nonstationary, cointegrated time series [BD08; PRC03; Sai07].

Since these tests are based on the cross-correlation function, they may miss nonlinear pairwise associations between the time series. If the time series are stationary and Gaussian, uncorrelated processes are, in fact, pairwise independent, and the tests listed above can be used to establish pairwise independence. Only few tests exist to establish pairwise independence in a non-Gaussian setting. Hong [Hon01] proposed a test statistic for pairwise independence of $\mathbf{X}$ and $\mathbf{Y}$ based on empirical characteristic functions instead of cross-correlations. The test of Kim and Lee [KL05] for the same problem uses a Cramér-von Mises statistic for this purpose, and was extended by Duchesne, Ghoudi, and Rémillard [DGR12] for nonlinear time series. Besserve et al. [BJL+11] and Besserve, Logothetis, and Schölkopf [BLS13] developed a widely applicable test statistic for pairwise independence based on a kernelized cross-spectral density (KCSD) operator, while Lacal and Tjøstheim [LT19] exploit local Gaussian correlations.

Recently, Khan and Khan [KK20] performed a comparative study of various tests for uncorrelated or pairwise independent processes, and found the approach of Atiq-ur-Rehman and Malik [AM14] to be superior to its competitors in many experiments.

Tests for independent processes

In general, pairwise independence of $\mathbf{X}$ and $\mathbf{Y}$ *does not* imply that the two time series are independent processes $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ in the

sense of Definition 1.3.4. However, if the time series are stationary and Gaussian, pairwise independence *does* imply independent processes, so that the tests for *uncorrelated processes* and *pairwise independent processes* listed above are, in fact, tests for *independent processes* in this case. Besserve, Logothetis, and Schölkopf [BLS13] derived very generic regularity conditions under which pairwise independence implies independent processes, without assuming normality. They use this result to extend the KCSD test for pairwise independence mentioned above [BJL+11] to a test for independent processes in the sense of Definition 1.3.4.

Zhang et al. [ZGS+08] proposed a test statistic based on the HSIC to test for independent processes when the time series jointly follow a k-th order *Markov process*, while the *model-free* tests based on permutation entropy [CGd11; MRM10] or the symbolic correlation integral [CMR+19] detect any statistical cross-dependence up to order k. Laumann, Kügelgen, and Barahona [LKB21] apply the HSIC to test for independent processes in a *nonstationary* setting, when multiple realizations of the time series are available.

Granger causality

A specific independence relation across time series that has received considerable attention in the past decades is known as Granger (non-) causality [BS11; Gew82; Gra69; PH77]. In the simplest bivariate form, the time series $\mathbf{Y}$ is said to *Granger cause* the time series $\mathbf{X}$ if and only if

$$x_t \not\perp\!\!\!\perp y_{t-\Delta}, \dots, y_{t-1} \mid x_{t-\Delta}, \dots, x_{t-1} \tag{1.31}$$

for a fixed lag Δ. Intuitively, Granger causality means that the time series $\mathbf{Y}$ contains additional information on the present value of the time series $\mathbf{X}$ that is not included in the past of $\mathbf{X}$ itself. The choice of independence relations reflects a *predictive scenario*, where one aims at predicting the present value of $\mathbf{X}$ from the past of both $\mathbf{X}$ and $\mathbf{Y}$. In fact, the classical approach to test for Granger causality [BS11] is to compare the residual variance of an estimated forecasting model for $\mathbf{X}$ based on its own past *only*, with that of a joint forecasting model based on the past of $\mathbf{X}$ *and* $\mathbf{Y}$. If the joint forecasting model has a significantly lower residual variance, the null hypothesis of Granger non-causality must be rejected.

The classical approach is to implement the test with VAR forecasting models, but more generic tests have been developed, *e.g.*, based on state-space models [BS15], kernel regressions [MPS08], neural networks [MV21], or nonparametric predictors [BKM96]. Since forecasts are only based on the conditional mean of x_t, other characteristics of its conditional distribution are ignored by these tests.

Therefore, Granger causality is often viewed as *predictive causality* or *causality in mean*. There are a few tests for Granger causality that take more characteristics of the conditional distribution into account [DP06; HJ94; NHK+11; TBE14]. Quite prominently, *transfer entropy* [KS02; Sch00] measures the conditional mutual information of x_t and $y_{t-k}, \ldots, y_{t-1}$ given $x_{t-k}, \ldots, x_{t-1}$, and thus provides the most generic test statistic for Granger causality [BBH+16; TBE14]. For some parametric models, transfer entropy yields the classical test statistic for Granger causality in mean [BBS09; BB12].

Tests for Granger causality from $\mathbf{E}$ to $\mathbf{X}$ are useful to assess whether event occurrences help forecasting the time series. However, the choice of (conditional) independence relations from Equation 1.31 does not reflect the information needs expressed in existing studies of event impacts in the case of singular events. All of the works discussed in Section 1.4.1 monitor the behavior of the time series over *windows of interest* before and after the event (the pre- and post-intervention phases). In contrast, Granger causality considers the behavior of the time series at a *single point in time*, while taking into account its own past and the past behavior of the event series. In this dissertation, we blend the probabilistic perspective of Granger causality with existing methods for singular event impacts to obtain a novel, probabilistic framework for event impact analysis.

Peak Event Coincidence Analysis | 2

This chapter is based on:

[SM20a] Erik Scharwächter and Emmanuel Müller. "Does Terrorism Trigger Online Hate Speech? On the Association of Events and Time Series." *Annals of Applied Statistics* 14(3), 2020. DOI: 10.1214/20-AOAS1338.

Perhaps the most illustrative way to analyze event impacts on stationary time series is to consider the association between event occurrences and the occurrence of *peaks* in the time series. In this chapter, we propose a novel statistical methodology to test, measure, and visualize precisely this association. Intuitively, a peak is a drastic, sudden and short-lived increase of the values of the time series. We show that if there is an unusually large number of *coincidences* between event occurrences and peaks, there is evidence for event impacts in the time series. The *coincidence rate*, *i.e.*, the number of coincidences normalized by the total number of events, serves as a measure of association between event occurrences and peaks in the time series. At last, a plot of these coincidence rates for peaks of increasing magnitudes provides a *visual summary* of this specific type of association.

Technically, we define a peak as the exceedance of a large threshold within the window of interest. We refine the notion of event impacts from Definition 1.3.5 so that it reflects our interest in peaks. We show that a test for our refined notion of event impacts can be implemented effectively within the framework of event coincidence analysis (ECA) [DSS+16]. ECA was originally developed to correlate point processes in continuous time. We formulate a discrete-time variant of ECA and derive all required distributions to enable analyses of peaks in stationary time series, with a special focus on serial dependencies and on peaks over multiple thresholds. Our derivations exploit a central result from extreme value theory (EVT) [Col01] and thus establish an interesting connection between event impact analysis, ECA and EVT. We demonstrate the utility of our approach by analyzing whether Islamist terrorist attacks in Western Europe and North America systematically trigger bursts of hate speech and counter-hate speech on Twitter.

2.1 Introduction

From the perspective of a human analyst peaks are often the most salient features of a time series. It is therefore quite natural to ask whether these peaks systematically coincide with a recurring event of interest. The relevance of peaks has been recognized already in early works on intervention analysis [BT75], where the goal is to model the effect of a singular event on a time series. At first glance, a focus on peaks may seem overly simplistic: Peaks are a

univariate concept that cannot easily be extended to multivariate time series. Furthermore, event occurrences may coincide with other patterns in the time series, *e.g.*, temporarily alter its variance or induce various anomalies. We argue that in many cases such patterns can be *reduced* to peaks after preprocessing the time series with a suitable feature transformation function. The feature transformation function must be designed or learned in such a way that it provides high values when the pattern of interest is observed, and low values when it is not observed.

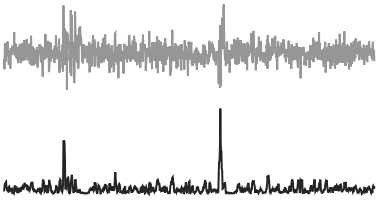

Figure 2.1: In the upper time series, the variance is temporarily increased at two points in time. The feature transformation described on the right results in the lower time series, where these points in time are clearly indicated by peaks.

For example, a useful feature transformation to study the effect of events on the *variance* of a time series would be the ratio between the variance in a short-term window and a long-term window of the time series. Increases in the variance of the original time series will then lead to peaks in the preprocessed time series, as in Figure 2.1. In case we expect events to trigger *anomalies* in the time series, we can preprocess the time series with an anomaly detection algorithm [MVS+15; RXW+19] that provides a score at every time point indicating how anomalous each observation is. If no suitable feature transformation is available, the methods developed later on in Chapter 3 are required.

In summary, we make the following contributions:

- We show that coincidences of peaks and events are indicators for event impacts in the sense of Definition 1.3.5.
- We implement a test for coincidences of peaks and events within the framework of ECA and analytically derive the null distribution of the ECA test statistic for this scenario, under mild assumptions on the time series.
- Since a single threshold may not be sufficient to comprehensively capture the association between event occurrences and peaks, we further derive the joint distribution of the ECA test statistic at multiple thresholds.
- We discuss two hypothesis tests for coincidences at multiple thresholds and propose a novel visualization of these coincidences via quantile-trigger rate plots (QTR plots).
- We validate our statistical methodology with a simulation study and use it to test the systematic association between terrorist attacks and online hate speech.

2.1.1 Problem statement

Let $\mathbf{X} = (x_1, \ldots, x_T)$ be a time series and $\mathbf{E} = (e_1, \ldots, e_T)$ be an event series. We assume that $\mathbf{X}$ and $\mathbf{E}$ are jointly stationary and that the event series is sparse, *i.e.*, $\Pr(e_t = 1) = \epsilon$ for a very small $\epsilon > 0$, so that $\sum_t e_t = N_{\mathbf{E}} \ll T$. On average, we have $L = \frac{1-\epsilon}{\epsilon} \gg 0$ time steps between two event occurrences.

Our goal is to detect event impacts in the sense of Definition 1.3.5. We exploit Lemma 1.3.1 and focus on specific properties of the time series by testing the association between event occurrences and a statistic $g(x_t, \ldots, x_{t+\Delta})$. In this chapter, we address the association between event occurrences and *peaks* in the time series. We assume that the time series has already been preprocessed in such a way that a focus on peaks is justified. Statistically, a focus on peaks translates to an interest in the *extremal properties* of the time series after event occurrences. Formally, we use the **maximum statistic**

maximum statistic

$$g(x_t, \ldots, x_{t+\Delta}) := \max(x_t, \ldots, x_{t+\Delta}) \tag{2.1}$$

to encode this interest, and test the following pair of hypotheses:

$$H_0 : \max(x_t, \ldots, x_{t+\Delta}) \perp\!\!\!\perp e_t \tag{2.2}$$

versus

$$H_1 : \max(x_t, \ldots, x_{t+\Delta}) \not\perp\!\!\!\perp e_t. \tag{2.3}$$

If we have evidence to reject the null hypothesis H_0 in favor of the alternative hypothesis H_1, we have evidence for event impacts in the sense of Definition 1.3.5.

The maximum statistic has two desirable properties: First, it is robust with respect to temporal delays. The statistic will take the same value regardless of whether a peak occurs at the beginning or end of the window of interest. Second, it is sensitive for very short-lived increases of the values of the time series by ignoring all but the largest observation.

The most straightforward way to implement a test for event impacts in the maximum statistic would be to estimate $F_{\max(x_t,\ldots,x_{t+\Delta})|e_t=1}$ and $F_{\max(x_t,\ldots,x_{t+\Delta})}$ from the data and check whether the two cdfs are identical. The downside of this approach is that due to the sparsity of the event series, the number of samples available to estimate the conditional distribution of the maximum statistic is potentially low. Instead, we propose a novel way to implement this test by counting the *number of coincidences* between event occurrences and peaks in the time series, where a peak is defined as a point in time where the maximum statistic exceeds a large threshold.

2.1.2 Related work

In the literature, coincidences are primarily used to measure the association between two event series, or between two **point processes**, *i.e.*, event series in continuous time. Most importantly for us, **event coincidence analysis** (ECA) [DSS+16] was developed to measure the association between two types of recurring events within a fixed window of interest. It has been used to assess whether floods

point processes

event coincidence analysis

systematically trigger epidemic outbreaks [DSS+16], whether natural disasters systematically trigger violent conflicts [SDD+16], and to assess many other coincidences [DDT+11; RWD+15; RBM+16; Sar18; SSH+16; SSD17; SWD+16]. Methodologically, it has been extended to conditional and joint coincidences [SSH+16] of events. We provide an introduction to ECA in Section 2.2.1, since it forms the basis of our test for event impacts in the maximum statistic.

event synchronization (ES)

The method of **event synchronization (ES)** [QKG02] is conceptually similar to ECA in that it counts coincidences between events. The main difference is that the window of interest for coincidences is not fixed but varies dynamically. While this makes the approach completely nonparametric, it also renders derivations of an analytical null distribution for the test statistic very hard to impossible. Odenweller and Donner [OD20] and Wolf et al. [WBB+20] demonstrate that ECA has clear advantages over ES in the presence of serially correlated events, but also that ES and ECA provide qualitatively similar results in other application scenarios.

Early measures to quantify the association of point processes have been studied in neuroscience for the analysis of neural spike trains [BKM04] and include cross-correlograms [Bro99], cross-intensity functions [Bri92], and the reliability statistic of Hunter and Milton [HM03]. A related measure from spatial statistics is Ripley's cross-K function [Dix02] that uses deviations from the expected number of coincidences within a certain radius under independence to measure event association. In the past decade, tests for Granger causality in point processes that transcend coincidences have been investigated thoroughly [BV18; EDD17; KPG+11; NRJ+09; XFZ16; ZPJ+20]. A recent alternative approach is the score-based likelihood ratio test of Galbraith, Smyth, and Stern [GSS20].

In contrast to all of the works listed above that operate on pairs of event series or point processes, we explicitly address the statistical association between an event series and extremal properties of a time series. Therefore, the approach developed in this chapter is closely related to measures and models for tail dependence of random variables [FJS05; YWZ19].

2.2 Methodology

Our key observation is that an unusually large number of coincidences between event occurrences and peaks in the time series is an indicator for event impacts in the maximum statistic defined in Equation 2.1. This observation allows us to implement a test for such event impacts using ECA.

This section is structured as follows. In Section 2.2.1, we provide a discrete-time formulation of ECA for pairs of event series that corresponds to the original continuous-time formulation for point processes [DSS+16]. In Section 2.2.2, we use our formulation of ECA to implement a test for event impacts in the maximum statistic. For this purpose, we derive a novel analytical null distribution for the ECA test statistic in our setting, *i.e.*, its distribution under the assumption that H_0 from Equation 2.2 holds. Next, in Section 2.2.3, we derive the joint null distribution for coincidences at multiple thresholds and describe two test procedures to assess statistical significance. At last, in Section 2.2.4, we complement our analytical results with a novel visualization of the association via quantile-trigger rate (QTR) plots.

2.2.1 Event coincidence analysis

ECA is a statistical methodology to assess whether two types of events are independent or whether one kind of event systematically triggers or precedes the other kind of event. The basic idea of ECA is to count how many times the two kinds of events coincide within a given window of interest, and to assess whether this number is statistically significant under an independence assumption.

Definition

Let $\mathbf{A} = (a_1, \ldots, a_T)$ and $\mathbf{B} = (b_1, \ldots, b_T)$ be two event series of length T with a *fixed* number of events $\sum_t a_t = N_{\mathbf{A}}$ and $\sum_t b_t = N_{\mathbf{B}}$. Furthermore, let $\Delta \in \mathbb{N}_0$ be the user-defined size of the window of interest. ECA measures the extent to which **B** events *precede* **A** events within the window of interest. We thus refer to **A** as the *lagging* and **B** the *leading* event series. ECA considers two possibilities to measure this extent: **trigger coincidences** and **precursor coincidences**. A trigger coincidence occurs whenever a **B** event triggers an **A** event within the window of interest, whereas a precursor coincidence occurs whenever an **A** event is preceded by a **B** event within the window of interest. In the special case $\Delta = 0$, the two concepts are identical. The two types of coincidences are illustrated in Figure 2.2 with $\Delta = 4$. In the example, there are three trigger coincidences and four precursor coincidences. A large number of trigger or precursor coincidences indicates a violation of an independence assumption that is yet to be specified, and—due to the temporal ordering—a possible causal link from **B** to **A**. The opposite direction can be analyzed by exchanging the labels.

trigger coincidences

precursor coincidences

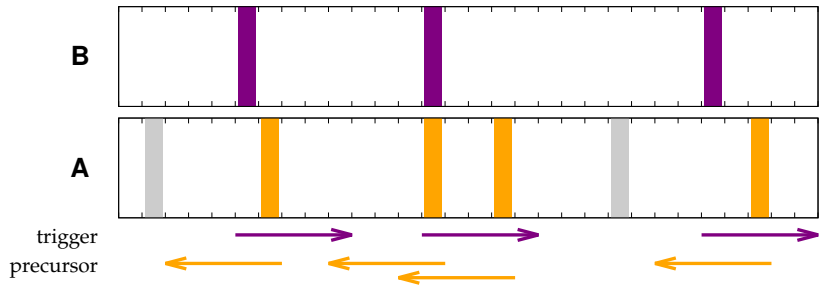

Figure 2.2: Trigger coincidences and precursor coincidences for two event series **A** and **B**, with a window of interest of size $\Delta = 4$.

number of trigger coincidences

Formally, the **number of trigger coincidences** is defined as

$$\mathrm{k_{tr}} = \mathrm{k_{tr}^{\Delta}}(\mathbf{B}, \mathbf{A}) := \sum_{t=1}^{T-\Delta} \mathrm{b}_t \cdot \left(\max_{\delta=0,\ldots,\Delta} \mathrm{a}_{t+\delta} \right) \tag{2.4}$$

number of precursor coincidences

and the **number of precursor coincidences** as

$$\mathrm{k_{pre}} = \mathrm{k_{pre}^{\Delta}}(\mathbf{B}, \mathbf{A}) := \sum_{t=\Delta+1}^{T} \mathrm{a}_t \cdot \left(\max_{\delta=0,\ldots,\Delta} \mathrm{b}_{t-\delta} \right). \tag{2.5}$$

The order of the function arguments **B** and **A** corresponds to the temporal ordering that is analyzed (and thus the potential causal direction). We omit the parameter Δ and the function arguments whenever they are clear from the context. The corresponding **coincidence rates** are given by normalizing the numbers of coincidences by the number of **B** events and **A** events, *i.e.*,

coincidence rates

$$\mathrm{r_{tr}} = \mathrm{r_{tr}^{\Delta}}(\mathbf{B}, \mathbf{A}) := \mathrm{k_{tr}^{\Delta}}(\mathbf{B}, \mathbf{A}) / N_{\mathbf{B}} \tag{2.6}$$

and

$$\mathrm{r_{pre}} = \mathrm{r_{pre}^{\Delta}}(\mathbf{B}, \mathbf{A}) := \mathrm{k_{pre}^{\Delta}}(\mathbf{B}, \mathbf{A}) / N_{\mathbf{A}}. \tag{2.7}$$

The coincidence rates serve as measures for the association of the two event series. In the example from Figure 2.2, we have observed rates $r_{\mathrm{tr}} = 1$ and $r_{\mathrm{pre}} = \frac{2}{3}$. A high *trigger coincidence rate* indicates that a large fraction of **B** events is followed by an **A** event. In other words, **B** events systematically trigger **A** events. A high *precursor coincidence rate* indicates that a large fraction of **A** events is preceded by a **B** event, *i.e.*, the occurrence of **A** events can be explained to a large degree by **B** events. The two measures are complementary and should be selected based on the research question.

Null distribution

The null hypothesis in ECA is that the event series **A** and **B** are independent processes; the alternative hypothesis is that **B** events systematically trigger or precede **A** events—depending on the research question under study. All quantities defined above are statistics computed from the event series **A** and **B** and thus them-

selves random variables with probability distributions induced by **A** and **B**. To assess whether an observed number of trigger coincidences or precursor coincidences is statistically significant under the null hypothesis, we must obtain the *null distributions* of these numbers, *i.e.*, their probability distributions under the assumption that the null hypothesis is true.

We focus on the null distribution of the number of trigger coincidences. For this purpose, we introduce the helper variables

$$\mathrm{a}_t^* := \max_{\delta=0,\ldots,\Delta} \mathrm{a}_{t+\delta} \tag{2.8}$$

for all $t = 1, \ldots, T - \Delta$ that indicate whether there is an **A** event in the window $t, \ldots, t + \Delta$. The helper variables are binary random variables and can thus be viewed as Bernoulli trials with a marginal success probability induced by the distribution of **A**. The helper variables allow rewriting Equation 2.4 as

$$\mathrm{k}_{\mathrm{tr}} = \sum_{t=1}^{T-\Delta} \mathrm{b}_t \cdot \mathrm{a}_t^* = \sum_{t:\mathrm{b}_t=1} \mathrm{a}_t^*, \tag{2.9}$$

which reveals that the number of trigger coincidences is a sum of $N_{\mathbf{B}}$ Bernoulli trials, where each summand is associated with an event occurrence in **B**. Sums over fixed numbers of *independent* and *identically distributed* Bernoulli trials follow binomial distributions. In general, however, two helper variables a_t^* and $\mathrm{a}_{t'}^*$, with $t \neq t'$ are neither identically distributed nor independent. Additional assumptions on the distribution of the event series **A** are required to derive the null distribution of k_{tr} analytically.

To date, the only scenario that has been treated analytically in the literature is the case where **A** is an iid Bernoulli process

or a Poisson process, the continuous-time analogue of a Bernoulli process

$$\mathrm{a}_t \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\pi_{\mathrm{a}}), \tag{2.10}$$

with success probability $\Pr(\mathrm{a}_t = 1) = \pi_{\mathrm{a}}$ for all t. In this case, the helper variables a_t^* are themselves *identically distributed* Bernoulli trials, with marginal success probability

$$\begin{aligned}
\pi_{\mathrm{a}^*} &:= \Pr(\mathrm{a}_t^* = 1) \\
&= 1 - \Pr(\mathrm{a}_t^* = 0) \\
&= 1 - \Pr(\mathrm{a}_t = 0, \ldots, \mathrm{a}_{t+\Delta} = 0) \\
&= 1 - \prod_{\delta=0,\ldots,\Delta} \Pr(\mathrm{a}_{t+\delta} = 0) \\
&= 1 - (1 - \pi_{\mathrm{a}})^{\Delta+1}.
\end{aligned} \tag{2.11}$$

Furthermore, any two helper variables a_t^* and $\mathrm{a}_{t'}^*$ are *independent* if the associated windows are non-overlapping, *i.e.*, if $|t - t'| > \Delta$. Therefore, under the additional assumption that the events in **B**

are separated by more than Δ time steps, so that their respective windows of interest are non-overlapping, the null distribution of k_{tr} for a fixed number of **B** events is the binomial

$$k_{tr} \sim \text{Binomial}(N_{\mathbf{B}}, \pi_{a^*}). \tag{2.12}$$

If **B** events are not strictly separated by more than Δ time steps, this null distribution still approximates the true null distribution as long as the probability to have two **B** events close to each other is very small.

The derivation of the null distribution for the number of precursor coincidences k_{pre} given a fixed number of **A** events is completely analogue, when assuming that **B** is also an iid Bernoulli process.

No such analytical derivations of the null distributions exist for event series other than iid Bernoulli processes (or in fact, point processes that are not Poisson processes). In any other case, Monte Carlo simulations are required to approximate the distributions, *e.g.*, using one of the methods described in Donges et al. [DSS+16].

Statistical test procedure

In ECA, the number of trigger coincidences $k_{tr}^{\Delta}(\mathbf{B}, \mathbf{A})$ is used as a test statistic to decide between the null hypothesis that **A** and **B** are independent processes, and the alternative hypothesis of a trigger relationship from **B** to **A**. If that number is unusually large, the null hypothesis is rejected in favor of the alternative hypothesis. Formally, the ***p*-value** for an observed number of trigger coincidences k_{tr} is defined as the probability of obtaining a test statistic value at least as large as the observed one:

p-value

$$p := \Pr(k_{tr} \geq k_{tr}). \tag{2.13}$$

The null hypothesis is rejected at the desired significance level α if $p < \alpha$. In the scenario discussed above, where **A** is an iid Bernoulli process and the $N_{\mathbf{B}}$ event occurrences in **B** are well separated, the p-value can be obtained directly from the probability mass function of the binomial distribution from Equation 2.12:

$$p = \sum_{k=k_{tr}}^{N_{\mathbf{B}}} \text{Binomial}(k; N_{\mathbf{B}}, \pi_{a^*}). \tag{2.14}$$

To apply this test in practice, we first estimate the success probability of the Bernoulli process **A**, *e.g.*, with the maximum likelihood (ML) estimator $\hat{\pi}_a := N_{\mathbf{A}}/T$. We then use the plug-in principle and estimate the success probability from Equation 2.11 for the binomial distribution as

Alternatively, we can estimate $\hat{\pi}_{a^*}$ from the observations of the helper variables a_t^* using its ML estimator.

$$\hat{\pi}_{a^*} := 1 - (1 - \hat{\pi}_a)^{\Delta+1}. \tag{2.15}$$

A test procedure based on the number of precursor coincidences can be obtained analogously.

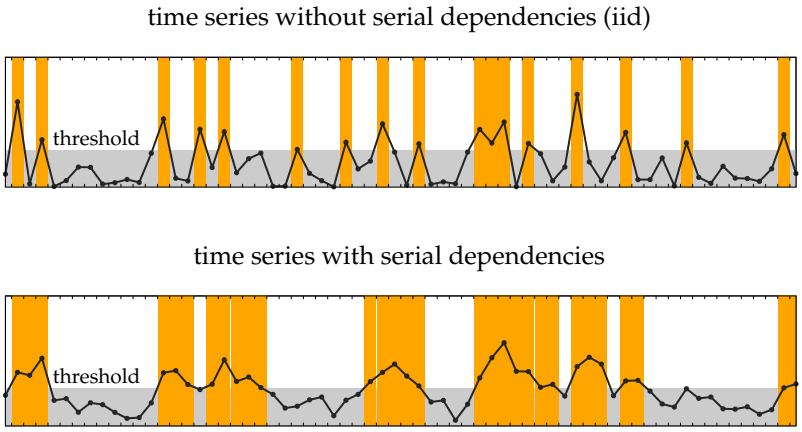

Figure 2.3: Threshold exceedance series (bars) for time series with and without serial dependencies (lines).

2.2.2 Peak coincidences

We now apply the framework of ECA described above to implement a test for event impacts in the maximum statistic as defined in Equation 2.1. For this purpose, we define a discretized notion of **peak** in the time series as an exceedance of a large threshold $\tau \in \mathbb{R}$. The **threshold exceedance series** is the event series obtained by applying the threshold at every time step:

peak

threshold exceedance series

$$\mathbf{A}_\tau(\mathbf{X}) := (\mathbf{1}_{x_1>\tau}, \ldots, \mathbf{1}_{x_T>\tau}) \tag{2.16}$$

The indicator function $\mathbf{1}_C$ is 1 if and only if the condition C is true, and 0 otherwise. The threshold exceedance series retains only information on the timing of the peaks, and disregards all other distributional characteristics of the time series. Figure 2.3 shows two example time series with their threshold exceedance series.

In the following, we show that when applying ECA to test for a trigger relationship between the event series and the threshold exceedance series, we implement a test for event impacts in the maximum statistic. Figure 2.3 illustrates two challenges that need to be addressed in this context: *serial dependencies* and *threshold selection*. Serial dependencies in the time series lead to clustering of events in the threshold exceedance series, so that the analytical results from Section 2.2.1 for Bernoulli processes cannot be applied. Furthermore, the choice of threshold has a strong impact on the results of the analysis, but is often not straightforward. In fact, the magnitude of the peak may vary from event to event: a full picture of the association between events and peaks can only be obtained when considering exceedances at multiple thresholds. We address the first challenge below by deriving a novel analytical null distribution for the number of trigger coincidences when the lagging event series is a threshold exceedance series. We address the second challenge in Section 2.2.3.

Definition

number of trigger coincidences

We obtain the **number of trigger coincidences** for a leading event series $\mathbf{E}$ and peaks in a lagging time series $\mathbf{X}$ by substituting the threshold exceedance series $\mathbf{A}_\tau(\mathbf{X})$ into Equation 2.4:

$$\begin{aligned} \mathrm{k}_{\mathrm{tr}} = \mathrm{k}_{\mathrm{tr}}^{\Delta,\tau}(\mathbf{E}, \mathbf{X}) &:= \mathrm{k}_{\mathrm{tr}}^{\Delta}(\mathbf{E}, \mathbf{A}_\tau(\mathbf{X})) \\ &= \sum_{t=1}^{T-\Delta} \mathrm{e}_t \cdot \left(\max_{\delta=0,\dots,\Delta} \mathbf{1}_{\mathrm{x}_{t+\delta}>\tau} \right). \end{aligned} \tag{2.17}$$

trigger coincidence rate

The **trigger coincidence rate** is obtained as before by normalizing the number of trigger coincidences by $N_{\mathbf{E}}$. It yields an interpretable measure of association between event occurrences and peaks in the time series. The number of precursor coincidences and the precursor coincidence rate are defined analogously. However, the derivations for the null distribution below are only valid for trigger coincidences, and a different strategy (and null hypothesis) must be applied for precursor coincidences.

Null distribution

Observe that our null hypothesis is more specific than the null hypothesis of independent processes employed in standard ECA.

We derive the distribution of the number of trigger coincidences k_{tr} under the assumption that the null hypothesis H_0 from Equation 2.2 holds, *i.e.*, that there are no event impacts in the maximum statistic. Same as before, we introduce binary helper variables

$$\mathrm{a}_t^* := \max_{\delta=0,\dots,\Delta} \mathbf{1}_{\mathrm{x}_{t+\delta}>\tau} \tag{2.18}$$

for all $t = 1, \dots, T-\Delta$ that now indicate whether there is a threshold exceedance in $\mathbf{X}$ in the window of interest $t, \dots, t+\Delta$. These helper variables are Bernoulli trials with a success probability induced by $\mathbf{X}$. Our key observation that enables an analytical derivation of the null distribution of k_{tr} is that we can swap the order of the max-operator and the indicator function in the helper variables,

$$\mathrm{a}_t^* = \mathbf{1}_{(\max_{\delta=0,\dots,\Delta} \mathrm{x}_{t+\delta})>\tau}. \tag{2.19}$$

This simple trick reveals that the marginal success probability of the binary helper variables can be obtained directly from the marginal distribution of the maximum statistic:

$$\begin{aligned}\pi_{a^*} &:= \Pr(a_t^* = 1)\\ &= \Pr\Big(\max_{\delta=0,\dots,\Delta} \mathbf{1}_{x_{t+\delta}>\tau} = 1\Big)\\ &= \Pr\Big(\mathbf{1}_{(\max_{\delta=0,\dots,\Delta} x_{t+\delta})>\tau} = 1\Big)\\ &= \Pr\Big(\max_{\delta=0,\dots,\Delta} x_{t+\delta} > \tau\Big)\\ &= 1 - \Pr\Big(\max_{\delta=0,\dots,\Delta} x_{t+\delta} \leq \tau\Big) \qquad (2.20)\end{aligned}$$

Probability distributions of maxima as in Equation 2.20 are studied in Extreme Value Theory (EVT). In fact, a central result from EVT is the **Extremal Types Theorem** (ETT), which states that the distribution of the maximum of many random variables approaches the **Generalized Extreme Value** (GEV) distribution. The most basic formulation of the ETT is for iid random variables:

Extremal Types Theorem

Generalized Extreme Value

Theorem 2.2.1 (ETT, Theorem 3.1.1 of Coles [Col01])

Let $x_1, \dots, x_n \overset{iid}{\sim} F$ *and* $z_n = \max_{i=1,\dots,n} x_i$. *If there exist constants* $a_n > 0$ *and* b_n *for* $n = 1, 2, \dots$ *such that*

$$\Pr\left(\frac{z_n - b_n}{a_n} \leq z\right) \longrightarrow G(z; \boldsymbol{\theta}) \text{ as } n \longrightarrow \infty$$

for a non-degenerate distribution function G *with parameter vector* $\boldsymbol{\theta}$, *then* G *is a member of the GEV family*

$$G(z; \boldsymbol{\theta}) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, *with parameters* $\boldsymbol{\theta} = (\xi, \mu, \sigma)$ *such that* $-\infty < \mu < \infty$, $\sigma > 0$ *and* $-\infty < \xi < \infty$.

The ETT also applies more generally to the maxima of stationary time series, as long as they fulfill a regularity condition that limits their long-range dependencies; see Coles [Col01, ch. 5.2] for technical details. We state the following intermediate result:

Limited long-range dependencies in the time series mean that distant observations are approximately independent, *e.g.*, $\Pr(x_t \leq \tau \wedge x_{t'} \leq \tau) \approx \Pr(x_t \leq \tau) \cdot \Pr(x_{t'} \leq \tau)$ if $|t - t'| > L$ for some large L.

Lemma 2.2.2 *Let* $\mathbf{X}$ *be a stationary time series that fulfills the conditions of the ETT such that* $\Pr(\max_{\delta=0,\dots,\Delta} x_{t+\delta} \leq \tau) \approx G(\tau; \boldsymbol{\theta}_\Delta)$. *The helper variables* a_t^* *defined in Equation 2.18 are then identically Bernoulli distributed with success probability approximated by*

$$\pi_{a^*} \approx 1 - G(\tau; \boldsymbol{\theta}_\Delta). \qquad (2.21)$$

The larger Δ, *the better the approximation by the GEV distribution.*

Proof. This is a direct consequence of Equation 2.20 and the ETT. The normalizing constants from the ETT disappear in the GEV parameter vector $\boldsymbol{\theta}_\Delta$ that depends on $\mathbf{X}$ and Δ. □

The regularity conditions of the ETT for stationary time series enforce limited long-range dependencies *within* $\mathbf{X}$ *only*. To obtain the distribution of k_{tr}, we need an additional regularity condition that limits long-range dependencies in the *joint process* $(\mathbf{E}, \mathbf{X})$; more precisely, we need to limit the long-range dependencies between $\mathbf{E}$ and peaks in $\mathbf{X}$. The intuition is that if events trigger peaks in the time series, the peak should appear within the window of interest just after an event occurrence—not much later. In other words, the probability of having a peak within the window $t, ..., t + \Delta$ should be largely unaffected by knowledge of an event occurrence at some time point $t' \ll t$. We formalize this regularity condition via approximate independence of the helper variables a^*_t:

We require that there exists some L such that for every subset of time points $\{t_1, ..., t_N\}$ with $|t_i - t_j| > L$ for all $i \neq j$, we have

$$\Pr\Big(\bigwedge_i a^*_{t_i} = 1 \mid \bigwedge_i e_{t_i} = 1 \Big) \approx \prod_i \Pr(a^*_{t_i} = 1 \mid e_{t_i} = 1). \tag{2.22}$$

If this is the case, a selection of helper variables associated with event occurrences can be viewed as approximately independent, as long as the event occurrences are separated by enough time steps. The key result of this section is the following:

Theorem 2.2.3 *Let* $\mathbf{X}$ *be a stationary time series that fulfills the conditions of the ETT. Let* $\mathbf{E}$ *be an event series such that the additional regularity condition of Equation 2.22 is fulfilled for some* L*. Furthermore, assume that the* $N_{\mathbf{E}}$ *events in* $\mathbf{E}$ *are separated by more than* L *time steps.*

Under H_0 *from Equation 2.2, the null distribution of the number of trigger coincidences from Equation 2.17 is given by the binomial distribution*

$$k_{tr}^{\Delta,\tau}(\mathbf{E}, \mathbf{X}) \sim \text{Binomial}(N_{\mathbf{E}}, \pi_{a^*}), \tag{2.23}$$

where π_{a^*} *is approximated by Equation 2.21.*

In other words, the theorem states that if there are no event impacts in the maximum statistic, the number of trigger coincidences follows a binomial distribution that can be estimated using the ETT.

Proof. The number of trigger coincidences is a sum of $N_{\mathbf{E}}$ Bernoulli trials, each associated with an event occurrence:

$$k_{tr} = \sum_{t=1}^{T-\Delta} e_t \cdot a^*_t = \sum_{t:e_t=1} a^*_t. \tag{2.24}$$

With the regularity condition from Equation 2.22 and sparsity of $\mathbf{E}$, these Bernoulli trials are approximately independent and identically distributed with success probability $\Pr(a^*_t = 1 \mid e_t = 1)$.

We can therefore model their sum by a binomial distribution with $N_{\mathbf{E}}$ trials and the given success probability. Under H_0, we have that $\Pr(\mathrm{a}_t^* = 1 \mid \mathrm{e}_t = 1) = \Pr(\mathrm{a}_t^* = 1) = \pi_{\mathrm{a}^*}$, which can be approximated by Equation 2.21 according to Lemma 2.2.2. □

We stress that these novel analytical results are valid in the presence of serial dependencies in the time series, where the previous analytical results from Section 2.2.1 for Bernoulli processes fail. Our derivations establish an interesting connection between ECA, EVT and the notion of event impacts from Definition 1.3.5.

Statistical test procedure

We adapt the ECA test procedure from Section 2.2.1 to test for event impacts in the maximum statistic. We use $\mathrm{k}_{\mathrm{tr}}^{\Delta,\tau}(\mathbf{E}, \boldsymbol{X})$ with fixed Δ and τ as a test statistic to decide between H_0 from Equation 2.2 and H_1 from Equation 2.3. The only difference to standard ECA is that we use the binomial model from Theorem 2.2.3 to compute the p-value for an observed number of trigger coincidences k_{tr}. If we reject H_0 in favor of H_1, we have evidence for event impacts in the wider sense of Definition 1.3.5.

To apply this test in practice, we use the ML estimates $\hat{\boldsymbol{\theta}}_\Delta$ for the parameters of the GEV distribution $G(\cdot; \boldsymbol{\theta}_\Delta)$. They are obtained by splitting the observed time series $\boldsymbol{X}$ into consecutive blocks of size $\Delta+1$, extracting the maximum from each block, and optimizing the likelihood function on these block maxima. Unfortunately, there is no analytical solution for the ML estimates of the GEV parameters, so that numerical optimization techniques are required; see Coles [Col01, ch. 3.3] for details. After estimating the GEV parameters, we use the plug-in principle and estimate the success probability π_{a^*} for the binomial model from Theorem 2.2.3 by evaluating the estimated GEV distribution at the threshold τ, *i.e.*,

$$\hat{\pi}_{\mathrm{a}^*} := 1 - G(\tau, \hat{\boldsymbol{\theta}}_\Delta). \tag{2.25}$$

Alternatively, we could estimate $\hat{\pi}_{\mathrm{a}^*}$ directly from the observations of the helper variables a_t^* using the Bernoulli ML estimator.

The common procedure in statistical hypothesis testing is to reject the *null hypothesis* if the observed test statistic value—in our case, the number of trigger coincidences—is so large that $p < \alpha$. It is important to keep in mind that the p-value depends not only on the null hypothesis, but on all other assumptions as well. The p-value may be small if *any of these assumptions* is violated [GSR+16], so that, in fact, any of these assumptions may potentially be rejected. Before applying the test procedure from above to test for event impacts, the assumptions of Theorem 2.2.3 should be verified carefully for the data at hand—using existing statistical methodology, any

available domain knowledge, and, in the end, human judgment. The key questions to answer beforehand are the following:

- Are the time series and the event series stationary?
- Are there no long-range dependencies within the time series?
- Is it reasonable to assume that the events have no long-range influence on the time series? More precisely, is it reasonable to assume that an event occurrence only influences the time series within its own window of interest, and never within the window of interest of a later event occurrence?

If the sequences are not stationary, the methodology described in this chapter is not applicable in the first place. If there are long-range dependencies in the time series, the ETT is not applicable, *i.e.*, the GEV distribution does not approximate the marginal success probabilities of the helper variables. At last, if events have long-range influence on the time series, the binomial distribution is not a suitable model for the number of trigger coincidences.

2.2.3 Multiple thresholds

The methodology proposed above depends on a threshold τ. In case a suitable threshold is unknown, or if a full picture of the association with peaks of various magnitudes is required, threshold exceedances at *multiple* thresholds have to be considered. Exceedances of multiple thresholds are highly dependent: if an observation exceeds any threshold τ, it also exceeds all lower thresholds. The numbers of trigger coincidences at multiple thresholds are thus dependent as well. We now derive the *joint null distribution* of the numbers of trigger coincidences at multiple thresholds. This enables joint analyses of multiple threshold exceedances and eliminates the need of selecting a single threshold.

Trigger coincidence processes

trigger coincidence process

Let $\boldsymbol{\tau} = (\tau_1, ..., \tau_M)$ be a sequence of increasing thresholds, *i.e.*, $\tau_1 < ... < \tau_M$. The **trigger coincidence process**

$$\mathbf{k}_{\text{tr}} = \mathbf{k}_{\text{tr}}^{\Delta,\boldsymbol{\tau}}(\mathbf{E}, \mathbf{X}) = \left(\mathrm{k}_{\text{tr}}^{\Delta,\tau_1}(\mathbf{E}, \mathbf{X}), ..., \mathrm{k}_{\text{tr}}^{\Delta,\tau_M}(\mathbf{E}, \mathbf{X})\right) \tag{2.26}$$

is the corresponding sequence of the numbers of trigger coincidences for all given thresholds τ_m. A trigger coincidence process is always monotonically decreasing. The **canonical** trigger coincidence process is given by the specific threshold sequence $\boldsymbol{\tau} = (\tau_1, ..., \tau_T) = (\mathrm{x}_{(1)}, ..., \mathrm{x}_{(T)})$, where $\mathrm{x}_{(t)}$ denotes the t-th order

canonical

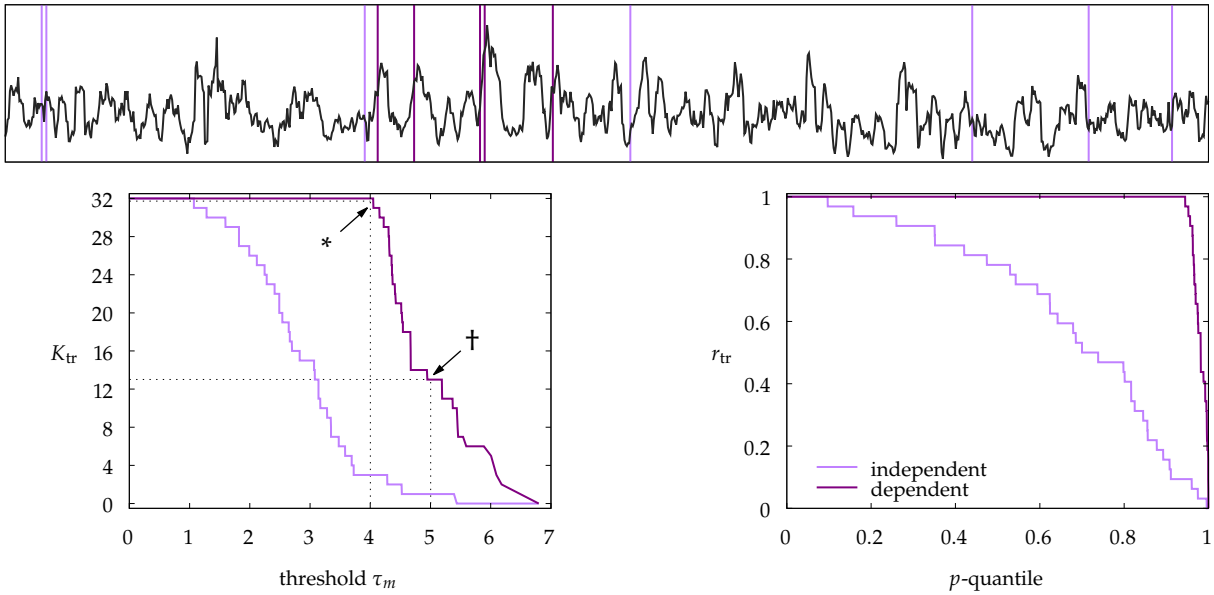

Figure 2.4: Canonical trigger coincidence processes (bottom left) and corresponding QTR plots (bottom right) with $\Delta = 7$ for a simulated time series and two event series (excerpts shown on top). We generated a time series of length $T = 4096$ from iid exponential random variables, applied a moving average (MA) filter of order 8, standardized and subtracted the minimum to obtain a time series $\boldsymbol{X}$ with serial dependencies. We then generated two event series $\boldsymbol{E}$: an independent and a dependent one. In the dependent case, we randomly sampled $N = 32$ time steps t from the time series where $x_t > 4$, and set $e_{t-4} = 1$ for these t. In the independent case, we distributed 32 events completely at random.

statistic of the time series, *i.e.*, $x_{(1)} < \ldots < x_{(T)}$. Trigger coincidence processes for other sequences of thresholds approximate the canonical trigger coincidence process.

In Figure 2.4 (bottom left), we visualize two canonical trigger coincidence processes by plotting the threshold values against the numbers of trigger coincidences; simulation details are in the caption. At low thresholds, large numbers of trigger coincidences are observed both for the dependent and the independent event series. For higher thresholds, the numbers of trigger coincidences for the dependent event series dramatically exceed the numbers of the independent event series. By construction, all 32 events in the dependent event series trigger an exceedance of the threshold 4; see the marker (∗). The threshold 5 is exceeded after 13 out of 32 events from the dependent event series, see the marker (†). For the independent series, the numbers are much lower in both cases.

Markov model

We consider the joint distribution of the trigger coincidence process $\mathbf{k}_{tr} = (k_{tr}^{\Delta,\tau_1}, \ldots, k_{tr}^{\Delta,\tau_M})$. The product rule yields

$$\Pr(\mathbf{k}_{tr}) = \Pr(k_{tr}^{\Delta,\tau_1}) \cdot \prod_{m=2}^{M} \Pr\left(k_{tr}^{\Delta,\tau_m} \mid k_{tr}^{\Delta,\tau_1}, \ldots, k_{tr}^{\Delta,\tau_{m-1}}\right). \qquad (2.27)$$

We have derived the marginal distribution $\Pr(k_{tr}^{\Delta,\tau})$ under the null hypothesis H_0 in Theorem 2.2.3 and now focus on the conditional distributions $\Pr(k_{tr}^{\Delta,\tau_m} \mid k_{tr}^{\Delta,\tau_1}, \ldots, k_{tr}^{\Delta,\tau_{m-1}})$.

Suppose there is an exceedance of the threshold τ' in $\mathbf{X}$ within the window of interest $t, \ldots, t + \Delta$. Using the helper variables from Equation 2.18, we write this condition as $a_t^*(\tau') = 1$, where we highlight in the notation that the helper variables are functions of the threshold. The probability that there is an exceedance of a threshold $\tau > \tau'$ in the same window of interest is given by

$$\Pr(a_t^*(\tau) = 1 \mid a_t^*(\tau') = 1) = \frac{\Pr(a_t^*(\tau) = 1)}{\Pr(a_t^*(\tau') = 1)} \tag{2.28}$$

$$\approx \frac{1 - G(\tau; \boldsymbol{\theta}_\Delta)}{1 - G(\tau'; \boldsymbol{\theta}_\Delta)}, \tag{2.29}$$

where we use Lemma 2.2.2 for the approximation. Equation 2.29 is therefore valid whenever the conditions of the lemma are met. Observations in the time series that do not exceed the threshold τ' cannot exceed the higher threshold τ, so that

Alternatively, the conditional success probability of the helper variables could be approximated with the generalized Pareto distribution by employing a peaks-over-threshold perspective [Col01].

$$\Pr(a_t^*(\tau) = 1 \mid a_t^*(\tau') = 0) = 0 \tag{2.30}$$

With these two results we can specify the conditional distribution for the number of trigger coincidences at threshold τ, given the number of coincidences at threshold τ':

Theorem 2.2.4 *Let $\mathbf{X}$ be a stationary time series that fulfills the conditions of the ETT. Let $\mathbf{E}$ be an event series such that the additional regularity condition of Equation 2.22 is fulfilled for some L for two thresholds τ and τ' with $\tau' < \tau$. Furthermore, assume that the $N_{\mathbf{E}}$ events in $\mathbf{E}$ are separated by more than L time steps.*

Under H_0 from Equation 2.2, the conditional null distribution of the number of trigger coincidences at threshold τ, given the number of coincidences k' at threshold τ', follows the binomial distribution

$$k_{tr}^{\Delta,\tau}(\mathbf{E}, \mathbf{X}) \mid k_{tr}^{\Delta,\tau'}(\mathbf{E}, \mathbf{X}) = k' \sim \text{Binomial}(k', \pi), \tag{2.31}$$

where π is approximated by Equation 2.29.

Proof. With the regularity condition from Equation 2.22 and sparsity of $\mathbf{E}$, we can model the conditional number of trigger coincidences by a binomial with k' trials and success probabilities $\Pr(a_t^*(\tau) = 1 \mid a_t^*(\tau') = 1, e_t = 1)$. Under H_0, we have

$$\begin{aligned} &\Pr(a_t^*(\tau) = 1 \mid a_t^*(\tau') = 1, e_t = 1) \\ &\qquad = \Pr(a_t^*(\tau) = 1 \mid a_t^*(\tau') = 1) =: \pi, \end{aligned} \tag{2.32}$$

which is approximated by Equation 2.29. □

For three thresholds $\tau > \tau' > \tau''$, the number of trigger coincidences at threshold τ is conditionally independent of the number at threshold τ'' given the number at threshold τ'. Therefore, the conditional distributions from Equation 2.27 can be simplified to a first-order Markov structure

$$\Pr(\mathrm{k}_{\mathrm{tr}}^{\Delta,\tau_m} \mid \mathrm{k}_{\mathrm{tr}}^{\Delta,\tau_1}, \ldots, \mathrm{k}_{\mathrm{tr}}^{\Delta,\tau_{m-1}}) = \Pr(\mathrm{k}_{\mathrm{tr}}^{\Delta,\tau_m} \mid \mathrm{k}_{\mathrm{tr}}^{\Delta,\tau_{m-1}}). \tag{2.33}$$

As a final result, we have that the joint distribution $\Pr(\mathbf{k}_{\mathrm{tr}} = \boldsymbol{k}_{\mathrm{tr}})$ of the trigger coincidence process for a sequence of thresholds $\boldsymbol{\tau}$ under the null hypothesis H_0 is fully described by Theorem 2.2.3 for the smallest threshold and Theorem 2.2.4 for all larger thresholds, when the number of events in $\mathbf{E}$ is fixed to $N_{\mathbf{E}}$.

Statistical test procedures

With our results from above and from Section 2.2.2, we can now devise two additional procedures to test for event impacts in the maximum statistic. These test procedures take a more holistic perspective on the association of event occurrences and peaks in the time series, in that they consider peaks over multiple thresholds instead of a single threshold only:

1. We can employ our test for *pointwise* exceedances of individual thresholds from Section 2.2.2 *multiple times* at all given thresholds and adjust the resulting p-values using standard methods for **multiple hypothesis testing** [DL07]. A potential shortcoming of this procedure is that the dependency structure of trigger coincidence processes is ignored.
   multiple hypothesis testing
2. We can compute the likelihood of the observed trigger coincidence process $\Pr(\mathbf{k}_{\mathrm{tr}} = \boldsymbol{k}_{\mathrm{tr}})$ with the distributions derived above and reject the null hypothesis if the *whole process* is unusually unlikely under the null hypothesis, in the sense specified below. This approach takes the full dependency structure into account, but requires Monte Carlo simulations. We refer to it as the **multiple threshold test**.
   multiple threshold test

For the second approach, we observe that the trigger coincidence process is a high-dimensional discrete random variable, where every single realization—even the mode of the distribution—has a very small likelihood. We have to assess whether the observed likelihood is *unusually small* with respect to the *distribution of the likelihood values* under the null hypothesis, *i.e.*, we treat the likelihood as a random variable. For numerical reasons, we work with the negative log-likelihood

$$\mathrm{s}(\mathbf{k}_{\mathrm{tr}}) = -\log \Pr(\mathbf{k}_{\mathrm{tr}}) \tag{2.34}$$

instead of the likelihood. Formally, we use s as our test statistic in the multiple threshold test and reject the null hypothesis H_0 at significance level α if the p-value $\Pr(\mathrm{s} \geq s) < \alpha$, where s is the observed value. We use Monte Carlo simulations to approximate this p-value. For this purpose, we generate R independent event series $\boldsymbol{E}'$ by randomly permuting the observed $\boldsymbol{E}$. For each independent event series, we determine the test statistic value s' and compute the Monte Carlo p-value [DH97] via $\hat{p} = \frac{1+|\{s'|s' \geq s\}|}{R+1}$.

2.2.4 Quantile-trigger rate plots

We conclude our methodological contributions by discussing means to visualize the association between event occurrences and peaks in the time series. Plots of trigger coincidence processes as in Figure 2.4 (bottom left) should help in visually assessing whether events in $\mathbf{E}$ systematically trigger peaks of various magnitudes in a time series $\mathbf{X}$ or not. However, the scales of the axes depend on the range of values in $\mathbf{X}$ and the number of events $N_{\mathbf{E}}$, which makes it difficult to visually recognize patterns. Furthermore, the absolute threshold value is not informative about the actual extremeness of a peak with respect to the bulk of the data. Therefore, we propose **quantile-trigger rate (QTR) plots** as a standardized visualization of trigger coincidence processes with normalized axes. In a QTR plot, the horizontal axis is normalized by using empirical p-quantiles from $\mathbf{X}$ instead of the absolute thresholds τ_m, while the vertical axis is normalized by using the trigger coincidence rate $\mathrm{r_{tr}}$ instead of the absolute number of trigger coincidences $\mathrm{k_{tr}}$.

quantile-trigger rate (QTR) plots

The QTR plot for the simulated example from above is shown in Figure 2.4 (bottom right). The most striking difference is that now the dependent curve appears more extreme, since the thresholds larger than 4 correspond to high empirical p-quantiles. Intuitively, the closer an observed trigger coincidence process to the top-right corner of the QTR plot, the more events coincide with threshold exceedances, at more extreme levels.

However, QTR plots have to be interpreted with care. The shape of a trigger coincidence process for an *independent* pair of event series and time series in a QTR plot depends on the statistical properties of the input data. For example, if $\mathbf{X}$ is an iid time series and $\mathbf{E}$ an iid Bernoulli process, the fraction of events that coincide with an exceedance of the empirical p-quantile of $\mathbf{X}$ with $\Delta = 0$ is exactly $1 - p$, and the trigger coincidence process is a straight line from $(0, 1)$ to $(1, 0)$ in the QTR plot. Figure 2.5 illustrates the impact of serial dependencies in $\mathbf{X}$ and increasing Δ on the shape of the trigger coincidence process *under independence* in a QTR plot. With increasing Δ, there are more trigger coincidences under

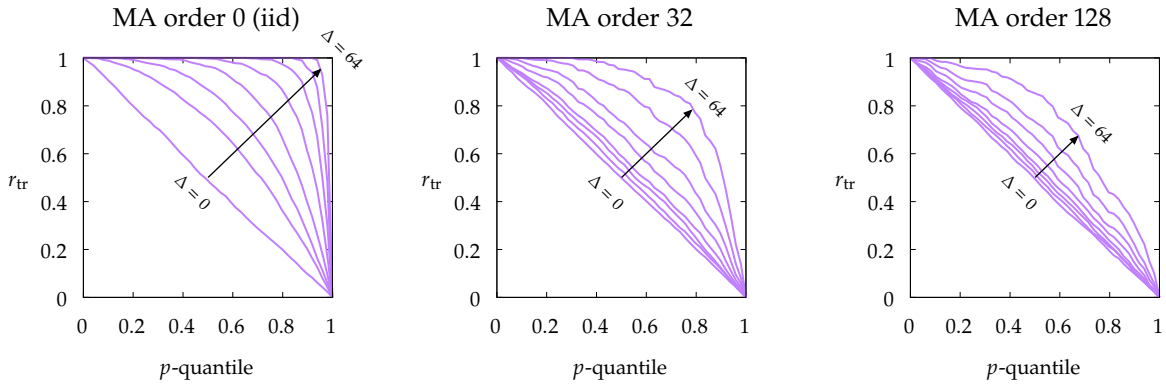

Figure 2.5: Expected QTR plots for three time series with different levels of serial dependencies (MA orders 0, 32, 128) and independent event series. For every MA order, we simulate a single time series $\boldsymbol{X}$ of length $T = 4096$ from the exponential moving average model described in Figure 2.4, and select 50 thresholds at equally spaced p-quantiles between 0 and 1. For every threshold τ and every $\Delta \in \{0, 1, 2, 4, 8, 16, 32, 64\}$, we estimate the expected trigger coincidence rate $r_{\text{tr}} = k_{\text{tr}}/N$ for an independent event series by simulating 100 independent event series $\boldsymbol{E}$ with $N = 32$ events and averaging the trigger coicidence rates over the 100 runs. For large Δ and τ, this expectation can be approximated by the expected value of the GEV-based binomial distribution from Theorem 2.2.3.

independence, and the lines in the QTR plot move towards the top-right corner of the plot. This effect is strongest for iid time series, but also occurs for time series with serial dependencies. Thus, a curve that bends towards the top-right corner of the QTR plot is necessary, but not sufficient to conclude a trigger relationship. We need one of the tests from Section 2.2.3 to assess whether the curve in a QTR plot is, in fact, unusual under the null hypothesis.

Given the intricacies of interpreting QTR plots, we suggest to use them primarily to put an observed curve into context by comparing it to the expected curve under the null hypothesis. An interesting direction for future research would be to make the curves in QTR plots comparable across different pairs of event series and time series by making them invariant to specific properties of the data. One possible way to achieve this is by plotting the ratios or differences between the observed trigger coincidence rates and the rates under independence, with suitable rescaling.

2.3 Experiments

The experimental part of this chapter is twofold. First, we validate our findings from above with Monte Carlo simulations. Second, we apply our methodology to a real-world problem, where we study the association between an event series of Islamist terrorist attacks and a time series that reflects hate speech on Twitter.

2.3.1 Simulations

Quality of the GEV-based binomial distribution

The central result from Section 2.2.2 is Theorem 2.2.3, which states that, under some regularity conditions, the null hypothesis H_0 from Equation 2.2 implies that the number of trigger coincidences for a single threshold approximately follows a binomial distribution with success probability obtained from the GEV distribution. This approximate result is useful specifically for the case of time series with serial dependencies, where the Bernoulli-based null distribution from standard ECA cannot be applied. We now demonstrate that the Bernoulli-based null distribution indeed fails to describe the empirically observed numbers of trigger coincidences for time series with serial dependencies, while our GEV-based null distribution accurately describes the observed data.

For this purpose, we simulate three time series with MA orders 0 (iid), 32 and 64 from the exponential time series model described earlier in Figure 2.4. For every time series, we simulate 1,000 independent pairs of event series with $N_{\mathbf{E}} = 32$ events, and record the numbers of trigger coincidences at the three thresholds $\tau \in \{3, 4, 5\}$ with $\Delta = 7$. For every time series and choice of threshold, we compare the empirically obtained (Monte Carlo) null distribution of the number of trigger coincidences with the two analytical null distributions. The three cumulative distribution functions are visualized in Figure 2.6 for every threshold and MA order. The visualizations clearly show that our GEV-based estimate closely follows the empirical distribution in all runs, while the Bernoulli-based estimate is only correct for iid time series. The results also demonstrate that—in these examples—a value of $\Delta = 7$ is large enough for the GEV approximation to be appropriate.

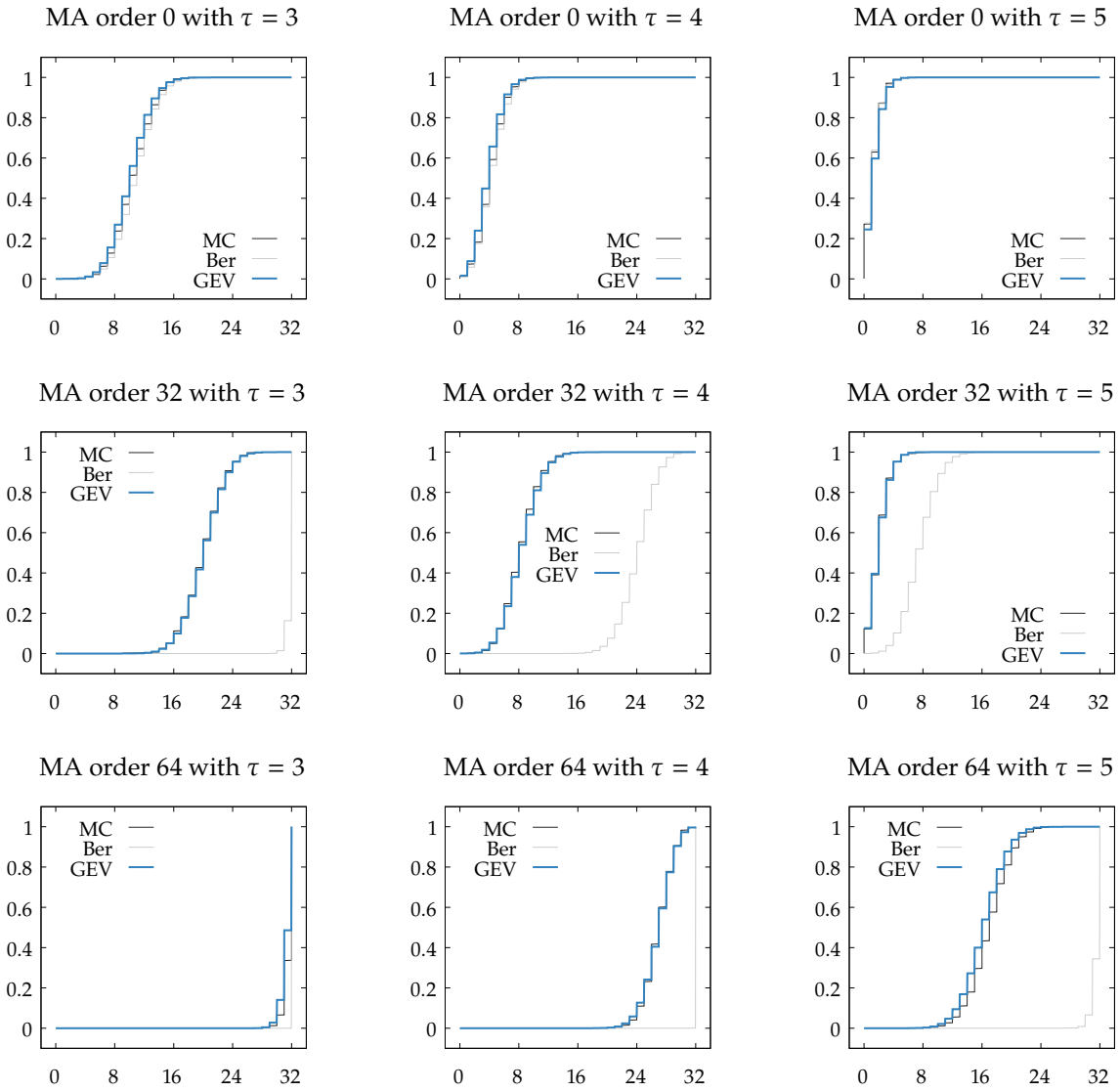

Figure 2.6: Cumulative distribution functions for the number of trigger coincidences under the null hypothesis H_0, obtained empirically by Monte Carlo simulations (MC), and analytically with the Bernoulli-based binomial distribution (Ber) and the GEV-based binomial distribution (GEV).

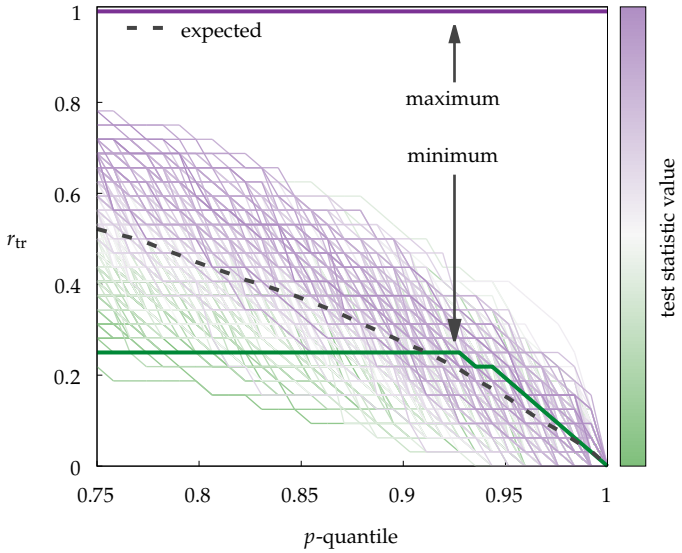

Figure 2.7: Simulated trigger coincidence processes under independence, colorized by the test statistic value, along with the processes that attain the theoretical minimum and maximum test statistic value.

Behavior of the multiple threshold test statistic

The central result from Section 2.2.3 is the Markov model for trigger coincidence processes based on Theorem 2.2.3 and Theorem 2.2.4. The Markov model yields the likelihood-based test statistic for our multiple threshold test. The underlying assumption is that larger values of the test statistic correspond with visually more "extreme" trigger coincidence processes in a QTR plot, *i.e.*, curves that bend towards the top-right corner of the QTR plot.

To confirm this assumption, we illustrate the test statistic values for a single simulated time series (MA order 8) and 1,000 independent event series (with 32 events). We use $\Delta = 7$ and 32 thresholds at equally spaced p-quantiles between 0.75 and 1 from the time series. All resulting trigger coincidence processes are plotted in Figure 2.7, colorized by their test statistic values. We also plot the trigger coincidence process with the highest (lowest) test statistic value that is theoretically possible; we obtain them by maximizing (minimizing) the test statistic over all possible processes with a dynamic programming approach. At last, we show the *marginally* expected trigger coincidence process at every threshold τ_m, *i.e.*, the value of $\mathrm{E}[\mathrm{k}_{\mathrm{tr}}^{\Delta,\tau_m}]$ obtained with the binomial distribution from Theorem 2.2.3. All simulated trigger coincidence processes are close to the marginally expected sequence; the more they bend towards the top-right corner of the plot, the higher the test statistic value. The trigger coincidence process with the highest possible test statistic value closely traces the top-right corner. This corresponds to our intuitive notion of the most unusual outcome, where all events trigger exceedances of the highest quantile.

2.3.2 Hate speech on Twitter

A recent publication by the United Nations Educational, Scientific and Cultural Organization (UNESCO) points out that the "character of hate speech online and its relation to offline speech and action are poorly understood" and that the "causes underlying the phenomenon and the dynamics through which certain types of content emerge, diffuse and lead—or not—to actual discrimination, hostility or violence" should be investigated more deeply [GGA+15]. The methodology proposed in this chapter enables such investigations; in particular, it enables analyses of the systematic relation between rare offline events and online hate speech.

Following recent studies by Burnap et al. [BWS+14] and Olteanu et al. [OCB+18], we analyze whether Islamist terrorist attacks systematically trigger bursts of hate speech and counter-hate speech on Twitter. We operationalize these speech acts by tracking usage of the hashtags `#stopislam` (anti-Muslim hate speech) and `#notinmyname` (Muslim counter-hate speech), as well as the Arabic keyword `kafir` (jihadist hate speech against "non-believers") over a period of three years (2015–2017). We correlate usage of these terms with severe terrorist incidents in Western Europe and North America in the same time period. If bursts of hate speech—peaks in the Twitter time series—coincide with terrorist attacks more often than expected under the null hypothesis, there is evidence for a systematic statistical relationship between the two.

Data

For a quantitative analysis of social media usage in reaction to Islamist terrorist attacks we have to operationalize these terms. We stress that our study design is not intended to provide definitive answers, but rather as a proof-of-concept that demonstrates how our statistical methodology can be applied to a research question from the social sciences. The selection of events and the definitions of hate speech and counter-hate speech from below can be criticized in many ways with regard to implicit biases and assumptions.

Islamist terrorist attacks. We obtained a comprehensive list of global terrorist attacks from the publicly available Global Terrorism Database (GTD) [Nat18]. We filtered the GTD for attacks that occurred in Western Europe and North America between January 2015 and December 2017, left at least 10 people wounded, and were conducted by the so-called *Islamic State of Iraq and the Levant* (ISIL), *Al-Qaida in the Arabian Peninsula* (AQAP), Jihadi-inspired extremists or Muslim extremists, according to the GTD. The resulting 17 severe Islamist terrorist attacks are shown in Table 2.1.

Table 2.1: Severe Islamist terrorist attacks in Western Europe and North America.

Date	City	Date	City
2015-01-07	Paris, France	2016-12-19	Berlin, Germany
2015-11-13	Paris, France	2017-03-22	London, UK
2015-12-02	San Bernardino, USA	2017-04-07	Stockholm, Sweden
2016-03-22	Brussels, Belgium	2017-05-22	Manchester, UK
2016-06-12	Orlando, USA	2017-06-03	London, UK
2016-07-14	Nice, France	2017-08-17	Barcelona, Spain
2016-07-24	Ansbach, Germany	2017-09-15	London, UK
2016-09-17	New York City, USA	2017-10-31	New York City, USA
2016-11-28	Columbus, USA		

Social media response. We retrieved time series of the global Twitter volume in the same time period (2015–2017) for the three keywords `#stopislam`, `#notinmyname` and `kafir` ("non-believer") that represent hate speech and counter-hate speech:

- The hashtag `#stopislam` has been observed in anti-Muslim hate speech before [MDA15; OCB+18] and has also received some media attention [Dew16; Hem16]. Many posts that contain the hashtag actually condemn its usage, so spikes in the volume should not be seen as pure bursts of hate speech. Yet, such condemnation is typically triggered by initial anti-Muslim posts. Due to the mixed usage, the magnitude of a spike is no indicator for the *extent* of online hate, only the *presence* of a spike is informative.
- The phrase "not in my name" is used by members of a group to express their disapproval of actions that are associated with that group or (perceived or actual) representatives of the group [ČB08; Tor06]. It was observed, for example, during global protests against the 2003 war of the US-led coalition against Iraq [Ben05], or more recently during protests sparked by the murder of a Muslim boy by Hindu nationalists in India 2017 [Kri17]. Most importantly for the present study, Muslim social media users have repeatedly used the hashtag after Islamist terrorist attacks [Dav14]. Due to the generic nature of the phrase, it cannot solely be viewed as Muslim counter-hate speech. Nonetheless, online social media posts that contain `#notinmyname` right after Islamist terrorist attacks are likely to convey a Muslim counter-hate message.
- The Arabic word `kafir` translates to the English word "non-believer." It is used by Muslim fundamentalists against other Muslims that do not adhere to the fundamentalist ideology [Alv14], and against non-Muslims [BM12], in both cases to justify their killing. The occurrence of the keyword `kafir` within online social media posts was recently shown to be a strong indicator for jihadist hate speech [DDV18]. We use male, female and plural forms (`kafir`—`kafirah`—`kuffar`) in Arabic script for the query.

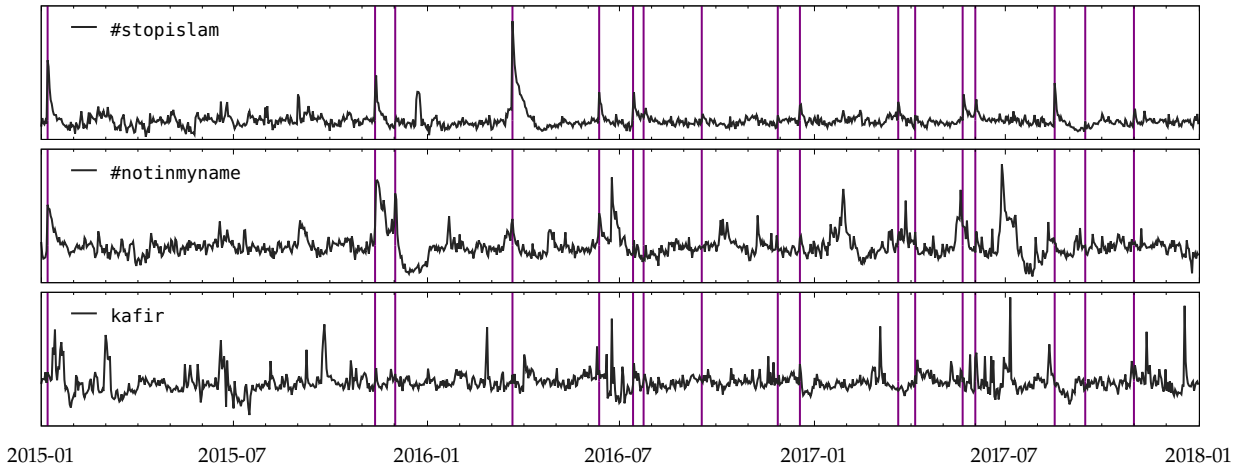

Figure 2.8: Daily Twitter volume of the keywords analyzed in this study. The vertical lines indicate dates of severe Islamist terrorist attacks in Western Europe and North America.

We used the ForSight platform by Crimson Hexagon/Brandwatch to retrieve daily time series of the global Twitter volume for our keywords. Posts with the keyword `RT` were excluded to ignore retweets. We preprocessed the original time series by taking the logarithm to base 2 and subtracting the running mean over the past 30 days to make them stationary. The global daily volume for all queries after preprocessing is shown in Figure 2.8, along with all Islamist terrorist attacks from Table 2.1.

`http://brandwatch.com/`

Experimental setup

The data described above spans a total of $T = 1,096$ days with $N_{\mathbf{E}} = 17$ events. We choose a window of interest with $\Delta = 7$ days to allow enough time for the news about the incidents to spread globally. For every social media time series $\boldsymbol{X}_i$, we estimate a GEV distribution G_i by splitting $\boldsymbol{X}_i$ into consecutive blocks of size $\Delta + 1$ and fitting the parameters of G_i to the block maxima by maximum likelihood estimation. We then select $M = 32$ thresholds $\boldsymbol{\tau}_i = (\tau_{i,1}, \ldots, \tau_{i,32})$ at equidistant p-quantiles between 0.75 and 1 from $\boldsymbol{X}_i$, and use the GEV distribution G_i to obtain the parameters of the binomial distributions from Theorem 2.2.3 and Theorem 2.2.4. We compute the observed trigger coincidence processes between the terrorist attack event series $\boldsymbol{E}$ and all social media time series $\boldsymbol{X}_i$, and obtain the respective test statistic values s_i for the multiple threshold test from our Markov model. To assess statistical significance, we compute Monte Carlo p-values for every time series with $R = 10{,}000$ simulated independent event series.

Model checking

The results of our statistical analysis are only reliable if the underlying assumptions are met. Before discussing the results, we verify that the GEV distributions used in our approach appropriately describe the observed block maxima from the time series. For this purpose, Figure 2.9 shows probability-probability plots (P-P plots) and quantile-quantile plots (Q-Q plots) that enable visual comparisons of the estimated GEV distributions G_i and the empirical distributions of the block maxima of the time series $\boldsymbol{X}_i$. When the model perfectly describes the observed data, all points in the plots reside on the diagonal lines.

We observe that the GEV distribution appears to be an appropriate model for the block maxima of all three time series. The goodness-of-fit is best on the `#stopislam` time series, and slightly worse on the other two time series. The Q-Q plot indicates that in all three cases the GEV distribution slightly underestimates the values of the highest quantiles. Note that we evaluate the GEV distributions only at the threshold locations to obtain estimates for the binomial success probabilities. The horizontal lines in the Q-Q plots show these threshold locations. It turns out that the vast majority of the thresholds lie within a region where the GEV distribution provides an appropriate fit to the observed data.

Results

QTR plots for the time series and event series under study are depicted in Figure 2.10, along with the Monte Carlo p-values obtained from the multiple threshold test. The plots also show the marginally expected trigger coincidence rates under the null hypothesis, and the marginal 95% percentiles to additionally assess *pointwise* exceedances of individual thresholds.

The analysis shows that Islamist terrorist attacks in Western Europe and North America *systematically* trigger bursts of anti-Muslim hate speech on Twitter (`#stopislam`, $\hat{p} = 0.0317$). 90% of Islamist terrorist attacks triggered an exceedance of the 0.85-quantile, and 60% of Islamist terrorist attacks even triggered an exceedance of the 0.95-quantile. Our results confirm the findings of previous quantitative studies [BWS+14; MDA15; OCB+18] with a novel statistical methodology and a larger study period.

On the other hand, our analysis *does not* provide evidence for a *systematic* association between Islamist terrorist attacks and peaks in jihadist hate speech (`kafir`, $\hat{p} = 0.2075$) or Muslim counter-hate speech (`#notinmyname`, $\hat{p} = 0.3561$) in the study period. We stress that individual terrorist attacks may still have triggered such a

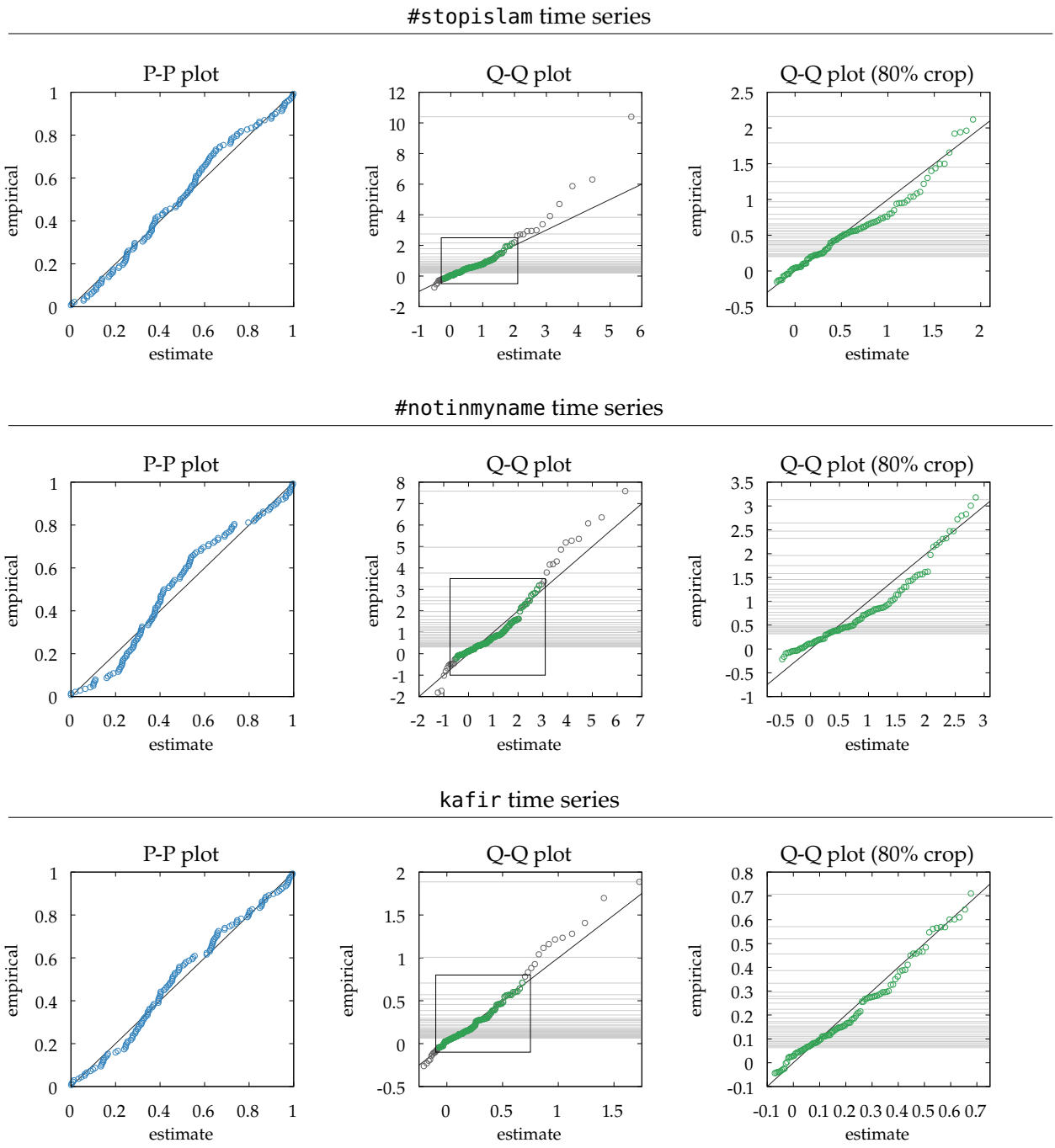

Figure 2.9: P-P plots and Q-Q plots to assess the goodness-of-fit of the GEV distribution to block maxima of the respective time series. The diagonal lines represent a perfect model fit. The right-most Q-Q plots are restricted to the empirical 0.1- to 0.9-quantiles, *i.e.*, the central 80% of the observed data. The horizontal lines in the Q-Q plots indicate the locations of the 32 thresholds used for the test.

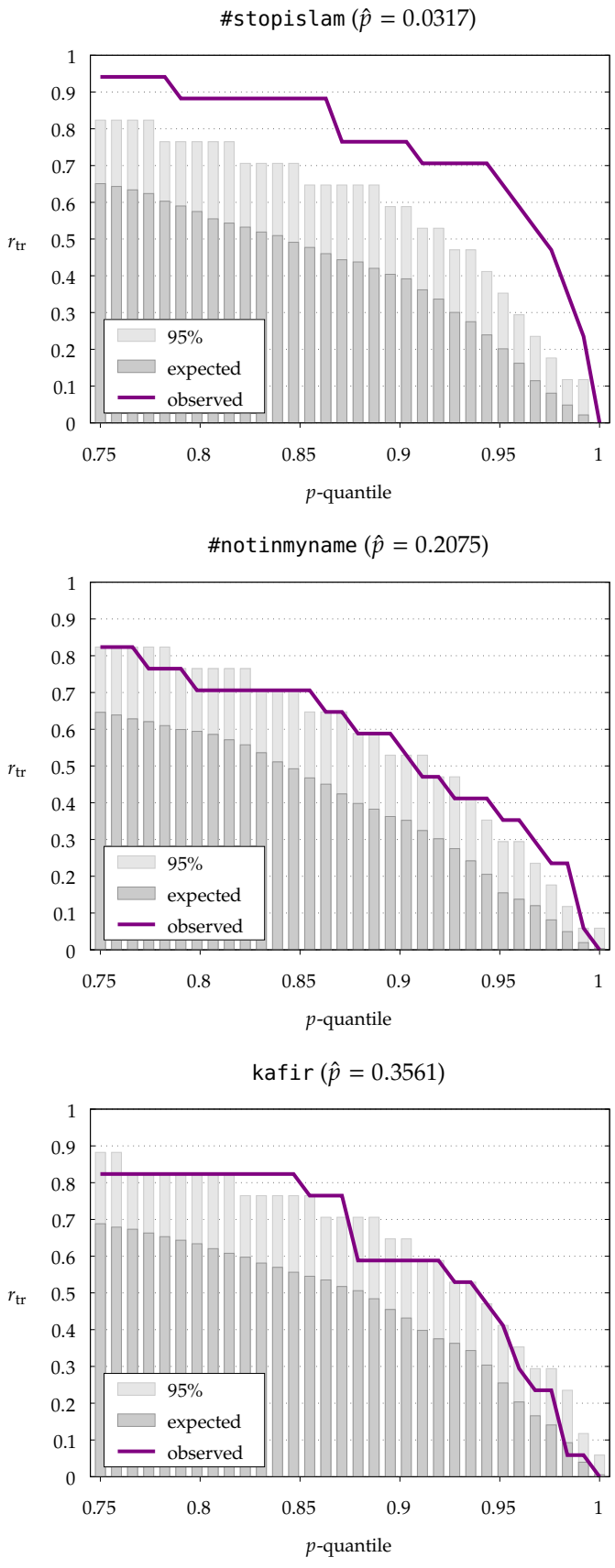

Figure 2.10: QTR plots for severe Islamist terrorist attacks and their online social media response.

social media response. Visual inspection of the data in Figure 2.8 suggests peaks in the hashtag `#notinmyname` for Islamist terrorist attacks before July 2016. Hashtag usage is typically subject to trends, so a systematic relationship can only be established for hashtags that are used consistently throughout the study period. The impact of an *individual* terrorist attack on the social media time series can be assessed, *e.g.*, with methods from Section 1.4.1.

Figure 2.10 shows that even for jihadist hate speech and Muslim counter-hate speech, the observed numbers of trigger coincidences fall above the *pointwise* 95% percentiles for some thresholds. Pointwise tests at these specific thresholds would reject the null hypothesis of independence at level $\alpha = 0.05$ on the basis of only a narrow perspective on the total association. The multiple threshold test thus decreases the dangers of data dredging at the cost of a lower sensitivity. To validate the results from the multiple threshold test, we computed all p-values for the pointwise tests at all thresholds and used different adjustment methods that control the family-wise error rate at level $\alpha = 0.05$: Bonferroni, single-step Šidák, step-down Holm in its original variant and in the Šidák variant [DL07]. For all multiple test adjustment methods, the results agree with our multiple threshold test.

Sensitivity analysis

To assess the stability of the results, we further experimented with different values of $\Delta = 4, \ldots, 16$ (ceteris paribus). We found that for $\Delta = 4, \ldots, 8$, the results of all tests on all time series are unchanged. For $\Delta = 9, \ldots, 14$, our multiple threshold test fails to reject the null hypothesis for the `#stopislam` time series, while the multiple pointwise test procedures still reject. For $\Delta = 15$, only the Šidák procedures reject the null hypothesis on `#stopislam`, while for $\Delta = 16$ no test procedure rejects the null hypotheses on any time series. Choosing a value of Δ that is longer than necessary thus reduces the sensitivity of the tests. We also varied the number of thresholds M between 8 and 64 (ceteris paribus), which did not change the outcome of any test. At last, we moved the thresholds upwards to more extreme levels by choosing equidistant p-quantiles from the ranges 0.85 to 1 and 0.95 to 1, respectively (ceteris paribus). The outcomes on the `#stopislam` and `kafir` time series remain unchanged, while our multiple threshold test now detects an additional trigger relationship for `#notinmyname` that is not detected by the multiple pointwise test procedures. Overall, the trigger relationship for `#stopislam` is very stable across all test procedures with different parametrizations, whereas the results on `#notinmyname` are inconclusive.

2.4 Conclusions

In this chapter, we have proposed a statistical methodology to study the association between event occurrences and peaks in a time series. For this purpose, we have formalized the notion of *event impacts in the maximum statistic* as a special case of event impacts in the sense of Definition 1.3.5. We showed that a test for event impacts in the maximum statistic can be implemented effectively within the framework of event coincidence analysis (ECA) by Donges et al. [DSS+16]. The major benefit of ECA is that it yields an interpretable measure of association between event occurrences and peaks in the time series: the *trigger coincidence rate*. We also proposed a novel visualization of this association via *quantile-trigger rate plots* (QTR plots). Technically, by using ECA, we avoid estimating the event-conditional distribution of the maximum statistic, which may be difficult in applications with few events. We validated our results with a simulation study and demonstrated the utility of our approach on a research question from the social sciences.

In our analytical derivations, we have used the extremal types theorem (ETT) and thus established a novel link between ECA and extreme value theory (EVT). We restricted our attention to the number of trigger coincidences for a leading event series and lagging peaks in a time series. For other research questions, one might be interested in the number of *precursor* coincidences in the same scenario, or in the *reverse* scenario with leading peaks in the time series and a lagging event series, or the association between peaks in *two time series*. We believe that EVT provides many more useful results that potentially improve our theoretical understanding of ECA applied on these problems. In particular, the threshold excess models based on the generalized Pareto distribution and the point process characterization of extremes [Col01] may fill some gaps in the theory of ECA that currently require practitioners to perform Monte Carlo simulations.

A potential downside of the approach described in this chapter is that the focus on peaks in the time series may not be suitable for all application scenarios. If we are interested in the association between event occurrences and other features of the time series, we must preprocess the time series with a function that transforms these features into peaks. In the next chapter, we discuss an approach for event impact analysis that does not require such transformations.

Multiple Two-Sample Testing | 3

This chapter is based on:

[SM20c] Erik Scharwächter and Emmanuel Müller. "Two-Sample Testing for Event Impacts in Time Series." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2020. DOI: 10.1137/1.9781611976236.2.

In the previous chapter, we studied the association between event occurrences and peaks in the time series. Here, we develop a statistical test to detect event impacts in a more general setting. This test is based on the observation that the joint independence relation that defines event impacts in Definition 1.3.5 can be split into multiple marginal independence relations. We propose to test the marginal independence relations with pairwise two-sample tests and combine the results with a multiple hypothesis testing approach. Our algorithm is highly computationally efficient and thus applicable to very long time series and event series. It requires only minimal regularity conditions that limit long-range dependencies similar to the regularity conditions seen in the previous chapter. Moreover, with a suitable two-sample test at hand, it can be applied to time series over arbitrary domains such as strings or graphs. We perform a large-scale simulation study with different types of event impacts to study the power and error rate of our multiple two-sample testing approach. Furthermore, we apply our test to analyze event impacts on household electricity meters in a smart home environment, and to analyze the impact of earthquake events on a Twitter time series.

3.1 Introduction

There are numerous ways in which a time series and an event series can be statistically associated. Some associations are easy to recognize by visual inspection of the data, others require advanced statistical methods to be uncovered. Figure 3.1 shows example pairs of event series and time series, where each pair is associated in a different way. In the simplest case, events lead to temporary changes of the mean of the time series, as illustrated in Figure 3.1 (first two rows). Every event occurrence induces the same pattern in the time series. The box plots on the right summarize the value distributions of the time series for $\delta = 0, \ldots, 15$ time steps after event occurrences. The box plots show that the means of the distributions fluctuate for a few time steps and then stabilize. However, events can have more subtle effects. In Figure 3.1 (third row), events temporarily increase the variance of the time series—as indicated by wider boxes and whiskers in the box plots. In Figure 3.1 (fourth row), events increase the risk of extreme observations from the tails of the distribution—as indicated by a larger number of outliers in the box plots.

Such visual analyses are limited to univariate numeric time series. If we consider multivariate numeric time series, or time series of graphs or strings, it is unclear how to proceed visually, and quantitative statistical methods are required.

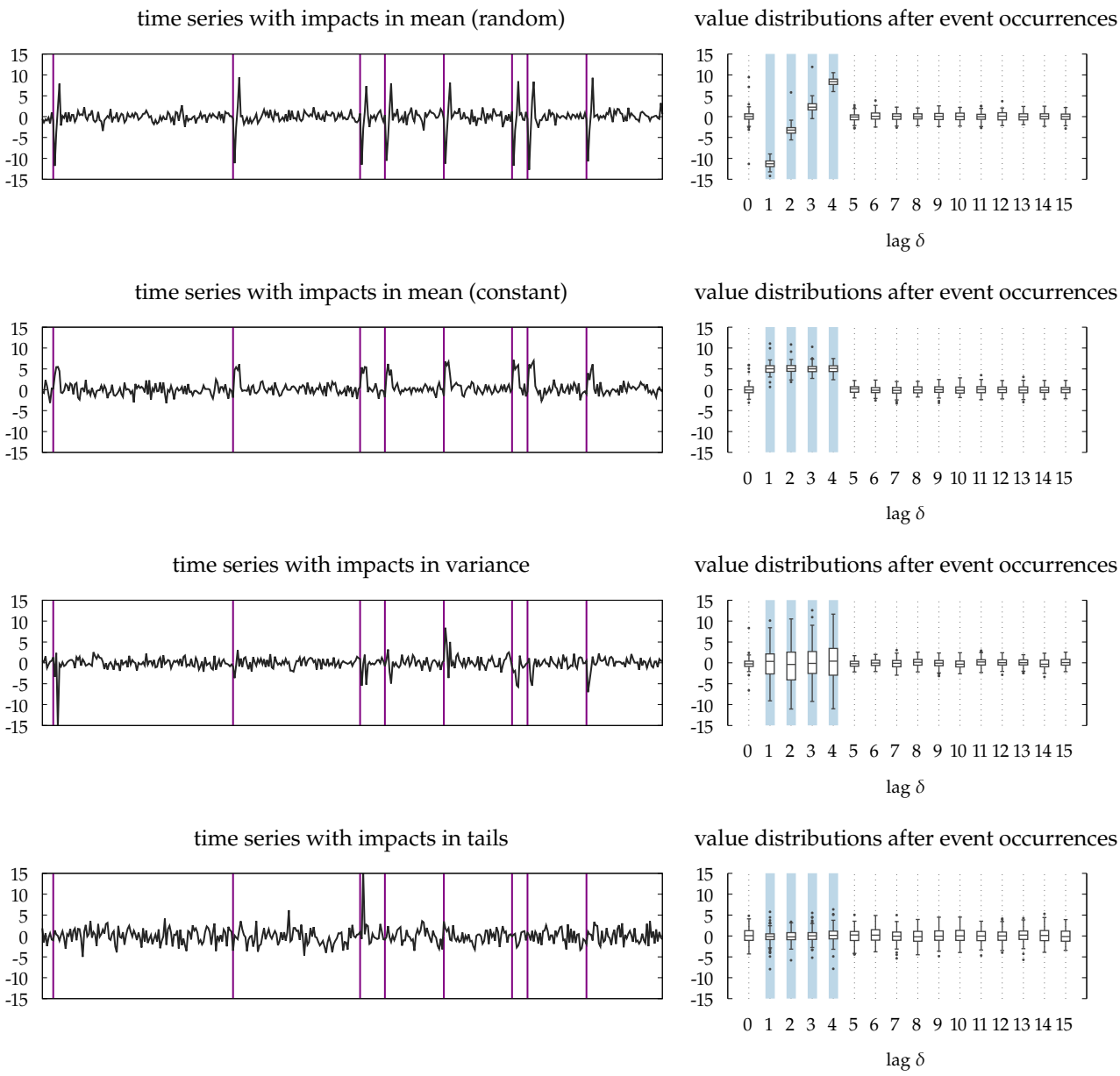

Figure 3.1: Different types of event impacts in a time series. The plots on the left show excerpts from the respective time series and event series; vertical lines indicate event occurrences. The box plots on the right summarize the observed value distributions at lags $\delta = 0, \ldots, 15$ after event occurrences, *i.e.*, all values $x_{t+\delta}$ where $e_t = 1$. The blue shading indicates lags with event impacts by construction. The models used to generate these examples are described in Section 3.3.1.

In Chapter 2, we argued that in many cases such event impacts can be reduced to peaks by preprocessing the time series with a suitable feature transformation function. In this chapter, we develop a more general test for event impacts that is directly applicable on a large variety of time series *without* requiring a feature transformation function. Formally, we test for event impacts by testing independence of the *marginal statistics* of the time series within the window of interest. In a nutshell, our algorithm tests whether the conditional distributions represented by the box plots in Figure 3.1 are all identical or not—without being restricted to univariate or even numeric time series. This is possible by leveraging recent advancements in kernel-based two-sample testing [GBR+12] that make our test applicable to time series from arbitrary domains augmented with a kernel function, including multivariate numeric, string or graph data, as in Figure 3.2.

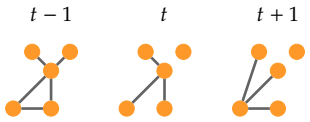

Figure 3.2: A time series of random graphs [SMD+16] can be analyzed with the approach developed in this chapter using a suitable graph kernel function [VSK+10].

With the focus on marginal independence, we restrict our attention to associations that affect at least one of the random variables within the window of interest marginally. Consequently, we lose the ability to detect event impacts that *exclusively* affect the *dependency structure* of the random variables within the window of interest. For example, two random variables within the window might be strongly correlated after an event occurrence, but uncorrelated when there was no event. This type of event impact cannot be detected by testing independence of the marginal statistics. In such a case, preprocessing with a feature transformation function that monitors the statistical associations of the random variables within the window of interest will still be required.

In summary, we make the following contributions:

- We show that marginal independences are an indicator for event impacts in the sense of Definition 1.3.5.
- We implement a simple and generic test for event impacts, under mild assumptions on the time series, via multiple pairwise two-sample tests of the conditional distributions at different lags after event occurrences.
- We propose realistic models for event impacts to analyze the performance of our test and two competing algorithms in a large-scale simulation study.
- We apply our approach to study event impacts in two real-world scenarios, including very long time series that could not be studied effectively before.

3.1.1 Problem statement

Let $\mathbf{X} = (x_1, \ldots, x_T)$ be a time series and $\mathbf{E} = (e_1, \ldots, e_T)$ be an event series. We assume that $\mathbf{X}$ and $\mathbf{E}$ are jointly stationary and that the event series is sparse, *i.e.*, $\Pr(e_t = 1) = \epsilon$ for a very small $\epsilon > 0$, so that $\sum_t e_t = N_{\mathbf{E}} \ll T$. On average, we have $L = \frac{1-\epsilon}{\epsilon} \gg 0$ time steps between two event occurrences.

Our goal is to detect event impacts in the sense of Definition 1.3.5. Same as in Chapter 2, we exploit Lemma 1.3.1 and focus on specific properties of the time series by testing the association between event occurrences and a statistic $g(x_t, \ldots, x_{t+\Delta})$. In this chapter, we address the association between event occurrences and the time series at *each individual lag* after an event occurrence. Formally, we use the **marginal statistics**

marginal statistics

$$g_\delta(x_t, \ldots, x_{t+\Delta}) := x_{t+\delta} \tag{3.1}$$

for all $\delta \in \{0, \ldots, \Delta\}$ and test the following hypotheses:

$$H_0 : \forall \delta \in \{0, \ldots, \Delta\}\ x_{t+\delta} \perp\!\!\!\perp e_t \tag{3.2}$$

versus

$$H_1 : \exists \delta \in \{0, \ldots, \Delta\}\ x_{t+\delta} \not\perp\!\!\!\perp e_t \tag{3.3}$$

If we have evidence to reject the null hypothesis H_0 in favor of the alternative hypothesis H_1, we have evidence for event impacts in the sense of Definition 1.3.5.

The pair of hypotheses from Equation 3.2 and Equation 3.3 can easily be generalized to time series over random vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ or arbitrary random elements $\mathbf{X} = (\xi_1, \ldots, \xi_T)$ like random graphs or random strings: The marginal statistics g_δ are always well-defined, and, as we discuss below, there are simple and effective ways to test their independence with two-sample tests.

3.1.2 Related work

Methodologically, our approach heavily relies on multiple two-sample testing. In **two-sample testing**, the problem is to decide whether two random samples come from the same probability distribution, or from different distributions. Some two-sample tests focus only on specific properties of the two samples. For example, *Student's two-sample t-test* [Stu08] and its variants [Hot31; Wel47] compare their means, while *F-tests* compare their variances [DS12]. In contrast, the *Kolmogorov-Smirnov* [Kol41; Was04] and the *Anderson-Darling* [Dar57; Pet76] two-sample tests compare the complete empirical distribution functions of univariate continuous samples. For multivariate continuous samples, nonparametric

two-sample testing

Two-sample tests based on specific properties of the samples are often more powerful for this specific property than generic tests on the full distribution, but—by design—fail to reject the null hypothesis if other properties of the samples diverge.

nearest-neighbor tests have been proposed [Hen88; Sch86]. In the case of categorical samples, a *test of homogeneity* based on the χ^2 distribution [DS12] can be used as a two-sample test. More recent works focus on testing means of high-dimensional random samples [CLX14; CQ10], testing random samples of graphs [GL18] and employing classifiers for two-sample testing [LO17]. In the past decade, *kernel-based approaches* to two-sample testing have been studied extensively [GBR+12; GBR+06; GHF+09; GSS+12; ZBG13]. They are applicable for arbitrary domains with a suitable kernel function, and have recently been combined with deep learning approaches [KKK+20; LXL+20]. At last, Ramdas, Trillos, and Cuturi [RTC17] provide an interesting unified perspective on several nonparametric two-sample tests, including kernel-based tests and the Kolmogorov-Smirnov test, via the *Wasserstein distance*.

3.2 Methodology

In the following, we provide a detailed exposition of the algorithm that we propose to test for marginal event impacts, *i.e.*, to decide between H_0 from Equation 3.2 and H_1 from Equation 3.3. We begin with a well-known, generic procedure for independence testing and continue with the necessary background to test independence of an event occurrence and multiple marginal statistics.

Marginal independence at a single lag. Independence of a mixed pair of random variables can be characterized by equality of all conditional cumulative distribution functions:

Theorem 3.2.1 (Theorem 15.11 of Wasserman [Was04]) *Let* x *and* k *be random variables, where* x *is continuous and* k *is discrete with outcomes* $1, \ldots, K$. *We have that* $\mathrm{x} \perp\!\!\!\perp \mathrm{k}$ *if and only if* $F_{\mathrm{x}|\mathrm{k}=1} = \ldots = F_{\mathrm{x}|\mathrm{k}=K}$.

For this reason, the individual independence relations between an event occurrence e_t and a marginal statistic $\mathrm{x}_{t+\delta}$ that appear in the null hypothesis H_0 from Equation 3.2 can be expressed equivalently via equality of the two conditional cdfs at lag δ,

$$\mathrm{x}_{t+\delta} \perp\!\!\!\perp \mathrm{e}_t \Leftrightarrow F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=0} = F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=1}. \tag{3.4}$$

This observation can be translated into a statistical test that operates on random samples from the two conditional cdfs at lag δ: If the statistical properties of the two random samples of $F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=0}$ and $F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=1}$ diverge significantly, the assumption that the two conditionals at lag δ are identical must be rejected. Since the marginal statistics are univariate, this test can be implemented easily with any of the two-sample tests listed in Section 3.1.2.

We observe that under the null hypothesis of marginal independence at lag δ, the conditional distributions from Equation 3.4 are identical to the marginal distribution of the time series:

$$F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=0} = F_{\mathrm{x}_{t+\delta}} = F_{\mathrm{x}_t} \tag{3.5}$$
$$F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=1} = F_{\mathrm{x}_{t+\delta}} = F_{\mathrm{x}_t}. \tag{3.6}$$

The first equality in each equation holds due to marginal independence, and the second equality in each equation holds due to the stationarity assumption.

Marginal independence at multiple lags. The complete null hypothesis H_0 from Equation 3.2 is composed of $\Delta + 1$ marginal independence relations in the form of Equation 3.4, one for each lag $\delta \in \{0, \ldots, \Delta\}$ within the window of interest. In this work, we propose to compare random samples *across different lags*. For this purpose, we exploit that under the null hypothesis of marginal independence at lags δ *and* δ', we have $F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=1} = F_{\mathrm{x}_t} = F_{\mathrm{x}_{t+\delta'}|\mathrm{e}_t=1}$. The null hypothesis H_0 from Equation 3.2 assumes marginal independence *at all lags* within the window of interest. Therefore,

$$H_0 \Rightarrow F_{\mathrm{x}_t|\mathrm{e}_t=1} = F_{\mathrm{x}_{t+1}|\mathrm{e}_t=1} = \ldots = F_{\mathrm{x}_{t+\Delta}|\mathrm{e}_t=1}. \tag{3.7}$$

This observation justifies our visual comparisons of the box plots from Figure 3.1 to assess event impacts.

As a result, we can test H_0 against H_1 by comparing random samples from the conditional distributions at different lags using *multiple* two-sample tests. If we reject H_0 due to diverging conditionals, we have evidence for event impacts in the sense of Definition 1.3.5.

It is important to note that the implication from Equation 3.7 does not hold in reverse, *i.e.*, the two statements are *not* equivalent. As a counterexample, consider the case where an event has exactly the same impact on all marginal statistics within the window of interest. In this case, the null hypothesis H_0 is violated, but the conditional distributions from Equation 3.7 will be identical. For this reason, the size of the window of interest used in the test should be larger than the duration of any potential event impact.

The biggest challenge when comparing random samples from the conditional distributions is sample construction. The reason is that we do not have access to *independent* random samples from the distributions as required for the two-sample tests. Instead, we must extract observations from the time series, which are potentially serially correlated or otherwise dependent. We discuss this challenge in Section 3.2.2 below. The fewer random samples we need for the tests, the better. However, the number of random samples should not be confused with the *size* of each random sample—the larger each random sample, the better for the tests.

3.2.1 Multiple test procedure

An overview of our test procedure is given in Algorithm 1. We call it **MEITEST** (Marginal Event Impact Test). This test is a modified version of the EITEST (Event Information Test) algorithm proposed in our previous work [SM20c] (Copyright © 2020 by SIAM). EITEST was derived from causation entropy [SB14], while MEITEST is derived here from event impact analysis in the sense of Definition 1.3.5. The difference between MEITEST and EITEST is the choice of conditional distributions in the two-sample tests.

MEITEST

The input is an observed time series $\boldsymbol{X} = (x_1, \ldots, x_T)$, an observed event series $\boldsymbol{E} = (e_1, \ldots, e_T)$ with $N = \sum e_t$ events, and the parameter Δ for the window of interest. The output of the algorithm is a p-value. If the p-value is smaller than the desired significance level α, we reject H_0 in favor of H_1. In line 3, random samples $\mathcal{T}_\delta$ from the conditional cdfs $F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=1}$ are constructed for all lags δ within the window of interest. In line 7, pairwise two-sample tests are performed for all of these random samples, where the output of each two-sample test is a p-value. In lines 11 and 12, the obtained p-values are adjusted for the multiple testing setting with Simes adjustments [DBW+10] to correctly control Type I errors. Details on sample construction and error control follow in Section 3.2.2 and Section 3.2.3.

The overall time complexity of the algorithm has the order

$$O(T + \Delta \cdot N + \Delta^2 \cdot \kappa(N) + \Delta \log \Delta), \tag{3.8}$$

where $\kappa(N)$ is the complexity of the underlying two-sample test. The first term in Equation 3.8 is incurred by extracting the time points of all N event occurrences from the event series. The second term comes from the construction of all $\Delta + 1$ random samples of size N by accessing the time series at the respective lags after event occurrences. The third term comes from the pairwise two-sample tests between all lags. Here, $\kappa(\cdot)$ is a function of N since all random samples $\mathcal{T}_\delta$ contain (at most) N observations. The last term is the cost of sorting the p-values for Simes adjustments.

Typically, we choose $\Delta \ll T$ for the window of interest, and we have $N \ll T$ due to the sparsity of the event series. The running time of the algorithm is thus dominated by a term that is linear in T, which makes MEITEST highly computationally efficient and applicable for very long time series and event series.

Algorithm 1: MEITEST

1 **for** $\delta = 0, \ldots, \Delta$ **do**
2 $\quad \mathcal{T}_\delta := \{x_{t+\delta} \mid e_t = 1\}$ with potential sparsification ;
3 **for** $i = 0, \ldots, \Delta - 1$ **do**
4 $\quad$ **for** $j = i + 1, \ldots, \Delta$ **do**
5 $\quad\quad p_{ij} :=$ TwoSampleTest$(\mathcal{T}_i, \mathcal{T}_j)$ with potential dissociation ;
6 $M := \Delta \cdot (\Delta + 1)/2$;
7 $p_{(1)}, \ldots, p_{(M)} :=$ SortIncreasing$(\{p_{ij} \mid i < j\})$;
8 **return** $\min_m \left\{ \frac{M}{m} \cdot p_{(m)} \right\}$;

3.2.2 Sample construction

The two-sample tests used in MEITEST require random samples from the conditional cdfs $F_{x_{t+\delta}|e_t=1}$ at all lags $\delta \in \{0, \ldots, \Delta\}$ within the window of interest. Principally, any observation $x_{t+\delta}$ from the time series is a realized value from the conditional cdf at lag δ, when $e_t = 1$ in the event series. In other words, all of these values from the time series are *identically distributed* with this conditional cdf. However, as noted earlier, these observations may be statistically associated due to serial dependencies within the time series—they are not, in general, *independent*. The observations within a random sample, however, must be identically distributed *and* independent for the statistical guarantees of the two-sample tests to apply. We need a regularity condition similar to the conditions in Chapter 2 to enforce that the observations within a random sample constructed from the time series are at least *approximately independent*.

Formally, we assume limited *marginal* long-range dependencies in the time series. We require that there exists some L such that for every subset of time points $\{t_1, \ldots, t_N\}$ with $|t_i - t_j| > L$ for all $i \neq j$, for $\delta \geq 0$, and for all outcomes $x_1, \ldots, x_N \in \mathbb{R}$, we have

$$\Pr\Big(\bigwedge_i \mathrm{x}_{t_i+\delta} < x_i \mid \bigwedge_i \mathrm{e}_{t_i} = 1 \Big) \approx \prod_i \Pr(\mathrm{x}_{t_i+\delta} < x_i \mid \mathrm{e}_{t_i} = 1). \quad (3.9)$$

If this regularity condition holds, and we assume that the event series $\boldsymbol{E}$ is so sparse that the N events are separated by more than L time steps, the set of observations $\mathcal{T}_\delta := \{x_{t+\delta} \mid e_t = 1\}$ approximates a random sample of $F_{x_{t+\delta}|e_t=1}$. If these assumptions do not hold, the sets constructed in line 3 of Algorithm 1 cannot be viewed as random samples. However, in the case where the regularity condition is fulfilled, but the event series is too dense, the sets can be **sparsified** by dropping observations that are too close to be seen as approximately independent.

sparsified

The regularity condition above and the potential sparsification are necessary to make the set $\mathcal{T}_\delta$ a random sample of $F_{\mathrm{x}_{t+\delta}|\mathrm{e}_t=1}$. However, a two-sample test for random samples $\mathcal{T}_\delta$ and $\mathcal{T}_{\delta'}$ as in line 7 of Algorithm 1 also requires the random samples to be independent *across* each other. In time series with serial dependencies, this is not the case. For example, if $\boldsymbol{X}$ is the realization of a first-order autoregressive process, the two observations x_t and x_{t+1} are correlated for all t. If there is an event at time step t, the sample construction scheme from above will place x_t in the set $\mathcal{T}_0$ and x_{t+1} in the set $\mathcal{T}_1$, thereby breaking the independence assumption across random samples. However, the two random samples can be **dissociated** *ad hoc* before performing the two-sample test, by passing observations from alternating event occurrences to the test. In the example above, x_t would be retained in $\mathcal{T}_0$, but x_{t+1} would not be retained in $\mathcal{T}_1$. For the next event occurrence at time step t', the observation $x_{t'}$ would not be retained in $\mathcal{T}_0$, but the observation $x_{t'+1}$ would be retained in $\mathcal{T}_1$, and so on, alternately.

dissociated

With the *ad hoc* dissociation procedure outlined above, the effective sample size used within each two-sample test is reduced by half. The procedure makes sure that the two random samples passed to the two-sample test are independent, but it does not make all random samples at all lags *pairwise independent*. To achieve pairwise independence, it would be necessary to retain only every $(\Delta + 1)$-st time step from each random sample. This reduces the effective sample size by a factor of $\Delta + 1$, which can drastically decrease the power of the two-sample tests for event series with relatively few event occurrences. If enough data is available, this approach would be a viable alternative to *ad hoc* dissociation.

3.2.3 Error control

In statistical hypothesis tests, the **false positive rate** (Type I error) is controlled at significance level α by rejecting the null hypothesis if and only if the p-value returned by the test is smaller than α. In standard statistical hypothesis testing problems (no multiple testing), the p-value is directly computed from a test statistic that collects evidence against the null hypothesis. The p-value in a standard statistical test is simply the probability of obtaining a test statistic value at least as extreme as the observed one, under the assumption that the null hypothesis is true.

false positive rate

For example, the number of trigger coincidences k_{tr} introduced in Chapter 2 was used as a test statistic to collect evidence against the null hypothesis of extremal independence from Equation 2.2.

When performing multiple two-sample tests, we have multiple null hypotheses, and we obtain multiple p-values: one for every two-sample test. Formally, we have the individual null hypotheses

$$H_0^{ij} : F_{\mathrm{x}_{t+i}|\mathrm{e}_t=1} = F_{\mathrm{x}_{t+j}|\mathrm{e}_t=1} \tag{3.10}$$

for all $0 \leq i < j \leq \Delta$, with alternative hypotheses

$$H_1^{ij} : F_{\mathrm{x}_{t+i}|\mathrm{e}_t=1} \neq F_{\mathrm{x}_{t+j}|\mathrm{e}_t=1}. \tag{3.11}$$

Under the *complete* null hypothesis H_0 from Equation 3.2 all of the individual null hypotheses are *simultaneously* true. If *any* of the individual null hypotheses H_0^{ij} is rejected in favor of H_1^{ij} by the respective two-sample test, the complete null hypothesis H_0 must be rejected in favor of the complete alternative hypothesis H_1 from Equation 3.3. The challenge of multiple hypothesis testing is that we have to control the false positive rate of the *complete* test at the significance level α—not the false positive rates of the individual tests. We cannot reject the individual null hypotheses if their respective p-values are smaller than α, since this approach would not control the false positive rate of the complete test.

There are several ways to define a false positive rate for multiple hypothesis testing [DBW+10]. In our case, we do not care *which* of the individual null hypotheses is false. In this scenario, the

family-wise error rate

family-wise error rate (FWER) is a suitable choice. Formally, let $\mathcal{H}_0 = \{H_0^{ij} \mid 0 \leq i < j \leq \Delta\}$ be the family of null hypotheses that we want to test, $\mathcal{T} \subseteq \mathcal{H}_0$ be the set of *true* null hypotheses and $\mathcal{R} \subseteq \mathcal{H}_0$ be the set of null hypotheses *rejected* by some procedure. The FWER is the probability that at least one of the true null hypotheses is rejected, *i.e.*, $\Pr(\mathcal{T} \cap \mathcal{R} \neq \emptyset)$ [DL07]. We use **Simes adjustments** [DBW+10] to guarantee $\Pr(\mathcal{T} \cap \mathcal{R} \neq \emptyset) < \alpha$ at the desired significance level α. Let $M := |\mathcal{H}_0| = \Delta \cdot (\Delta + 1)/2$ be the total number of pairwise two-sample tests, and $p_{(1)}, \ldots, p_{(M)}$ be the p-values returned by the tests, ordered increasingly. We reject the complete null hypothesis H_0 if $p_{(m)} < \frac{m}{M}\alpha$ for *any* $m = 1, \ldots, M$. The corresponding adjusted p-value for the complete test is then obtained from the individual p-values via $p := \min_m\{\frac{M}{m}p_{(m)}\}$.

Simes adjustments

3.3 Experiments

We evaluate MEITEST against the standard Granger causality test based on linear predictive models (GC) [Gra69] and a nonparametric test for non-zero transfer entropy (TE) [Sch00]. We perform a large-scale simulation study, where we assess the performance of all approaches on coupled pairs of time series and event series, generated by different event impact models. We also generate uncoupled pairs by randomly permuting the event series after generating a coupled pair. To assess the detection performance, we report true positive rates and false positive rates. At last, we demonstrate the utility of our test with two real-life applications.

Evaluation measures. A true positive is a coupled pair of time series and event series, generated by any of the event impact models described below, that is correctly detected as being coupled. A false positive is an uncoupled pair that is falsely detected as being coupled. The corresponding **true positive rate** (TPR, power) and **false positive rate** (FPR) are obtained by normalizing over the total number of coupled and uncoupled pairs, respectively. TPR should ideally be close to 1, whereas the FPR should be upper bounded by the significance level α that was chosen for the test.

true positive rate

false positive rate

Setup. We set the significance level to $\alpha = 0.05$. In MEITEST, we use a window of interest with a maximum lag of $\Delta = 32$. We report results with the Kolmogorov-Smirnov (KS) two-sample test [Was04], and the Maximum Mean Discrepancy (MMD) two-sample test [GBR+12] with the default RBF kernel with median heuristic and the gamma approximation to the null distribution. Furthermore, we report results based on Welch's two-sample t-test (TT) [Wel47] that is sensitive only for differences in the means of the two samples. For GC, we use a history of length 32 for consistency with the window of interest employed in MEITEST. In contrast, for TE, we use a history of length 1 only, since larger histories required significantly more running time (from a few hours to more than two weeks) and actually lowered the performance of TE; possibly due to estimation issues of the high-dimensional conditional distributions. For a fair comparison across all algorithms, we parametrize all event impact models such that events have impacts at lag 1.

We implemented MEITEST and the underlying two-sample tests in Python. The source code can be found on `https://github.com/diozaka/eitest`. For the experiments with GC, we used the R package `lmtest`. For TE, we used the R package `RTransferEntropy`.

`https://cran.r-project.org/package=lmtest`

`https://cran.r-project.org/package=RTransferEntropy`

3.3.1 Simulations

We first describe the four event impact models used for evaluation and then report the performances of all approaches. In the first two models, events have impact on the mean of the time series, in the third they modulate its variance, while in the fourth they alter the tails of its distribution. Examples from the models are illustrated in Figure 3.1. In all experiments, we first generate an event series of length T with N event occurrences by sampling N time steps $t_1, ..., t_N$ without replacement and setting $e_{t_n} = 1$ for these time steps. Then, we sample a time series given the event series using the models described below. All models below induce finite event impacts, and the parameter q determines their **order**, *i.e.*, their temporal duration. The parameter r in the models controls the difficulty of the detection problem in different ways.

order

Impacts in mean. We modulate the mean of the time series by a moving average model [Ham94] of order $q \in \mathbb{N}$ that uses the event series as innovations, with additive noise:

$$x_t = \sum_{j=1}^{q} \phi_j e_{t-j} + z_t. \tag{3.12}$$

The weights $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q) \in \mathbb{R}^q$ determine the shape of the event impacts and $z_t \overset{\text{iid}}{\sim} \text{Normal}(0, 1)$ is normally distributed. In this model, every event has the same deterministic impact on the time series and overlapping impacts simply add up. In the *random mean model*, we sample $\boldsymbol{\phi}$—for each coupled pair individually—from the isotropic normal distribution $\text{Normal}(\mathbf{0}, r_m \cdot \boldsymbol{I})$, where r_m is the signal-to-noise ratio between event impacts and noise term. In the *constant mean model*, we set $\boldsymbol{\phi} = (r_m, \ldots, r_m) \in \mathbb{R}^q$ so that events lead to a constant mean shift over q time steps. In both cases, larger values of r_m make the detection problem easier.

Impacts in variance. We modulate the variance of the time series by sampling from a normal distribution with variance that depends on whether there was an event within the last $q \in \mathbb{N}$ time steps:

$$x_t \mid \max_{q'=1,\ldots,q} e_{t-q'} = 1 \sim \text{Normal}(0, r_v) \tag{3.13}$$

$$x_t \mid \max_{q'=1,\ldots,q} e_{t-q'} = 0 \sim \text{Normal}(0, 1) \tag{3.14}$$

The more r_v deviates from 1, the stronger the event impacts, and the easier the detection problem. By construction, this event impact model alters *only* the variance of the distribution, and no other properties. In particular, the mean remains unchanged.

Impacts in tails. We modulate the tail behavior of the time series by sampling either from a normal distribution (light tails) or from Student's t-distribution (heavy tails), depending on whether there was an event occurrence within the last $q \in \mathbb{N}$ time steps:

$$x_t \mid \max_{q'=1,\ldots,q} e_{t-q'} = 1 \sim \text{Student-t}(r_t) \tag{3.15}$$

$$x_t \mid \max_{q'=1,\ldots,q} e_{t-q'} = 0 \sim \text{Normal}\left(0, \frac{r_t}{r_t - 2}\right) \tag{3.16}$$

The parameter $r_t > 0$ specifies the degrees of freedom for Student's t-distribution. A random variable $z \sim \text{Student-t}(r_t)$ with $r_t > 2$ has $E[z] = 0$ and $\text{Var}[z] = \frac{r_t}{r_t - 2}$. Therefore, our model for event impacts in tails does not alter the mean or variance of the time series when $r_t > 2$. In the tail impact model, events increase the risk of extremely large or small observations in the time series.

The tail impact model simulates the associations analyzed with ECA in Chapter 2 in a systematic way.

For $r_t \longrightarrow \infty$, Student's t-distribution approximates a normal distribution, and the detection problem will be harder. Detection of event impacts is easiest when $|r_t - 2| \longrightarrow 0$.

Benchmark and results. Our default parametrization for the event series is $T = 8192$, with $N = 64$ events in case of the mean and variance impact models, and $N = 512$ for the tail impact model. We need more events in the tail impact model since extreme values are rare even in a heavy-tailed distribution. The default impact order is $q = 4$ in all models. In the random mean model, we use the default signal-to-noise ratio $r_m = 1$; in the constant mean model, we use a default level shift of $r_m = 0.5$. In the variance model, the default variance is $r_v = 8$. For the tail impact model, we set the default degrees of freedom to $r_t = 3$. We change the detection difficulty by varying all parameters from these default values. For every parametrization, we generate 100 pairs of coupled event series and time series and 100 uncoupled pairs.

Figure 3.3 shows the true positive rates of all competing tests; the corresponding false positive rates are depicted in Figure 3.4. MEITEST outperforms or is on par with all competitors almost across the whole model space that we explore. The TT and KS variants of MEITEST slightly outperform the MMD variant on impacts on mean, but MMD drastically outperforms the other variants on impacts in variance and tails. Clearly, the TT variant cannot detect event impacts that do not alter the mean of the time series. These results suggest that a two-sample test for the mean or other specific properties of the random samples that we assume to be affected by event occurrences provides higher detection rates than generic tests if this assumption is met. Without prior knowledge of the nature of the event impacts, the MMD test should be favored over the KS test.

As expected, MEITEST fails to detect identical event impacts if the impact order matches the size of the window of interest, *i.e.*, if $q \approx \Delta$. At last, we observe that all tests approximately control the false positive rates at the desired significance level $\alpha = 0.05$. There is a slight tendency of MMD to overreject, *i.e.*, its false positive rates are slightly larger than α. Since we do not observe this behavior in KS, we suspect this behavior is due to the gamma approximation to the null distribution of the MMD test statistic.

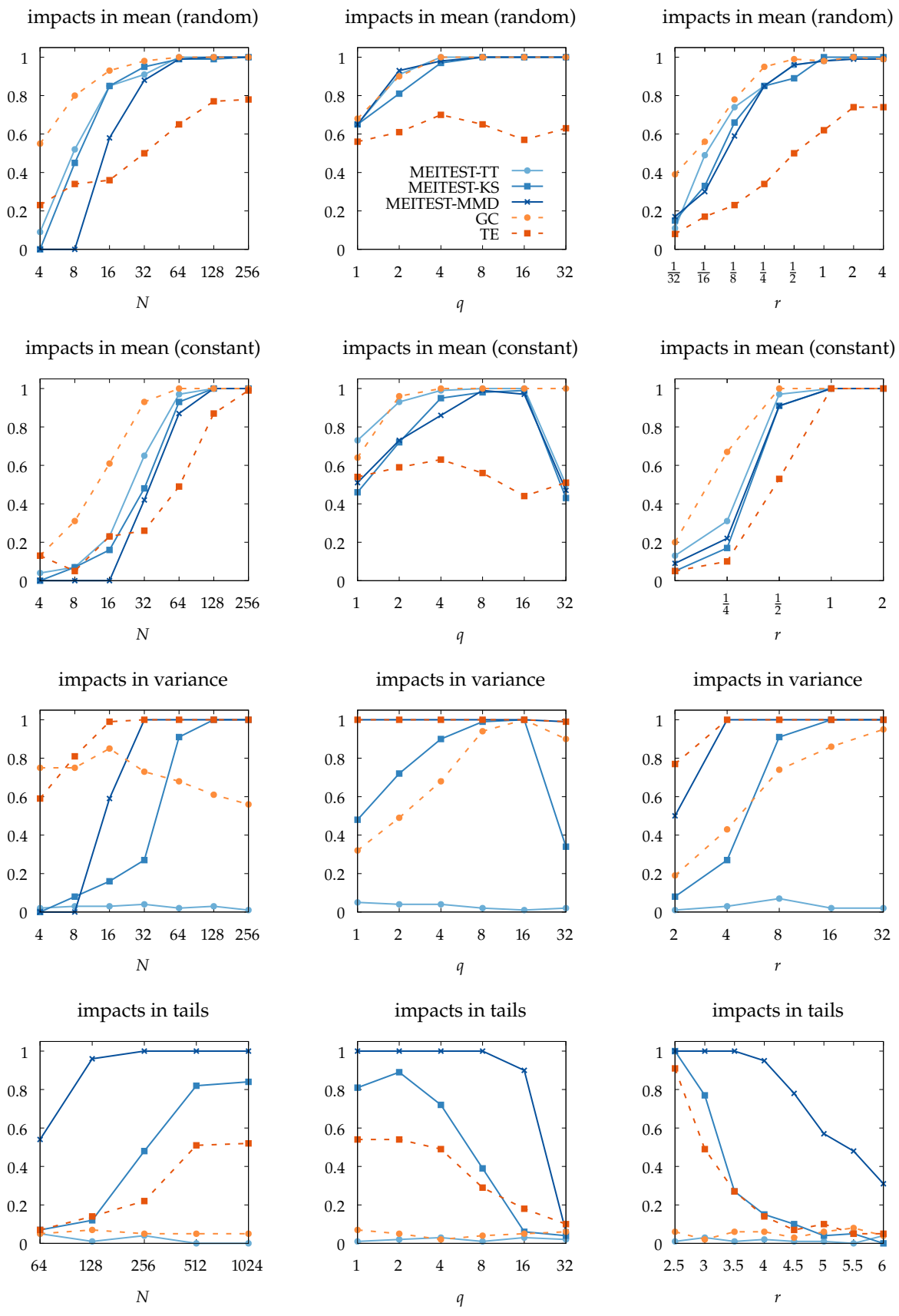

Figure 3.3: TPR of MEITEST, Granger causality (GC) and transfer entropy (TE) under different event impact models.

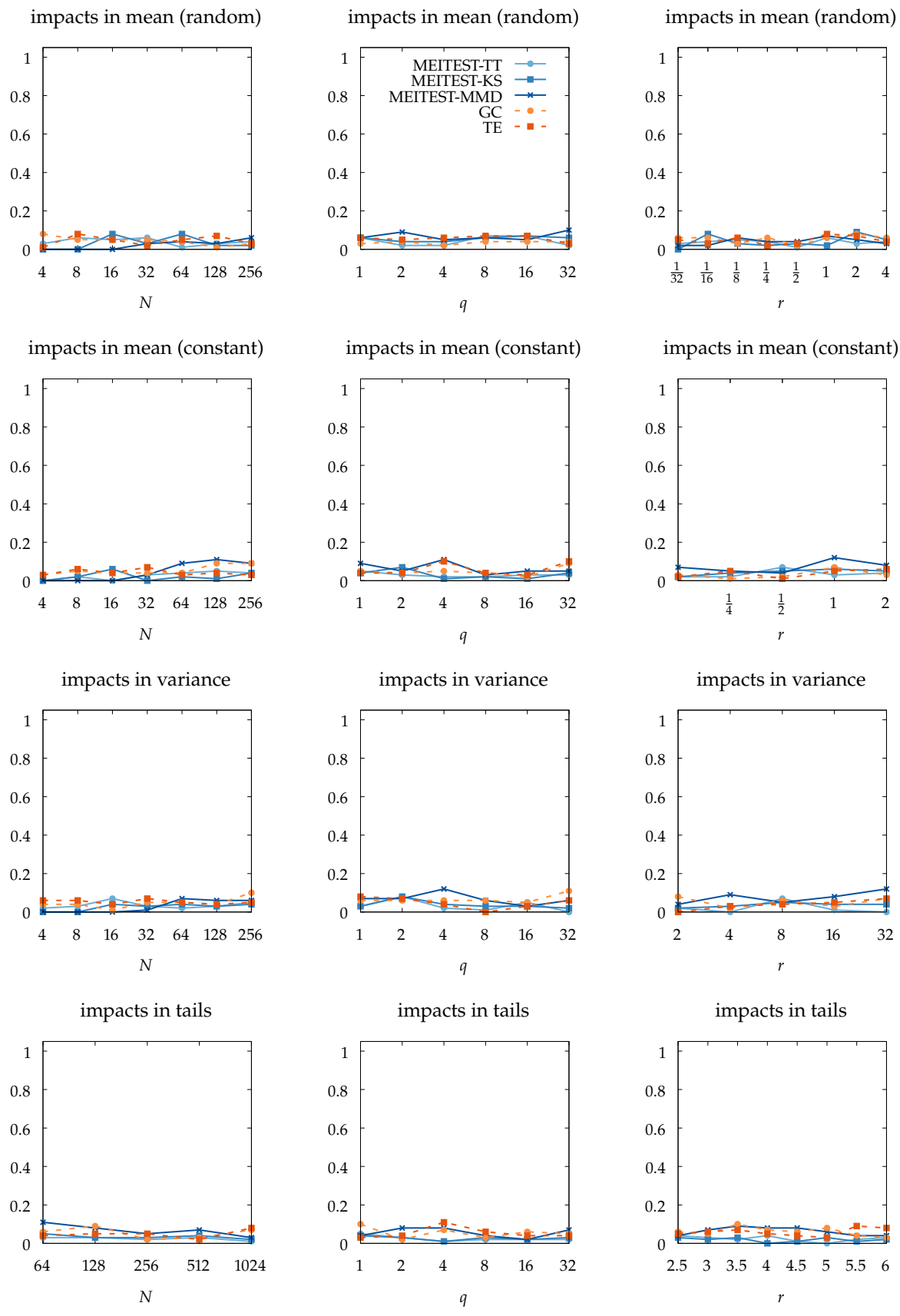

Figure 3.4: FPR of MEITEST, Granger causality (GC) and transfer entropy (TE) under different event impact models.

3.3.2 Electricity monitoring

In our first application, we demonstrate the utility of MEITEST for household electricity monitoring in a smart home environment. Specifically, we analyze the effect of turning on the clothes washer on various electricity meters in a residential house.

Data. For the experiment, we use the publicly available Almanac of Minutely Power dataset (AMPds) [MEB+16]. The dataset contains two years of minutely electricity, water and natural gas measurements from a residential house in Canada. We focus on electricity consumption, which was recorded using 21 physical meters placed at various locations in the building to separately measure the consumption of different household appliances (clothes washer, clothes dryer, dishwasher, *etc.*), rooms (bedroom, home office, garage, *etc.*), and the whole house consumption. Each time series contains 1,051,200 measurements. We extract 413 clothes washing events from the clothes washer electricity (CWE) meter. We are thus dealing with a very long time series and a very sparse event series where $\Pr(e_t = 1) \approx 0.0004$, or approximately $L \approx 2{,}500$ time steps between two event occurrences. An excerpt of the resulting event series is depicted in Figure 3.5 along with the clothes washer meter (CWE, top) and the whole house meter (WHE, bottom) between April 4th, 2012 and April 7th, 2012. The different scales of the y-axes indicate the low signal to noise ratio of the clothes washer impacts within the whole house time series, which makes the detection problem hard.

Results. In all experiments, we set the maximum lag in the window of interest to $\Delta = 120$ minutes (2 hours). The p-values obtained on all meters are shown in Table 3.1. Results that are significant at level $\alpha = 0.05$ (unadjusted) are shaded. Since the time series are very long, neither GC nor TE terminated within one hour and had to be aborted. The MMD-based test rejects on all instances where the TT- and KS-based tests reject, and some more. This behavior indicates that the MMD variant of MEITEST is more powerful than the other variants for generic event impacts.

On the other hand, it could also be a consequence of the tendency of MMD to overreject. A conservative user would trust a result only if at least two variants of MEITEST agree.

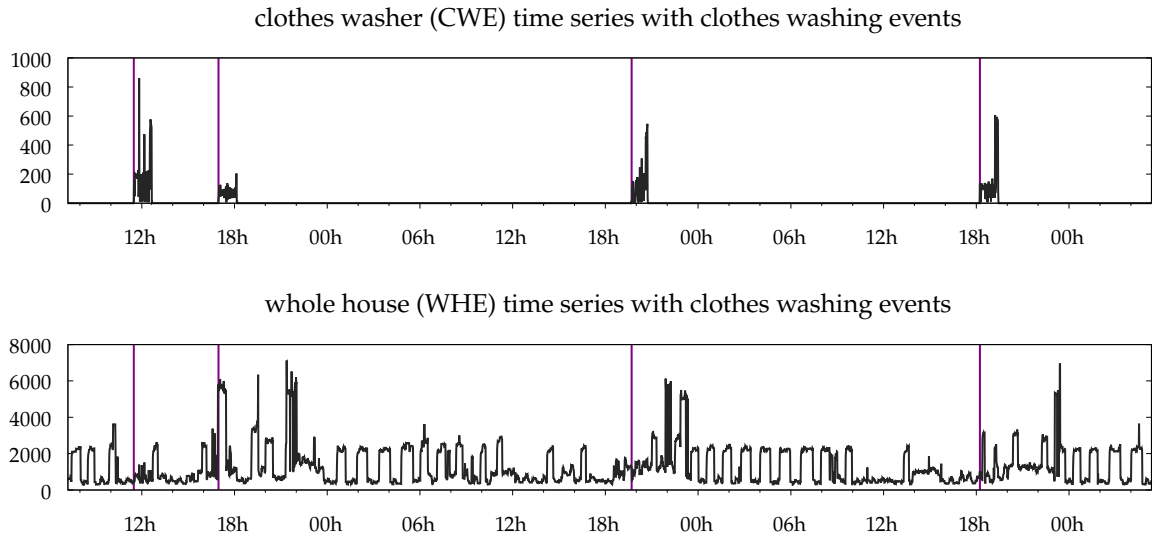

Figure 3.5: Clothes washer and whole house electricity consumption (excerpts); vertical lines are clothes washing events. We ignore the apparent periodicity in the WHE time series that, strictly speaking, makes it nonstationary.

meter	MEITEST-TT	MEITEST-KS	MEITEST-MMD
WHE	< 0.0001	< 0.0001	< 0.0001
RSE	0.9996	1.0000	0.9334
GRE	0.7376	1.0000	0.9718
MHE	< 0.0001	< 0.0001	< 0.0001
B1E	1.0000	1.0000	0.0012
BME	0.9998	0.6959	< 0.0001
CWE	< 0.0001	< 0.0001	< 0.0001
DWE	0.9998	1.0000	0.0341
EQE	0.9980	1.0000	0.9447
FRE	0.9998	1.0000	0.4338
HPE	0.2049	1.0000	0.0622
OFE	0.6952	1.0000	0.6240
UTE	1.0000	1.0000	0.0002
WOE	0.9999	1.0000	0.5589
B2E	< 0.0001	0.0045	< 0.0001
CDE	< 0.0001	< 0.0001	< 0.0001
DNE	1.0000	1.0000	0.3284
EBE	1.0000	1.0000	0.0068
FGE	0.9999	1.0000	0.9300
HTE	< 0.0001	< 0.0001	< 0.0001
OUE	< 0.0001	< 0.0001	< 0.0001
TVE	0.4342	1.0000	0.0004
UNE	0.0271	0.0270	< 0.0001

Table 3.1: AMPds p-values for all electricity meters as returned by MEITEST. The p-values are adjusted internally for multiple testing at all pairs of lags, but not externally across meters and two-sample tests. Shaded rows are significant at level $\alpha = 0.05$ for at least one of the employed two-sample tests.

Despite the low signal to noise ratio, all variants correctly identify an association between the clothes washer and the whole house meter (WHE). The tests also consistently identify associations in several other meters, *e.g.*, the clothes dryer meter (CDE). Since the time series are univariate, we can visualize the event impacts using box plots. These visualizations are shown in Figure 3.6. They nicely illustrate the diversity of the event impacts that can be detected with our approach. In particular, they illustrate that events do not always lead to peaks in the time series.

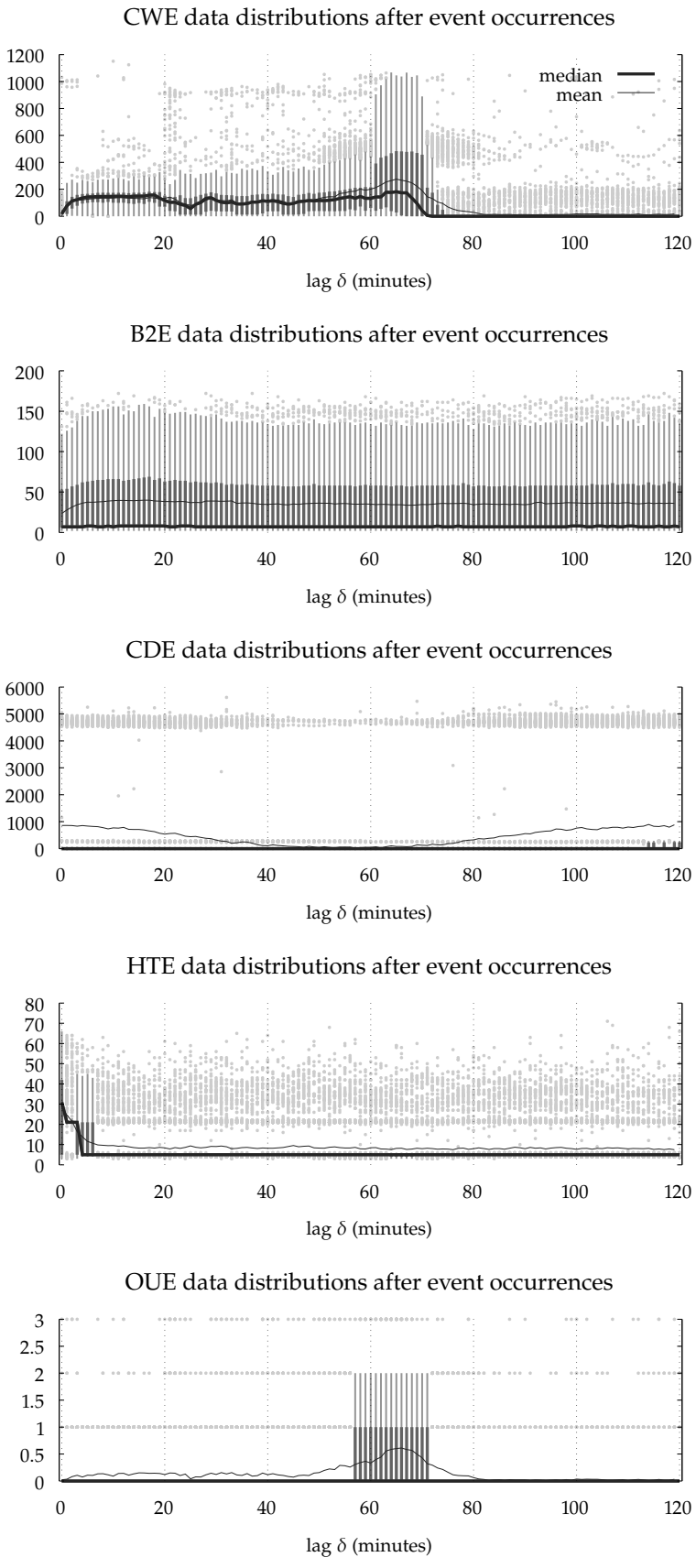

Figure 3.6: Box plots of the data distributions at lags $\delta = 0, \ldots, 120$ after event occurrences for several electricity meters, along with mean and median lines. Thick boxes mark the lower and upper quartiles of the distributions, thin boxes mark whiskers, and individual points are outliers.

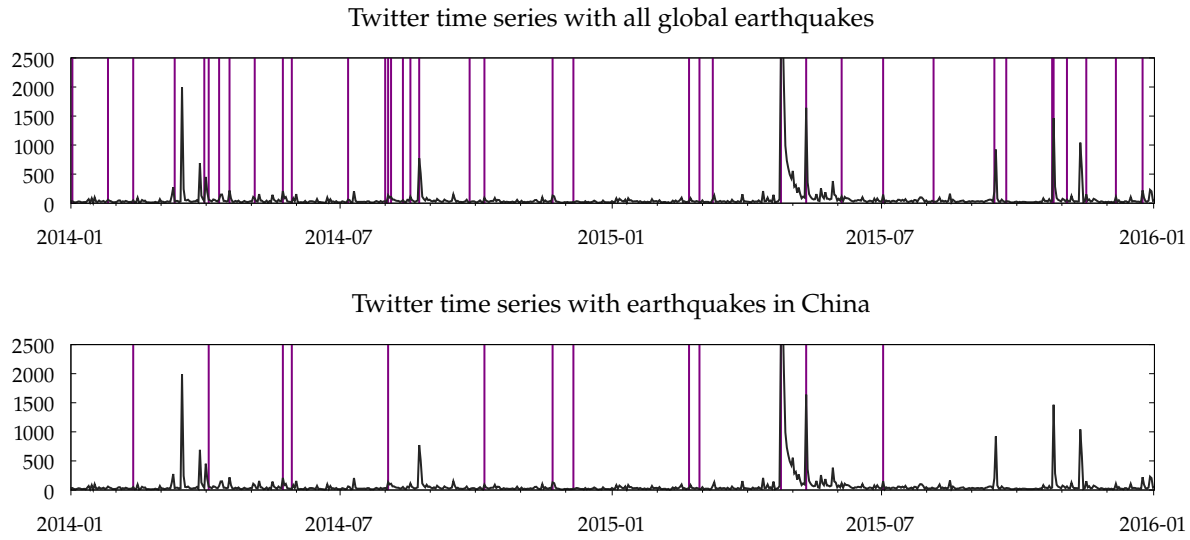

Figure 3.7: Volume of the keyword "earthquake" in German Twitter (excerpt), along with two earthquake event series.

3.3.3 Earthquakes on Twitter

At last, we analyze the coupling between earthquakes and German social media usage. Since social media reactions often come in bursts of posts, we expect that events temporarily fatten the tails of the time series. We first test whether daily usage of the keyword "earthquake" in German Twitter is influenced by the occurrence of severe earthquakes worldwide. We then focus specifically on earthquakes that hit China, the country with the largest number of disastrous earthquakes in the time period we study.

Data. We obtained time series of the daily number of tweets posted in Germany that contain the keyword `earthquake`, translated into more than 30 languages, between 2010 and 2017 (2,557 days), using the ForSight platform by Crimson Hexagon/Brandwatch. For the daily earthquake event series, we used the publicly available Emergency Events Database (EM-DAT) provided by the Centre for Research on the Epidemiology of Disasters (CRED) and extracted all severe earthquakes in the same time period. We created two event series: the first containing all earthquakes globally (162 events), the second containing only earthquakes in China (40 events). Excerpts from the two pairs are depicted in Figure 3.7.

`http://brandwatch.com/`

`http://emdat.be/`

Results. We set the maximum lag in the window of interest to $\Delta = 7$ days. According to all variants of MEITEST (TT, KS, and MMD), the event series with all global earthquakes is coupled with German Twitter activity: We obtain $p = 0.0069$ with TT, and $p < .0001$ with KS and MMD. This result matches the intuition

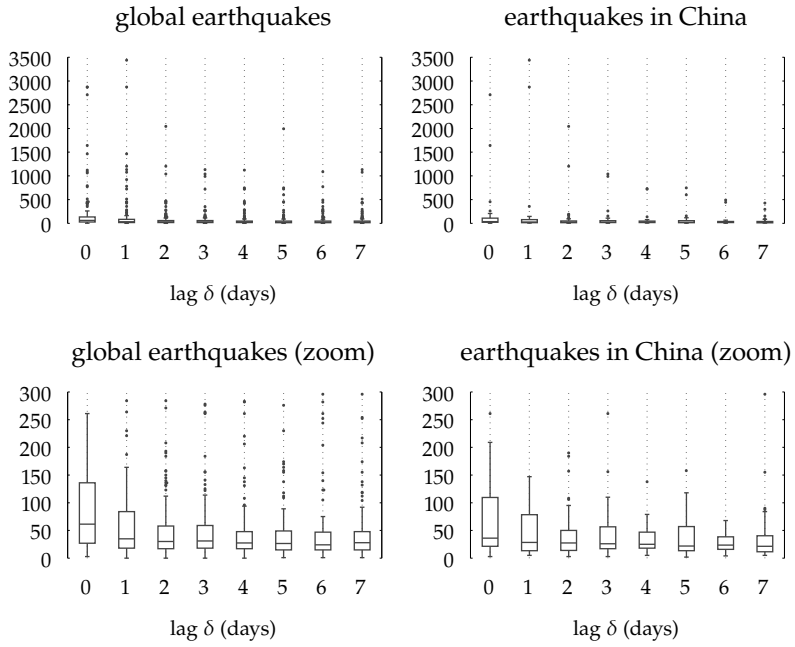

Figure 3.8: Box plots of the data distributions at lags $\delta = 0, \ldots, 7$ after event occurrences, for the Twitter time series and two earthquake event series (global and China).

that there should be an association between the series. GC does not detect an association ($p = 0.1919$). When it comes to the event series with earthquakes in China, the TT and KS variants of our test do not have enough evidence for a statistical association (TT: $p = 0.6154$, KS: $p = 0.6039$), whereas the MMD variant finds an association with $p = 0.0254$. GC now detects an association as well ($p = 0.0090$), and thus contradicts its earlier result. TE provides inconsistent results on both tasks: the test delivers largely fluctuating p-values when run repeatedly. Overall, the results on earthquakes in China are inconclusive. Visualizations of the conditional distributions for both event series can be found in Figure 3.8.

3.4 Conclusions

In this chapter, we have proposed a simple and versatile test for event impacts that is sensitive for violations of marginal independence at all lags within the window of interest. Our approach reduces the test for event impacts to a multiple two-sample testing problem, which makes it applicable to time series over arbitrary domains, as long as a two-sample test is available for this domain. The main challenge when applying our test in practice is the construction of valid random samples of the conditional distributions from time series with serial dependencies. We have worked out the assumptions on the time series and the event series that are required for effective sample construction. In a nutshell, the time series and the event series must satisfy a regularity condition that

limits the long-range dependencies between event occurrences and the marginal statistics of the time series. We proposed sparsification and an *ad hoc* dissociation scheme to further reduce the dependencies within and between random samples.

In our large-scale simulation study, we have shown that our test detects a large variety of event impacts with high sensitivities and low error rates. The test can be fine-tuned to event impacts that affect only specific properties of the time series (like its mean), by choosing a two-sample test that is sensitive only towards the respective properties of the random samples. Since our test focuses on the marginal statistics within the window of interest, it cannot—by design—detect event impacts that *only* affect the dependency structure of the random variables within the window of interest. In this case, a feature transformation that extracts information on the dependency structure is required before performing the test.

An alternative to the multiple two-sample testing approach proposed in this chapter is to perform a single K-sample test on all random samples together. The most prominent approach is the *analysis-of-variance* (ANOVA) procedure to test equality of all means [War14], and some tests for the complete empirical distribution functions exist as well [SS87]. The literature on K-sample testing is less active than for two-sample testing, but may enable tests for marginal event impacts with even higher power than our multiple two-sample testing approach.

The advantage of the marginal independence testing approach over the ECA-based approach from Chapter 2 is that it is applicable out-of-the-box for time series over different domains and for different types of event impacts. The disadvantage is that the detection of event impacts *per se* does not contribute to much understanding of the nature of these event impacts. The focus on peaks in Chapter 2 enabled using the trigger coincidence rate as an intuitive measure to quantify the degree of association between the event series and the time series. The choice of the two-sample test in this chapter can at least shed some light on the properties of the time series that are affected by event occurrences. In the following chapter, we propose an alternative way to study this association, by modeling the behavior of the time series within the window of interest with a versatile probabilistic model.

4 Time Warping Impact Models

This chapter is based on:

[SM22] Erik Scharwächter and Emmanuel Müller. "Discrete Probabilistic Models for Time Warping." In: *Manuscript in review*, 2022.

Once a statistical association between the event series and the time series is established with the tests from Chapter 2 or Chapter 3, we can proceed to describe the nature of this association in more detail. The trigger coincidence rates and QTR plots introduced in Section 2.2 are first steps in this direction, in that they measure and visualize the strength of the association in an interpretable way. However, they are useful only in the case where events trigger threshold exceedances in univariate time series. Box plots of the data distributions at individual lags within the window of interest as in Figure 3.1, Figure 3.6 and Figure 3.8 are another way of visualizing the impact of events in univariate time series, but they neglect information about serial dependencies between lags. In this chapter, we propose an alternative way to describe the impact of events in univariate time series using probabilistic models.

Formally, we model the *deviant behavior* of the time series, *i.e.*, the conditional distribution of the window of interest $(x_t, \dots, x_{t+\Delta})$ from the time series $\mathbf{X}$ given that $e_t = 1$ in the event series $\mathbf{E}$,

$$(x_t, \dots, x_{t+\Delta}) \mid e_t = 1 \sim F_t, \tag{4.1}$$

where F_t is the joint cdf at time step t that needs to be specified. Due to stationarity, we have that $F_t = F$ for all t. As in the previous chapters, we assume that the long-range dependencies in the time series are limited and that the event series is sparse, so that the windows of interest associated with different event occurrences are approximately independent.

Let $\boldsymbol{X}$ and $\boldsymbol{E}$ be (finite) realizations of the time series and the event series, respectively, and let t_i for $i = 1, \dots, N$ denote the time points of the $N = \sum_t e_t$ event occurrences in $\boldsymbol{E}$. We treat the subsequences $\boldsymbol{x}_i = (x_{t_i}, \dots, x_{t_i+\Delta})$ of length $T' = \Delta + 1$ as approximately iid realizations of a multivariate random sample $\mathbf{x}_i \overset{\text{iid}}{\sim} F$ for $i = 1, \dots, N$. In the following, we proceed with this simplified notation.

Our model family is based on the observation that event impacts often follow the same generic pattern across all event occurrences, but the pattern may come at different delays or be otherwise temporally distorted. This can be observed, for example, in the box plots for the CWE data distributions in Figure 3.6. The peaks in electricity consumption may occur somewhere between lag 45 and lag 80, but most of the time they occur between lags 60 and 70. An illustration of temporal distortion is given in Figure 4.1.

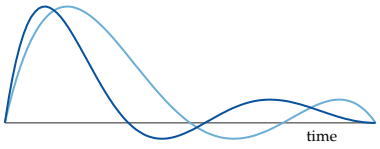

Figure 4.1: Two subsequences that follow the same generic pattern but are subject to temporal distortion.

A suitable probabilistic model for the subsequences must be able to temporally align the subsequences and identify the underlying prototypical pattern. This can be achieved by first applying an existing time warping algorithm [SC78] for alignment and then modeling the aligned subsequences. However, by separating the alignment step from the modeling step, we lose information on the stochastic mechanism that generates the temporal distortions. Instead, we propose a novel family of discrete-time probabilistic models with an embedded temporal distortion mechanism. Our model family learns a low-dimensional prototype for the underlying pattern that is stochastically stretched in time to produce a set of structurally similar but temporally misaligned sequences.

From now on, we refer to subsequences of the original time series simply as *sequences*. From a modeling perspective, it does not matter where these sequences come from.

In contrast to previous work, we place emphasis on different choices for the stochastic warping mechanism. While previous work is restricted by Markov assumptions, our model family allows employing a large variety of distributions for the warping mechanism. We instantiate our model family with Markov, multinomial and Dirichlet-compound multinomial warping distributions to demonstrate its modeling capacity, and provide a generic Monte Carlo Expectation-Maximization algorithm for inference. We empirically study various characteristics of these model instantiations and show that they yield state-of-the-art performance in structural averaging and preserve relevant features for classification. We developed our model family with applications in event impact analysis in mind, but it is useful whenever modeling short sequences that are subject to temporal distortions. Therefore, in the following, we provide a generic treatment of the model family without a special focus on event impact analysis.

4.1 Introduction

Invariance to time warping is a key feature of many state-of-the-art learning algorithms for sequences. The seminal dynamic time warping (DTW) distance was defined more than four decades ago [SC78] and its underlying principles have since spurred numerous adaptations and improvements. In a nutshell, time warping refers to a—possibly nonuniform—transformation of the time axis of the input. The large majority of works treat time warping as a *pairwise* problem: time warping distances like the DTW distance search for optimal alignments between two or more sequences by jointly transforming their time axes to maximize similarity (see Figure 4.2 for an example). Although very successful, these solutions are often *ad hoc* and heuristic in nature: they do not establish a deeper understanding of the stochastic mechanism that generates the *individual* warped sequences from a prototypical pattern.

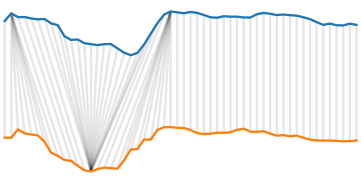

Figure 4.2: Optimal alignment between two short sequences according to the DTW algorithm.

In this work, we shed more light on the genesis of warped sequences by developing a family of generative models for time warping. The key idea of our model family is that a set of *structurally similar* but *temporally misaligned* sequences is generated by stochastically stretching a *low-dimensional prototype* in time. Models from our family thus jointly solve the problems of sequence alignment and low-dimensional approximation on the time axis. We observe that a nonuniform transformation of the time axis can be formulated as the product of a *random warping matrix* and the prototype vector. A member of our model family is defined by specifying a prior distribution over warping matrices. Our work generalizes and unifies existing discrete-time warping models based on monotone Markov chains or semi-Markov models [CGS03; GS00; HTR03; LNR+05]. We view the warping prior as a probabilistic alternative to the hard constraints typically employed to regularize DTW distances [RK04; SC78; THF+18], and to the deterministic warping functions in curve registration [KG92].

Our models for time warping are useful in many ways: (1) When trained on an observed set of structurally similar sequences, the prototype is recovered and provides a concise structural average of the dataset. (2) After training, the models can be used to align any input sequence to an arbitrary reference time axis without solving a new optimization problem. (3) For any input sequence, the models yield piecewise constant approximations to the observed data. (4) The models can be integrated into probabilistic learning architectures such as classification with Bayes classifiers or clustering via mixture modeling.

We focus on learning *low-dimensional* prototypes that retain and amplify relevant information for the downstream task. However, our work can easily be adapted to *higher-dimensional* prototypes, as in the Continuous Profile Model of Listgarten et al. [LNR+05]. Although we do not explicitly focus on periodic sequences or sequences that are only partially observed, the model family and MCEM algorithm also have the capacity to describe the generative process of such data with appropriate warping priors and corresponding Monte Carlo sampling schemes.

4.1.1 Related work

Literature on time warping is vast. It ranges from approaches that make classical DTW distance computation more efficient [KP00; PDM16], less constrained [AZ18; RK04], or differentiable [CB17], to approaches that focus on completely different dynamics in the input sequences [KP01; ZT09]. An important line of research addresses the extension of DTW from pairwise alignment towards joint alignment of multiple sequences [KMP19; WMP+16; ZT12].

Recently, recurrent neural networks have been shown to implement a form of time warping [TO18]. However, most related are works that incorporate time warping into statistical models. They can be divided into continuous time models and discrete time models.

Continuous-time models. *Curve registration* [KG92; RL98; Sil95] is a statistical formulation of time warping for functional data analysis, where the input data is a set of curves in continuous time [RS05]. Curve registration has received considerable attention in the statistics community in the past two decades [GS04; KR08; KSW11; LWT19; WK19; WG99; WED+19]. The basic idea is to estimate continuous monotonic time warping functions that minimize a misfit criterion over the input curves. The *deformable motifs* model [SDK07] and *congealing* [Lea06; MHL12; MRL09], are hybrid approaches that model temporal transformations for discrete-time input in a continuous-time space. In contrast, we focus on direct modeling of time warping in *discrete time*, which matches the DTW approach and can be formulated with methods from multivariate statistics.

Discrete-time models. Few methods were proposed to model time transformations in discrete time. In the simplest case, sequences are shifted by random global offsets [CGM+03]. An early method to model nonuniform transformations of discrete time builds on semi-Markov and segmental Markov models [GS00], but the authors only provide a heuristic training procedure. Several authors refined this idea and developed Markov models to capture nonuniform time transformations [CGS03; HTR03; KS06; KSL04; LNR+05]. The warping component of our model family is *more generic* and can be instantiated with a Markov prior—among others. Our model family thus overcomes the restriction to geometric duration distributions inherent in Markov models, and the restriction to independent state durations in semi-Markov models. A downside of this generality is that we cannot provide an efficient Baum-Welch algorithm, but need Monte Carlo methods for estimation.

Structural averaging. Many of the statistical models mentioned above provide a temporal alignment of the sequences that allows computing a structural average. There are other approaches for alignment and structural averaging of sequences that are not based on statistical models. Most of them directly optimize pairwise DTW distances [BFF+18; GMT+96; PKG11; WG97] or some variation thereof, such as *generalized time warping* [ZT12], *graphical time warping* [WMP+16], and *trainable time warping* [KMP19]. We include some of them as competitors in our evaluation.

4.2 Methodology

4.2.1 Discrete warping

prototype

Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_L) \in \mathbb{R}^L$ be a **prototype** of length L. We generate a warped sequence $\boldsymbol{x}$ of length $T' \geq L$ by stretching the prototype $\boldsymbol{\mu}$ in time. The prototype is stretched by repeating every entry μ_l exactly n_l times, where $n_l \geq 1$ is a positive integer that denotes the number of repetitions of the l-th entry of the prototype. The resulting sequence is piecewise constant and has length $T' = \sum n_l$. A transformation is **uniform** if $n_1 = ... = n_L$, and **nonuniform** if $n_l \neq n_{l'}$ for some l and l'. Nonuniform transformations lead to temporal distortion of the prototype.

uniform

nonuniform

Discrete time transformations, uniform or nonuniform, can be formulated as a *linear operation* on the prototype vector: Let $\boldsymbol{A} \in \{0,1\}^{T' \times L}$ be the discrete **warping matrix**

warping matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \Big\} n_1 \quad \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & 0 \end{pmatrix} \Big\} n_2 \quad \begin{pmatrix} & \ddots & & & \ddots & \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \Big\} n_L \tag{4.2}$$

The warped sequence $\boldsymbol{x}$ is given by the product

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\mu} = (\underbrace{\mu_1, ..., \mu_1}_{n_1}, \underbrace{\mu_2, ..., \mu_2}_{n_2}, ..., \underbrace{\mu_L, ..., \mu_L}_{n_L}) \in \mathbb{R}^{T'}. \tag{4.3}$$

The warping matrix $\boldsymbol{A}$ is fully determined by the vector $(n_1, ..., n_L)$. We can thus conveniently switch between the vector representation and the matrix representation of the warping operation, $\boldsymbol{A} \equiv (n_1, ..., n_L)$. The total number of $T' \times L$ warping matrices is given by the number of L-compositions of the integer T', and can be obtained from the binomial coefficient $\binom{T'-1}{L-1}$ [Sta11]. The inner product of a warping matrix $\boldsymbol{A}'\boldsymbol{A} = \text{diag}(\sum_t a_{t1}, ..., \sum_t a_{tL}) = \text{diag}(n_1, ..., n_L)$ has full rank since $n_l \geq 1$ for all l.

4.2.2 Probabilistic warping

For example, in applications in event impact analysis, we would model all subsequences within the windows of interest of size $T' = \Delta + 1$ after N event occurrences.

We model N random sequences $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of the same length T'. We assume that all of these sequences were generated by warping the same (fixed, but unknown) prototype $\boldsymbol{\mu}$ of length L to length T' with random warping matrices $\mathbf{A}_i$ and additive noise, *i.e.*, $\mathbf{x}_i = \mathbf{A}_i\boldsymbol{\mu} + \mathbf{z}_i$. We view the warping matrices $\mathbf{A}_i$ as iid random matrices with prior pmf $\Pr(\mathbf{A}_i = \boldsymbol{A}; \boldsymbol{\psi}) := p(\boldsymbol{A}; \boldsymbol{\psi})$ that depends on the parameter $\boldsymbol{\psi}$. Furthermore, we assume isotropic normal noise $\mathbf{z}_i$ with zero mean and variance σ^2, such that

$$\mathbf{x}_i \mid \mathbf{A}_i = \boldsymbol{A} \overset{\text{iid}}{\sim} \text{Normal}(\boldsymbol{A}\boldsymbol{\mu}, \sigma^2\boldsymbol{I}), \tag{4.4}$$

where $\boldsymbol{I}$ is the $T' \times T'$ identity matrix. The warping matrices are latent variables in our model. The marginal pdf of $\mathbf{x}_i$ thus decomposes into the mixture density

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\boldsymbol{A}} f_{\text{Normal}}(\boldsymbol{x}; \boldsymbol{A}\boldsymbol{\mu}, \sigma^2\boldsymbol{I}) \cdot p(\boldsymbol{A}; \boldsymbol{\psi}) \tag{4.5}$$

with parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \boldsymbol{\psi})$. Since we use an isotropic normal distribution for the noise term, the serial dependencies within the sequence are completely captured by the prior distribution of the warping matrices. The key modeling question is how to parametrize the prior pmf of the warping matrices $p(\boldsymbol{A}; \boldsymbol{\psi})$. Due to the equivalence of the representations, any discrete multivariate distribution over vectors of positive integers $\mathbf{n} = (\mathrm{n}_1, \ldots, \mathrm{n}_L)$ is a valid prior distribution for the random warping matrix $\mathbf{A}$. Since the length of the observed sequences is fixed to T', the support of the prior is constrained to vectors of positive integers $\boldsymbol{n} = (n_1, \ldots, n_L)$ that sum to T'. In the following, we demonstrate how to instantiate this model family with Markovian, multinomial, and Dirichlet-compound multinomial warping priors.

Markovian time warping

In the Markovian time warping model, the numbers of repetitions $\mathrm{n}_1, \ldots, \mathrm{n}_L$ for each entry in the prototype are obtained from a monotonic Markov chain that sequentially traverses L states over T' time steps and stays in each state l for exactly $\mathrm{n}_l \geq 1$ time steps. Let $\phi_l \in (0, 1]$ for $l = 1, \ldots, L-1$ be the probability to switch from state l to state $l+1$, and $1 - \phi_l$ be the probability to stay in state l. If the chain has reached state L, it stays in this state with probability 1. We obtain the prior pmf

$$p(\boldsymbol{n}; \boldsymbol{\phi}) = \frac{1}{Z(\boldsymbol{\phi})} \prod_{l=1}^{L-1} (1 - \phi_l)^{n_l - 1} \phi_l, \tag{4.6}$$

where $\boldsymbol{\phi} = (\phi_1, ..., \phi_{L-1})$ are the trainable parameters. Equation 4.6 reveals that the monotonic Markov model is a product of geometric duration distributions. It needs to be renormalized to yield a valid pmf over the finite set of $T' \times L$ warping matrices. The normalizing constant $Z(\boldsymbol{\phi}) := Z_{T',L}(\boldsymbol{\phi})$ can be computed by dynamic programming using

$$Z_{t,l}(\boldsymbol{\phi}) = \phi_{l-1} \sum_{n=1}^{t-l+1} (1 - \phi_{l-1})^{n-1} Z_{t-n,l-1}(\boldsymbol{\phi}) \tag{4.7}$$

with recursion start $Z_{t,1} = 1$ for $t = 1, ..., T'$. Although Markov models have been used for warping fixed-length sequences in the past, the above distribution has not been described in this context before. Existing works apply models that are normalized over sequences of arbitrary length. For consistency with previous works, we ignore the normalization term for inference. We write $\mathbf{A}_i \overset{\text{iid}}{\sim} \text{MkWarp}(T', L, \boldsymbol{\phi})$ if the prior distribution of the warping matrices $\mathbf{A}_i \equiv (\mathrm{n}_{i1}, ..., \mathrm{n}_{iL})$ is given by the (unnormalized) Markovian distribution from Equation 4.6.

Multinomial time warping

In the multinomial time warping model, the numbers of repetitions $\mathrm{n}_1, ..., \mathrm{n}_L$ jointly follow the multinomial distribution

$$(\mathrm{n}_1 - 1, ..., \mathrm{n}_L - 1) \sim \text{Multinomial}(T' - L, \boldsymbol{\pi}), \tag{4.8}$$

with repetition probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_L)$, $\pi_l \in (0, 1)$, $\sum_l \pi_l = 1$. The multinomial distribution allows outcomes with counts of 0. By modeling $\mathrm{n}_l - 1$ instead of n_l as multinomial, we make sure that all count values are strictly larger than zero. The total number of *additional* repetitions to draw from the multinomial distribution is $T' - L$. We write $\mathbf{A}_i \overset{\text{iid}}{\sim} \text{MWarp}(T', L, \boldsymbol{\pi})$ if the prior distribution of the warping matrices $\mathbf{A}_i \equiv (\mathrm{n}_{i1}, ..., \mathrm{n}_{iL})$ is given by the multinomial distribution from Equation 4.8.

Dirichlet-compound multinomial time warping

Finally, we instantiate our model family with a Dirichlet-compound multinomial distribution [JKB97]. From a Bayesian perspective, this distribution takes the repetition probabilities $\boldsymbol{\pi}$ of the multinomial distribution as the outcome of a symmetric Dirichlet trial $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha)$, and marginalizes over all possible values $\boldsymbol{\pi}$. In the Dirichlet-compound multinomial time warping model, the numbers of repetitions $\mathrm{n}_1, ..., \mathrm{n}_L$ jointly follow the distribution

$$(\mathrm{n}_1 - 1, ..., \mathrm{n}_L - 1) \sim \text{Dirichlet-Multinomial}(T' - L, \alpha). \tag{4.9}$$

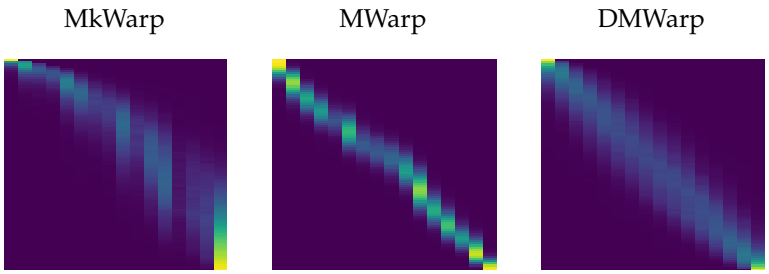

Figure 4.3: Expected warping matrices after training on ECG200 (class -1, $T' = 96$, $L = 16$).

We write $\mathbf{A}_i \overset{\text{iid}}{\sim} \text{DMWarp}(T', L, \alpha)$ if the warping matrices $\mathbf{A}_i \equiv (\mathrm{n}_{i1}, \ldots, \mathrm{n}_{iL})$ follow the Dirichlet-compound multinomial distribution from Equation 4.9. The concentration parameter $\alpha > 0$ of the compound distribution is usually treated as a hyper-parameter that is known a priori. It controls the variability of the resulting warping matrices: For $\alpha \longrightarrow 0$, few entries from the prototype are stretched strongly, while other values are not stretched at all. For $\alpha \longrightarrow \infty$, the prototype is stretched linearly with high probability. The symmetry assumption implies that for any α the expected warping matrix is linear. If prior knowledge on the warping process is available, an asymmetric Dirichlet-compound distribution with a bias towards nonlinear transformations may be more suitable.

Visualization

Some intuitions about the behavior of the three warping priors can be developed by looking at the expected warping matrices $\mathrm{E}_{\mathbf{A};\boldsymbol{\psi}}[\mathbf{A}] \in [0,1]^{T' \times K}$ under the respective prior pmfs $p(\boldsymbol{A}; \boldsymbol{\psi})$ with (hyper-) parameters $\boldsymbol{\psi}$. A visualization can be found in Figure 4.3. For the visualization, we performed training on the ECG200 dataset (class -1, $T' = 96$, $L = 16$) from the UCR Time Series Classification Archive [DBK+18]. We use the MCEM procedure described in Section 4.2.4 for training and estimate the expected values with importance sampling. We observe that the three priors have very different ways to distribute the warping uncertainty. The multinomial warping prior MWarp has the lowest variability and induces a strong bias towards the warping matrices observed during training, while the Dirichlet-compound warping prior DMWarp has the highest variability with a bias towards linear warping. The Markovian prior MkWarp combines bias towards the warping matrices observed during training with a high variability. Interestingly, in this example, it has a rather low variability in the beginning of the warping process, and increasing variability towards the end of the warping process.

4.2.3 Using the model

The probabilistic warping model yields quantities for sequence approximation, alignment, and structural averaging. They are visualized in Figure 4.4. All quantities require expected warping matrices either from the prior distribution $\Pr_{\mathbf{A};\boldsymbol{\psi}}$ or from the posterior distribution given an observed sequence $\Pr_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}}$.

- $\hat{\boldsymbol{x}}_i := \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}}[\mathbf{A}_i]^\dagger \cdot \boldsymbol{x}_i \in \mathbb{R}^L$ is the **low-dimensional representation** of an observed sequence $\boldsymbol{x}_i$ obtained by inverting the expected warping under the posterior distribution. We use $\boldsymbol{B}^\dagger = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'$ to denote the pseudo-inverse of $\boldsymbol{B}$. — low-dimensional representation
- $\bar{\boldsymbol{x}}_i := \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}}[\mathbf{A}_i] \cdot \hat{\boldsymbol{x}}_i \in \mathbb{R}^{T'}$ is the **piecewise constant approximation** of an observed sequence, obtained by mapping the low-dimensional representation back to the original time axis, again using the expected warping matrix under the posterior distribution. The piecewise constant approximation may be smoothed at the transitions between the levels due to variability in the posterior distribution $\Pr_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}}$. — piecewise constant approximation
- In fact, any $\tilde{\boldsymbol{x}}_i := \tilde{\boldsymbol{A}} \cdot \hat{\boldsymbol{x}}_i \in \mathbb{R}^{T'}$ is a piecewise constant representation of an observed sequence, where the warping matrix $\tilde{\boldsymbol{A}}$ maps the low-dimensional representation back to some time axis. If the same **reference warping matrix** $\tilde{\boldsymbol{A}}$ is used for all sequences, the sequences are **aligned** to a common time axis. Unless otherwise noted, we use the prior expectation $\tilde{\boldsymbol{A}} := \mathrm{E}_{\mathbf{A};\boldsymbol{\psi}}[\mathbf{A}]$ as the reference warping matrix. Due to variability in the prior distribution, the result will typically be strongly smoothed. With $\tilde{\boldsymbol{A}}_j := \mathrm{E}_{\mathbf{A}_j|\mathbf{x}_j=\boldsymbol{x}_j;\boldsymbol{\theta}}[\mathbf{A}_j]$ for some j, all $\boldsymbol{x}_i$ are aligned to the observed sequence $\boldsymbol{x}_j$. — reference warping matrix; aligned

The reference warping matrix can also be used to obtain a structural average $\tilde{\boldsymbol{x}}$ of a dataset, by warping the prototype $\boldsymbol{\mu}$ to the common time axis, $\tilde{\boldsymbol{x}} := \tilde{\boldsymbol{A}} \cdot \boldsymbol{\mu}$. If smoothing is undesired, the posterior and prior expectations in the equations above can be replaced with the posterior mode $\arg\max_{\boldsymbol{A}}\{\Pr(\mathbf{A}_i = \boldsymbol{A} \mid \mathbf{x}_i = \boldsymbol{x}_i; \boldsymbol{\theta})\}$ and prior mode $\arg\max_{\boldsymbol{A}}\{\Pr(\mathbf{A} = \boldsymbol{A}; \boldsymbol{\psi})\}$. With this approach, we obtain hard (Viterbi) alignments between time points in the sequences and entries in the prototype.

4.2.4 Inference

In practice, the parameters of the models described above are unknown and must be estimated from observed data before they can be used for approximation, alignment and structural averaging.

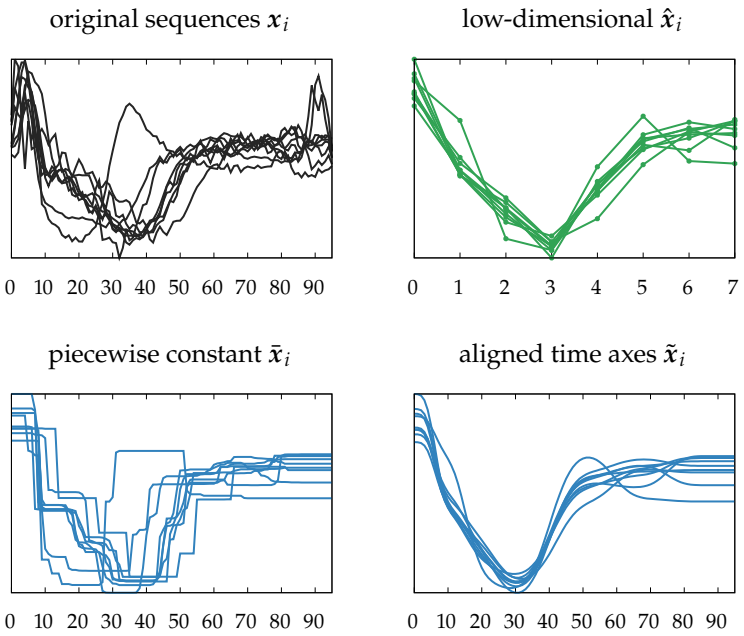

Figure 4.4: Various quantities obtained with the MWarp model on ECG200 (class -1, $T' = 96$, $L = 8$).

Parameter estimation

maximum likelihood estimation

We perform **maximum likelihood estimation** on the observed sequences $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ to recover the underlying prototype $\boldsymbol{\mu} \in \mathbb{R}^L$, the noise variance $\sigma^2 > 0$, and the parameters of the warping prior $\boldsymbol{\psi}$. Analytical maximization of the *incomplete-data* log-likelihood

$$\mathrm{LL}(\boldsymbol{\theta}) := \sum_{i=1}^{N} \log f(\boldsymbol{x}_i; \boldsymbol{\theta}) \tag{4.10}$$

In the incomplete-data log-likelihood function, only the sequences themselves are observed, not the (latent) warping matrices that produced these sequences.

with the pdf f from Equation 4.5 is impossible, since we marginalize over the warping matrices inside the logarithm. Instead, we use the expectation-maximization algorithm (**EM algorithm**) and iteratively maximize the *expected complete-data* log-likelihood, where we substitute the latent warping matrices with random matrices. The expected value is then taken with respect to the conditional distribution $\Pr_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\mathrm{old}}}$ using an initial guess of the parameters $\boldsymbol{\theta}^{\mathrm{old}}$. In our case, the expected complete-data log-likelihood is a function of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \boldsymbol{\psi})$ and given by

EM algorithm

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{i=1}^{N} \Big\{ & \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\mathrm{old}}} \left[\log f_{\mathrm{Normal}}(\boldsymbol{x}_i; \mathbf{A}_i\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \right] \\ & + \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\mathrm{old}}} \left[\log p(\mathbf{A}_i; \boldsymbol{\psi}) \right] \Big\}. \end{aligned} \tag{4.11}$$

When Q is maximized with respect to $\boldsymbol{\theta}$, a new set of parameters $\hat{\boldsymbol{\theta}}$ is obtained. The new parameters are guaranteed to have an *incomplete-data* log-likelihood $\mathrm{LL}(\hat{\boldsymbol{\theta}}) \geq \mathrm{LL}(\boldsymbol{\theta}^{\mathrm{old}})$. This procedure is repeated with $\boldsymbol{\theta}^{\mathrm{old}} := \hat{\boldsymbol{\theta}}$ until the EM algorithm converges to a stationary point of the log-likelihood function.

Prototype and noise variance

In our model family, the EM objective function from Equation 4.11 simplifies to a least squares problem. When taking the derivative of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\mu}$ and setting it to zero, we obtain

$$\hat{\mu}_l = \frac{\sum_i \sum_t \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl}] x_{it}}{\sum_i \sum_t \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl}]} \tag{4.12}$$

for $l = 1, \ldots, L$. The maximizing argument for the variance σ^2 is

$$\begin{aligned} \hat{\sigma}^2 = \frac{1}{NT'} \Bigg(& \sum_i \sum_t x_{it}^2 \\ & - 2 \cdot \sum_i \sum_t \sum_l \mu_l \, \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl}] x_{it} \\ & + \sum_i \sum_t \sum_l \mu_l^2 \, \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl}] \Bigg). \end{aligned} \tag{4.13}$$

It turns out that both parameter updates depend on the warping prior only via entries of the posterior expectations $\mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i,\boldsymbol{\theta}^{\text{old}}}[\mathbf{A}_i]$ and no other expected values.

Parameters of the warping priors

For the (unnormalized) Markovian time warping model, the state change probabilities $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{L-1})$ with $\phi_l \in (0, 1)$ have to be updated as well. From Equation 4.11, we obtain the standard estimates for geometric distributions:

$$\hat{\phi}_l = \left(\frac{1}{N} \sum_i \sum_t \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl}] \right)^{-1}. \tag{4.14}$$

In the multinomial time warping model, we additionally estimate the repetition probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)$. With a Lagrange multiplier to ensure $\sum_l \pi_l = 1$ we obtain

$$\hat{\pi}_l = \frac{\sum_i \sum_t \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl}] - N}{\sum_i \sum_t \sum_{l'} \mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}^{\text{old}}}[\mathrm{a}_{itl'}] - LN}. \tag{4.15}$$

Again, the parameter updates can be computed from entries of the posterior expectations $\mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}^{\text{old}}}[\mathbf{A}_i]$ and no other expected values. The Dirichlet-compound multinomial time warping model does not have extra parameters to estimate.

Computing the expectations

For all model instantiations presented here, the only expected values required for EM estimation are entries of the posterior expectation $\mathrm{E}_{\mathbf{A}|\mathbf{x}=\boldsymbol{x};\boldsymbol{\theta}}[\mathbf{A}]$. In our model family, the pmf of the posterior distribution is

$$p_{\mathbf{A}|\mathbf{x}=\boldsymbol{x}}(\boldsymbol{A};\boldsymbol{\theta}) = \frac{f_{\text{Normal}}(\boldsymbol{x};\boldsymbol{A}\boldsymbol{\mu},\sigma^2\boldsymbol{I}) \cdot p(\boldsymbol{A};\boldsymbol{\psi})}{f(\boldsymbol{x};\boldsymbol{\theta})} \tag{4.16}$$

with $f(\boldsymbol{x};\boldsymbol{\theta})$ from Equation 4.5. This pmf can be evaluated analytically for the monotonic Markov prior using the standard forward-backward algorithm [Rab89]. For many other priors, exact evaluation is infeasible, and the posterior expectations must be approximated with Monte Carlo methods.

importance sampling

We propose to estimate the posterior expectations with **importance sampling**. We use the prior distribution with pmf $p(\boldsymbol{A};\boldsymbol{\psi})$ as the proposal distribution, and the numerator from above,

$$p^*_{\mathbf{A}|\mathbf{x}=\boldsymbol{x}}(\boldsymbol{A};\boldsymbol{\theta}) := f_{\text{Normal}}(\boldsymbol{x};\boldsymbol{A}\boldsymbol{\mu},\sigma^2\boldsymbol{I}) \cdot p(\boldsymbol{A};\boldsymbol{\psi}) \tag{4.17}$$

as the unnormalized target function. This function is proportional to the posterior pmf up to the denominator $f(\boldsymbol{x};\boldsymbol{\theta})$ that is constant with respect to $\boldsymbol{A}$. The posterior expectation is then estimated as a weighted average of the warping matrices sampled from the proposal distribution, with weights obtained from the unnormalized target function and the proposal pmf [RC04]. More precisely, for every $i = 1, \ldots, N$, we sample S warping matrices $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(S)}$ from the proposal and approximate the posterior expectation via

$$\mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}}[\mathbf{A}_i] \approx \sum_{s=1}^{S} \left(\frac{w_i^{(s)}}{\sum_{s'} w_i^{(s')}} \right) \boldsymbol{A}^{(s)} \tag{4.18}$$

with

$$w_i^{(s)} := \frac{p^*_{\mathbf{A}|\mathbf{x}=\boldsymbol{x}_i}(\boldsymbol{A}^{(s)};\boldsymbol{\theta})}{p(\boldsymbol{A}^{(s)};\boldsymbol{\psi})} = f_{\text{Normal}}(\boldsymbol{x}_i;\boldsymbol{A}^{(s)}\boldsymbol{\mu},\sigma^2\boldsymbol{I}). \tag{4.19}$$

The importance sampling approach may be inefficient if the posterior distribution is very different from the proposal distribution. In some cases, it may be more efficient to use a uniform sampling distribution, or to use Markov chain Monte Carlo (MCMC) methods instead. The full **Monte Carlo EM (MCEM)** procedure for inference is shown in Algorithm 2.

Monte Carlo EM (MCEM)

Algorithm 2: MCEM estimation

repeat
  compute $\mathrm{E}_{\mathbf{A}_i|\mathbf{x}_i=\boldsymbol{x}_i;\boldsymbol{\theta}^{\text{old}}}[\mathbf{A}_i]$ for $i = 1, \ldots, N$;
  compute $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$ as in Equation 4.12 and Equation 4.13 ;
  optionally, compute update for prior parameters $\hat{\boldsymbol{\psi}}$;
  update $\boldsymbol{\theta}^{\text{old}} := (\hat{\boldsymbol{\mu}}, \hat{\sigma}^2, \hat{\boldsymbol{\psi}})$
until *convergence*;

4.3 Experiments

We now study the three instantiations of our model (MkWarp, MWarp and DMWarp) empirically. Our main goal is to assess whether there are tangible differences in the behavior of the three warping priors for downstream tasks. We evaluate the effect of the prototype length L on their representative power in terms of reconstruction error, and the impact of L on the final alignments. At last, we apply our models within a classification task.

We trained our models on several sequence datasets from the UCR Time Series Classification Archive 2018 [DBK+18]. The UCR archive contains sequence datasets from various domains, each separated into two or more classes. We handpicked datasets that visually exhibit strong temporal misalignments, as we designed our models to capture the generative process of such data. We use the ECG200 dataset as a running example to demonstrate the properties of our model family. For every training set, class, and value of L, we run the MCEM algorithm for 20 iterations with 10 randomized restarts. We average our performance measures over all restarts.

ECG200 is particularly interesting from the perspective of event impact analysis, since it contains sequences extracted from a long ECG time series at heart beat events.

Prototypes are initialized by sampling from the standard normal distribution. In the Markovian time warping model, we initialize the state transition probabilities with $\phi_l := 1 - \frac{L}{T'}$. In case of multinomial time warping, the repetition probabilities are initialized with $\pi_l := \frac{1}{L}$. For Dirichlet-compound multinomial warping, we set $\alpha := 1$. Monte Carlo estimates for the expectations in Algorithm 2 are computed from 1,000 random samples.

4.3.1 Representative power

We first illustrate the representative power of our model family on the ECG200 dataset (class -1, $T' = 96$). Figure 4.5 shows the first sequence from the training dataset along with its piecewise constant approximation $\bar{\boldsymbol{x}}_i$ (see Section 4.2.3) obtained from MWarp, with $L \in \{2, 4, 8\}$. Clearly, a larger number of entries in the prototype L allows a more fine-grained approximation to the original sequence. If L is chosen too small, salient features of the sequence are lost.

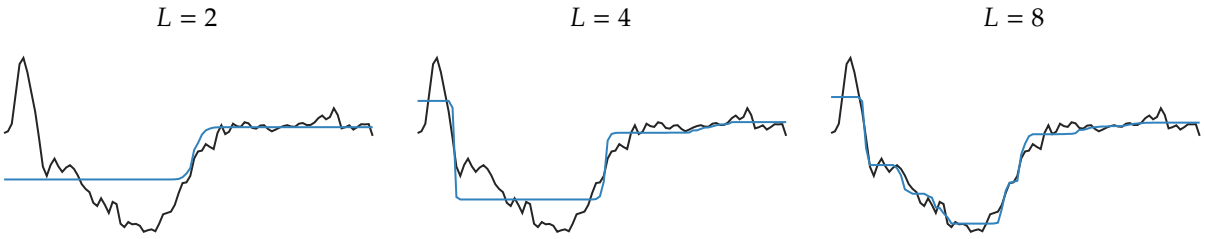

Figure 4.5: Representative power of the MWarp model on ECG200 for different prototype lengths L. Black lines depict the original sequences $\boldsymbol{x}_i$, blue lines the respective approximations $\bar{\boldsymbol{x}}_i$.

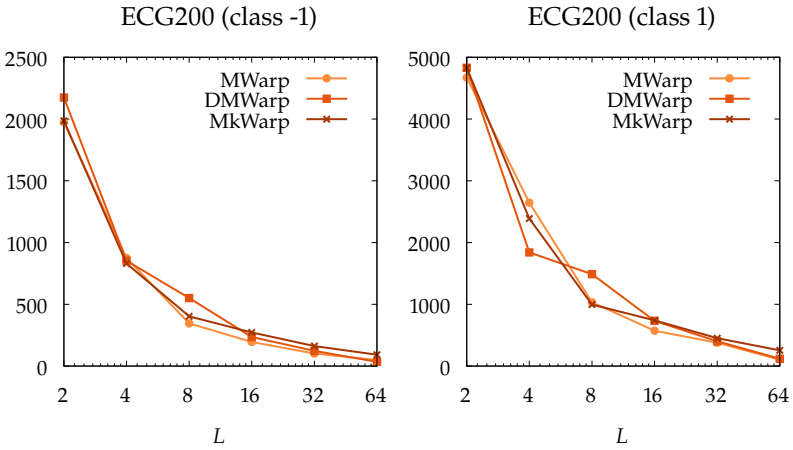

Figure 4.6: Sum of squared reconstruction errors (SSE) on ECG200 data by prototype length L.

With only $L = 8$ entries in the prototype, the approximation closely follows the original sequence of length $T' = 96$.

sum of squared errors (SSE)

To assess the representative power quantitatively, we computed the **sum of squared errors (SSE)** of the piecewise constant approximation $\text{SSE} = \sum_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|^2$ on both classes of the ECG200 training data for all three instantiations of our model family with various values of L. Figure 4.6 shows that the SSE drops when increasing L from 2 to 8, but then converges with only minor improvements in SSE for higher values of L.

The three warping priors perform equally well in reconstructing the datasets. We observe this result on all datasets that we experimented with, with only minor differences in the reconstruction performance of the priors.

4.3.2 Alignment quality

We now evaluate the performance of the model family in computing alignments. In contrast to previous models for sequence alignment, we align the input sequences not only by transforming the time axes, but by jointly transforming the time axes *and* reducing their temporal resolutions via the piecewise constant approximations

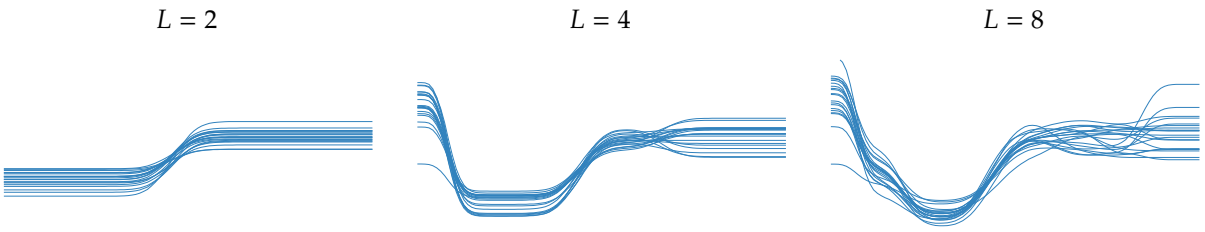

Figure 4.7: Alignment results $\tilde{\boldsymbol{x}}_i$ of MWarp on ECG200, using $\tilde{\boldsymbol{A}} = \mathrm{E}_{\mathbf{A};\psi}[\mathbf{A}]$ as the reference warping matrix.

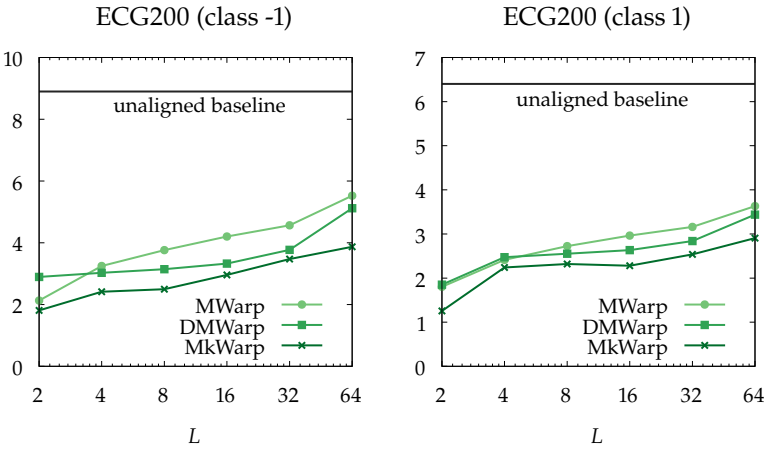

Figure 4.8: Pairwise Euclidean distances D between the aligned sequences $\tilde{\boldsymbol{x}}_i$ with $\tilde{\boldsymbol{A}} = \mathrm{E}_{\mathbf{A};\psi}[\mathbf{A}]$.

described above. Generally, aligning coarse approximations is a simpler problem than aligning fine-grained approximations. We illustrate this observation in Figure 4.7, again using MWarp on the ECG200 training data (class -1). When the dataset is reduced to only two levels ($L = 2$), the jump from the first level to the second level can easily be aligned across all sequences, and the two levels have a low variability on the vertical axis across all sequences. As the approximations contain more detail ($L \in \{4, 8\}$), some of the aligned segments show a higher variability in the levels, which indicates alignment errors.

To measure the alignment performance of our models quantitatively, we compute the **pairwise Euclidean distances** between all sequences after alignment and obtain their average, *i.e.*,

pairwise Euclidean distances

$$D = \frac{2}{N(N-1)} \sum_{i<j} d(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j). \tag{4.20}$$

Results can be found in Figure 4.8 using the prior expectation $\tilde{\boldsymbol{A}} = \mathrm{E}_{\mathbf{A};\psi}[\mathbf{A}]$ as the reference warping matrix. For comparison, the average Euclidean distance on the *raw unaligned* ECG200 training data is $D = 8.9$ for class -1, and $D = 6.4$ for class 1.

Our models improve over these baseline values for all choices of L by a large margin. The pairwise distances grow with L,

which confirms that the alignment problem becomes more difficult when the models have to retain more details from the sequences. The sequences are standardized to have mean 0 over time, so the optimal value $D = 0$ is achieved when all sequences are represented by a constant line at level 0. This corresponds to a model with a prototype of length $L = 1$. For a fair evaluation, D should only be compared across models with the same representational power—in our case, with the same prototype lengths L. When fixing the prototype length L, the Markovian model MkWarp provides alignments with lower pairwise distances than DMWarp on ECG200, which in turn provides lower distances than MWarp.

The differences in the pairwise Euclidean distances for the different warping priors may be explained by the different levels of smoothing in the aligned sequences that they induce. A warping prior that distributes the alignment uncertainty over more time steps leads to a stronger smoothing in the aligned sequences than a prior with lower alignment uncertainty. The visualizations in Figure 4.3 suggest that the MkWarp and DMWarp priors have a higher alignment uncertainty. The question is whether the differences in the pairwise Euclidean distances are *only* due to the differences in smoothing, or whether the alignments have also improved.

For comparison, Figure 4.9 shows the pairwise Euclidean distances after alignment with the alternative reference warping matrix $\tilde{\boldsymbol{A}}_1 = \mathrm{E}_{\mathbf{A}_1|\mathbf{x}_1=\boldsymbol{x}_1;\boldsymbol{\theta}}[\mathbf{A}_1]$. The differences between the model instantiations are smaller now, which confirms that the different levels of smoothing may obscure the evaluation measure. The MkWarp model still outperforms the other models, but the DMWarp model does not appear superior to the MWarp model anymore.

This reference warping matrix aligns all observed sequences to the first observed sequence $\boldsymbol{x}_1$.

At last, Figure 4.10 shows the pairwise Euclidean distances between the low-dimensional representations $\hat{\boldsymbol{x}}_i$ before they are projected to a common time axis by a reference warping matrix. The differences in alignment quality become even less tangible in the low-dimensional representations. In fact, optimal alignment in the low-dimensional space is implicit in the Gaussian part of the likelihood function from Equation 4.5, since the maximum likelihood estimator for the prototype is the average of the low-dimensional projections. Overall, the results on alignment quality remain inconclusive: The choice of the prior distribution *does* have an effect on the aligned sequences according to our evaluation measure, but we cannot clearly state whether this difference in the evaluation measure actually reflects a change in the alignment quality.

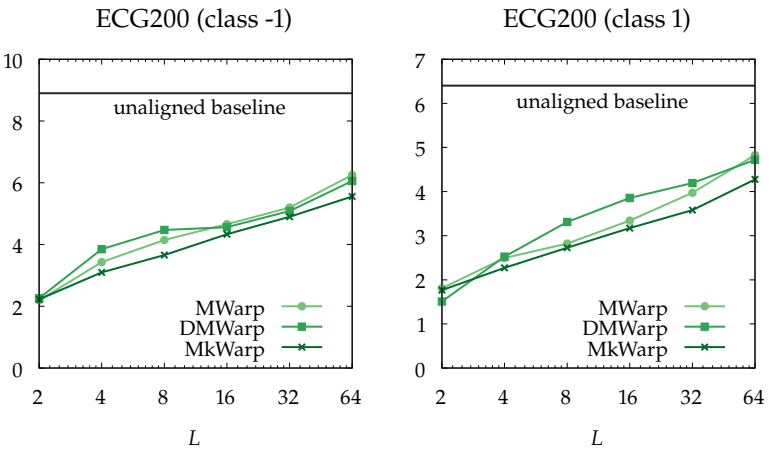

Figure 4.9: Pairwise Euclidean distances D between the aligned sequences $\tilde{\boldsymbol{x}}_i$ with the reference warping matrix $\tilde{\boldsymbol{A}}_1 = \mathrm{E}_{\mathbf{A}_1|\mathbf{x}_1=\boldsymbol{x}_1;\boldsymbol{\theta}}[\mathbf{A}]$.

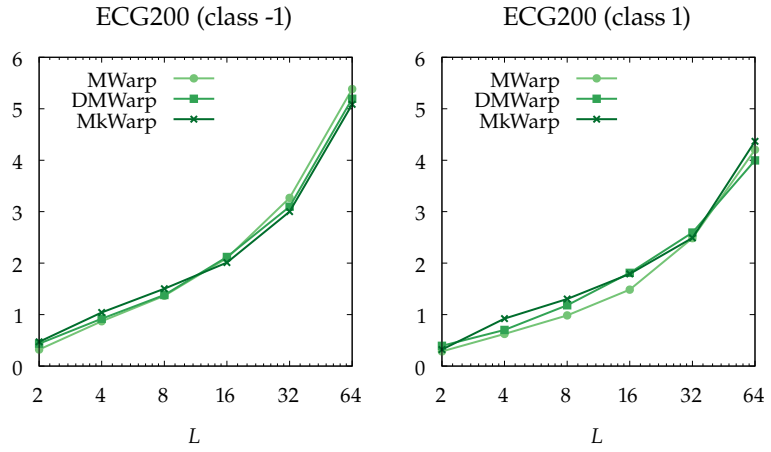

Figure 4.10: Pairwise Euclidean distances D between the low-dimensional projections $\hat{\boldsymbol{x}}_i$. These values cannot be compared with the unaligned baselines due to the different dimensionalities.

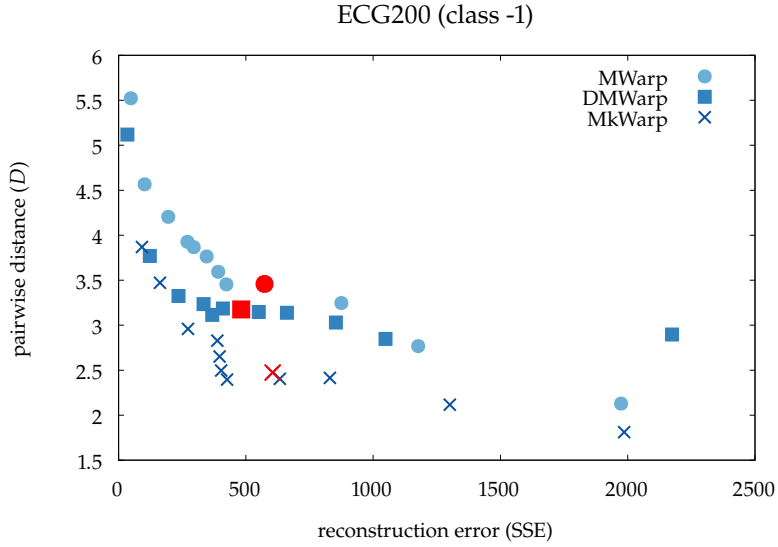

Figure 4.11: Reconstruction error (SSE) against pairwise distance after alignment (D) for all three priors. Every point corresponds to a specific choice of L for that model. Red markers highlight the value of L that yielded the highest expected complete-data log-likelihood during training.

4.3.3 Model selection

The experiments in Section 4.3.1 and Section 4.3.2 show that the goals encoded in the two performance measures (good approximation and good alignment) are conflicting. By reconstruction error, larger values of L are better. By alignment quality, smaller values of L are better. The question is how to choose a good trade-off between these two goals. For this purpose, we plot the reconstruction error against the pairwise distances in Figure 4.11 for various values of L and our three warping priors. Note that we optimize neither reconstruction error, nor alignment quality directly during training, but instead find the optimal model parameters for every L by maximum likelihood. For every warping prior, we thus highlight the choice of L that yielded the highest expected complete-data log-likelihood (ECLL) on the training data.

We observe that the ECLL finds a trade-off between the two goals. The models assign the highest likelihoods to parametrizations that allow for a decent reconstruction *and* decent alignment of the time series. In the example, the best values for ECG200 are $L = 6$ for MWarp and $L = 7$ for DMWarp and MkWarp.

If we view our problem as a multiobjective optimization problem with the two objectives SSE and D, Figure 4.11 suggests that MkWarp is consistently closer to the Pareto frontier than MWarp and DMWarp.

4.3.4 Classification

Finally, we evaluate the performance of our model family in downstream classification tasks from the UCR Time Series Classification Archive [DBK+18]. We include several state-of-the-art methods for sequence alignment in our experiments: Generalized Time Warping (GTW) [ZT12], Soft-DTW (sDTW) [CB17], Trainable Time Warping (TTW) [KMP19], and Temporal Transformer Networks (TTN) [LWT19]. We employ our own models and the competitors in two kinds of classifiers:

1. **Bayes classifiers:** Every class is modeled separately by fitting our probabilistic time warping models to the data. As probabilistic competitors, we model classes by normal distributions with isotropic covariance structure (NormIso) and full covariance structure (NormFull). Our models have $O(L)$ free parameters per class, NormIso has $O(T)$ free parameters, and NormFull $O(T^2)$.
2. **Nearest centroid classifiers (NCC):** Every class is represented by a centroid, *i.e.*, a structural average computed for each class from the data. A new instance is assigned to the class with the closest centroid according to the DTW distance. NCC classifiers have $O(T)$ parameters per class.

Table 4.1: Classification performance (test accuracy)

		ECG200	CBF	EthanolLevel	HandOutlines	Haptics	InlineSkate	Trace	Coffee
Bayes	MkWarp	0.69	**0.96**	0.27	0.68	0.37	0.24	**0.82**	0.64
Bayes	MWarp	**0.76**	0.92	0.34	**0.79**	**0.42**	0.21	0.76	0.91
Bayes	DMWarp	0.72	0.94	0.27	0.69	0.37	0.24	0.79	0.60
Bayes	NormIso	**0.76**	0.79	0.29	0.76	0.39	0.16	0.58	**1.00**
Bayes	NormFull	0.57	0.34	**0.60**	0.68	0.31	**0.26**	0.62	0.93
NCC	MkWarp	0.73	0.76	0.25	0.65	0.30	0.18	0.54	0.54
NCC	MWarp	**0.79**	0.80	**0.33**	**0.89**	**0.42**	0.18	0.56	**0.96**
NCC	DMWarp	0.76	0.72	0.26	0.64	0.30	**0.21**	**0.62**	0.71
NCC	TTN	0.70	0.58	0.25	0.67	0.24	0.16	0.43	0.46
NCC	sDTW	0.78	0.80	0.26	0.81	0.38	0.19	0.49	0.93
NCC	GTW	0.73	0.61	0.25	0.45	0.28	0.20	0.52	0.71
NCC	TTW	0.77	**0.82**	0.28	0.81	0.40	0.20	0.55	0.92

Experimental setup

We use the train/test split from the UCR archive for training and evaluation. We implemented our model family and the simple Bayes classifiers in Python. For the NCC classifiers, we used different approaches to compute structural averages for each class. We obtained the source codes of GTW and TTW provided online from the authors of TTW*. For TTW, we stick to the recommendation of the original authors and set the order parameter to the values 4, 8, 16 and 32. We run the optimization algorithms for TTW and GTW over 100 iterations. For sDTW, we used the implementation from its original authors†, and set the exponent γ to the values 0.25, 0.5, 1, and 2. We compute structural averages from their barycenter algorithm over 50 iterations. We implemented the TTN architecture in Python using PyTorch‡, with one convolutional layer with 4 filters and a kernel size of 9, followed by a fully connected layer, tanh nonlinearities, and dropout ($p = 0.2$). For a fair competition, we did not include the TTN in a supervised classification architecture as the original authors, but instead use the unsupervised loss function from [WED+19] to align the time series for every class separately. We use the Adam optimizer with a learning rate $\eta = 0.0001$ over 500 iterations for training.

Results

Classification accuracies can be found in Table 4.1. For methods that depend on the choice of hyperparameters (our models, TTW, and sDTW), we print the best result obtained using that method. The classifiers built from our model family are highly competitive,

* https://github.com/soheil-khorram/TTW
† https://github.com/mblondel/soft-dtw
‡ https://pytorch.org/

and provide the best results on many datasets. Among the Bayes classifiers, our model family has by far the fewest parameters, and still outperforms the competitors. The gain in accuracy is especially strong for the CBF and Trace datasets, which contain simple shapes that are highly misaligned. Among our three model instantiations, there is no warping prior that strictly dominates the other priors in terms of Bayes classification accuracy. This observation challenges the widespread belief that Markovian models are sufficient. Among the NCC classifiers, the multinomial prior MWarp yields the overall best results. We believe that the difference in performance to the other models from our family is mostly due to the lower variability in the prior expectation $\mathrm{E}_{\mathbf{A};\psi}[\mathbf{A}]$ of MWarp as visualized in Figure 4.3. When prior expectations are used as reference transformations to compute structural averages, the multinomial prior contains more and potentially discriminative information on the warpings observed during training.

The Coffee dataset does not exhibit temporal misalignments, and the simple Bayes classifier with isotropic normal distributions achieves perfect classification results. All other datasets were handpicked by visual inspection for the presence of distinct shapes with temporal misalignments, *before* conducting the experiment. We posit that the difference in performance to the state-of-the-art approaches for time warping is due to their inability to distinguish noise from signal in the warping process. We do not claim that our probabilistic time warping models outperform the competitors in general. However, the experiment suggests that our models indeed preserve discriminative features of the input sequences.

4.3.5 Alignments and averages

At last, we visualize aligned sequences and structural averages computed with MWarp on datasets from the classification experiment in Figure 4.12. To compute the alignments, we selected the prototype lengths that attain the maximal expected complete-data log-likelihood (ECLL) after training. The models successfully align the visually salient features of most sequences. Some high-frequency components are not represented by the low-dimensional prototypes and thus lost, most notably on EthanolLevel.

4.4 Conclusions

We have presented a generic model family for discrete time warping that can be instantiated with many prior distributions over the warping matrices. Our model family generalizes existing work that is restricted by Markov assumptions. In fact, it allows exploiting

the large variety of discrete multivariate distributions [JKB97] to encode complex dependencies in the warping matrices. We have shown that maximum likelihood parameter estimation for our model family simplifies to a least squares problem, and provided a versatile Monte Carlo EM algorithm that is widely applicable for many warping priors. Our experiments show that the choice of warping prior indeed has impact on the model performance in downstream tasks. When applied for classification, our models outperform state-of-the-art competitors. We believe that the MWarp model is most suitable to provide a concise summary of the data, as the multinomial warping prior seems to retain more discriminative information than the other two priors we considered. This is particularly relevant for applications in event impact analysis, where we are interested in visualizing the prototypical pattern of the deviant behavior after event occurrences.

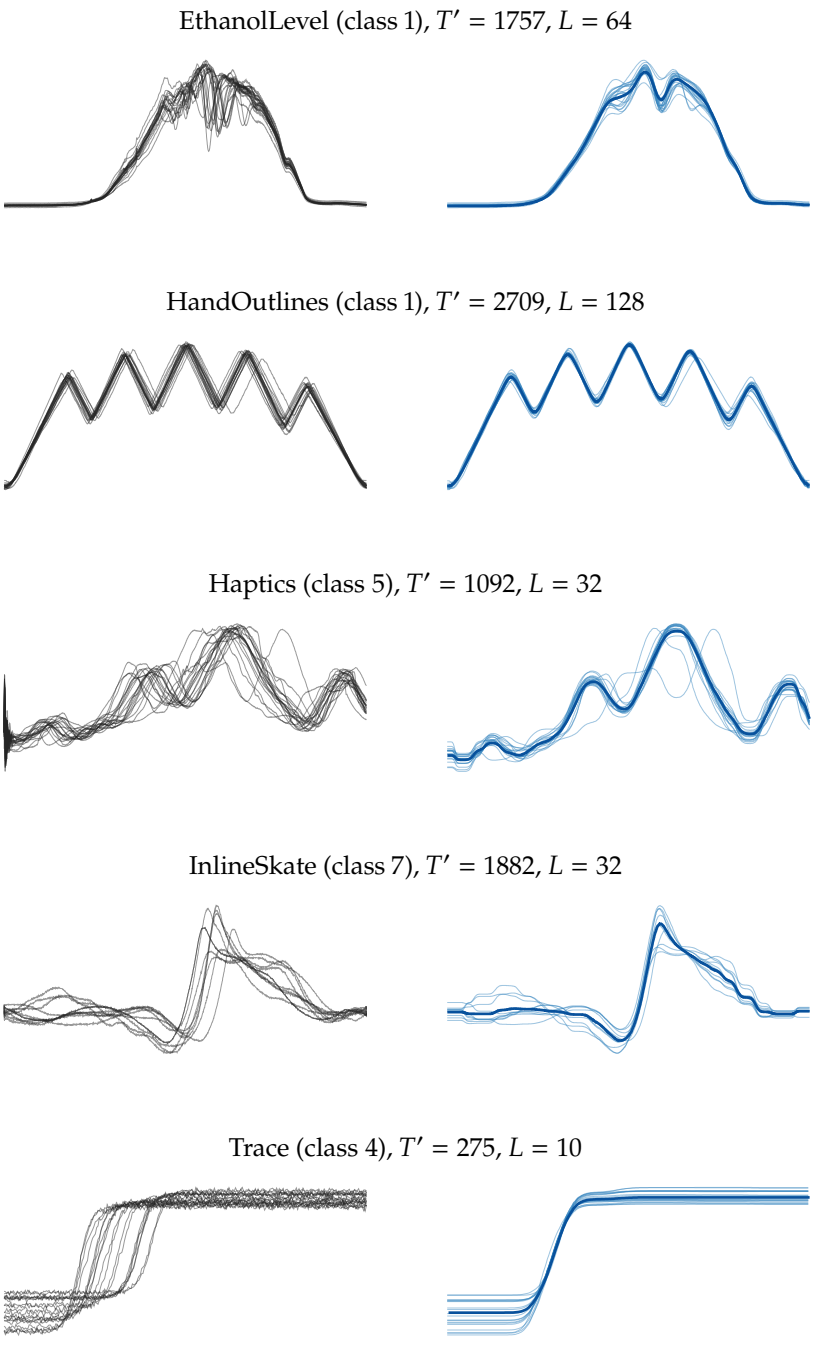

Figure 4.12: *Left:* sequences $\boldsymbol{x}_i$ before alignment; *right:* sequences $\tilde{\boldsymbol{x}}_i$ after alignment with MWarp, along with the structural averages $\tilde{\boldsymbol{x}}$.

Statistical Evaluation of Anomaly Detectors

5

This chapter is based on:

[SM20b] Erik Scharwächter and Emmanuel Müller. "Statistical Evaluation of Anomaly Detectors for Sequences." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Mining and Learning from Time Series (KDD MiLeTS)*, 2020.

We conclude the discussion of our methods for stationary time series by pointing out important practical implications for the evaluation of anomaly detection algorithms for time series. In Chapter 2, we have presented the trigger coincidence rate and precursor coincidence rate as measures for the *association* of event occurrences and peaks in a stationary time series. In a wider machine learning context, these measures can be viewed as time-tolerant versions of the classical measures of precision and recall routinely used for the *evaluation* of binary classifiers, event detection and anomaly detection algorithms.

In this chapter, we focus on applications of these measures for the evaluation of point-based anomaly detection algorithms in time series. Although precision and recall are widely used for evaluating anomaly detectors in sequential settings, their statistical properties in these settings are poorly understood. In the following, we demonstrate that the statistical perspective on these measures developed in Chapter 2 for event impact analysis also helps understanding the behavior of precision and recall as evaluation measures in sequential settings.

We first formalize notions of precision and recall with temporal tolerance analogously to the trigger coincidence rate and precursor coincidence rate seen earlier in this work. We argue that these measures can be derived from time-tolerant confusion matrices that also yield time-tolerant variants of many other standard performance measures. However, care has to be taken to preserve interpretability of these measures under temporal tolerance. We perform a statistical simulation study to demonstrate that precision and recall potentially overestimate the performance of a detector, when computed with temporal tolerance. Our statistical perspective on the evaluation problem allows us to compute null distributions for the two evaluation measures—and any other measure derived from a confusion matrix—to assess the statistical significance of reported results under an independence assumption. While statistical significance does not mean that the reported performance of a detector is practically useful, lack of statistical significance means that the reported results are not better than random guessing, even if the nominal performance value is apparently large. Our analysis reveals that, when developing an algorithm for point-based anomaly detection in time series, we really seek an anomaly scoring function that maximizes the statistical association between the ground-truth label sequence and peaks in the anomaly score.

5.1 Introduction

Anomaly detection in sequential data is a highly active research topic [BJR17; MVS+15; RXW+19; SFT+17; SLZ+19; XFC+18]. Precision and recall are two widespread measures to evaluate the performance of anomaly detectors, both for iid data and for sequential data. An important characteristic of sequential data is that the decisions of a detector can be imprecise [AM17; AM18] without impairing its practical utility: if an anomaly at time step t is detected at time step $t \pm \Delta$ for some small lag Δ, this is still a useful result for many applications. Recently, Tatbul et al. [TLZ+18] pointed out that the classical precision and recall measures, when applied to sequential detection problems, may misrepresent the performance of the detector. They introduced novel precision and recall measures for *range-based* anomaly detection. However, the problem persists even for *point-based* anomalies, where the ground-truth anomaly label is a single time step. In this chapter, we study in detail time-tolerant notions of precision and recall for *point-based* anomaly detection in sequential data, with a special focus on the statistical properties of these measures. We provide a generic problem statement for anomaly detection and classical measures for precision and recall below. In Section 5.2, we define notions of precision and recall with temporal tolerance similar to the coincidence rates for event impact analysis from Chapter 2. At last, we study the statistical properties of these measures in Section 5.3.

Some authors, including the author of this work, prefer the term "event detection" if the ground-truth labels correspond to some external event of interest. However, for the evaluation with precision and recall, it does not matter whether the labels are assigned based on external information outside of the time series or based on properties of the time series itself.

5.1.1 Anomaly detection

We use a very generic formulation of the anomaly detection problem to capture a wide spectrum of approaches with our analysis. We are given an input time series $\mathbf{X} = (x_1, \ldots, x_T)$ of length T over an arbitrary domain. Furthermore, we are given an **anomaly scoring function** $z_t(\mathbf{X}) \in \mathbb{R}$ to compute a time series of anomaly scores $\mathbf{Z} = (z_1, \ldots, z_T)$ from the input time series. If the observation x_t is likely an anomaly, the anomaly score z_t should be high; if the observation x_t appears normal, z_t should be low. An anomaly is predicted at time step t if the anomaly score is larger than some predefined threshold, $z_t \geq \tau$. The exact notion of what constitutes an anomaly is domain-specific and should be reflected in the choice of the anomaly scoring function. Anomaly detectors of this type are widely used across many disciplines. For example, Wiedermann et al. [WRD+16] use the clustering coefficient as an anomaly score for dynamic networks to detect El Niño events in climate data, and Earle, Bowden, and Guy [EBG11] use an energy transient score [WAY+98] to detect earthquakes from Twitter data.

anomaly scoring function

http://twitter.com/

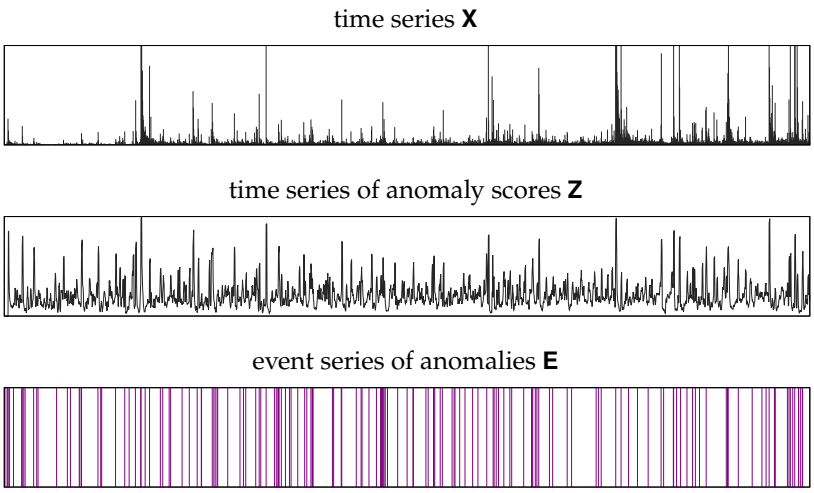

Figure 5.1: The running example: Anomaly detection on Twitter data for earthquake detection.

We use the problem of earthquake detection on Twitter as the running example in this work. Figure 5.1 (top row) shows the daily volume of tweets that were posted in Germany between 2010 and 2017 and contain the word "earthquake," translated to various languages. The plot also shows all severe earthquakes that occurred globally in the same time period (bottom row). We obtained the Twitter data using the ForSight platform by Crimson Hexagon/Brandwatch, and the earthquake data from the International Disaster Database EM-DAT, provided by the Centre for Research on the Epidemiology of Disasters. Our goal is to evaluate whether an anomaly detector on the Twitter time series has the potential to detect earthquakes globally. For this purpose, we stick to Earle, Bowden, and Guy [EBG11] and use the energy transient score as the anomaly score (middle row), *i.e.*, we set $z_t = \mathrm{STA}_t / (\mathrm{LTA}_t + 1)$, where STA is the short-term average of the input time series over the past 3 days, and LTA is the long-term average over the past 14 days. The energy transient score reacts to drastic changes in the level of the time series, while being robust with respect to the absolute levels of these changes.

`http://brandwatch.com/`

`http://emdat.be/`

This is the same data that we used for the experiments with MEITEST in Section 3.3.3. While the focus in Chapter 3 was to establish a statistical association between the Twitter time series **X** and the earthquake occurrences in **E**, the focus here is how to evaluate the performance of the anomaly score **Z** for detection.

5.1.2 Evaluation measures

While the anomaly score encodes the *feature* of interest that the anomaly detector should react to, the detection threshold τ controls the **precision** and **recall** of the anomaly detector. Let $\mathbf{E} = (e_1, \ldots, e_T)$ be the ground-truth event series of anomalies, with value $e_t = 1$ if there is an actual anomaly at time step t, and $e_t = 0$ if there is no anomaly. In our running example, actual anomalies correspond to severe earthquakes captured within EM-DAT.

precision

recall

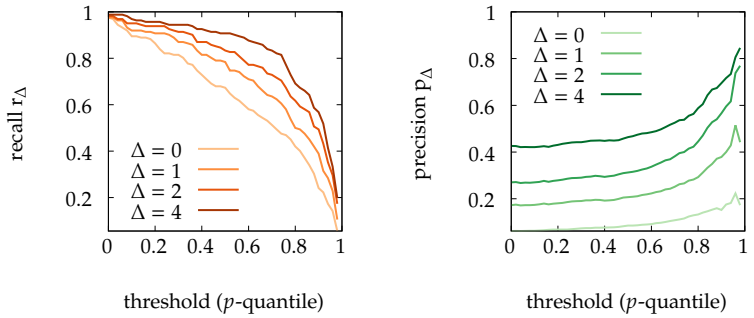

Figure 5.2: Precision and recall by threshold and tolerance Δ.

Classical precision p_0 and recall r_0 are defined as

$$p_0 = \frac{\sum_t e_t \cdot \mathbf{1}[z_t \geq \tau]}{\sum_t \mathbf{1}[z_t \geq \tau]} \tag{5.1}$$

$$r_0 = \frac{\sum_t e_t \cdot \mathbf{1}[z_t \geq \tau]}{\sum_t e_t}. \tag{5.2}$$

true positives

The numerator is the number of **true positives**, *i.e.*, the number of time steps that are correctly predicted as anomalous, while the denominator is either the number of predicted anomalies or the number of actual anomalies. There is no temporal tolerance in the classical definition of precision and recall.

The relationship between the threshold and precision and recall for the earthquake detection problem is visualized in Figure 5.2. The values for p_0 and r_0 correspond to the lines labeled $\Delta = 0$ in the plots. The detection thresholds on the horizontal axis are chosen from the p-quantiles of the time series of anomaly scores. The results are not particularly good: we can obtain acceptable recall values at low detection thresholds, but the cost is an unacceptably low precision. Furthermore, we observe that recall is a monotonically decreasing function of the detection threshold τ: the number of true positives in the numerator decreases with increasing τ while the denominator stays constant. Precision, on the other hand, is a non-monotone function of the detection threshold, since both the numerator and the denominator change with τ.

The question is what to conclude from these bad results: Did we use an inappropriate anomaly scoring function that should be replaced by some other function? Or is our evaluation measure too strict, as it does not allow temporal tolerance in the detection? In the following, we study the effect of using evaluation measures with temporal tolerance on the reported performance values.

5.2 Methodology

5.2.1 Precision and recall

In sequential data, a predicted anomaly can often be considered a true positive if there is an actual anomaly *close to* the predicted time point. The higher the temporal tolerance, the higher the number of true positives, and the higher will be *both* precision and recall. We formalize measures for **precision** p_Δ and **recall** r_Δ with symmetric temporal tolerance Δ by relaxing the classical definitions as follows:

precision

recall

$$p_\Delta = \frac{\sum_t \left(\max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right) \cdot \mathbf{1}[z_t \geq \tau]}{\sum_t \mathbf{1}[z_t \geq \tau]} \tag{5.3}$$

$$r_\Delta = \frac{\sum_t e_t \cdot \left(\max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)}{\sum_t e_t}. \tag{5.4}$$

These measures are symmetric variants of the trigger coincidence rate and precursor coincidence rate from ECA [DSS+16] as defined in Section 2.2.2. If $\Delta = 0$, the time-tolerant measures are equivalent to the classical measures for precision and recall. If $\Delta > 0$, the definition of a true positive in the numerator changes. In fact, there are now two different types of **true positives**: In the case of precision, a true positive is a predicted anomaly at time step t with an actual anomaly within the range $[t - \Delta, t + \Delta]$. In the case of recall, a true positive is an actual anomaly at time step t with a predicted anomaly within the range $[t - \Delta, t + \Delta]$.

We use a different notation here to emphasize how these measures relax classical precision and recall.

true positives

Figure 5.2 shows the impact of various choices for the temporal tolerance Δ on the measured values for precision and recall. Depending on the choice of the threshold and the temporal tolerance, the reported values for precision and recall vary drastically. Even for moderate temporal tolerances, the detection results appear much better than before. In contrast to our earlier results with the classical measures, the results with time-tolerant measures lead us to the conclusion that the anomaly scoring function described above applied on the Twitter time series provides a decent way to detect severe earthquakes globally. Which conclusion is correct? We will shed more light on this question in Section 5.3.

5.2.2 Confusion matrices

The extension of precision and recall to time-tolerant measures via relaxed notions of *true positives* is intuitive, but has some subtleties. In fact, the two measures are computed from two distinct **confusion matrices**, where temporal tolerance is allowed *either* in the ground-truth time steps *or* in the predicted time steps.

confusion matrices

Table 5.1: Confusion matrix

	AA	AnA	
PA	TP	FP	$\sum$
PnA	FN	TN	$\sum$
	$\sum$	$\sum$	T

The general structure of a confusion matrix is shown on the left. It contains the numbers of observations that fall into the four categories true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), along with marginal sums. The row and column headings define the marginal conditions: actual anomaly (AA), actually no anomaly (AnA), predicted anomaly (PA), predicted no anomaly (PnA). The confusion matrix partitions the observations so that every observation falls into exactly one category. Many performance measures can be computed from confusion matrices [Pow07], typically by normalizing individual entries by marginal sums. The measures are interpretable because all entries and marginals have straightforward interpretations.

When allowing temporal tolerance in a confusion matrix, the result must still be a partition with interpretable entries and marginals. Table 5.2 and Table 5.3 show confusion matrices obtained when allowing temporal tolerance either in the ground-truth time steps or the predicted time steps. Both confusion matrices partition the observations, but some entries and marginals are hard to interpret. Some of the measures usually computed from confusion matrices are therefore uninformative. The precision from Equation 5.3 is the TP entry from Table 5.2 (PA-AAΔ) normalized by the marginal row sum (PA), whereas the recall from Equation 5.4 is given by the TP entry from Table 5.3 (PAΔ-AA) normalized the marginal column sum (AA). In both cases, the TP entries and normalization terms are interpretable and yield informative evaluation measures.

5.2.3 Null distributions

They key benefit of the statistical perspective on the evaluation problem that we put forward here is that we can treat all entries from the confusion matrices above, and all evaluation measures derived thereof, as random variables. For this purpose, we assume that the time series and the event series of ground-truth labels are, in fact, realizations of two stochastic processes. The evaluation measures are then statistics computed from these processes, with probability distributions induced by the joint distribution of the processes. This perspective allows us to establish the statistical significance of observed values of these measures, under the null hypothesis that the ground-truth labels and the time series are independent—in the same way as we established statistical significance of the number of trigger coincidences in Chapter 2.

If we find that a reported value for, *e.g.*, the precision or the recall is statistically significant under the null hypothesis, we have statistical evidence that the anomaly scoring function *systematically* captures aspects of the time series that are encoded in the ground-truth anomaly labels. In other words, we have evidence that the

Table 5.2: Relaxed confusion matrix for sequential data, with tolerance in ground-truth

	AAΔ	AnAΔ	
PA	$\sum_t \left(\max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right) \mathbf{1}[z_t \geq \tau]$	$\sum_t \left(1 - \max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right) \mathbf{1}[z_t \geq \tau]$	$\sum_t \mathbf{1}[z_t \geq \tau]$
PnA	$\sum_t \left(\max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right) (1 - \mathbf{1}[z_t \geq \tau])$	$\sum_t \left(1 - \max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right) (1 - \mathbf{1}[z_t \geq \tau])$	$\sum_t (1 - \mathbf{1}[z_t \geq \tau])$
	$\sum_t \left(\max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right)$	$\sum_t \left(1 - \max_{t'=t-\Delta}^{t+\Delta} e_{t'}\right)$	T

AAΔ: actual anomaly with tolerance Δ, AnAΔ: actually no anomaly with tolerance Δ, PA: predicted anomaly, PnA: predicted no anomaly

Table 5.3: Relaxed confusion matrix for sequential data, with tolerance in predictions

	AA	AnA	
PAΔ	$\sum_t e_t \left(\max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)$	$\sum_t (1 - e_t) \left(\max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)$	$\sum_t \left(\max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)$
PnAΔ	$\sum_t e_t \left(1 - \max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)$	$\sum_t (1 - e_t) \left(1 - \max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)$	$\sum_t \left(1 - \max_{t'=t-\Delta}^{t+\Delta} \mathbf{1}[z_{t'} \geq \tau]\right)$
	$\sum_t e_t$	$\sum_t (1 - e_t)$	T

AA: actual anomaly, AnA: actually no anomaly, PAΔ: predicted anomaly with tolerance Δ, PnAΔ: predicted no anomaly with tolerance Δ

performance of the anomaly detector is *better than random guessing*. This sounds like a minimal goal to achieve, but is not at all trivial: The probability distributions of the evaluation measures depend on (1) the statistical properties of the time series, (2) the choice of threshold, and (3) the temporal tolerance. Therefore, a reported performance measure is *uninformative* unless we have knowledge of its probability distribution under the null hypothesis. We simply cannot say, by looking at Figure 5.2, whether the performance of the anomaly detector is good or not, for any choice of Δ. We need the probability distributions for this purpose.

The analytical results derived by Donges et al. [DSS+16] for ECA provide probability distributions of the two types of true positives from Table 5.2 and Table 5.3 (PA-AAΔ and PAΔ-AA), under the assumption that the ground-truth anomalies and the predicted anomalies follow independent Bernoulli processes. In this case, both quantities follow simple binomial distributions. Our analytical results from Section 2.2.2 provide the distribution of PAΔ-AA in the more general case where the anomaly score is a stationary process with limited long-range dependencies. Unfortunately, there are no analogous derivations for any other entry from the time-tolerant confusion matrices. In this chapter, we do not use the existing analytical results, but perform Monte Carlo simulations to estimate the required probability distributions without potentially limiting assumptions on the data generating processes.

5.3 Experiments

In the following, we study the behavior of the probability distributions of precision and recall under the assumption of independent processes, within our running example of earthquake detection. For this purpose, we simulate 10,000 independent "ground-truth" event series $\boldsymbol{E}'$ by randomly permuting the actual ground-truth event series $\boldsymbol{E}$. In doing so, we keep the number of ground-truth anomalies constant and assume that they follow a Bernoulli process. We believe that this assumption is reasonable for ground-truth anomalies, which typically occur rarely and are not clustered. We compute time-tolerant confusion matrices and the time-tolerant measures of precision and recall for all permuted event series $\boldsymbol{E}'$ with the anomaly score $\boldsymbol{Z}$ to obtain Monte Carlo estimates of their respective probability distributions.

5.3.1 Visualizations

First, we visualize the precision and recall values obtained from a subset of 100 random permutations for various temporal tolerances and thresholds in Figure 5.3. The visualization also shows the performance measures observed on the non-permuted ground-truth event series.

The observed precision and recall values on the non-permuted ground-truth are generally higher than the values from the randomly permuted event series, especially at larger thresholds. This confirms that the anomaly score contains useful information on earthquake occurrences. However, when the temporal tolerance is increased, the gap between the simulated and the observed performance values tends to shrink: the performance values on the permuted event series increase to a stronger degree than the performance values on the actual ground-truth. The consequence is that reported performance values, in particular when computed with a high temporal tolerance, may not reflect the actual performance of the anomaly detector. In the worst case, one might conclude that the anomaly score allows detection of anomalies that are *statistically independent* of the anomaly score.

5.3.2 Using the distributions

The simulations clearly show that assessment of the statistical significance of the observed performance measures is imperative. For this purpose, we fix the temporal tolerance to $\Delta = 2$ and set the threshold τ to the empirical 0.9-quantile of the anomaly score $\boldsymbol{Z}$. We observe PAΔ-AA = 80 (recall $r_\Delta = 0.49$) and PA-AAΔ = 145

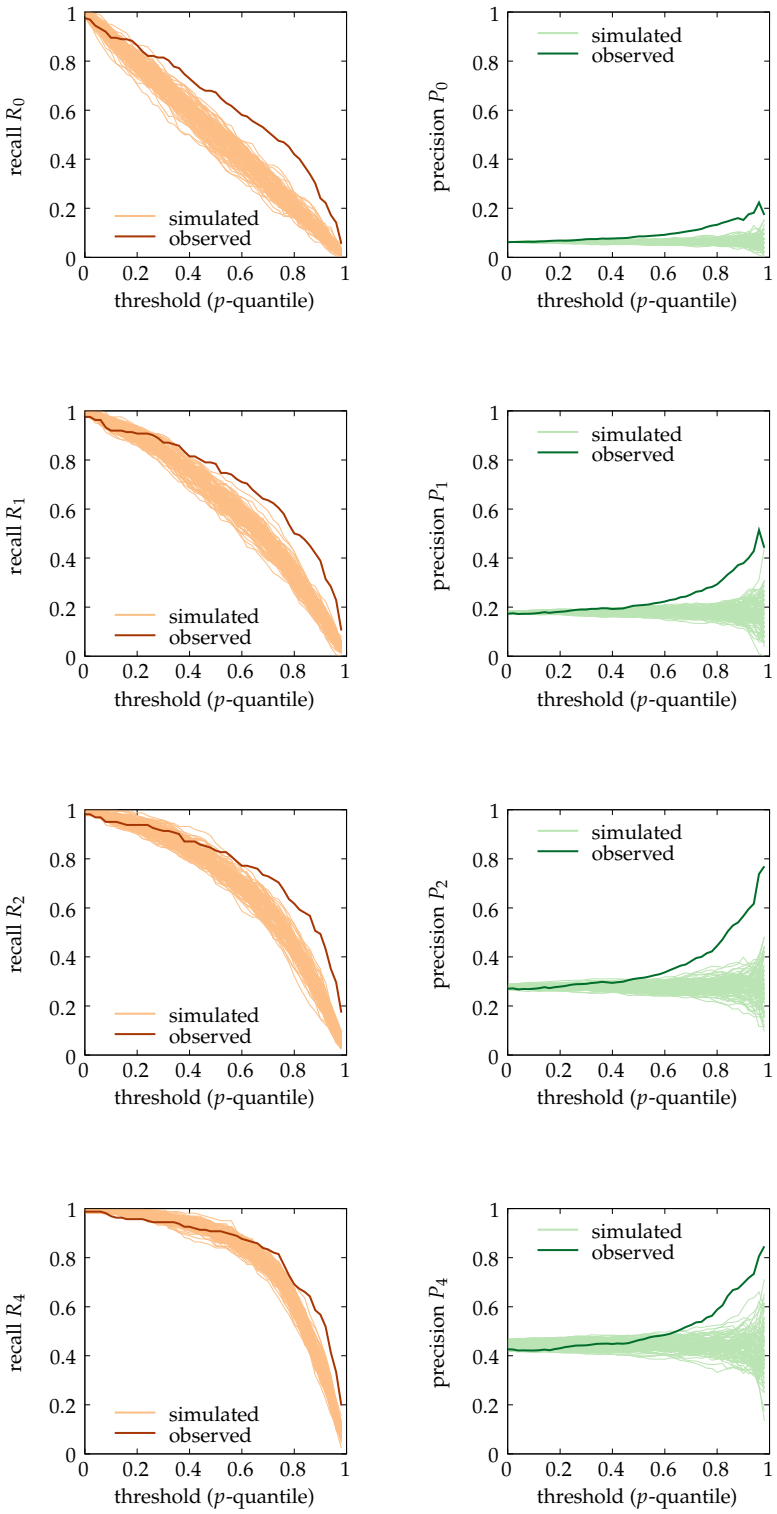

Figure 5.3: Simulated and observed values for the precision P_Δ and recall R_Δ, for $\Delta \in \{0, 1, 2, 4\}$ and thresholds at various p-quantiles.

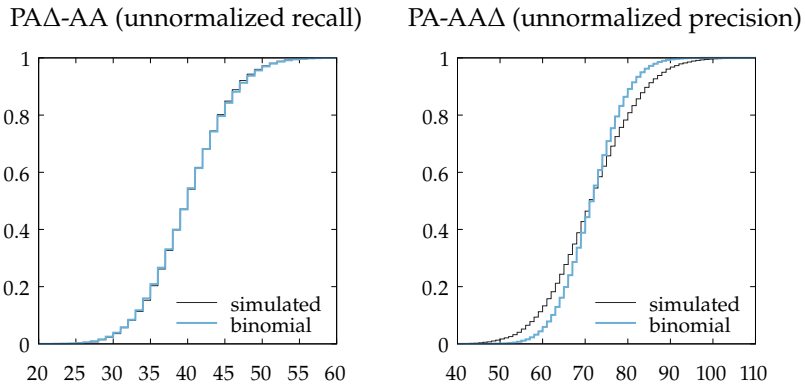

Figure 5.4: Cumulative distribution functions of the two types of true positives required to compute precision and recall, with $\Delta = 2$ and τ set to the empirical .9-quantile of $\mathbf{Z}$.

(precision $p_\Delta = 0.56$) on the non-permuted ground-truth from our example. To assess the statistical significance of the reported numbers, we now have a closer look at the null distributions for the performance measures obtained in the Monte Carlo simulations for this specific parametrization.

Figure 5.4 (simulated) shows the empirical cumulative distribution functions for the numbers of true positives obtained from all 10,000 Monte Carlo simulations for the specific choice of Δ and τ mentioned above. These distributions summarize vertical slices (at the 0.9-quantile) of the corresponding plots in the third row of Figure 5.3, without normalization. Given the simulated distributions, we can easily compute Monte Carlo p-values [DH97] for the numbers of true positives: The Monte Carlo p-value is the fraction of simulations with a true positive value at least as high as the value reported on the actual ground-truth event series. Since *all* simulated values were smaller than the reported values (80 and 145, respectively), we have a highly significant $p < 0.0001$ for both precision and recall for this choice of Δ and τ.

If we compare the reported values of the performance measures with the expected values and variances of the Monte Carlo distributions under the null hypothesis of independence, we also have a means to assess *how good* the reported performance is. We obtain an expected value of 39.87 ± 5.27 for PAΔ-AA and an expected value of 71.42 ± 10.14 for PA-AAΔ, where the range is one standard deviation. The reported values of 80 and 145, respectively, are well outside this range and about twice as good as random guessing.

The analytical null distributions derived in the literature [DSS+16] and in Chapter 2 are all binomial. To complete our analysis, we now check whether the Monte Carlo simulations also yield binomial distributions in our running example. Figure 5.4 (binomial) shows the cumulative distribution functions of binomial random variables when the binomial success probabilities are estimated from our Monte Carlo simulations. The plots suggest that the true positive PAΔ-AA for the recall follows a binomial distribution, whereas the

true positive PA-AAΔ for the precision seems to be overdispersed with respect to the binomial distribution. We have repeated the experiment with different thresholds and temporal tolerances and observed the same behavior across all experiments.

5.4 Conclusions

Statistical evaluation of point-based anomaly detectors with time-tolerant notions of precision and recall is surprisingly similar to the methodology of peak event coincidence analysis described in Chapter 2. We have shown that the time-tolerant measures for precision and recall are computed from two distinct time-tolerant confusion matrices. These time-tolerant confusion matrices can, in principle, be used to derive time-tolerant variants of other well-known measures. When interpreting the input time series and the ground-truth event series as realizations of two stochastic processes, we can estimate the probability distributions for all these measures under the null hypothesis of independence. We have presented a simple way to estimate these null distributions with Monte Carlo simulations, by randomly permuting the ground-truth labels. The null distributions allow assessing the statistical significance of reported results, and provide a means to put the reported results into context by comparing them with the expected values of the respective measures under independence.

Since the probability distributions of the evaluation measures vary with the statistical properties of the anomaly score $\mathbf{Z}$, the detection threshold τ, and the temporal tolerance Δ, we stress that a reported performance value is *completely uninformative* unless its null distribution is provided. Therefore, we believe that providing null distributions for reported precision and recall values should become a community standard in anomaly detection for sequential data. While Monte Carlo simulations are sufficient in most practical cases, future theoretical work should improve the analytical understanding of the null distributions required for this task.

At last, we believe that the statistical perspective on the evaluation of anomaly detectors provides interesting links between *association measures* for event impact analysis and *evaluation measures* for anomaly detectors. In fact, any of the evaluation measures derived from the confusion matrices in Table 5.2 and Table 5.3 could be used as an association measure for event impact analysis. While *true positives*, *i.e.*, the numbers of trigger coincidences and precursor coincidences from Chapter 2, indicate excitatory associations, *false positives* and *false negatives* indicate inhibitory associations between the event series and the time series. Therefore, these quantities can also be used meaningfully in future tests for event impacts.

6 Differentiable Segmentation

This chapter is based on:

[SLM21] Erik Scharwächter, Jonathan Lennartz, and Emmanuel Müller. "Differentiable Segmentation of Sequences." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

The methods proposed in the previous chapters for event impact analysis are all based on the assumption that the time series and the event series are jointly stationary. The notion of event impacts from Definition 1.3.5, the independence tests and association measures developed in Chapter 2 and Chapter 3, as well as the probabilistic warping model from Chapter 4 all require that the distribution of the time series does not change over time. In this final chapter, we focus on event impact analysis for *nonstationary* time series. Given that there are countless ways in which nonstationarity can emerge, we restrict our attention to one of the simplest possible types of nonstationarity one can imagine: a *segmented* time series, where the data-generating process changes its dynamics at specific points in time, but remains stationary within each segment. Another type of nonstationarity—the case of an *integrated* time series with a stochastic trend—has been discussed earlier in Section 1.3.2.

There are two straightforward ways to study event impacts in segmented time series. Given the locations of the change points, we can apply any of the methods introduced in this work *per segment* and make statements about event impacts on the level of (stationary) segments. For example, in Figure 6.1, event occurrences systematically lead to peaks in the second segment of the time series, but not in the other segments. Alternatively, we can test whether the occurrence of change points *itself* is statistically associated with the occurrence of events. In this case, we could apply, *e.g.*, the standard ECA approach of Donges et al. [DSS+16] for event series described in Section 2.2.1, or any other method designed to correlate pairs of event series, and assess to what extent events systematically trigger or precede change points.

The key problem that needs to be solved in both cases is to estimate the locations of the change points. In the following, we propose a solution to this problem that is based on **segmented models** [Mug03]. Segmented models are widely used to describe nonstationary sequential data with discrete change points. They are well-suited for event impact analysis, since they provide not only the locations of the change points, but also a description of the data-generating process within each segment, in the form of a probabilistic model. These descriptions can be used to assess the nature of the event impact individually for each change point, by comparing the models before and after a change.

segmented models

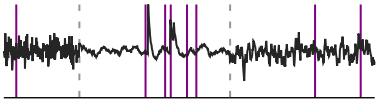

Figure 6.1: Time series with three segments and temporary event impacts only in the second segment.

Same as the probabilistic time warping model from Chapter 4, we study segmented models with applications in event impact analysis in mind, but they can be used in many other contexts. In the following, we provide a generic treatment of the subject without a special focus on event impact analysis.

6.1 Introduction

Nonstationarity is a classical challenge in the analysis of various types of data. A common source of nonstationarity is the presence of change points, where the data-generating process switches its dynamics from one regime to another regime. In some applications, the *detection* of change points is of primary interest, since they may indicate important tipping points in the phenomenon under study [ACH19; BN93; BT65; LXD+15; MJ14; Pag54]. Other applications require *models* for the dynamics within each segment, which may yield more insights into the phenomenon under study and enable predictions. A plethora of segmented models for regression analysis [ADL+16; BP03; Haw76; Ler80; MC70; Mug03] and time series analysis [AH13; DLR06; DXS+16; Ham90] have been proposed in the literature, where the segmentation is either in the data dimensions or the index set. In this work, we study segmented models for *sequential data*, where the data-generating process changes its dynamics at specific *points in time*.

Estimation of classical segmented models requires solving a mixed discrete-continuous optimization problem, where the segmentation is the discrete part and all other model parameters are continuous. Several non-standard estimation algorithms have been developed that are highly specialized for their specific model assumptions. Unfortunately, the dependence on non-standard algorithms for estimation makes it hard to integrate segmented models with highly expressive deep learning architectures that critically depend on *gradient-based* optimization. Therefore, we propose a *relaxed variant* of segmented models that enables joint estimation of all model parameters, including the segmentation function, with gradient descent. Our relaxed model enables analyses of event impacts for segmented time series under very expressive data-generating processes.

In summary, we make the following contributions:

- We formulate a continuous relaxation of segmented models for sequential data that can be estimated with standard algorithms for gradient descent. Our model includes the important class of segmented generalized linear models as a special case, which makes it highly versatile.

- We show that discrete segmentation functions can be replaced by continuous warping functions during the estimation process. As a result, the learnable warping functions proposed recently for sequence alignment [LWT19; WED+19] can be employed within our relaxed segmented model.
- We develop a novel family of warping functions based on the two-sided power (TSP) distribution [KD04; VK02] that is specifically designed for the segmentation task. TSP-based warping functions are differentiable, have simple closed-form expressions that allow fast evaluation, and their parameters correspond with change points.
- We use our approach to model the spread of COVID-19 with Poisson regression, apply it on a change point detection task, and learn classification models with concept drift. The experiments show that our approach effectively solves all these tasks with a standard algorithm for gradient descent.

6.1.1 Problem statement

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)$ be a sequence of T observations, and let $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_T)$ be an additional sequence of covariates used to predict these observations. We assume that observations and covariates are vectors, but they can be scalar-valued as well. The **data-generating process (DGP)** of $\mathbf{X}$ given $\mathbf{Z}$ is time-varying and follows a **segmented model** with $K \ll T$ segments on the time axis. Let τ_k denote the beginning of segment k. We assume that

data-generating process (DGP)

segmented model

$$\mathbf{x}_t \mid \mathbf{z}_t = z_t \overset{\text{iid}}{\sim} f_{\text{DGP}}(z_t, \boldsymbol{\theta}_k), \text{ if } \tau_k \leq t < \tau_{k+1}, \tag{6.1}$$

where the DGP in segment k is parametrized by $\boldsymbol{\theta}_k$. For example, in a segmented Gaussian autoregressive time series of order h, the vector of covariates is $\mathbf{z}_t = (\mathrm{x}_{t-h}, ..., \mathrm{x}_{t-1}, 1)$ and the DGP satisfies $\mathrm{x}_t \mid \mathbf{z}_t = z_t \overset{\text{iid}}{\sim} \text{Normal}(z_t' \boldsymbol{\theta}_k, \sigma^2)$. Although our notation in Equation 6.1 implies a probabilistic DGP, our formalism equally applies to fully deterministic models, which can be viewed as probabilistic models with degenerate distributions.

We express the segmentation of the time axis by a segmentation function $\zeta : \{1, ..., T\} \longrightarrow \{1, ..., K\}$ that maps each time point t to a segment identifier k such that $\zeta(t) = k$ for all $\tau_k \leq t < \tau_{k+1}$. By design, the segmentation function is monotonically increasing with boundary constraints $\zeta(1) = 1$ and $\zeta(T) = K$. Equation 6.1 can now be rewritten as $\mathbf{x}_t \mid \mathbf{z}_t = z_t \overset{\text{iid}}{\sim} f_{\text{DGP}}(z_t, \boldsymbol{\theta}_{\zeta(t)})$. We denote all segment-wise parameters by $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$. The problem is to find a segmentation function ζ as well as segment-wise parameters $\boldsymbol{\Theta}$ that minimize a **loss function** $\mathrm{L}(\zeta, \boldsymbol{\Theta})$, for example, the negative log-likelihood of the observations $\mathbf{X}$.

loss function

6.1.2 Related work

The model from Equation 6.1 is typically studied for change point detection [BW20; TOV20] and modeling of nonstationary time series [Cai94; DLR06; GS99; KL01; RH08; SHJ+16], but is general enough to capture classification models with concept drift [GŽB+13] and segmented generalized linear models [Mug03]. Classical estimation algorithms exploit the fact that model estimation within a segment is often straightforward when the segmentation is known. These approaches decouple the search for an optimal segmentation ζ algorithmically from the estimation of the parameters $\boldsymbol{\Theta}$:

We use the term *sequential data* instead of the term *time series* to reflect this generality in applications.

$$\min_{\zeta,\boldsymbol{\Theta}} \mathrm{L}(\zeta, \boldsymbol{\Theta}) = \min_{\zeta} \min_{\boldsymbol{\Theta}} \mathrm{L}(\zeta, \boldsymbol{\Theta}). \tag{6.2}$$

Various algorithmic search strategies have been explored for the outer minimization of ζ, including grid search [Ler80], dynamic programming [BP03; Haw76], hierarchical clustering [MC70] and other greedy algorithms [ADL+16], some of which come with provable optimality guarantees. These algorithms are often tailored to a specific class of models like piecewise linear regression, and do not generalize beyond. Moreover, the use of non-standard optimization techniques in the outer minimization hinders the integration of such models with highly expressive deep learning architectures, which usually require joint optimization of all model parameters with gradient descent.

6.2 Methodology

We propose a continuous relaxation of the segmented model from Equation 6.1 to enable joint estimation of all parameters with gradient descent. In a nutshell, we replace the discrete segmentation function ζ with a differentiable warping function γ in the segmented model definition.

6.2.1 Relaxed segmented models

We begin by relaxing the segmented model definition to

$$\mathbf{x}_t \mid \mathbf{z}_t = z_t \overset{\text{iid}}{\sim} f_{\text{DGP}}\left(z_t, \hat{\boldsymbol{\theta}}_t\right), \tag{6.3}$$

where we substitute the actual parameter $\boldsymbol{\theta}_k$ of the DGP at time step t in segment k by the predictor $\hat{\boldsymbol{\theta}}_t$. The predictor $\hat{\boldsymbol{\theta}}_t$ is a weighted sum over the individual segment parameters,

$$\hat{\boldsymbol{\theta}}_t := \sum_k \hat{w}_{kt} \boldsymbol{\theta}_k, \tag{6.4}$$

with weights $\hat{w}_{kt} \in [0,1]$ such that $\sum_k \hat{w}_{kt} = 1$ for all t. The weights implicitly define a soft alignment matrix that aligns time steps t and segment identifiers k. For each segmentation function ζ, we can construct an alignment matrix by setting $\hat{w}_{kt} = 1$ if and only if $\zeta(t) = k$. This leads to the original segmented model from Equation 6.1. In our relaxed model, we employ continuous predictors $\hat{\zeta}_t \in [1, K]$ for the values $\zeta(t) \in \{1, \ldots, K\}$. We define the alignment weight $\hat{w}_{kt}$ for segment k and time step t via the difference between the continuous predictor $\hat{\zeta}_t$ and k:

Alignment matrices are no different from the warping matrices seen in Chapter 4. The former align time steps with segments, the latter align time steps with prototype entries.

$$\hat{w}_{kt} := \max\left(0, 1 - \left|\hat{\zeta}_t - k\right|\right) \tag{6.5}$$

The smaller the difference in Equation 6.5, the closer the alignment weight $\hat{w}_{kt}$ will be to 1. With this choice of weights, when $k \leq \hat{\zeta}_t \leq k+1$, the predictor $\hat{\boldsymbol{\theta}}_t$ from Equation 6.4 will be a *linear interpolation* of the parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k+1}$. Higher-order interpolations can be achieved by redefining the weights accordingly.

Warping functions for segmentation

The key question is how to effectively parametrize the continuous predictors $\hat{\zeta}_t$ for the discrete segmentation function. We observe that the continuous analogue of a monotonically increasing segmentation function is a **warping function** [RL98]. Warping functions describe monotonic alignments between closed continuous intervals. Formally, the function $\gamma : [0,1] \longrightarrow [0,1]$ is a warping function if it is monotonically increasing and satisfies the boundary constraints $\gamma(0) = 0$ and $\gamma(1) = 1$. In our relaxed model, we obtain the continuous predictors $\hat{\zeta}_t$ from such a warping function. More precisely, we transform a warping function γ into a sequence of continuous predictors by evaluating γ at T evenly-spaced grid points on the unit interval $[0,1]$ and rescaling the result to the domain $[1, K]$. Let $u_t = (t-1)/(T-1)$ for $t = 1, \ldots, T$ be a grid on the unit interval $[0,1]$. We define the predictors for the segmentation function as

warping function

Similar transformations of warping functions have recently been applied for time series alignment with neural networks [WED+19].

$$\hat{\zeta}_t := 1 + \gamma(u_t) \cdot (K-1). \tag{6.6}$$

The continuous predictors are now fully determined by the warping function γ. An example segmentation function and predictors based on warping functions are shown in Figure 6.2. Apparently, ideal warping functions for segmentation are piecewise constant. The following definition formalizes this observation:

Definition 6.2.1 *The warping function* $\gamma : [0,1] \longrightarrow [0,1]$ ***exactly represents*** *the segmentation function* $\zeta : \{1, \ldots, T\} \longrightarrow \{1, \ldots, K\}$ *with*

exactly represents

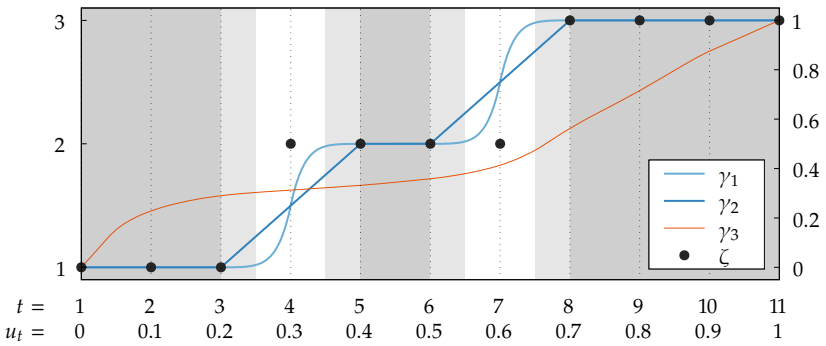

Figure 6.2: Example segmentation function $\zeta(t)$ and warping functions $\gamma_i(u)$. The shaded regions are piecewise constant in γ_1 and γ_2, respectively; γ_3 is strictly increasing.

respect to the unit grid $(u_t)_{t=1}^T$ *if for all* $t = 1, \ldots, T$ *we have*

$$\gamma(u_t) = \frac{k-1}{K-1} \Leftrightarrow \zeta(t) = k. \tag{6.7}$$

Exact representation entails that the predictors $\hat{\zeta}_t$ defined in Equation 6.6 satisfy $\hat{\zeta}_t = \zeta(t)$ for all t. Clearly, exact representation can only be achieved with piecewise-constant warping functions.

Relaxed optimization problem

The relaxed segmented model described above is parametrized by the segment parameters $\boldsymbol{\Theta}$ and the warping function γ. In general, the loss $\mathrm{L_R}$ under the relaxed segmented model is a lower bound for the original loss from Equation 6.2,

$$\min_{\zeta, \boldsymbol{\Theta}} \mathrm{L}(\zeta, \boldsymbol{\Theta}) \geq \min_{\gamma, \boldsymbol{\Theta}} \mathrm{L_R}(\gamma, \boldsymbol{\Theta}). \tag{6.8}$$

Therefore, we propose to solve the original discrete-continuous optimization problem by minimizing the relaxed loss. To obtain a discrete segmentation function from the optimal warping function, the predictors $\hat{\zeta}_t$ can simply be rounded the nearest integers.

In practice, we cannot easily minimize over the space of all warping functions, but have to optimize over a suitable *family* of warping functions. If the family of warping functions is *differentiable* with respect to its parameters, the relaxed model can be estimated with gradient descent. The right-hand side of Equation 6.8 is a continuous relaxation of the discrete-continuous optimization problem on the left-hand side *only* if the family of warping functions used for optimization can represent segmentation functions exactly.

Several families of warping functions have been proposed in the literature and can principally be employed within our model. [Aik91; CSS10; DFH18; FHB+15; GS04; GG04; KSW11; LWT19; RL98; WED+19]. However, none of them contains piecewise-constant functions. Therefore, none of them can exactly represent segmentation functions, which means that the relaxed optimization problem

is—strictly speaking—not a relaxation of the original problem. Below, we define a novel family of piecewise-constant warping functions that is tailored specifically for the segmentation task, with only one parameter per change point.

6.2.2 TSP-based warping functions

Warping functions are similar to cdfs [LWT19]. Cdfs are monotonically increasing, right-continuous, and normalized over their domain [Was04]. If their support is bounded to $[0, 1]$, they satisfy the same boundary constraints as warping functions. Therefore, we can exploit the vast literature on statistical distributions to define and characterize families of warping functions. Our family of warping functions is based on the two-sided power (TSP) distribution [KD04; VK02].

Background: Two-sided power distribution

TSP distribution

The **TSP distribution** models continuous random variables with support $[a, b] \subset \mathbb{R}$. In its most illustrative form, its pdf is unimodal with power-law decay on both sides. Formally, the pdf is

$$f_{\text{TSP}}(u; a, m, b, n) = \begin{cases} \frac{n}{b-a}\left(\frac{u-a}{m-a}\right)^{n-1}, & \text{for } a < u \leq m \\ \frac{n}{b-a}\left(\frac{b-u}{b-m}\right)^{n-1}, & \text{for } m \leq u < b \\ 0, & \text{elsewhere,} \end{cases} \tag{6.9}$$

with $a \leq m \leq b$. The parameters a and b define the boundaries of the support, m is the mode (anti-mode) of the distribution, and $n > 0$ is the power parameter that tapers the distribution. The triangular distribution [JKB94] is the special case with $n = 2$. In the following, we restrict our attention to the unimodal regime with $a < m < b$ and $n > 1$. In this case, the cdf is given by

$$F_{\text{TSP}}(u; a, m, b, n) = \begin{cases} 0, & \text{for } u \leq a \\ \frac{m-a}{b-a}\left(\frac{u-a}{m-a}\right)^{n}, & \text{for } a \leq u \leq m \\ 1 - \frac{b-m}{b-a}\left(\frac{b-u}{b-m}\right)^{n}, & \text{for } m \leq u \leq b \\ 1, & \text{for } b \leq u. \end{cases} \tag{6.10}$$

three-parameter TSP distribution

For convenience, we introduce a **three-parameter TSP distribution** with support restricted to *subintervals* of $[0, 1]$ located around the mode. This variant of the distribution is fully specified by the mode $m \in (0, 1)$, the width $w \in (0, 1]$ of the subinterval, and the power $n > 1$. Depending on the mode and the width, the distribution is symmetric or asymmetric. Intuitively, the three-parameter TSP

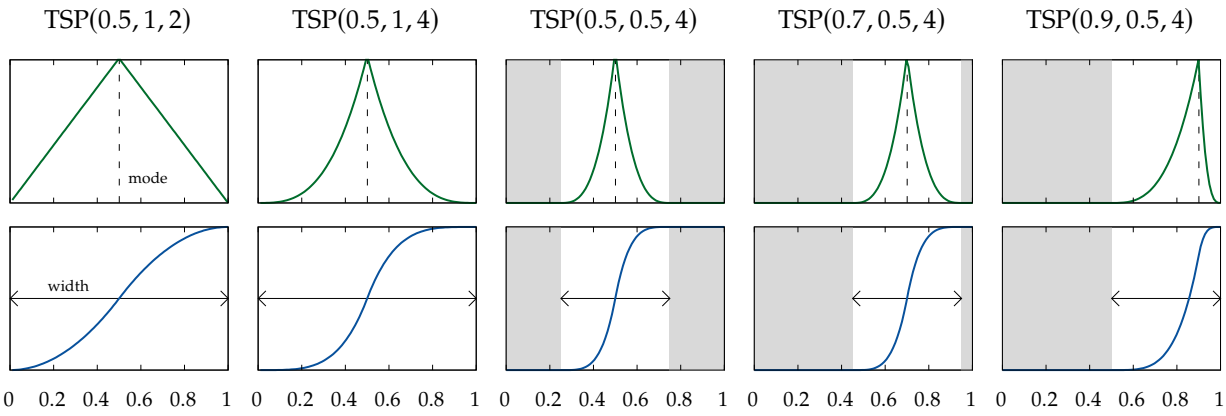

Figure 6.3: Three-parameter variant of the two-sided power distribution TSP(m, w, n) on the interval $[0, 1]$. Dashed lines denote the modes m, arrows the widths w; shaded regions have probability zero. *Top row:* probability density function. *Bottom row:* cumulative distribution function.

distribution describes a symmetric two-sided power kernel of window size w that is located at m and becomes asymmetric only if a symmetric window would exceed the domain $[0, 1]$.

We denote the three-parameter TSP distribution as TSP(m, w, n) and write $f_{\text{TSP}}(u; m, w, n)$ and $F_{\text{TSP}}(u; m, w, n)$ for its pdf and cdf, respectively. Illustrations of the pdf and cdf of the three-parameter TSP distribution for various parametrizations can be found in Figure 6.3. The original TSP parameters a and b are given by

$$a = \max\left(0, \min\left(1 - w, m - \frac{w}{2}\right)\right), \tag{6.11}$$

$$b = \min(1, a + w), \tag{6.12}$$

and yield a unimodal regime. The TSP distribution is a peaked alternative to the beta distribution. An advantage of the TSP distribution over the beta distribution is that its pdf and cdf have closed form expressions that are easy to evaluate computationally. Moreover, they are differentiable almost everywhere.

Warping with mixtures of TSP distributions

TSP-based warping function

We define the **TSP-based warping function** $\gamma_{\text{TSP}} : [0, 1] \longrightarrow [0, 1]$ for K segments as a mixture distribution of $K - 1$ three-parameter TSP distributions. Mixtures of unimodal distributions have step-like cdfs that approximate segmentation functions. We use uniform mixture weights, and treat the width w and power n of the TSP component distributions as hyperparameters. The components differ only in their modes $\boldsymbol{m} = (m_1, \ldots, m_{K-1})$ with $m_k \in (0, 1)$:

$$\gamma_{\text{TSP}}(u; \boldsymbol{m}) := \frac{1}{K - 1} \sum_k F_{\text{TSP}}(u; m_k, w, n). \tag{6.13}$$

We constrain the modes to be strictly increasing, so that γ_{TSP} is identifiable. If the windows around two consecutive modes m_{k-1} and m_k are non-overlapping, then $\gamma_{\text{TSP}}(u; m) = \frac{k-1}{K-1}$ between these windows. Furthermore, $\gamma_{\text{TSP}}(u; m) = 0$ before the first window and $\gamma_{\text{TSP}}(u; m) = 1$ after the last window. Therefore, the family of TSP-based warping functions contains piecewise-constant functions. The functions γ_1 and γ_2 in Figure 6.2 are examples of TSP-based warping functions. In fact, any segmentation function can be exactly represented by a TSP-based warping function:

Lemma 6.2.1 *For every segmentation function ζ, there is a TSP-based warping function γ_{TSP} that exactly represents ζ.*

Proof. We construct γ_{TSP} by placing the $K-1$ modes m_k at the locations of the $K-1$ change points in ζ (projected to the unit grid), and choose a window size w not larger than the grid resolution.

Formally, let τ_{k+1} be beginning of the $(k+1)$-th segment, such that $\zeta(\tau_{k+1} - 1) = k$ and $\zeta(\tau_{k+1}) = k+1$. We place the k-th mode of γ_{TSP} at the beginning of the $(k+1)$-th segment, *i.e.*, we set $m_k := (u_{\tau_{k+1}-1} + u_{\tau_{k+1}})/2$ for all $k = 1, \ldots, K-1$, and use a window size $w < 1/(T-1)$. The power $n > 1$ can be chosen freely. This choice of γ_{TSP} satisfies the conditions from Definition 6.2.1. □

The constructive proof reveals that the modes $\boldsymbol{m} = (m_1, \ldots, m_{K-1})$ of a TSP-based warping function γ_{TSP} correspond with change points in the segmentation function ζ.

When using the family of TSP-based warping functions within the relaxed segmented model, the continuous optimization problem from Equation 6.8 is a proper relaxation of the original discrete-continuous optimization problem. Furthermore, since TSP-based warping functions are differentiable, the relaxed model can be estimated with gradient descent.

6.2.3 Model architecture

We have described all components of the relaxed segmented model. It can use any differentiable family of warping functions to approximate a segmentation function with gradient descent. An overview of the complete model architecture *with TSP-based warping functions* is provided in Table 6.1. To simplify the estimation problem, we rewrite the TSP modes as a normalized cumulative sum,

$$\hat{m}_k := \frac{\sum_{k' \leq k} \exp(\mu_{k'})}{\sum_{k'=1}^{K} \exp(\mu_{k'})} \quad \text{for } k = 1, \ldots, K-1 \tag{6.14}$$

Table 6.1: The relaxed segmented model with TSP-based warping functions.

data-generating process	$\mathbf{x}_t \mid \mathbf{z}_t = \boldsymbol{z}_t \overset{\text{iid}}{\sim} f_{\text{DGP}}\left(\boldsymbol{z}_t, \hat{\boldsymbol{\theta}}_t\right)$	$t = 1, \ldots, T$
parameter predictors	$\hat{\boldsymbol{\theta}}_t := \sum_k \boldsymbol{\theta}_k \max\left(0, 1 - \left\lvert\hat{\zeta}_t - k\right\rvert\right)$	$t = 1, \ldots, T$
segmentation predictors	$\hat{\zeta}_t := 1 + \gamma_{\text{TSP}}\left(\frac{t-1}{T-1}; \hat{\boldsymbol{m}}\right) \cdot (K-1)$	$t = 1, \ldots, T$
TSP mode predictors	$\hat{m}_k := \frac{\sum_{k' \leq k} \exp(\mu_{k'})}{\sum_{k'=1}^{K} \exp(\mu_{k'})}$	$k = 1, \ldots, K-1$
segment parameters	$\boldsymbol{\theta}_k$	$k = 1, \ldots, K$
TSP parameters	μ_k	$k = 1, \ldots, K$

with unconstrained real parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. The transformation of the parameters guarantees that the modes $\hat{\boldsymbol{m}} = (\hat{m}_1, \ldots, \hat{m}_{K-1})$ are strictly increasing and come from the interval $(0, 1)$. The warping function is now overparametrized, since the transformation is invariant to additive terms in the parameters $\boldsymbol{\mu}$. This issue can be resolved by enforcing $\mu_1 := 0$.

The learnable parameters of this architecture are $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ for the DGP and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ for the warping function. The hyperparameters are the number of segments $1 < K \ll T$, and the window size $w \in (0, 1]$ and power $n > 1$ of the TSP distributions. This architecture is a concatenation of simple functions that are either fully differentiable or differentiable almost everywhere. Therefore, all parameters can be learned jointly using gradient descent. As noted in Section 6.2.1, a hard segmentation ζ of the input sequence can be obtained at any time during or after training by rounding $\hat{\zeta}_t$ to the nearest integers.

For effective training with gradient descent, the window size of γ_{TSP} should initially be *larger* than the sampling resolution of the unit grid, $w > 1/(T-1)$, to allow the loss to back-propagate across segment boundaries. The window size can be interpreted as the *receptive field* of the individual TSP components. The window size can be tapered down to $w \leq 1/(T-1)$ over the training epochs to obtain a warping function that exactly represents a segmentation function in the sense of Definition 6.2.1.

6.3 Experiments

Our relaxed segmented model is highly versatile and can be employed for many different tasks. In Section 6.3.1, we study the performance of our approach in estimating a segmented generalized linear model (GLM) [Mug03] on COVID-19 data. In Section

6.3.2, we evaluate our approach on a change point detection benchmark against various competitors. In Section 6.3.3, we apply it on a streaming classification benchmark with concept drift. At last, Section 6.3.4 illustrates potential future applications of our model for discrete representation learning.

We implemented our model in Python* using PyTorch† and optimize the parameters with Adam [KB15]. We employ three different families of warping functions in our relaxed model: nonparametric (NP) [LWT19], CPA-based (CPAb) [WED+19], and our TSP-based functions (TSPb). Source codes for the model and all experiments can be found online at `https://github.com/diozaka/diffseg`.

6.3.1 Poisson regression

Recent work has applied segmented Poisson regression to model COVID-19 case numbers [KGB+20; MSP20]. We follow Küchenhoff et al. [KGB+20] and model daily time series of *newly reported* COVID-19 cases during the first wave of the pandemic in Germany in the year 2020. We obtained official data from Robert Koch Institute‡, the German public health institute. A visualization of the reported data in Figure 6.4 (right plot, bars) reveals nonstationary growth rates and weekly periodicity. Therefore, we use time and a day-of-week indicator as covariates in the model. We tie the coefficients for the day-of-week indicators across all segments, while the daily growth rates and the bias terms differ in every segment.

We estimate a standard segmented Poisson regression model with the baseline algorithm by [Mug03], and our relaxed models (TSPb, CPAb, NP) with gradient descent. We estimate the baseline model and the relaxed models with $K = 2, 4, 8, 16, 32$, and 64 segments. The true number of segments is unknown in this task. Additional details and results can be found in Section 6.A.1.

Figure 6.4 (left plot) shows the goodness-of-fit (log-likelihood) of all models. TSPb consistently reaches the performance of the baseline algorithm. Moreover, TSPb consistently outperforms the other families of warping functions in this experiment. The improvement over NP is particularly large, which indicates that the "nonparametric" warping functions of Lohit, Wang, and Turaga [LWT19] are harder to train than the parametric families. CPAb performs similar to TSPb, but the training time is much longer due to the more complex mathematical operations involved.

* `https://python.org/`
† `https://pytorch.org/`
‡ `https://www.rki.de/`

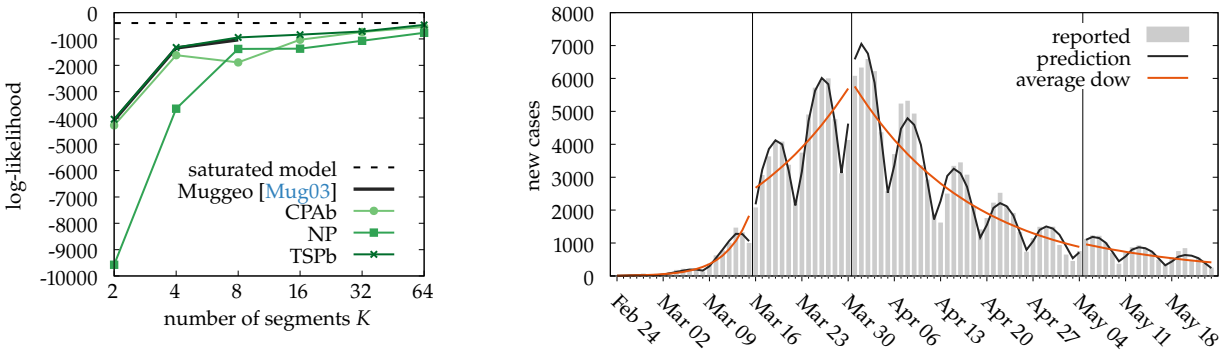

Figure 6.4: Segmented Poisson regression results on COVID-19 case numbers in Germany (2020).

saturated model

We observe that the goodness-of-fit generally grows with the number of segments K and approaches the performance of a **saturated model** (one Poisson distribution per data point). For $K > 8$, the baseline of Muggeo [Mug03] terminates without a model estimate. Figure 6.4 (right plot) visualizes the best TSPb model for $K = 4$ (blue line). We also provide smoothed predictions where the average day of week (dow) effect is incorporated into the bias term to highlight the change of the growth rate from segment to segment (purple line).

In the context of event impact analysis, we could now correlate the epidemic change points with the implementation of lockdown policies and assess the respective event impacts by comparing the growth rates from above from segment to segment.

The three change points are located at 2020-03-16, 2020-03-31, and 2020-05-05. The baseline algorithm of Muggeo [Mug03] detects consistent change points at 2020-03-16 (±0 days), 2020-03-30 (−1 day), and 2020-05-01 (−4 days). Since the reported data is not iid within a segment (it is only conditionally iid given the covariates), other algorithms for change point detection cannot be applied as competitors on this task. Overall, the experiment shows that our model architecture allows effective training of segmented generalized linear models using gradient descent, in particular, when employed with TSPb warping functions.

6.3.2 Change point detection

In the next experiment, we evaluate our relaxed segmented model on a change point detection task with simulated data. Our experimental design exactly follows Arlot, Celisse, and Harchaoui [ACH19]. All details are given in Section 6.A.2. We sample random sequences of length $T = 1000$ with 10 change points at predefined locations. For every segment in every sequence, a distribution is chosen randomly from a set of predefined distributions, and observations within that segment are sampled independently from that distribution. We follow scenario 1 of Arlot, Celisse, and Harchaoui [ACH19], where all predefined distributions have different means and/or variances. An example is shown in Figure 6.5 (left).

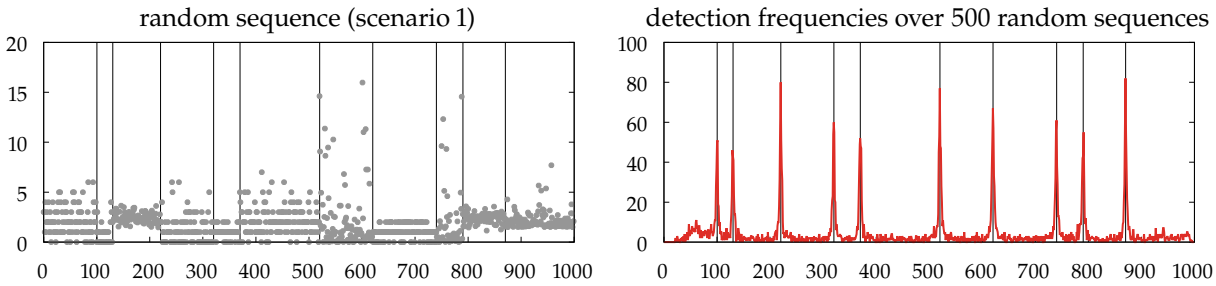

Figure 6.5: Change point detection task of Arlot, Celisse, and Harchaoui [ACH19] with detection results from our model.

Table 6.2: Empirical change point detection results.

algorithm	sensitivity	#CPs (mean±std)	d_{hdf} (mean±std)	d_{fro} (mean±std)	reference
random	-	10	127.8 ± 45.5	3.3 ± 0.2	*baseline*
DP≥1	$\mu\sigma$	10	753.3 ± 120.8	4.1 ± 0.3	[TOV20]
DP≥10	$\mu\sigma$	10	123.6 ± 170.7	**2.1 ± 0.7**	[TOV20]
BinSeg	$\mu\sigma$	10.0 ± 3.1	122.0 ± 97.7	2.5 ± 0.4	[SK74]
PELT	$\mu\sigma$	86.2 ± 26.7	100.0 ± 26.9	8.7 ± 1.5	[KFE12]
NP (*ours*)	$\mu\sigma$	10	88.4 ± 35.7	2.3 ± 0.4	*this work*
CPAb (*ours*)	$\mu\sigma$	10	88.4 ± 32.3	2.7 ± 0.3	*this work*
TSPb (*ours*)	$\mu\sigma$	10	**82.5 ± 32.3**	2.3 ± 0.3	*this work*
E-Divisive	*	5.9 ± 1.9	162.0 ± 108.2	2.2 ± 0.4	[MJ14]
KCP	*	8.6 ± 1.4	67.3 ± 55.1	1.4 ± 0.5	[ACH19]
KCpE	*	10	**33.8 ± 37.6**	**1.2 ± 0.6**	[HC07]

sensitivity: $\mu\sigma$ = mean/variance only, * = full distribution.
#CPs: number of change points; if equal to 10, #CPs is a parameter, otherwise, it is inferred automatically.

We sample $N = 500$ such sequences, with change points at the same locations across all samples. We apply our relaxed segmented model to estimate the change points. We model the data-generating process by a normal distribution with different means and variances in every segment. This design choice makes our approach sensitive *only* to changes in the means and variances of the observed data, and no other distributional characteristics.

We fit our model individually to all N sequences to obtain individual estimates for the change points. Figure 6.5 (right) shows how many times a specific time step was detected as a change point by our approach (with TSPb warping functions). We observe clear peaks at the correct change point locations, which indicates that our model successfully recovers the original segmentations.

Table 6.2 summarizes the empirical detection performance of our approach (TSPb, CPAb and NP) and various competitors, including a baseline where change points are drawn randomly without replacement. It shows the Hausdorff distance d_{hdf} and Frobenius distance d_{fro} between the true segmentations and the detected segmentations (the lower the better).

Among the approaches that detect changes in the mean and variance, our model performs best in terms of d_{hdf} and second best in terms of d_{fro}, regardless of the choice of warping function. Note that all approaches in this group optimize the same objective function (the likelihood of the data under a normal distribution). The dynamic programming approaches (DP$\geq\ell$) [TOV20] exactly find the optimal solution for a predefined number of change points, with a minimum segment length of ℓ. PELT [KFE12] finds the optimal solution with an arbitrary number of change points, while BinSeg [SK74] approximates that solution. Although our gradient-based method finds suboptimal solutions in terms of the likelihood, it produces results with the lowest segmentation costs. This indicates a regularizing effect of the relaxed segmented model that avoids degenerate segmentations. Our approach is only outperformed by algorithms for kernel-based change point detection [ACH19; HC07] that are sensitive towards all distributional characteristics.

6.3.3 Concept drift

A key novelty of our segmented model architecture is that it allows joint training of the segmentation *and any other model component* using gradient descent. As a proof of concept, we apply our model on a classification problem with concept drift [GŽB+13]. The model has to learn the points in time when the target concepts change, and a useful feature transformation for the task. We use the insect stream benchmark of Souza et al. [SRM+20] for this purpose. The task is to classify insects into 6 different species using 33 features collected from an optical sensor. The challenge is that these species behave differently when the air temperature (which is not included as a feature) changes. The benchmark contains multiple data streams, where the air temperature is controlled in different ways (incremental, abrupt, incremental-gradual, incremental-abrupt-reoccurring, incremental-reoccurring). The classifier must adapt the learned concepts depending on the current air temperature.

We employ our relaxed segmented model for softmax regression with the cross entropy loss to obtain segmentations of the data streams. We focus on the five data streams with balanced classes from the benchmark, and measure performance by the classification accuracy. We fit models with $K = 2, 4, 8, \ldots, 128$ segments using TSPb warping functions. We transform the input instances with a fully connected layer followed by a ReLU nonlinearity before passing them to the segmented model. The feature transformation is shared across all segments. We jointly learn the parameters of the segmented model and the feature transformation with gradient descent. Details can be found in Section 6.A.3.

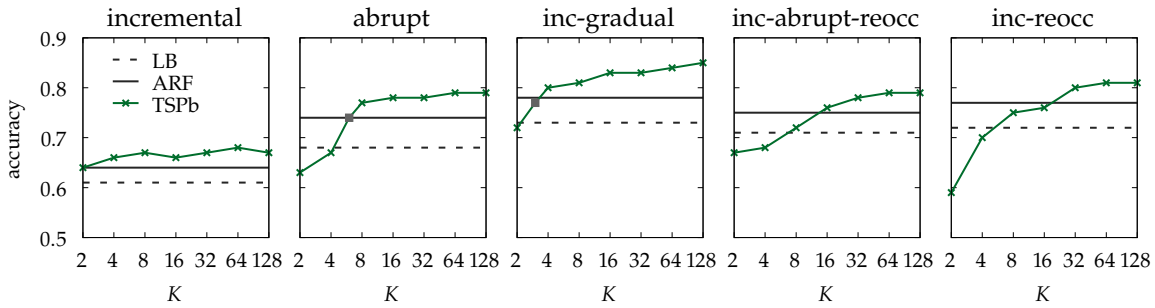

Figure 6.6: Classification performance on the insect stream benchmark of Souza et al. [SRM+20].

Results are visualized in Figure 6.6. In addition to our own results, we include the prequential accuracies of the strongest competitors reported by Souza et al. [SRM+20]: Leveraging Bagging (LB) [BHP10] and Adaptive Random Forests (ARF) [GBR+17].

Our model reaches and outperforms the strongest competitors on all streams from the benchmark, if the number of segments K is large enough to accommodate the concept drift present in the stream. Only the data stream with *abrupt* concept drift satisfies our modeling assumption of a piecewise stationary data generating process. It consists of six segments with constant air temperatures within each segment. The *incremental-gradual* stream has three segments, with mildly varying temperatures in the outer segments and mixed temperatures in the inner segment. The black boxes in their plots show the performances achieved with our approach for $K = 6$ and $K = 3$ segments, respectively. The results indicate that these are in fact the minimum numbers of segments required to obtain competitive performance on these data streams.

6.3.4 Representation learning

At last, we apply our model on a speech signal to showcase its potential for discrete representation learning on the level of phonemes. We assume that the speech signal—represented by a sequence of 12-dimensional MFCC vectors—is piecewise constant within a phoneme. We model it by a minimal DGP with no covariates that simply copies the parameter vectors to the output. See Section 6.A.4 for a complete model description. We fit the model to a single utterance ("choreographer", 10 phonemes) from the TIMIT corpus [GLF+93] by minimizing the mean squared error, and obtain the result visualized in Figure 6.7.

Although the simple DGP does not capture all dynamics of the speech signal, 7 out of 9 phoneme boundaries were correctly identified, with a time tolerance of 20 ms. This minimal experiment suggests that relaxed segmented models, when combined with more powerful DGPs, may be useful for discrete representation

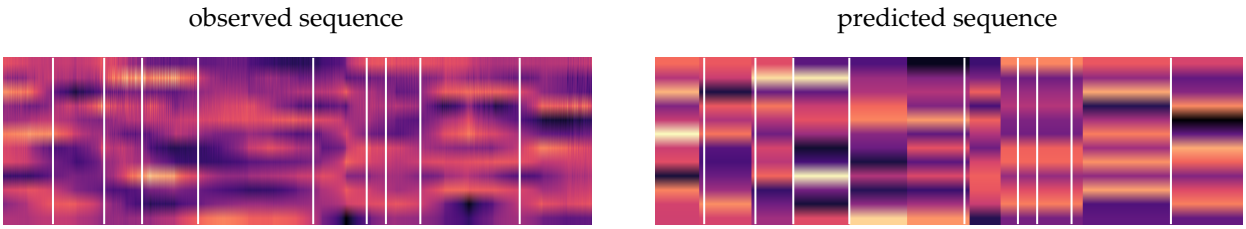

Figure 6.7: Model fit on the utterance "choreographer" with true phoneme boundaries (vertical lines).

learning [FHL+19; OVK17; Rol17], in particular for learning segmental embeddings [CMS+19; KDS16; KSK+20; WLL18]. In fact, our relaxed segmented model may be part of a larger model architecture, where the covariates $\mathbf{z}_t$ and the parameters $\boldsymbol{\theta}_k$ come from some upstream computational layer, and the outputs $\mathbf{x}_t$ are passed on to the next computational layer with an arbitrary downstream loss function. We interpret $\mathbf{z}_t$ as *covariates* and $\boldsymbol{\theta}_k$ as *parameters* to be consistent with prior work on segmented models. It is more accurate to interpret $\mathbf{z}_t$ as *temporal variables* that differ for every time step t, and $\boldsymbol{\theta}_k$ as *segmental variables* that differ for every segment k. The DGP combines the information from both types of variables to produce an output for every time step.

6.4 Conclusions

We have described a novel approach to learn segmented models for nonstationary sequential data with discrete change points. Our relaxed segmented model formulation can use any family of continuous warping functions to approximate a discrete segmentation function. If the family of warping functions is differentiable, our model can be trained with gradient descent. We have introduced the novel family of TSP-based warping functions designed specifically for the segmentation task: it is differentiable, contains piecewise-constant functions that exactly represent segmentation functions, its parameters directly correspond to segment boundaries, and it is simple to evaluate computationally.

The most immediate limitation of our relaxed segmented model that it shares with classical segmented models [BP03; Mug03] is that the number of segments K is a hyperparameter and needs to be chosen prior to model fitting. This issue can be resolved *externally* with any model selection criterion [DLR06]. However, due to our differentiable formulation, future work may also perform model selection *internally* within a single objective function that is optimized with gradient descent, *e.g.*, using the differentiable architecture search approach of Liu, Simonyan, and Yang [LSY19].

A key advantage of our relaxed segmented model formulation is that it enables the integration of state-of-the-art deep learning architectures with segmented models, which makes it highly versatile. The only structure that our model imposes *per se* is that the sequence under study is segmented, and that the individual observations are conditionally iid given some covariates. The experiments on a diverse set of tasks demonstrate the high modeling capacities of our approach, when combined with a suitable model for the data-generating process within each segment.

Due to these modeling capacities, our approach is well-suited for applications in event impact analysis for nonstationary time series. If we believe that an observed time series can be split into stationary segments, we can use our approach to estimate a generic and highly expressive segmented model for the data-generating process. After estimating the model, we can correlate the resulting change points with the event series and compare the data-generating processes between two consecutive segments to assess the nature of the event impact. Alternatively, we can apply any of the methods outlined in this work for event impact analysis in stationary time series within each segment individually.

6.A Details on experiments

We provide additional details, formal model descriptions, and some additional results for the experiments from Section 6.3 below.

6.A.1 Poisson regression

Data

We obtained official data on COVID-19 cases in Germany from Robert Koch Institute (RKI), the German public health institute. The data is publicly available under an open data license. For every day in the study period, we aggregate all new cases *reported* on that day. Due to the delays between the actual infection of a patient and the time the infection is reported to the health authorities, the new cases reported on a specific day contain new infections from several days before.

"Fallzahlen in Deutschland" by Robert Koch Institut (RKI), open data license "Data licence Germany – attribution – Version 2.0" (dl-de/by-2-0), URL: https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74

Segmented Poisson regression model

Poisson regression is a generalized linear model (GLM) for count data, where the data generating process is modeled by a Poisson distribution and the linear predictor is transformed with the logarithmic link function [MN89].

Let x_t denote the number of newly reported cases at time t. Furthermore, let z_t^{Tu}, z_t^{We}, z_t^{Th}, z_t^{Fr}, z_t^{Sa} and z_t^{Su} denote binary day-of-week indicators. If the time step t is a Monday, all indicators are 0. For all other days, exactly one indicator is set to 1. We use the following vector of covariates, with a bogus covariate 1 for the bias terms:

$$z_t = [1, t, z_t^{\text{Tu}}, z_t^{\text{We}}, z_t^{\text{Th}}, z_t^{\text{Fr}}, z_t^{\text{Sa}}, z_t^{\text{Su}}] \tag{6.15}$$

In our segmented Poisson regression model, the bias terms and the daily growth rates (the parameters associated with the covariates 1 and t) differ in every segment, while the parameters for the day-of-week indicators are tied across all segments, *i.e.*, they are independent of the segment identifier k:

$$\boldsymbol{\theta}_k = [\theta_{k,1}, \theta_{k,2} \mid \theta^{\text{Tu}}, \theta^{\text{We}}, \theta^{\text{Th}}, \theta^{\text{Fr}}, \theta^{\text{Sa}}, \theta^{\text{Su}}] \tag{6.16}$$

In this model, the bias terms $\theta_{k,1}$ control the base rates of newly reported cases *on a Monday* within every segment k, while the parameters for the day-of-week indicators are global scaling factors that control the relative increase or decrease of the base rates on

the other days-of-week *with respect to Monday*. Overall, we have the following segmented Poisson regression model:

$$\begin{aligned}\hat{x}_t &= \mathrm{E}_{\mathrm{x}_t|\mathbf{z}_t=\mathbf{z}_t}[\mathrm{x}_t] \\ &:= \exp\left(\theta_{k,1} + \theta_{k,2}\cdot t + \sum_{\text{day}\in\{\text{Tu,We,Th,Fr,Sa,Su}\}} \theta^{\text{day}} z_t^{\text{day}}\right), \end{aligned} \tag{6.17}$$

if $\zeta(t) = k$. The training objective is to minimize the negative log-likelihood under a Poisson distribution:

$$\mathrm{L}(\zeta, \boldsymbol{\Theta}) := -\sum_t \log \mathrm{Poisson}(x_t; \hat{x}_t) \tag{6.18}$$

In our TSP-based warping functions (TSPb), we set the window size to $w = 0.5$ and the power to $n = 16$. In the CPA-based warping functions (CPAb) of Weber et al. [WED+19], we set the dimensionality of the underlying velocity fields to K, the number of segments to learn. This choice makes the function family flexible enough to approximate warping functions with $K-1$ discrete steps, using the minimum number of parameters necessary. Note that the "nonparametric" warping functions (NP) of Lohit, Wang, and Turaga [LWT19] have $T-1$ parameters, where T is the sequence length. We perform training with Adam with a learning rate of $\eta = 0.01$ for a total of 10000 training epochs. In the last 2048 epochs, we round the predictors $\hat{\zeta}_t$ to the nearest integers to obtain a hard segmentation function $\zeta(t)$. We perform 10 restarts of the training procedure with random initialization and keep the model with the best fit for evaluation (highest log-likelihood).

Competitor

We use the reference implementation of Muggeo [Mug03] from the R `segmented` package. As pointed out in Section 6.1.1, our segmented model architecture learns a segmentation of the *index set* $t = 1, \ldots, T$. The segmented models by Muggeo [Mug03] learn a segmentation in the *domain of one (or more) of the covariates*. Since we use the index t as a covariate, we can configure the algorithm of Muggeo [Mug03] to segment the covariate t, which makes the two models equivalent.

Ablation study

Figure 6.8 shows the goodness-of-fit (log-likelihood, the higher the better) obtained with our relaxed segmented model (TSPb) using different settings of the hyperparameters. The performance is quite robust with respect to the width and power of the TSP

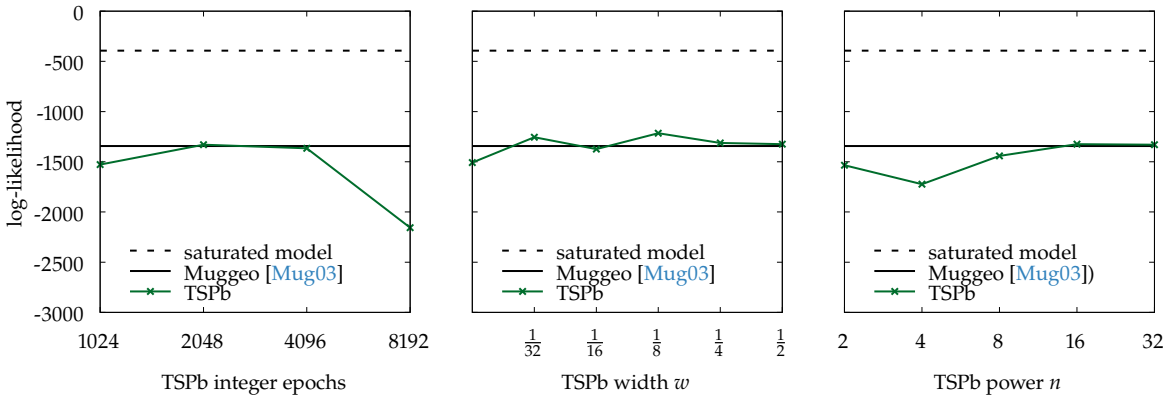

Figure 6.8: Ablation study for TSPb hyperparameters (segmented Poisson regression); default values for the fixed hyperparameters are: 2048 integer epochs, $w = 0.5$, and $n = 16$, respectively.

components. The number of integer epochs (with hard segmentations) has impact on the estimation performance. We hypothesize that the learning problem is simpler with soft segmentations, so that gradient descent moves towards a better region of the loss function. During integer epochs, the parameters within a segment are fine-tuned within the region found during the soft epochs.

6.A.2 Change point detection

Data-generating process

We follow scenario 1 of Arlot, Celisse, and Harchaoui [ACH19] to sample $N = 500$ random sequences of length $T = 1000$ with a total of 10 change points ($K = 11$ segments). The change points are located at time steps 100, 130, 220, 320, 370, 520, 620, 740, 790, and 870. We define a change point as the *beginning* of a new segment. The data-generating process is described in Algorithm 3, where $\mathcal{P}$ is a set of predefined probability distributions. We have

$$\begin{aligned}\mathcal{P} := \{ &\text{Binomial}(n = 10, p = 0.2),\\ &\text{NegativeBinomial}(n = 3, p = 0.7),\\ &\text{Hypergeometric}(M = 10, n = 5, N = 2),\\ &\text{Normal}(\mu = 2.5, \sigma^2 = 0.25),\\ &\text{Gamma}(a = 0.5, b = 5),\\ &\text{Weibull}(a = 2, b = 5),\\ &\text{Pareto}(a = 3, b = 1.5)\},\end{aligned} \tag{6.19}$$

where a is the shape parameter and b is the scale parameter in the case of Gamma, Weibull, and Pareto. For every algorithm in the evaluation, we create a new sample of 500 sequences.

Algorithm 3: Data-generating process of Arlot, Celisse, and Harchaoui [ACH19].

for *all sequences* $n = 1, \ldots, N$ **do**
  for *all segments* $k = 1, \ldots, K$ **do**
    sample a distribution F_{nk} uniformly from $\mathcal{P} \setminus \{F_{n,k-1}\}$;
    sample $\mathrm{x}_{nt} \overset{\text{iid}}{\sim} F_{nk}$ for all time steps t in segment k ;

Segmented normal model

We employ our relaxed segmented model with the assumption that the data-generating process within each segment is a *normal distribution* with its own mean and variance,

$$\mathrm{x}_t \overset{\text{iid}}{\sim} \text{Normal}(\mu_k, \sigma_k^2), \quad \text{if } \zeta(t) = k. \tag{6.20}$$

With this design choice, we can detect changes in the mean and variance between the segments, but no other distributional characteristics. The training objective is the negative log-likelihood:

$$\mathrm{L}(\zeta, \boldsymbol{\Theta}) := -\sum_t \log \text{Normal}(x_t; \mu_{\zeta(t)}, \sigma^2_{\zeta(t)}) \tag{6.21}$$

We ensure a positive variance throughout training by estimating the logarithm of the variance, *i.e.*, the parameter vector within a segment k is given by $\boldsymbol{\theta}_k = [\mu_k, \log\left(\sigma_k^2\right)]$.

For TSPb, we set the window size to $w = 0.125$ and the power to $n = 16$. For CPAb, we set the dimensionality of the velocity field to K. NP has no hyperparameters. We perform training with Adam with a learning rate of $\eta = 0.1$ for a total of 300 epochs with 100 integer epochs. We perform 10 restarts with random initialization and keep the model with the best fit for evaluation.

Competitors

For the experiments with the dynamic programming approaches (DP $\geq \ell$), we use the implementation of Truong, Oudre, and Vayatis [TOV20] from the Python `ruptures` module, using the normal cost function with a minimum segment size of ℓ, 10 change points, and no subsampling. For the binary segmentation approach (BinSeg) by Scott and Knott [SK74] and the Pruned Exact Linear Time (PELT) method of Killick, Fearnhead, and Eckley [KFE12], we use the R `changepoint` package (`cpt.meanvar`, normal test statistic, MBIC penalty [ZS07], minimum segment size 2, maximum number of change points 100). The experiments with the E-divisive approach

[MJ14] were conducted with the R `ecp` package (`e.divisive`, significance level 0.05, 199 random permutations, minimum, minimum segment size 2, moment index 1). For the experiments with the kernel-based approaches KCP [ACH19] and KCpE [HC07], we use the Python package `chapydette` kindly provided by Jones and Harchaoui [JH20], with the default Gaussian-Euclidean kernel, bandwidth 0.1, and minimum segment size 2.

Performance measures

We use the same evaluation measures as Arlot, Celisse, and Harchaoui [ACH19] and follow their definitions. Let ζ and ζ' be two segmentations of a sequence of length T. Let $\mathcal{T}$ and $\mathcal{T}'$ be the corresponding sets of change points. The Hausdorff distance is the largest distance between any change point from one segmentation and its nearest neighbor from the other segmentation:

$$d_{\text{hdf}}(\zeta, \zeta') := \max \left\{ \max_{1 \le i \le |\mathcal{T}|} \min_{1 \le j \le |\mathcal{T}'|} |\tau_i - \tau'_j|, \max_{1 \le j \le |\mathcal{T}|} \min_{1 \le i \le |\mathcal{T}'|} |\tau_i - \tau'_j|, \right\} \tag{6.22}$$

The Frobenius distance between two segmentations [LAB14] is defined as the Frobenius distance between the *rescaled equivalence matrix* representations of the segmentations:

$$d_{\text{fro}}(\zeta, \zeta') := \|\boldsymbol{M}^{\zeta} - \boldsymbol{M}^{\zeta'}\|_F, \tag{6.23}$$

$$\text{where} \quad m^{\zeta}_{t,t'} = \frac{\mathbf{1}_{\zeta(t)=\zeta(t')}}{\sum_{t''} \mathbf{1}_{\zeta(t)=\zeta(t'')}}. \tag{6.24}$$

The Frobenius distance penalizes over-segmentation more strongly than the Hausdorff distance.

Ablation study

We experimented with different values for the TSPb hyperparameters; Table 6.3 shows the results. The detection performance is robust towards the choice of hyperparameters.

6.A.3 Concept drift

Benchmark data

The insect stream benchmark is described in detail in Souza et al. [SRM+20]. The five data streams with balanced class distributions that we consider here have the following lengths. *incremental*:

power n	width w	d_{hdf} (mean±std)	d_{fro} (mean±std)
16	0.5	78.8 ± 31.9	2.4 ± 0.3
16	0.25	83.3 ± 35.1	2.3 ± 0.3
16	0.125	82.5 ± 32.3	2.3 ± 0.3
16	0.0625	85.0 ± 36.1	2.4 ± 0.3
4	0.125	83.0 ± 34.4	2.4 ± 0.3
8	0.125	79.5 ± 31.3	2.3 ± 0.4
16	0.125	82.5 ± 32.3	2.3 ± 0.3
32	0.125	80.4 ± 30.1	2.3 ± 0.3

Table 6.3: Ablation study for TSPb hyperparameters (change detection task).

57,018; *abrupt*: 52,848; *incremental-gradual*: 24,150; *incremental-abrupt-reoccurring*: 79,986; *incremental-reoccurring*: 79,986 (sic).

Softmax regression model

Softmax regression is a standard multi-class classification model, where the data-generating process is modeled by a categorical distribution with probabilities computed from the softmax function. Let $(x_t, \boldsymbol{y}_t)$ denote a training instance from the insect stream benchmark, where $x_t \in \{1, ..., C\}$ is the target class label and $\boldsymbol{y}_t$ is the raw observation. We learn a more informative representation $\boldsymbol{z}_t$ of the observation $\boldsymbol{y}_t$ by passing it through a linear layer with output dimension $D = 8$, followed by a ReLU nonlinearity. We denote this feature transformation by $\boldsymbol{z}_t := g_{\boldsymbol{\phi}}(\boldsymbol{y}_t)$, where $\boldsymbol{\phi}$ are the learnable parameters of the transformation. The covariates $\boldsymbol{z}_t$ are used within a segmented softmax regression model to predict the targets x_t,

$$\mathrm{x}_t \mid \boldsymbol{z}_t = \boldsymbol{z}_t \overset{\text{iid}}{\sim} \text{Categorical}(\text{Softmax}(\boldsymbol{\Theta}_k \boldsymbol{z}_t)), \quad \text{if } \zeta(t) = k. \tag{6.25}$$

In this model, $\boldsymbol{\Theta}_k \in \mathbb{R}^{C \times D}$ is a matrix, so that $\boldsymbol{\Theta}_k \boldsymbol{z}_t \in \mathbb{R}^C$ contains unnormalized classification scores for every class c in segment k. The softmax function transforms the scores into normalized class probabilities. The feature transformation $g_{\boldsymbol{\phi}}$ is shared across all segments, while the parameters of the linear predictor $\boldsymbol{\Theta}_k$ change from segment to segment. With this design, we learn a feature transformation that is useful for the classification task in general, while taking concept drift in the label associations into account. The training objective is to minimize the negative log-likelihood under the categorical distribution, more commonly known as the cross-entropy loss

$$\mathrm{L}(\zeta, \boldsymbol{\Theta}, \boldsymbol{\phi}) = -\sum_t \log \text{Categorical}(x_t; \text{Softmax}(\boldsymbol{\Theta}_{\zeta(t)} \boldsymbol{z}_t)) \tag{6.26}$$

$$= -\sum_t \left(\left[\boldsymbol{\Theta}_{\zeta(t)} \boldsymbol{z}_t \right]_{x_t} - \log \sum_c \exp \left[\boldsymbol{\Theta}_{\zeta(t)} \boldsymbol{z}_t \right]_c \right). \tag{6.27}$$

We employ TSPb warping functions with window size $w = 0.125$ and power $n = 16$. We perform training with Adam with a learning rate of $\eta = 0.1$ for a total of 300 epochs, with 100 integer epochs. We perform 10 restarts with random initialization and keep the model with the best fit for evaluation.

6.A.4 Representation learning

Piecewise constant model

We assume the piecewise constant, deterministic DGP

$$\boldsymbol{x}_t := \boldsymbol{\theta}_k, \quad \text{if } \zeta(t) = k, \tag{6.28}$$

and minimize the mean squared error of the output

$$\mathrm{L}(\zeta, \boldsymbol{\Theta}) := \sum_t \|\boldsymbol{x}_t - \boldsymbol{\theta}_k\|^2. \tag{6.29}$$

We fit our relaxed segmented model with TSPb warping functions ($K = 10$ segments) with window size $w = 0.125$ and power $n = 16$. We perform training with Adam with a learning rate of $\eta = 0.1$ for a total of 300 epochs with 100 integer epochs. We perform 10 restarts with random initialization and keep the model with the best fit for evaluation.

7 Summary and Outlook

In this dissertation, we proposed the first framework to systematically study the impacts of *recurring events* on a time series. The key innovation of our work is to view the time series *and* the event series as stochastic processes, so that event impacts can be formulated in terms of statistical associations and probability distributions. Our framework merges the probabilistic perspective of Granger causality [BS11; Gra69] with existing methods to analyze the impacts of a *singular* event [BSI19; LCL+21; WAG+21]. Intuitively, in the case of randomly recurring events, we say that there are event impacts in a time series if the behavior of the time series within a specific window of interest around an event is not statistically independent of the occurrence of the event. We formalized and characterized this novel notion of event impacts in Chapter 1, to lay a solid foundation for our subsequent contributions.

In the main part of this work, we used our framework to *test*, *measure*, and *model* event impacts in various ways. We provide summaries of our contributions in each of these categories below. Despite our efforts, event impact analysis is still in its infancy, and much work remains to be done. Some of the most pressing open questions are discussed after the respective summaries.

7.1 Tests

The primal question we addressed in Chapter 2 and Chapter 3 was how to develop powerful and computationally efficient test procedures to detect event impacts in a stationary time series. In both chapters, we introduced special cases of event impacts that are easy to interpret and facilitate the detection problem. If a test for these types of event impacts rejects the null hypothesis, this implies event impacts in the general sense of Chapter 1.

In Chapter 2 (Peak Event Coincidence Analysis), we focused on the association between event occurrences and the *occurrence of peaks*, *i.e.*, drastic, sudden and short-lived increases of the values of the time series. We argued that a focus on peaks is quite generic, since many other features of interest can be transformed into peaks by preprocessing the time series with a suitable feature transformation function. We proposed to capture this association formally by testing independence between the maximum statistic of the window of interest and the occurrence of events.

We showed that Event Coincidence Analysis (ECA) [DSS+16] can be used to implement this test. We count how many times an event occurrence coincides with a peak in the time series, *i.e.*, how many times the maximum statistic exceeds a user-defined threshold after an event occurrence. If the number of coincidences is unusually large with respect to an independence assumption, we have evidence for event impacts. Our key contribution in this chapter lies in the derivation of an approximate null distribution for the number of coincidences under independence. For this derivation, we made use of the Extremal Types Theorem, a central result from Extreme Value Theory [Col01]. The ECA-based test strategy depends on the choice of a threshold that defines peaks. In situations where that choice is difficult, it is sensible to test for event impacts using *multiple* thresholds. For this reason, we also provide an approximate joint null distribution for the numbers of coincidences at multiple thresholds, and describe two additional test algorithms to handle multiple thresholds. We validated our derivations in a simulation study and applied our test to assess the impacts of terrorist attacks on Twitter time series.

In Chapter 3 (Multiple Two-Sample Testing), we followed a different test strategy. We argued that many types of event impacts can be detected by considering only *marginal associations* between event occurrences and the behavior of the time series at *individual lags* within the window of interest, instead of the joint behavior of that window at all lags. With the focus on marginal associations, we lose the ability to detect event impacts in the dependency structure of the window of interest. The benefit is that we can detect many types of marginal event impacts in time series over arbitrary domains without applying a feature transformation first. We proposed to capture these associations formally by testing independence between all marginal statistics of the window of interest and the occurrence of events.

Subsequently, we described a novel multiple two-sample testing approach that implements a test for marginal event impacts by comparing the probability distributions across various lags after event occurrences. The choice of two-sample test within our algorithm determines which properties of the distributions are compared, *e.g.*, their means, variances, or complete distribution functions. The latter is possible even for non-numeric data by leveraging recent advancements in kernel-based two-sample testing [GBR+12]. We evaluated the performance of our algorithm against two competitors by performing a large-scale simulation study with a selection of exemplary models for event impacts, and applied it on two real-world datasets.

Future research

The tests from Chapter 2 and Chapter 3 are valid for large classes of stationary time series, but many time series seen in applications are, in fact, **nonstationary**. We discussed two types of nonstationary time series in Chapter 1 that can be studied with our approaches: integrated time series and segmented time series. However, with the existing methods, event impact analysis for other types of nonstationary time series is not possible yet. A feasible direction for future work on nonstationary time series is the scenario where multiple paired realizations of the time series and the event series are available. This scenario may occur, for example, in industrial plants where a repetitive manufacturing process is monitored and subject to events. Tests for event impacts in this scenario could identify whether the occurrence of an event *at a specific time step* has impact on the time series. Similar nonstationary scenarios are currently studied in tests for independent processes [LKB21].

In our test procedures, we assumed that the event series is sparse, so that we can view the windows of interest after individual event occurrences as approximately independent. When facing a **dense event series**, our procedures are not suitable anymore, because the windows of interest are too close to each other or overlap. The sparsification scheme outlined in Chapter 3 may alleviate this problem, but means that a lot of data must be dropped. In fact, with a dense event series, our notions of typical behavior and deviant behavior of the time series are not useful anymore. Instead, we must try to capture event impacts differently using novel test strategies. For example, we may be interested in whether the *number of events* within some interval has influence on the behavior of the time series, or whether the *specific configuration of events* within some interval has influence. These questions have a tendency towards Granger causality [Gra69], and depart from previous approaches to study impacts of singular events.

Finally, in the scenario that we primarily discuss in this dissertation, with a stationary time series and a sparse event series, it is possible to compute **alternative statistics** from the window of interest (other than the maximum and marginal statistics that we use), and implement novel tests for event impacts using these statistics. For example, we could devise a statistic that focuses on the dependency structure within the window of interest. In the end, the choice of statistic is highly application dependent, and new statistics may directly incorporate some computations that we currently regard as preprocessing steps for specific use cases.

7.2 Measures

A secondary question we could only touch in this work was how to measure the association between an event series and a time series in a meaningful way. In Chapter 2, we borrowed the *trigger coincidence rate* from ECA to quantify this association, and computed the fraction of events that trigger a peak in the time series. We also proposed *quantile-trigger rate plots* as a standardized visualization of this measure for peaks at multiple thresholds. Our subsequent discussion of anomaly detectors from Chapter 5 (Statistical Evaluation of Anomaly Detectors) revealed that measures typically used to *evaluate* anomaly detection algorithms may serve as additional measures to quantify the *association* between an event series and a time series. The fundamental challenge that we observed was how to establish the statistical significance of the reported quantities. As a workaround, we proposed a simple Monte Carlo simulation scheme to obtain the null distribution of any measure under independence by shuffling the event series.

Future research

Generally, **analytical derivations** of the null distributions are preferable to Monte Carlo simulations. Therefore, a natural follow-up to our work could focus on these derivations for other measures and specific families of time series and event series. However, statistical significance does not imply practical significance: as pointed out in Chapter 5, an observed value may be considered as statistically significant even if it hardly differs from its expected value under the null distribution. An interesting direction for future work is the development of association measures for event impact analysis that satisfy the properties of **effect sizes** [KP12]. For example, in the context of peak event coincidence analysis, a positive effect size value could indicate an excitatory relationship between events and peaks, a negative effect size value an inhibitory relationship, and a value of 0 the lack of relationship.

7.3 Models

The last question that we covered in this dissertation was how to specify probabilistic models that support event impact analysis. In Chapter 4 (Time Warping Impact Models), we developed a model family for the deviant behavior of a stationary univariate time series in the presence of events. The key idea of our model family is that the individual event occurrences induce the same prototypical pattern in the time series, but this pattern comes

at different delays or is temporally distorted. Our model family captures the data-generating process of the deviant behavior by learning a low-dimensional prototype and a stochastic warping mechanism that explains the temporal distortions. We showed how to instantiate our model family with different warping priors and derived a generic Monte Carlo Expectation-Maximization algorithm for inference. We evaluated the performance of the warping priors in terms of representative power, alignment quality, and discriminative power, on a various datasets.

Finally, in Chapter 6 (Differentiable Segmentation), we focused on a specific nonstationary case. We discussed how to perform event impact analysis when the time series follows a segmented model, where the data-generating process changes at specific points in time, but remains stationary between two change points. We proposed a relaxed variant of the segmented model that enables joint estimation of the change points and the parameters within each segment using standard gradient descent. Our relaxation allows complex models for the data-generating process that could not previously be employed within segmented models, like deep neural networks. We demonstrate the modeling capacity of our relaxed formulation by applying it on a diverse selection of tasks, from Poisson regression to change point detection, classification, and representation learning. After fitting our relaxed segmented model to a nonstationary time series, we can use any of our methods for event impact analysis in the stationary case *individually* within each segment, or correlate the *occurrence of change points* with the occurrence of events, or both.

Future research

We introduced our probabilistic warping model in Chapter 4 for *univariate* stationary time series, but the observations that led to its development equally hold for the **multivariate case**. An additional challenge that arises in the multivariate case is that an underlying prototype may not only be temporally distorted, but also subject to transformations in its data dimensions. The effect is that the deviant behavior after each event occurrence may appear very different from event to event, when in fact it is generated from the same lower-dimensional (both in time and data dimensions) prototype. The state-of-the-art solution for this problem is Canonical Time Warping (CTW) [TNZ+16; ZD16; ZT09], which combines Dynamic Time Warping with Canonical Correlation Analysis. A useful extension of our probabilistic warping model for the multivariate case would include an additional latent transformation of the data dimensions of the multivariate prototype.

Generally, we believe that model-based approaches for event impact analysis are a promising direction for future work. In particular, **event-driven time series models** [CVJ16; YHH+16] that capture the data-generating process of the complete time series using information on event occurrences are useful for forecasting the future behavior of a time series that is subject to event impacts.

7.4 Final note

At last, possibly one of the most pressing and most difficult challenges that is yet to be solved is **causal inference** for event impact analysis. The methods we have proposed in this dissertation are suitable to discover and capture specific statistical associations between event occurrences and the behavior of the time series. They do not immediately allow for causal statements. Adapting the notion of Granger causality, an event series has *causal impact* on a time series if the event series has impact on the time series under the condition that "all the knowledge in the universe available at that time" [Gra80] is considered. We leave this endeavor open for future work, and hope that our contributions in this dissertation provide a useful basis to start from.

Bibliography

[Aba05] Alberto Abadie. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72, 2005, pp. 1–19 (cit. on p. 19).

[ADH11] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. "Synth: An R package for Synthetic Control Methods in Comparative Case Studies." *Journal of Statistical Software* 42(13), 2011 (cit. on p. 20).

[ADH+10] Alberto Abadie, Alexis Diamond, Hainmueller, and Jens. "Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program." *Journal of the American Statistical Association* 105(490), 2010, pp. 493–505 (cit. on p. 20).

[AG03] Alberto Abadie and Javier Gardeazabal. "The economic costs of conflict: A case study of the Basque country." *American Economic Review* 93(1), 2003, pp. 113–132 (cit. on pp. 1, 20).

[AZ18] Abubakar Abid and James Zou. "Autowarp: Learning a Warping Distance from Unlabeled Time Series Using Sequence Autoencoders." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2018 (cit. on p. 81).

[Abr80] Bovas Abraham. "Intervention analysis and multiple time series." *Biometrika* 67(1), 1980, pp. 73–78 (cit. on p. 19).

[ADL+16] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. "Fast algorithms for segmented regression." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2016 (cit. on pp. 114, 116).

[AM17] Roy J. Adams and Benjamin M. Marlin. "Learning time series detection models from temporally imprecise labels." In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017 (cit. on p. 102).

[AM18] Roy J. Adams and Benjamin M. Marlin. "Learning time series segmentation models from temporally imprecise labels." In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018 (cit. on p. 102).

[Aik91] Kiyoaki Aikawa. "Speech recognition using time-warping neural networks." In: *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 1991 (cit. on p. 118).

[Alv14] Hayat Alvi. "The Diffusion of Intra-Islamic Violence And Terrorism: The Impact of the Proliferation of Salafi/Wahhabi Ideologies." *Middle East Review of International Affairs* 18(2), 2014, pp. 38–50 (cit. on p. 50).

[And03] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. John Wiley & Sons, Inc., 2003 (cit. on p. 4).

[ACH19] Sylvain Arlot, Alain Celisse, and Zaïd Harchaoui. "A Kernel Multiple Change-point Algorithm via Model Selection." *Journal of Machine Learning Research* 20, 2019, pp. 1–56 (cit. on pp. 114, 124–126, 132–134).

[AC85] Orley Ashenfelter and David Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *The Review of Economics and Statistics* 67(4), 1985, pp. 648–660 (cit. on p. 19).

[AM14] Atiq-ur-Rehman and Muhammad Irfan Malik. "The modified R a robust measure of association for time series." *Electronic Journal of Applied Statistical Analysis* 7(1), 2014, pp. 1–13 (cit. on p. 23).

[AH13] Alexander Aue and Lajos Horváth. "Structural breaks in time series." *Journal of Time Series Analysis* 34(1), 2013, pp. 1–16 (cit. on p. 114).

[BP03] Jushan Bai and Pierre Perron. "Computation and analysis of multiple structural change models." *Journal of Applied Econometrics* 18(1), 2003, pp. 1–22 (cit. on pp. 114, 116, 128).

[BBS09] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. "Granger causality and transfer entropy are equivalent for gaussian variables." *Physical Review Letters* 103(23), 2009 (cit. on p. 25).

[BB12] Lionel Barnett and Terry Bossomaier. "Transfer Entropy as a Log-Likelihood Ratio." *Physical Review Letters* 109(138105), 2012 (cit. on p. 25).

[BS15] Lionel Barnett and Anil K Seth. "Granger causality for state-space models." *Physical Review E* 040101, 2015 (cit. on p. 24).

[BM12] Jamie Bartlett and Carl Miller. "The edge of violence: Towards telling the difference between violent and non-violent radicalization." *Terrorism and Political Violence* 24(1), 2012, pp. 1–21 (cit. on p. 50).

[BK19] Jozef Baruník and Tobias Kley. "Quantile coherency: A general measure for dependence between cyclical economic variables." *The Econometrics Journal* 22(2), 2019, pp. 131–152 (cit. on p. 22).

[BN93] M Basseville and Igor V Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., 1993 (cit. on p. 114).

[BKM96] David Bell, Jim Kay, and Jim Malley. "A non-parametric approach to non-linear causality testing." *Economics Letters* 51, 1996, pp. 7–18 (cit. on p. 24).

[Ben05] W. Lance Bennett. "Social Movements beyond Borders: Organization, Communication, and Political Capacity in Two Eras of Transnational Activism." In: *Transnational Protest and Global Activism*, 2005. Ed. by Donatella Della Porta and Sidney Tarrow. Rowman & Littlefield Publishers, Inc., 2005, pp. 203–226 (cit. on p. 50).

[Ber83] Brock B. Bernstein. "An Optimum Sampling Design and Power Tests for Environmental Biologists." *Journal of Environmental Management* 16, 1983, pp. 35–43 (cit. on p. 20).

[BDM04] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics* 119(1), 2004, pp. 249–275 (cit. on p. 20).

[BJL+11] Michel Besserve, Dominik Janzing, Nikos K. Logothetis, and Bernhard Schölkopf. "Finding dependencies between frequencies with the kernel cross-spectral density." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011 (cit. on pp. 23, 24).

[BLS13] Michel Besserve, Nikos K Logothetis, and Bernhard Schölkopf. "Statistical analysis of coupled time series with Kernel Cross-Spectral Density operators." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2013 (cit. on pp. 23, 24).

[BHP10] Albert Bifet, Geoff Holmes, and Bernhard Pfahringer. "Leveraging bagging for evolving data streams." In: *Lecture Notes in Computer Science (ECML PKDD)*, 2010 (cit. on p. 127).

[BBH+16] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T. Lizier. *An Introduction to Transfer Entropy*. Springer International Publishing Switzerland, 2016 (cit. on p. 25).

[BSI19] Christian Bottomley, J. Anthony G. Scott, and Valerie Isham. "Analysing Interrupted Time Series with a Control." *Epidemiologic Methods* 8(1), 2019, pp. 1–10 (cit. on pp. 19, 137).

[BD08] Chafik Bouhaddioui and Jean-Marie Dufour. "Tests for Non-Correlation of Two Infinite-Order Cointegrated Vector Autoregressive Series." *Journal of Applied Probability & Statistics* 3(1), 2008, pp. 77–94 (cit. on p. 23).

[BR06] Chafik Bouhaddioui and Roch Roy. "A generalized portmanteau test for independence of two infinite-order vector autoregressive series." *Journal of Time Series Analysis* 27(4), 2006, pp. 505–544 (cit. on p. 23).

[BJR+16] George Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis*. 5th ed. John Wiley & Sons, Inc., 2016 (cit. on pp. 2, 10, 14, 16, 19, 22).

[BT65] George Box and George Tiao. "A Change in Level of a Non-Stationary Time Series." *Biometrika* 52(1/2), 1965, pp. 181–192 (cit. on pp. 1, 18, 19, 114).

[BT75] George Box and George Tiao. "Intervention Analysis with Applications to Economic and Environmental Problems." *Journal of the American Statistical Association* 70(349), 1975, pp. 70–79 (cit. on pp. 1, 19, 27).

[BS11] Steven L Bressler and Anil K Seth. "Wiener-Granger Causality: A well established methodology." *NeuroImage* 58(2), 2011, pp. 323–329 (cit. on pp. 24, 137).

[BJR17] Robert A Bridges, Jessie D Jamieson, and Joel W Reed. "Setting the threshold for high throughput detectors." In: *Proceedings of the IEEE International Conference on Big Data (IEEE BigData)*, 2017 (cit. on p. 102).

[BFF+18] Markus Brill, Till Fluschnik, Vincent Froese, Brijnesh Jain, Rolf Niedermeier, and David Schultz. "Exact mean computation in dynamic time warping spaces." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2018 (cit. on p. 82).

[Bri92] David R. Brillinger. "Nerve Cell Spike Train Data Analysis: A Progression of Technique." *Journal of the American Statistical Association* 87(418), 1992, pp. 260–271 (cit. on p. 30).

[BGK+15] Kay H. Brodersen, Fabian Galluser, Jim Koehler, Nicolas Remy, and Steven L. Scott. "Inferring Causal Impact Using Bayesian Structural Time-Series Models." *The Annals of Applied Statistics* 9(1), 2015, pp. 247–274 (cit. on p. 20).

[Bro99] Carlos D. Brody. "Correlations Without Synchrony." *Neural Computation* 11, 1999, pp. 1537–1551 (cit. on p. 30).

[BKM04] Emery N. Brown, Robert E. Kass, and P. Mitra. "Multiple neural spike train data analysis: state-of-the-art and future challenges." *Nature Neuroscience* 7(5), 2004, pp. 456–461 (cit. on p. 30).

[BV18] Kailash Budhathoki and Jilles Vreeken. "Causal Inference on Event Sequences." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2018 (cit. on p. 30).

[BW20] Gerrit J. J. van den Burg and Christopher K. I. Williams. "An Evaluation of Change Point Detection Algorithms." *arXiv* 2003.06222(stat.ML), 2020 (cit. on p. 116).

[BWS+14] Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. "Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack." *Social Network Analysis and Mining* 4(206), 2014, pp. 1–14 (cit. on pp. 49, 52).

[CMR+19] M. Victoria Caballero-Pintado, Mariano Matilla-García, Jose M. Rodríguez, and Manuel Ruiz Marín. "Two Tests for Dependence (of Unknown Form) between Time Series." *Entropy* 21(9), 2019, p. 878 (cit. on p. 24).

[Cai94] Jun Cai. "A Markov Model of Switching-Regime ARCH." *Journal of Business and Economic Statistics* 12(3), 1994, pp. 309–316 (cit. on p. 116).

[CLX14] T. Tony Cai, Weidong Liu, and Yin Xia. "Two-sample test of high dimensional means under dependence." *Journal of the Royal Statistical Society B* 76(2), 2014, pp. 349–372 (cit. on p. 61).

[CS63] Donald T. Campbell and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin Company, 1963 (cit. on pp. 1, 19).

[CGd11] Jose S. Cánovas, Antonio Guillamón, and María del Carmen Ruíz. "Using permutations to detect dependence between time series." *Physica D: Nonlinear Phenomena* 240(14-15), 2011, pp. 1199–1204 (cit. on p. 24).

[CFH+89] Stephen R. Carpenter, Thomas M. Frost, Dennis Heisey, and Timothy K. Kratz. "Randomized Intervention Analysis and the Interpretation of Whole-Ecosystem Experiments." *Ecological Society of America* 70(4), 1989, pp. 1142–1152 (cit. on p. 20).

[ČB08] Sabina Čehajić and Rupert Brown. "Not in My Name: A Social Psychological Study of Antecedents and Consequences of Acknowledgment of In-Group Atrocities." *Genocide Studies and Prevention* 3(2), 2008, pp. 195–211 (cit. on p. 50).

[CVJ16] Sunandan Chakraborty, Ashwin Venkataraman, and Srikanth Jagabathula. "Predicting Socio-Economic Indicators using News Events." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016 (cit. on p. 142).

[CQ10] Song Xi Chen and Ying Li Qin. "A two-sample test for high-dimensional data with applications to gene-set testing." *Annals of Statistics* 38(2), 2010, pp. 808–835 (cit. on p. 61).

[CRK11] Marc Chesney, Ganna Reshetar, and Mustafa Karaman. "The impact of terrorism on financial markets: An empirical study." *Journal of Banking and Finance* 35(2), 2011, pp. 253–267 (cit. on pp. 1, 17).

[CRK19] Mathieu Chevalier, James C. Russell, and Jonas Knape. "New measures for evaluation of environmental perturbations using Before-After-Control-Impact analyses." *Ecological Applications* 29(2), 2019 (cit. on p. 20).

[CHW16] Lianhua Chi, Bo Han, and Yun Wang. "Open Problem: Accurately Measuring Event Impacts on Time Series." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Mining and Learning from Time Series (KDD MiLeTS)*, 2016 (cit. on pp. 17, 18).

[CSH+16] Lianhua Chi, Saket Sathe, Bo Han, and Yun Wang. "A Novel Method for Assessing Event Impacts on Event-Driven Time Series." In: *Proceedings of the IEEE International Conference on Data Mining, Workshops (ICDMW)*, 2016 (cit. on pp. 12, 17, 18).

[CMS+19] Jan Chorowski, Ricard Marxer, Guillaume Sanchez, and Antoine Laurent. "Unsupervised Neural Segmentation and Clustering for Unit Discovery in Sequential Data." In: *NeurIPS Workshop on Perception as Generative Reasoning*, 2019 (cit. on p. 128).

[CAM+19] Alec P. Christie, Tatsuya Amano, Philip A. Martin, Gorm E. Shackelford, Benno I. Simmons, and William J. Sutherland. "Simple study designs in ecology produce inaccurate estimates of biodiversity responses." *Journal of Applied Ecology* 56(12), 2019, pp. 2742–2754 (cit. on p. 20).

[CGM+03] Darya Chudova, Scott Gaffney, Eric Mjolsness, and Padhraic Smyth. "Translation-invariant mixture models for curve clustering." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003 (cit. on p. 82).

[CGS03] Darya Chudova, Scott Gaffney, and Padhraic Smyth. "Probabilistic models for joint clustering and time-warping of multidimensional curves." In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003 (cit. on pp. 81, 82).

[CG14] Kacper Chwialkowski and Arthur Gretton. "A Kernel Independence Test for Random Processes." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2014 (cit. on p. 22).

[CSS10] Gerda Claeskens, Bernard W. Silverman, and Leen Slaets. "A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 72(5), 2010, pp. 673–694 (cit. on p. 118).

[Col01] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London, Ltd, 2001 (cit. on pp. 2, 27, 37, 39, 42, 56, 138).

[CSB+16] Mary M. Conner, W. Carl Saunders, Nicolaas Bouwes, and Chris Jordan. "Evaluating impacts using a BACI design, ratios, and a Bayesian approach with a focus on restoration." *Environmental Monitoring and Assessment* 188(555), 2016, pp. 1–14 (cit. on p. 20).

[CTM96] F. H. J. Crome, M. R. Thomas, and L. A. Moore. "A novel Bayesian approach to assessing impacts of rain forest logging." *Ecological Applications* 6(4), 1996, pp. 1104–1123 (cit. on p. 20).

[CB17] Marco Cuturi and Mathieu Blondel. "Soft-DTW: a Differentiable Loss Function for Time-Series." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2017 (cit. on pp. 81, 96).

[DV03] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. 2nd ed. Springer-Verlag New York Inc., 2003 (cit. on pp. 2, 10).

[Dar57] D. A. Darling. "The Kolmogorov-Smirnov, Cramer-von Mises Tests." *The Annals of Mathematical Statistics* 28(4), 1957, pp. 823–838 (cit. on p. 60).

[DBK+18] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. *The UCR Time Series Archive*. URL: https://www.cs.ucr.edu/%7B~%7Deamonn/time%7B%5C_%7Dseries%7B%5C_%7Ddata%7B%5C_%7D2018/, 2018 (cit. on pp. 86, 91, 96).

[Dav14] Helen Davidson. *The Guardian: 'Not in my name' campaign organiser warns of danger of political rhetoric*. URL: https://www.theguardian.com/australia-news/2014/oct/03/not-in-my-name-campaign-organiser-warns-of-danger-of-political-rhetoric, 2014 (cit. on p. 50).

[DLR06] Richard A Davis, Thomas C M Lee, and Gabriel A. Rodriguez-Yam. "Structural Break Estimation for Nonstationary Time Series Models." *Journal of the American Statistical Association* 101(473), 2006, pp. 223–239 (cit. on pp. 114, 116, 128).

[DHL+16] Richard A. Davis, Schot H. Holan, Robert Lund, and Nalini Ravishanker. *Handbook of Discrete-Valued Time Series*. CRC Press, Taylor & Francis Group, LLC, 2016 (cit. on pp. 2, 10).

[DH97] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997 (cit. on pp. 44, 110).

[DDV18] Tom De Smedt, Guy De Pauw, and Pieter Van Ostaeyen. *Automatic Detection of Online Jihadist Hate Speech*. Tech. rep. CLiPS Research Center, University of Antwerp, 2018 (cit. on p. 50).

[DS12] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. 4th ed. Pearson Education, Inc., 2012 (cit. on pp. 60, 61).

[DFH18] Nicki Skafte Detlefsen, Oren Freifeld, and Soren Hauberg. "Deep Diffeomorphic Transformer Networks." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 (cit. on p. 118).

[Dew16] Caitlin Dewey. *Washington Post: #StopIslam Twitter-trended for all the right reasons*. URL: https://www.washingtonpost.com/news/the-intersect/wp/2016/03/22/stopislam-twitter-trended-for-all-the-right-reasons/), 2016 (cit. on p. 50).

[DP06] Cees Diks and Valentyn Panchenko. "A new statistic and practical guidelines for nonparametric Granger causality testing." *Journal of Economic Dynamics and Control* 30(9-10), 2006, pp. 1647–1669 (cit. on p. 25).

[DXS+16] Jie Ding, Yu Xiang, Lu Shen, and Vahid Tarokh. "Multiple Change Point Analysis: Fast Implementation And Strong Consistency." In: *ICML Anomaly Detection Workshop*, 2016 (cit. on p. 114).

[Dix02] Philip M. Dixon. "Ripley's K Function." *Encyclopedia of Environmetrics* 3, 2002, pp. 1796–1803 (cit. on p. 30).

[DBW+10] Alex Dmitrienko, Frank Bretz, Peter H. Westfall, James Troendle, Brian L. Wiens, Ajit C. Tamhane, and Jason C. Hsu. "Multiple Testing Methodology." In: *Multiple Testing Problems in Pharmaceutical Statistics*, 2010. Ed. by Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz. Chapman and Hall/CRC, 2010 (cit. on pp. 63, 66).

[DDT+11] J. F. Donges, R. V. Donner, M. H. Trauth, N. Marwan, H.-J. Schellnhuber, and J. Kurths. "Nonlinear detection of paleoclimate-variability transitions possibly related to human evolution." *Proceedings of the National Academy of Sciences* 108(51), 2011, pp. 20422–20427 (cit. on p. 30).

[DSS+16] J. F. Donges, C.-F. Schleusner, J. F. Siegmund, and R. V. Donner. "Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks." *European Physics Journal Special Topics* 225, 2016, pp. 471–487 (cit. on pp. 2, 27, 29–31, 34, 56, 105, 107, 110, 113, 138).

[DEW19] M. A. M. M. van Dortmont, S. van den Elzen, and J. J. van Wijk. "ChronoCorrelator: Enriching events with time series." In: *Proceedings of the Eurographics Conference on Visualization (EuroVis)*, 2019 (cit. on p. 17).

[DGR12] Pierre Duchesne, Kilani Ghoudi, and Bruno Rémillard. "On testing for independence between the innovations of several time series." *The Canadian Journal of Statistics* 40(3), 2012 (cit. on p. 23).

[DR03] Pierre Duchesne and Roch Roy. "Robust Tests For Independence of Two Time Series." *Statistica Sinica* 13(3), 2003, pp. 827–852 (cit. on p. 23).

[DL07] Sandrine Dudoit and Mark J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer Science+Business Media, LLC, 2007 (cit. on pp. 43, 55, 66).

[EBG11] Paul Earle, Daniel Bowden, and Michelle Guy. "Twitter earthquake detection: Earthquake monitoring in a social world." *Annals of Geophysics* 54(6), 2011, pp. 708–715 (cit. on pp. 102, 103).

[EDD17] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. "Graphical Modeling for Multivariate Hawkes Processes with Nonparametric Link Functions." *Journal of Time Series Analysis* 38(2), 2017, pp. 225–242 (cit. on p. 30).

[ER97] Khalid El Himdi and Roch Roy. "Tests for noncorrelation of two multivariate ARMA time series." *The Canadian Journal of Statistics* 25(2), 1997, pp. 233–256 (cit. on p. 23).

[ERD03] Khalid El Himdi, Roch Roy, and Pierre Duchesne. "Tests for Non-Correlation of Two Multivariate Time Series: A Nonparametric Approach." *Mathematical Statistics and Applications: Festschrift for Constance von Eeden* 42, 2003, pp. 397–416 (cit. on p. 23).

[ES93] Walter Enders and Todd Sandler. "The Effectiveness of Antiterrorism Policies: A Vector-Autoregression-Intervention Analysis." *The American Political Science Review* 87(4), 1993, pp. 829–844 (cit. on p. 19).

[FN09] Marcelo Fernandes and Breno Néri. "Nonparametric Entropy-Based Tests of Independence Between Stochastic Processes." *Econometric Reviews* 29(3), 2009, pp. 276–306 (cit. on p. 22).

[FF10] Konstantinos Fokianos and Roland Fried. "Interventions in INGARCH processes." *Journal of Time Series Analysis* 31(3), 2010, pp. 210–225 (cit. on p. 19).

[FF12] Konstantinos Fokianos and Roland Fried. "Interventions in log-linear Poisson autoregression." *Statistical Modelling* 12(4), 2012, pp. 299–322 (cit. on p. 19).

[FHL+19] Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. "SOM-VAE: Interpretable discrete representation learning on time series." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019 (cit. on p. 128).

[FJS05] Gabriel Frahm, Markus Junker, and Rafael Schmidt. "Estimating the tail-dependence coefficient: Properties and pitfalls." *Insurance: Mathematics and Economics* 37, 2005, pp. 80–100 (cit. on p. 30).

[FHB+15] Oren Freifeld, Soren Hauberg, Kayhan Batmanghelich, and John W. Fisher. "Highly-expressive spaces of well-behaved transformations: Keeping it simple." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015 (cit. on p. 118).

[FLE+14] Roland Fried, Tobias Liboschik, Hanan Elsaied, Stella Kitromilidou, and Konstantinos Fokianos. "On Outliers and Interventions in Count Time Series following GLMs." *Austrian Journal of Statistics* 43(3), 2014, pp. 181–193 (cit. on p. 19).

[GS04] Scott Gaffney and Padhraic Smyth. "Joint probabilistic curve clustering and alignment." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2004 (cit. on pp. 82, 118).

[GGA+15] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering Online Hate Speech*. United Nations Educational, Scientific and Cultural Organization (UNESCO), 2015 (cit. on p. 49).

[GSS20] Christopher Galbraith, Padhraic Smyth, and Hal S. Stern. "Quantifying the association between discrete event time series with applications to digital forensics." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 183(3), 2020, pp. 1005–1027 (cit. on p. 30).

[Gal13] Robert G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013 (cit. on p. 4).

[GŽB+13] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. "A survey on concept drift adaptation." *ACM Computing Surveys* 1(1), 2013 (cit. on pp. 116, 126).

[GLF+93] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Philadelphia, 1993 (cit. on p. 127).

[GS00] Xianping Ge and Padhraic Smyth. "Deformable Markov Model Templates for Time-Series." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2000 (cit. on pp. 81, 82).

[GG04] Daniel Gervini and Theo Gasser. "Self-modelling warping functions." *Journal of the Royal Statistical Society B* 66(4), 2004, pp. 959–971 (cit. on p. 118).

[Gew81] John Geweke. "A comparison of tests of the independence of two covariance-stationary time series." *Journal of the American Statistical Association* 76(374), 1981, pp. 363–373 (cit. on p. 23).

[Gew82] John Geweke. "Measurement of Linear Dependence and Feedback between Multiple Time Series." *Journal of the American Statistical Association*, 1982 (cit. on p. 24).

[GL18] Debarghya Ghoshdastidar and Ulrike von Luxburg. "Practical methods for graph two-sample testing." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2018 (cit. on p. 61).

[GMS81] D. Gillings, D. Makuc, and E. Siegel. "Analysis of interrupted time series mortality trends: An example to evaluate regionalized perinatal care." *American Journal of Public Health* 71(1), 1981, pp. 38–46 (cit. on p. 19).

[Gla72] Gene V. Glass. "Estimating the Effects of Intervention Into a Non-stationary Time-Series." *American Educational Research Journal* 9(3), 1972, pp. 463–477 (cit. on p. 19).

[GBR+17] Heitor M. Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfharinger, Geoff Holmes, and Talel Abdessalem. "Adaptive random forests for evolving data stream classification." *Machine Learning* 106(9-10), 2017, pp. 1469–1495 (cit. on p. 127).

[GG61] Leo A. Goodman and Yehuda Grunfeld. "Some Nonparametric Tests for Comovements between Time Series." *Journal of the American Statistical Association* 56(293), 1961, pp. 11–26 (cit. on p. 22).

[Gra69] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37(3), 1969, pp. 424–438 (cit. on pp. 2, 24, 66, 137, 139).

[Gra80] C. W.J. Granger. "Testing for causality. A personal viewpoint." *Journal of Economic Dynamics and Control* 2(C), 1980, pp. 329–352 (cit. on p. 142).

[GSR+16] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." *European Journal of Epidemiology* 31(4), 2016, pp. 337–350 (cit. on p. 39).

[GBR+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J Smola. "A Kernel Two-Sample Test." *Journal of Machine Learning Research* 13, 2012, pp. 723–773 (cit. on pp. 59, 61, 67, 138).

[GBR+06] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. "A kernel method for the two-sample-problem." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2006 (cit. on p. 61).

[GBS+05] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. "Measuring statistical dependence with Hilbert-Schmidt norms." In: *Lecture Notes in Computer Science (Algorithmic Learning Theory)*, 2005 (cit. on p. 22).

[GHF+09] Arthur Gretton, Zaid Harchaoui, Kenji Fukumizu, and Bharath K. Sriperumbudur. "A Fast, Consistent Kernel Two-Sample Test." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2009 (cit. on p. 61).

[GSS+12] Arthur Gretton, Bharath Sriperumbudur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, and Kenji Fukumizu. "Optimal kernel choice for large-scale two-sample tests." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2012 (cit. on p. 61).

[GMT+96] Lalit Gupta, Dennis L. Molfese, Ravi Tammana, and Panagiotis G Simos. "Nonlinear Alignment and Averaging for Estimating the Evoked Potential." *IEEE Transactions on Biomedical Engineering* 43(4), 1996, pp. 348–356 (cit. on p. 82).

[GS99] Valery Guralnik and Jaideep Srivastava. "Event detection from time series data." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1999 (cit. on p. 116).

[HP91] Marc Hallin and Madan L. Puri. *Rank Tests for Time Series Analysis*. Tech. rep. IMA Preprint Series #876, 1991 (cit. on p. 22).

[HS05] Marc Hallin and Abdessamad Saidi. "Testing non-correlation and non-causality between multivariate ARMA time series." *Journal of Time Series Analysis* 26(1), 2005, pp. 83–105 (cit. on p. 23).

[HS07] Marc Hallin and Abdessamad Saidi. "Optimal tests of noncorrelation between multivariate time series." *Journal of the American Statistical Association* 102(479), 2007, pp. 938–951 (cit. on p. 23).

[Ham90] James D. Hamilton. "Analysis of time series subject to changes in regime." *Journal of Econometrics* 45(1-2), 1990, pp. 39–70 (cit. on p. 114).

[Ham94] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994 (cit. on pp. 16, 68).

[HLO+16] Heejoon Han, Oliver Linton, Tatsushi Oka, and Yoon Jae Whang. "The cross-quantilogram: Measuring quantile dependence and testing directional predictability between time series." *Journal of Econometrics* 193(1), 2016, pp. 251–270 (cit. on p. 22).

[HC07] Zaïd Harchaoui and Olivier Cappé. "Retrospective Multiple Change-Point Estimation With Kernels." In: *IEEE Workshop on Statistical Signal Processing (SSP)*, 2007 (cit. on pp. 125, 126, 134).

[Hau76] Larry D. Haugh. "Checking the Independence of Two Covariance-Stationary Time Series: A Univariate Residual Cross-Correlation Approach." *Journal of the American Statistical Association* 71(354), 1976, pp. 378–385 (cit. on p. 23).

[Haw76] Douglas M. Hawkins. "Point Estimation of the Parameters of Piecewise Regression Models." *Journal of the Royal Statistical Society C* 25(1), 1976 (cit. on pp. 114, 116).

[Hem16] Chris Hemmings. *The Independent: From #StopIslam to Allison Pearson and Katie Hopkins, the social media response to Brussels has been shocking*. URL: https://www.independent.co.uk/voices/from-stopislam-to-allison-pearson-and-katie-hopkins-the-social-media-response-to-brussels-has-been-a6946116, 2016 (cit. on p. 50).

[Hen88] Norbert Henze. "A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences." *The Annals of Statistics* 16(2), 1988, pp. 772–783 (cit. on p. 61).

[HJ94] Craig Hiemstra and Jonathan D. Jones. "Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation." *The Journal of Finance* 49(5), 1994, pp. 1639–1664 (cit. on p. 25).

[Hon96] Yongmiao Hong. "Testing for independence between two covariance stationary time series." *Biometrika* 83(3), 1996, pp. 615–625 (cit. on p. 23).

[Hon01] Yongmiao Hong. "Testing for independence between two stationary time series via the empirical characteristic function." *Annals of Economics and Finance* 2(1), 2001, pp. 123–164 (cit. on p. 23).

[Hot31] Harold Hotelling. "The Generalization of Student's Ratio." *The Annals of Mathematical Statistics* 2(3), 1931 (cit. on p. 60).

[HTR03] Nicholas P. Hughes, Lionel Tarassenko, and Stephen J. Roberts. "Markov Models for Automated ECG Interval Analysis." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2003 (cit. on pp. 81, 82).

[HM00] Bradley E. Huitema and Joseph W. McKean. "Design specification issues in time-series intervention models." *Educational and Psychological Measurement* 60(1), 2000, pp. 38–58 (cit. on p. 19).

[HM03] John D. Hunter and John G. Milton. "Amplitude and frequency dependence of spike timing: Implications for dynamic regulation." *Journal of Neurophysiology* 90(1), 2003, pp. 387–394 (cit. on p. 30).

[JR19] Carsten Jentsch and Lena Reichmann. "Generalized binary time series models." *Econometrics* 7(4), 2019, pp. 1–26 (cit. on p. 10).

[JKB94] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions, vol. 2*. John Wiley & Sons, Inc., 1994 (cit. on p. 119).

[JKB97] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Discrete multivariate distributions*. John Wiley & Sons, Inc., 1997 (cit. on pp. 85, 99).

[JH20] Corinne Jones and Zaïd Harchaoui. "End-to-end Learning for Retrospective Change-point Estimation." In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2020 (cit. on p. 134).

[KS02] A. Kaiser and T. Schreiber. "Information transfer in continuous processes." *Physica D: Nonlinear Phenomena* 166(1-2), 2002, pp. 43–62 (cit. on p. 25).

[KQP+16] Janani Kalyanam, Mauricio Quezada, Barbara Poblete, and Gert Lanckriet. "Prediction and characterization of high-activity events in social media triggered by real-world news." *PLoS ONE* 11(12), 2016, pp. 1–13 (cit. on pp. 1, 17).

[KP12] Ken Kelley and Kristopher J. Preacher. "On effect size." *Psychological Methods* 17(2), 2012, pp. 137–152 (cit. on p. 140).

[KP00] Eamonn Keogh and Michael Pazzani. "Scaling up dynamic time warping for datamining applications." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2000 (cit. on p. 81).

[KP01] Eamonn Keogh and Michael Pazzani. "Derivative Dynamic Time Warping." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2001 (cit. on p. 81).

[KK20] Waqar Muhammad Khan and Asad ul Islam Khan. "Most stringent test of independence for time series." *Communications in Statistics - Simulation and Computation* 49(11), 2020, pp. 2808–2826 (cit. on p. 23).

[KMP19] Soheil Khorram, Melvin G. McInnis, and Emily Mower Provost. "Trainable Time Warping: Aligning Time-series in the Continuous-time Domain." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019 (cit. on pp. 81, 82, 96).

[KFE12] R. Killick, P. Fearnhead, and I. a. Eckley. "Optimal detection of changepoints with a linear computational cost." *Journal of the American Statistical Association* 107(500), 2012, pp. 1590–1598 (cit. on pp. 125, 126, 133).

[KL05] Eunhee Kim and Sangyeol Lee. "A test for independence of two stationary infinite order autoregressive processes." *Annals of the Institute of Statistical Mathematics* 57(1), 2005, pp. 105–127 (cit. on p. 23).

[KPG+11] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N. Brown. "A Granger causality measure for point process models of ensemble neural spiking activity." *PLoS Computational Biology* 7(3), 2011 (cit. on p. 30).

[KS06] Seyoung Kim and Padhraic Smyth. "Segmental hidden Markov models with random effects for waveform modeling." *Journal of Machine Learning Research* 7, 2006, pp. 945–969 (cit. on p. 82).

[KSL04] Seyoung Kim, Padhraic Smyth, and Stefan Luther. "Modeling waveform shapes with random effects segmental hidden Markov models." In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004 (cit. on p. 82).

[KB15] Diederik P. Kingma and Jimmy Lei Ba. "Adam: A method for stochastic optimization." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015 (cit. on p. 123).

[KKK+20] Matthias Kirchler, Shahryar Khorasani, Marius Kloft, and Christoph Lippert. "Two-sample testing using deep learning." In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020 (cit. on p. 61).

[KG92] Alois Kneip and Theo Gasser. "Statistical Tools to Analyze Data Representing a Sample of Curves." *The Annals of Statistics* 20(3), 1992, pp. 1266–1305 (cit. on pp. 81, 82).

[KR08] Alois Kneip and James O. Ramsay. "Combining registration and fitting for functional models." *Journal of the American Statistical Association* 103(483), 2008, pp. 1155–1165 (cit. on p. 82).

[KY86] Paul D. Koch and Shie Shien Yang. "A method for testing the independence of two time series that accounts for a potential pattern in the cross-correlation function." *Journal of the American Statistical Association* 81(394), 1986, pp. 533–544 (cit. on p. 23).

[KL01] Jens Kohlmorgen and Steven Lemm. "A dynamic HMM for on-line segmentation of sequential data." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2001 (cit. on p. 116).

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. The MIT Press, 2009 (cit. on p. 4).

[Kol41] A. Kolmogoroff. "Confidence Limits for an Unknown Distribution Function." *The Annals of Mathematical Statistics* 12(4), 1941, pp. 461–463 (cit. on p. 60).

[KDS16] Lingpeng Kong, Chris Dyer, and Noah A. Smith. "Segmental recurrent neural networks." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016 (cit. on p. 128).

[KDS+15] Evangelos Kontopantelis, Tim Doran, David A. Springate, Iain Buchan, and David Reeves. "Regression based quasi-experimental approach when randomisation is not an option: Interrupted time series analysis." *BMJ* 350(h2750), 2015 (cit. on p. 19).

[KD04] Samuel Kotz and Johan René van Dorp. *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific Publishing Co. Pte. Ltd., 2004 (cit. on pp. 115, 119).

[KSK+20] Felix Kreuk, Yaniv Sheena, Joseph Keshet, and Yossi Adi. "Phoneme Boundary Detection using Learnable Segmental Features." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020 (cit. on p. 128).

[Kri17] Murali Krishnan. *Deutsche Welle: 'Not in my name' - Indians protest vigilante attacks on Muslims*. URL: https://www.dw.com/en/not-in-my-name-indians-protest-vigilante-attacks-on-muslims/a-39456019, 2017 (cit. on p. 50).

[KGB+20] Helmut Küchenhoff, Felix Günther, Andreas Bender, and Michael Höhle. *Analyse der Epidemischen Covid-19 Kurve in Bayern durch Regressionsmodelle mit Bruchpunkten*. Tech. rep. Munich, Germany: Ludwigs-Maximilians-Universität München, Germany, 2020 (cit. on p. 123).

[KSW11] Sebastian Kurtek, Anuj Srivastava, and Wei Wu. "Signal estimation under random time-warpings and nonlinear signal alignment." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2011 (cit. on pp. 82, 118).

[LT19] Virginia Lacal and Dag Tjøstheim. "Estimating and Testing Nonlinear Local Dependence Between Two Time Series." *Journal of Business and Economic Statistics* 37(4), 2019, pp. 648–660 (cit. on p. 23).

[LAB14] Rémi Lajugie, Sylvain Arlot, and Francis Bach. "Large-margin metric learning for constrained partitioning problems." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2014 (cit. on p. 134).

[LKB21] Felix Laumann, Julius von Kügelgen, and Mauricio Barahona. "Kernel two-sample and independence tests for non-stationary random processes." *Engineering Proceedings* 5(31), 2021 (cit. on pp. 24, 139).

[Lea06] Erik G. Learned-Miller. "Data Driven Image Manifolds through Continous Joint Alignment." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(2), 2006 (cit. on p. 82).

[Ler80] P. M. Lerman. "Fitting Segmented Regression Models by Grid Search." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29(1), 1980, pp. 77–84 (cit. on pp. 114, 116).

[LCL+21] Lihua Li, Meaghan S. Cuerden, Bian Liu, Salimah Shariff, Arsh K. Jain, and Madhu Mazumdar. "Three statistical approaches for assessment of intervention effects: A primer for practitioners." *Risk Management and Healthcare Policy* 14, 2021, pp. 757–770 (cit. on pp. 19, 137).

[LXD+15] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. "M-Statistic for Kernel Change-Point Detection." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2015 (cit. on p. 114).

[LKF+14] Tobias Liboschik, Pascal Kerschke, Konstantinos Fokianos, and Roland Fried. "Modelling interventions in INGARCH processes." *International Journal of Computer Mathematics* 93(4), 2014, pp. 640–657 (cit. on p. 19).

[Lin15] Ariel Linden. "Conducting interrupted time-series analysis for single- and multiple-group comparisons." *Stata Journal* 15(2), 2015, pp. 480–500 (cit. on p. 19).

[LA11] Ariel Linden and John L. Adams. "Applying a propensity score-based weighting model to interrupted time series data: Improving causal inference in programme evaluation." *Journal of Evaluation in Clinical Practice* 17(6), 2011, pp. 1231–1238 (cit. on p. 19).

[LNR+05] Jennifer Listgarten, Radford M. Neal, Sam T. Roweis, and Andrew Emili. "Multiple Alignment of Continuous Time Series." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2005 (cit. on pp. 81, 82).

[LXL+20] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and D. J. Sutherland. "Learning Deep Kernels for Non-Parametric Two-Sample Tests." In: *International Conference on Machine Learning (ICML)*, 2020 (cit. on p. 61).

[LSY19] Hanxiao Liu, Karen Simonyan, and Yiming Yang. "DARTS: Differentiable architecture search." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019 (cit. on p. 128).

[LWT19] Suhas Lohit, Qiao Wang, and Pavan Turaga. "Temporal Transformer Networks: Joint Learning of Invariant and Discriminative Time Warping." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019 (cit. on pp. 82, 96, 115, 118, 119, 123, 131).

[LCG17] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. "Interrupted time series regression for the evaluation of public health interventions: A tutorial." *International Journal of Epidemiology* 46(1), 2017, pp. 348–355 (cit. on p. 19).

[LCG18] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. "The use of controls in interrupted time series studies of public health interventions." *International Journal of Epidemiology* 47(6), 2018, pp. 2082–2093 (cit. on p. 19).

[LSG18] James Lopez Bernal, S. Soumerai, and Antonio Gasparrini. "A methodological framework for model selection in interrupted time series studies." *Journal of Clinical Epidemiology* 103, 2018, pp. 82–91 (cit. on p. 19).

[LO17] David Lopez-Paz and Maxime Oquab. "Revisiting classifier two-sample tests." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017 (cit. on p. 61).

[LLL+14] Chen Luo, Jian-Guang Lou, Qingwei Lin, Qiang Fu, Rui Ding, Dongmei Zhang, and Zhe Wang. "Correlating events with time series for incident diagnosis." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014 (cit. on pp. 1, 17, 18).

[MSK+19] Lovish Madaan, Ankur Sharma, Praneet Khandelwal, Shivank Goel, Parag Singla, and Aaditeshwar Seth. "Price forecasting & anomaly detection for agricultural commodities in India." In: *Proceedings of the Conference on Computing and Sustainable Societies (COMPASS)*, 2019 (cit. on p. 17).

[MDA15] Walid Magdy, Kareem Darwish, and Norah Abokhodair. "Quantifying Public Response towards Islam on Twitter after Paris Attacks." *arXiv* 1512.04570(cs.SI), 2015 (cit. on pp. 50, 52).

[MEB+16] Stephen Makonin, Bradley Ellert, Ivan V. Bajić, and Fred Popowich. "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014." *Scientific Data* 3, 2016, pp. 1–12 (cit. on p. 72).

[MVS+15] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. "Long Short Term Memory Networks for Anomaly Detection in Time Series." In: *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2015 (cit. on pp. 28, 102).

[MV21] Ričards Marcinkevičs and Julia E. Vogt. "Interpretable Models for Granger Causality Using Self-Explaining Neural Networks." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021 (cit. on p. 24).

[MPS08] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. "Kernel method for nonlinear Granger causality." *Physical Review Letters* 144103, 2008 (cit. on p. 24).

[MRM10] Mariano Matilla-García, José Miguel Rodríguez, and Manuel Ruiz Marín. "A symbolic test for testing independence between time series." *Journal of Time Series Analysis* 31(2), 2010, pp. 76–85 (cit. on p. 24).

[MHL12] Marwan A. Mattar, Allen R. Hanson, and Erik G. Learned-Miller. "Unsupervised Joint Alignment and Clustering using Bayesian Nonparametrics." In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012 (cit. on p. 82).

[MRL09] Marwan A. Mattar, Michael G. Ross, and Erik G. Learned-Miller. "Nonparametric Curve Alignment." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009 (cit. on p. 82).

[MJ14] David S. Matteson and Nicholas A. James. "A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data." *Journal of the American Statistical Association* 109(505), 2014, pp. 334–345 (cit. on pp. 114, 125, 134).

[MN89] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. 2nd. Chapman and Hall/CRC, 1989 (cit. on p. 130).

[MMM+80] David McDowall, Richard McCleary, Errol E. Meidinger, and Richard A. (jr.) Hay. *Interrupted Time Series Analysis*. SAGE Publications, Inc., 1980 (cit. on p. 19).

[MC70] Victor E. McGee and Willard T. Carleton. "Piecewise regression." *Journal of the American Statistical Association* 65(331), 1970, pp. 1109–1124 (cit. on pp. 114, 116).

[Mug03] Vito M.R. Muggeo. "Estimating regression models with unknown break-points." *Statistics in Medicine* 22(19), 2003, pp. 3055–3071 (cit. on pp. 113, 114, 116, 122–124, 128, 131, 132).

[MSP20] Vito M.R. Muggeo, Gianluca Sottile, and Mariano Porcu. *Modelling COVID-19 outbreak: Segmented Regression to Assess Lockdown Effectiveness*. Tech. rep. ResearchGate, 2020 (cit. on p. 123).

[Mur00] Paul A. Murtaugh. "Paired Intervention Analysis in Ecology." *Journal of Agricultural, Biological, and Environmental Statistics* 5(3), 2000, pp. 280–292 (cit. on p. 20).

[Mur02] Paul A. Murtaugh. "On Rejection Rates of Paired Intervention Analysis." *Ecology* 83(6), 2002, pp. 1752–1761 (cit. on p. 20).

[Mur03] Paul A. Murtaugh. "On Rejection Rates of Paired Intervention Analysis: Reply." *Ecology* 84(10), 2003, pp. 2799–2802 (cit. on p. 20).

[Nat18] National Consortium for the Study of Terrorism and Responses to Terrorism (START). *Global Terrorism Database [2014-2017]. Retrieved from https://www.start.umd.edu/gtd*. 2018 (cit. on p. 49).

[NRJ+09] Aatira G. Nedungadi, Govindan Rangarajan, Neeraj Jain, and Mingzhou Ding. "Analyzing multiple spike trains with nonparametric granger causality." *Journal of Computational Neuroscience* 27, 2009, pp. 55–64 (cit. on p. 30).

[NHK+11] Yoshihiko Nishiyama, Kohtaro Hitomi, Yoshinori Kawasaki, and Kiho Jeong. "A consistent nonparametric test for nonlinear causality." *Journal of Econometrics* 165(1), 2011, pp. 112–127 (cit. on p. 25).

[OD20] Adrian Odenweller and Reik V. Donner. "Disentangling synchrony from serial dependency in paired-event time series." *Physical Review E* 101(5), 2020 (cit. on p. 30).

[OCB+18] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. "The Effect of Extremist Violence on Hateful Speech Online." In: *Proceedings of the International Conference on Web and Social Media (ICWSM)*, 2018 (cit. on pp. 49, 50, 52).

[OVK17] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural discrete representation learning." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2017 (cit. on p. 128).

[Pag54] E. S. Page. "Continuous Inspection Schemes." *Biometrika* 41(1/2), 1954 (cit. on p. 114).

[Par18] Kun Il Park. *Fundamentals of probability and stochastic processes with applications to communications*. Springer International Publishing AG, 2018 (cit. on p. 4).

[PKG11] François Petitjean, Alain Ketterlin, and Pierre Gançarski. "A global averaging method for dynamic time warping, with applications to clustering." *Pattern Recognition* 44(3), 2011, pp. 678–693 (cit. on p. 82).

[Pet76] A. N. Pettitt. "A Two-Sample Anderson-Darling Rank Statistic." *Biometrika* 63(1), 1976, pp. 161–168 (cit. on p. 60).

[PRC03] Dinh Tuan Pham, Roch Roy, and Lyne Cédras. "Tests for non-correlation of two cointegrated ARMA time series." *Journal of Time Series Analysis* 24(5), 2003, pp. 553–577 (cit. on p. 23).

[PH77] David A. Pierce and Larry D. Haugh. "Causality in temporal systems: Characterizations and a survey." *Journal of Econometrics* 5(3), 1977, pp. 265–293 (cit. on p. 24).

[Pow07] David M. W. Powers. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Tech. rep. Adelaide, Australia: Flinders University of South Australia, 2007 (cit. on p. 106).

[PDM16] Thomas Prätzlich, Jonathan Driedger, and Meinard Müller. "Memory-restricted Multiscale Dynamic Time Warping." In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016 (cit. on p. 81).

[QKG02] R. Quian Quiroga, T. Kreuz, and P. Grassberger. "Event synchronization: A simple and fast method to measure synchronicity and time delay patterns." *Physical Review E* 66(4), 2002 (cit. on p. 30).

[Rab89] Lawrence R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77(2), 1989, pp. 257–286 (cit. on p. 90).

[RTC17] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. "On wasserstein two-sample testing and related families of nonparametric tests." *Entropy* 19(2), 2017, pp. 1–15 (cit. on p. 61).

[RWD+15] A. Rammig, M. Wiedermann, J. F. Donges, F. Babst, W. Von Bloh, D. Frank, K. Thonicke, and M. D. Mahecha. "Coincidences of climate extremes and anomalous vegetation responses: Comparing tree ring patterns to simulated productivity." *Biogeosciences* 12(2), 2015, pp. 373–385 (cit. on p. 30).

[RL98] James O. Ramsay and Xiaochun Li. "Curve registration." *Journal of the Royal Statistical Society B* 60(2), 1998, pp. 351–363 (cit. on pp. 82, 117, 118).

[RS05] James O. Ramsay and Bernard W. Silverman. *Functional data analysis*. 2nd. Springer Science+Business Media, LLC, 2005 (cit. on p. 82).

[RK04] Chotirat Ann Ratanamahatana and Eamonn Keogh. "Making Time-series Classification More Accurate Using Learned Constraints." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2004 (cit. on p. 81).

[RXW+19] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. "Time-Series Anomaly Detection Service at Microsoft." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019 (cit. on pp. 28, 102).

[RBM+16] Aljoscha Rheinwalt, Niklas Boers, Norbert Marwan, and Jürgen Kurths. "Non-linear time series analysis of precipitation events using regional climate networks for Germany." *Climate Dynamics* 46, 2016, pp. 1065–1074 (cit. on p. 30).

[RF15] Michael W. Robbins and Thomas J. Fisher. "Cross-Correlation Matrices for Tests of Independence and Causality Between Two Multivariate Time Series." *Journal of Business and Economic Statistics* 33(4), 2015, pp. 459–473 (cit. on p. 23).

[RC04] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. 2nd ed. Springer Science+Business Media Inc., 2004 (cit. on p. 90).

[RH08] Joshua W. Robinson and Alexander J. Hartemink. "Non-stationary dynamic bayesian networks." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2008 (cit. on p. 116).

[Rol17] Jason Tyler Rolfe. "Discrete variational autoencoders." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017 (cit. on p. 128).

[Ros06] Jeffrey S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific Publishing Co. Pte. Ltd., 2006 (cit. on p. 4).

[SHJ+16] Ardavan Saeedi, Matthew Hoffman, Matthew Johnson, and Ryan Adams. "The Segmented iHMM: A simple, efficient hierarchical infinite HMM." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2016 (cit. on p. 116).

[Sai07] Abdessamad Saidi. "Consistent testing for non-correlation of two cointegrated ARMA time series." *The Canadian Journal of Statistics* 35(1), 2007, pp. 169–188 (cit. on p. 23).

[SC78] Hiroaki Sakoe and Seibi Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26(1), 1978, pp. 43–49 (cit. on pp. 80, 81).

[SDK07] Suchi Saria, Andrew Duchi, and Daphne Koller. "Discovering Deformable Motifs in Continuous Time Series Data." In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007 (cit. on p. 82).

[Sar18] Nicholas V. Sarlis. "Statistical significance of earth's electric and magnetic field variations preceding earthquakes in Greece and Japan revisited." *Entropy* 20(8), 2018 (cit. on p. 30).

[SLM21] Erik Scharwächter, Jonathan Lennartz, and Emmanuel Müller. "Differentiable Segmentation of Sequences." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021 (cit. on p. 113).

[SM20a] Erik Scharwächter and Emmanuel Müller. "Does Terrorism Trigger Online Hate Speech? On the Association of Events and Time Series." *Annals of Applied Statistics* 14(3), 2020, pp. 1285–1303. DOI: 10.1214/20-AOAS1338 (cit. on p. 27).

[SM20b] Erik Scharwächter and Emmanuel Müller. "Statistical Evaluation of Anomaly Detectors for Sequences." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Mining and Learning from Time Series (KDD MiLeTS)*, 2020 (cit. on p. 101).

[SM20c] Erik Scharwächter and Emmanuel Müller. "Two-Sample Testing for Event Impacts in Time Series." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2020. DOI: 10.1137/1.9781611976236.2 (cit. on pp. 57, 63).

[SM22] Erik Scharwächter and Emmanuel Müller. "Discrete Probabilistic Models for Time Warping." In: *Manuscript in review*, 2022 (cit. on p. 79).

[SMD+16] Erik Scharwächter, Emmanuel Müller, Jonathan Donges, Marwan Hassani, and Thomas Seidl. "Detecting change processes in dynamic networks by frequent graph evolution rule mining." In: *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2016 (cit. on p. 59).

[Sch86] Mark F Schilling. "Multivariate Two-Sample Tests Based on Nearest Neighbours." *Journal of the American Statistical Association* 81(395), 1986, pp. 799–806 (cit. on p. 61).

[SDD+16] Carl-Friedrich Schleussner, Jonathan F. Donges, Reik V. Donner, and Hans Joachim Schellnhuber. "Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries." *Proceedings of the National Academy of Sciences* 113(33), 2016, pp. 9216–9221 (cit. on p. 30).

[SS87] F. W. Scholz and M. A. Stephens. "K-sample Anderson–Darling tests." *Journal of the American Statistical Association* 82(399), 1987, pp. 918–924 (cit. on p. 77).

[Sch00] Thomas Schreiber. "Measuring information transfer." *Physical Review Letters* 85(2), 2000, pp. 461–464 (cit. on pp. 25, 66).

[Sch89] Gabi Schulgen. "A measure of similarity for response curves based on ranks." *Statistics in Medicine* 8(11), 1989, pp. 1401–1411 (cit. on p. 22).

[SK74] A. J. Scott and M. Knott. "A Cluster Analysis Method for Grouping Means in the Analysis of Variance." *Biometrics* 30(3), 1974, pp. 507–512 (cit. on pp. 125, 126, 133).

[SL21] Byeongchan Seong and Kiseop Lee. "Intervention analysis based on exponential smoothing methods: Applications to 9/11 and COVID-19 effects." *Economic Modelling* 98, 2021 (cit. on p. 19).

[Sha03] Jun Shao. *Mathematical Statistics*. 2nd ed. Springer Science+Business Media, LLC, 2003 (cit. on p. 4).

[Sha09] Xiaofeng Shao. "A generalized portmanteau test for independence between two stationary time series." *Economic Theory* 25, 2009, pp. 195–210 (cit. on p. 23).

[SSH+16] Jonatan F. Siegmund, Tanja G. M. Sanders, Ingo Heinrich, Ernst van der Maaten, Sonia Simard, Gerhard Helle, and Reik V. Donner. "Meteorological Drivers of Extremes in Daily Stem Radius Variations of Beech, Oak, and Pine in Northeastern Germany: An Event Coincidence Analysis." *Frontiers in Plant Science* 7(733), 2016 (cit. on p. 30).

[SSD17] Jonatan F. Siegmund, Nicole Siegmund, and Reik V. Donner. "CoinCalc—A new R package for quantifying simultaneities of event series." *Computers and Geosciences* 98, 2017, pp. 64–72 (cit. on p. 30).

[SWD+16] Jonatan F. Siegmund, Marc Wiedermann, Jonathan F. Donges, and Reik V. Donner. "Impact of temperature and precipitation extremes on the flowering dates of four German wildlife shrub species." *Biogeosciences*, 2016 (cit. on p. 30).

[SFT+17] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet. "Anomaly detection in streams with extreme value theory." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017 (cit. on p. 102).

[SHA+20] Ben Silver, Xinyue He, Steve R. Arnold, and Dominick V. Spracklen. "The impact of COVID-19 control measures on air quality in China." *Environmental Research Letters* 15(084021), 2020 (cit. on p. 1).

[Sil95] Bernard W. Silverman. "Incorporating Parametric Effects into Functional Principal Components Analysis." *Journal of the Royal Statistical Society B* 57(4), 1995, pp. 673–689 (cit. on p. 82).

[Sim77] Dean Keith Simonton. "Cross-Sectional Time-Series Experiments: Some Suggested Statistical Analyses." *Psychological Bulletin* 84(3), 1977, pp. 489–502 (cit. on p. 19).

[Smi14] Eric P. Smith. "BACI Design." *Wiley StatsRef: Statistics Reference Online*, 2014 (cit. on p. 20).

[SRM+20] Vinicius M.A. Souza, Denis M. dos Reis, André G. Maletzke, and Gustavo E.A.P.A. Batista. "Challenges in benchmarking stream learning algorithms with real-world data." *Data Mining and Knowledge Discovery* 34(6), 2020, pp. 1805–1858 (cit. on pp. 126, 127, 134).

[Sta11] Richard P. Stanley. *Enumerative Combinatorics, Volume 1*. 2nd ed. Cambridge University Press, 2011 (cit. on p. 83).

[Ste03] Allan Stewart-Oaten. "On rejection rates of paired intervention analysis: Comment." *Ecology* 84(10), 2003, pp. 2795–2802 (cit. on p. 20).

[SBO92] Allan Stewart-Oaten, James R. Bence, and Craig W. Osenberg. "Assessing Effects of Unreplicated Perturbations: No Simple Solutions." *Ecology* 73(4), 1992, pp. 1396–1404 (cit. on p. 20).

[SMP86] Allan Stewart-Oaten, William W. Murdoch, and Keith R. Parker. "Environmental Impact Assessment: "Pseudoreplication" in Time?" *Ecology* 67(4), 1986, pp. 929–940 (cit. on pp. 1, 20).

[Stu08] Student. "The Probable Error of a Mean." *Biometrika* 6(1), 1908, pp. 1–25 (cit. on p. 60).

[SLZ+19] Ya Su, Rong Liu, Youjian Zhao, Wei Sun, Chenhao Niu, and Dan Pei. "Robust anomaly detection for multivariate time series through stochastic recurrent neural network." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019 (cit. on p. 102).

[SB14] Jie Sun and Erik M. Bollt. "Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings." *Physica D: Nonlinear Phenomena* 267, 2014, pp. 49–57 (cit. on p. 63).

[TBE14] Abderrahim Taamouti, Taoufik Bouezmarni, and Anouar El Ghouch. "Nonparametric estimation and inference for conditional density based Granger causality measures." *Journal of Econometrics* 180(2), 2014, pp. 251–264 (cit. on p. 25).

[TO18] Corentin Tallec and Yann Ollivier. "Can recurrent neural networks warp time?" In: *International Conference on Learning Representations (ICLR)*, 2018 (cit. on p. 82).

[THF+18] Chang Wei Tan, Matthieu Herrmann, Germain Forestier, Geoffrey I. Webb, and François Petitjean. "Efficient search of the best warping window for Dynamic Time Warping." In: *Proceedings of the SIAM International Conference on Data Mining (SIAM SDM)*, 2018 (cit. on p. 81).

[TLZ+18] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. "Precision and recall for time series." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2018 (cit. on p. 102).

[TBB+05] Antonio Terlizzi, Lisandro Benedetti-Cecchi, Stanislao Bevilacqua, Simonetta Fraschetti, Paolo Guidetti, and Marti J. Anderson. "Multivariate and univariate asymmetrical analyses in environmental impact assessment: a case study of Mediterranean subtidal sessile assemblages." *Marine Ecology Progress Series* 289, 2005, pp. 27–42 (cit. on p. 20).

[TKO+17] Lauric Thiault, Laëtitia Kernaléguen, Craig W. Osenberg, and Joachim Claudet. "Progressive-Change BACIPS: a flexible approach for environmental impact assessment." *Methods in Ecology and Evolution* 8(3), 2017, pp. 288–296 (cit. on p. 20).

[Tjø96] Dag Tjøstheim. "Measures of dependence and tests of independence." *Statistics* 28(3), 1996, pp. 249–284 (cit. on p. 22).

[TOS18] Dag Tjøstheim, Håkon Otneim, and Bård Støve. "Statistical dependence: Beyond Pearson's rho." *arXiv* 1809.10455(math.ST), 2018 (cit. on p. 22).

[Tor06] Simon Tormey. "'Not in my name': Deleuze, Zapatismo and the critique of representation." *Parliamentary Affairs* 59(1), 2006, pp. 138–154 (cit. on p. 50).

[TNZ+16] George Trigeorgis, Mihalis A. Nicolaou, Stefanos Zafeiriou, and Björn Schuller. "Deep Canonical Time Warping." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016 (cit. on p. 141).

[TOV20] Charles Truong, Laurent Oudre, and Nicolas Vayatis. "Selective review of offline change point detection methods." *Signal Processing* 167, 2020 (cit. on pp. 116, 125, 126, 133).

[Und92] A. J. Underwood. "Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world." *Journal of Experimental Marine Biology and Ecology* 161(2), 1992, pp. 145–178 (cit. on p. 20).

[Und94] A. J. Underwood. "On Beyond BACI: Sampling Designs that Might Reliably Detect Environmental Disturbances." *Ecological Applications* 4(1), 1994, pp. 3–15 (cit. on p. 20).

[VK02] J. René Van Dorp and Samuel Kotz. "The standard two-sided power distribution and its properties: With applications in financial engineering." *American Statistician* 56(2), 2002, pp. 90–99 (cit. on pp. 115, 119).

[VSK+10] S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. "Graph Kernels." *Journal of Machine Learning Research* 11, 2010, pp. 1201–1242 (cit. on p. 59).

[WSZ+02] A. K. Wagner, S. B. Soumerai, F. Zhang, and D. Ross-Degnan. "Segmented regression analysis of interrupted time series studies in medication use research." *Journal of Clinical Pharmacy and Therapeutics* 27(4), 2002, pp. 299–309 (cit. on p. 19).

[WK19] Heiko Wagner and Alois Kneip. "Nonparametric registration to low-dimensional function spaces." *Computational Statistics and Data Analysis* 138, 2019, pp. 49–63 (cit. on p. 82).

[WLL18] Yu-Hsuan Wang, Hung-yi Lee, and Lee Lin-shan. "Segmental Audio Word2Vec: Representing Utterances as Sequences of Vectors with Applications in Spoken Term Detection." In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018 (cit. on p. 128).

[WG97] Kongming Wang and Theo Gasser. "Alignment of curves by dynamic time warping." *The Annals of Statistics* 25(3), 1997, pp. 1251–1276 (cit. on p. 82).

[WG99] Kongming Wang and Theo Gasser. "Synchronizing sample curves nonparametrically." *The Annals of Statistics* 27(2), 1999, pp. 439–460 (cit. on p. 82).

[WMP+16] Yizhi Wang, David J. Miller, Kira Poskanzer, Yue Wang, Lin Tian, and Guoqiang Yu. "Graphical Time Warping for Joint Alignment of Multiple Curves." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2016 (cit. on pp. 81, 82).

[War14] Russell T. Warne. "A primer on multivariate analysis of variance (MANOVA) for behavioral scientists." *Practical Assessment, Research and Evaluation* 19(17), 2014, pp. 1–10 (cit. on p. 77).

[Was04] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science+Business Media, LLC, 2004 (cit. on pp. 4, 60, 61, 67, 119).

[WO20] Norio Watanabe and Fumiaki Okihara. "On GARCH Models with Temporary Structural Changes." In: *Data Analysis and Applications 1*, 2020. Ed. by Christos H. Skiadas and James R. Bozeman. ITSE Ltd, 2020, pp. 91–104 (cit. on p. 19).

[WAG+21] Hannah S. Wauchope, Tatsuya Amano, Jonas Geldmann, Alison Johnston, Benno I. Simmons, William J. Sutherland, and Julia P.G. Jones. "Evaluating Impact Using Time-Series Data." *Trends in Ecology and Evolution* 36(3), 2021, pp. 196–205 (cit. on pp. 1, 20, 137).

[WED+19] Ron Shapira Weber, Matan Eyal, Nicki Skafte Detlefsen, Oren Shriki, and Oren Freifeld. "Diffeomorphic Temporal Alignment Nets." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2019 (cit. on pp. 82, 97, 115, 117, 118, 123, 131).

[Wei18] Christian H. Weiss. *An Introduction to Discrete-valued Time Series*. John Wiley & Sons, Ltd., 2018 (cit. on pp. 2, 10).

[Wel47] B. L. Welch. "The generalisation of Student's problems when several different population variances are involved." *Biometrika* 34(1-2), 1947, pp. 28–35 (cit. on pp. 60, 67).

[WRD+16] Marc Wiedermann, Alexander Radebach, Jonathan F. Donges, Jürgen Kurths, and Reik V. Donner. "A climate network-based index to discriminate different types of El Niño and La Niña." *Geophysical Research Letters* 43(13), 2016, pp. 7176–7185 (cit. on p. 102).

[WAY+98] Mitchell Withers, Richard Aster, Christopher Young, Judy Beiriger, Mark Harris, Susan Moore, and Julian Trujillo. "A comparison of select trigger algorithms for automated global seismic phase and event detection." *Bulletin of the Seismological Society of America* 88(1), 1998, pp. 95–106 (cit. on p. 102).

[WBB+20] Frederik Wolf, Jurek Bauer, Niklas Boers, and Reik V. Donner. "Event synchrony measures for functional climate network analysis: A case study on South American rainfall dynamics." *Chaos* 30(3), 2020 (cit. on p. 30).

[XFC+18] Haowen Xu, Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, and Dan Pei. "Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications." In: *Proceedings on the World Wide Web Conference (WWW)*, 2018 (cit. on p. 102).

[XFZ16] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. "Learning Granger Causality for Hawkes Processes." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2016 (cit. on p. 30).

[YWZ19] Xing Yan, Qi Wu, and Wen Zhang. "Cross-Sectional Learning of Extremal Dependence Among Financial Assets." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2019 (cit. on p. 30).

[YS81] Mark C.K. Yang and Jack F. Schreckengost. "Difference Sign Test for Comovements Between Two Time Series." *Communications in Statistics - Theory and Methods* 10(4), 1981, pp. 355–369 (cit. on p. 22).

[YHH+16] Wei Yang, Ai Han, Yongmiao Hong, and Shouyang Wang. "Analysis of crisis impact on crude oil prices: a new approach with interval time series modelling." *Quantitative Finance* 16(12), 2016, pp. 1917–1928 (cit. on p. 142).

[YH86] A. Yassouridis and E. Hansert. "Equidirection: A Measure of Similarity Among Time Series." *Biometrical Journal* 28(6), 1986, pp. 747–758 (cit. on p. 22).

[ZBG13] Wojciech Zaremba, Matthew Blaschko, and Arthur Gretton. "B-tests: Low Variance Kernel Two-Sample Tests." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2013 (cit. on p. 61).

[ZS07] Nancy R. Zhang and David O. Siegmund. "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data." *Biometrics* 63(1), 2007, pp. 22–32 (cit. on p. 133).

[ZPJ+20] Wei Zhang, Thomas Kobber Panum, Somesh Jha, Prasad Chalasani, and David Page. "CAUSE: Learning Granger Causality from Event Sequences using Attribution Methods." In: *Proceedings of the International Conference on Machine Learning (ICML)*, 2020 (cit. on p. 30).

[ZGS+08] Xinhua Zhang, Arthur Gretton, Le Song, and Alex Smola. "Kernel Measures of Independence for non-iid Data." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2008 (cit. on p. 24).

[ZLW08] Xun Zhang, K. K. Lai, and Shou Yang Wang. "A new approach for crude oil price analysis based on Empirical Mode Decomposition." *Energy Economics* 30(3), 2008, pp. 905–918 (cit. on p. 17).

[ZYW+09] Xun Zhang, Lean Yu, Shouyang Wang, and Kin Keung Lai. "Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method." *Energy Economics* 31(5), 2009, pp. 768–778 (cit. on pp. 1, 17).

[ZD16] Feng Zhou and Fernando De La Torre. "Generalized Canonical Time Warping." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(2), 2016, pp. 279–294 (cit. on p. 141).

[ZT09] Feng Zhou and Fernando de la Torre. "Canonical Time Warping for Alignment of Human Behavior." In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2009 (cit. on pp. 81, 141).

[ZT12] Feng Zhou and Fernando de la Torre. "Generalized Time Warping for Multi-modal Alignment of Human Motion." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012 (cit. on pp. 81, 82, 96).