Ph.D. Dissertation

Some Advances in Multivariate Statistical Modeling

Máté Baranyi

Supervisor:

Marianna Bolla D.Sc.

Full professor
Department of Stochastics, Institute of Mathematics,
Budapest University of Technology and Economics

MŰEGYETEM 1782

BME

2022

Contents

Preface

My dissertation provides an outlook on the research activities that I have been involved in during my Ph.D. studies. This work revolves around three seemingly distinct topics of statistics: graphical models, nonparametric statistics, and time series analysis, and some connections are drawn between the seemingly distant topics too. I have been introduced to graphical models by my supervisor, Prof. Marianna Bolla, and wrote my Master's thesis [1] around this topic too. This research direction resulted in multiple publications later on: a survey [J1] with my supervisor and a fellow Ph.D. student, Fatma Abdelkhalek; a research article on our novel algorithm [J2] with my supervisor, and an applied paper [C1] with other co-authors. Nonparametric statistics is a topic I have heard about initially during my masters. Then, during my doctoral studies, it unwittingly became one of my main research topics through publications and applied R&D projects. Two of my (otherwise kind of unrelated) research papers [J2, J3] use the theory and methods of nonparametric regression. These articles are detailed in the dissertation. The third main area of statistics I have been introduced to is multidimensional time series analysis. I familiarized myself with the topic while I assisted the writing process of the book [2] of my supervisor and her co-author, Tamás Szabados. This also resulted in a research article [J4] with them and Fatma Abdelkhalek, backed by an R&D project. Based ob this, we wrote two other applied papers with the help of an industrial partner [C4, C5].

Over the years, I have been introduced to many different topics of statistics, data analysis, and machine learning, therefore the list of my publications (on page 103) looks quite diverse. My dissertation is not aimed to give a complete portfolio of my research activities. Some research directions related to deep learning [C3], educational data science [C1], and R&D projects with industrial partners are left out.

I would like to express my gratitude to my supervisor, Prof. Marianna Bolla. She helped me develop my academic and research skills in various ways. She inspired me to work hard, and taught me the importance of maintaining high morals while doing so. I am also grateful for the autonomy, that allowed me to conduct individual research too next to our joint work.

I also thank all my co-authors for our collaboration, without them I would have evolve as a researcher very differently. I especially thankful to Roland Molontay for the joint work, and all the advice, guidance throughout my studies. I am indebted to the Department of Stochastics for providing a stimulating working environment. Last but not least, I owe a great debt of gratitude to my family, friends, partner, and the people in my life for their patience and constant support.

Introduction

In this dissertation, various aspects of multivariate statistical modeling are discussed. Multivariate statistical methods are extensively studied and applied across all domains of science. Characterizing the connections among multiple variables of an experiment or research question is essential for a lot of well-known topics: regression, dimension reduction, hypothesis testing, and data analysis in general. This work focuses on three important subdivisions of statistics, namely, graphical models, nonparametric regression, and time series analysis.

Chapter 1 introduces a novel regression method, called Iterated Conditional Expectation (ICE) algorithm, operating on specific graphical models, namely probabilistic DAGs (directed acyclic graphs) and regression graphs. The chapter is based on an article of similar title [J2], written with my supervisor. The topic is at the intersection of nonparametric statistics and probabilistic graphical models (or simply just graphical models). Therefore, in Section 1.1, the latter is studied in a general setting, largely based on my Master's thesis [1] and a joint survey paper with my supervisor [J1]. Various parts of this introduction to graphical models come up also in Chapter 2, and additional material (not explicitly used in the dissertation) is available in Appendix A. Generally, the purpose of graphical models is to provide a clearer, computationally more efficient way to look at multidimensional probability spaces by representing them with simple graph structures. The aim of these representations is to give us a picture of the dependencies (*edges*) and conditional independencies that hold between the variables (*vertices*). Independence does not often occur in complex systems. However, conditional independence does often arise and can lead to significant representational and computational savings. This is a generalization of independence, where two pieces of a system become independent once we observe a third piece. In the theory of graphical models, the concept of conditional independence plays an important role. In Section 1.1.1, we introduce the notation and some of the theoretical background that is essential to understand the rest of the work. Later on, we describe examples of graphical models and try to point out connections between the different types. Two of the most commonly used models are the probabilistic DAGs (also called Bayesian networks) and the Markov random fields (sometimes Markov networks). We start with DAGs in Section 1.1.2, which are widely used models in the case of directed graph representations. The main properties of these were summarized in the '80s by Judea Pearl [3] and others. Then we continue with Markov random fields in Section 1.1.4. These are so common among undirected representations that some use (incorrectly) the term undirected graphical model for these. The Markov random field was introduced as the general setting for the Ising model, also in the '80s. In Section 1.1.6, chain graph models are introduced, as an example of mixed graphical models. They are mixed in the sense that the graph representation contains both directed and undirected edges. Different types of chain graphs are discussed, some of them, mainly the so-called regression graphs, play an important role in Chapter 1. Then, we continue with our novel algorithm.The proposed ICE algorithm is a regression technique with multiple outputs, and it is applicable in environments with missing data as well. The output variables are estimated one by one in an iterative fashion with nonparametric regression methods. The method assumes the presence of a graphical model compatible with the variables in the experiment. We describe the technique (Section 1.2) and prove its (mean-square) consistency (Section 1.3) over probabilistic DAGs (Section 1.1.2) and regression graphs (Section 1.1.7). The algorithm is more general, but these two types of graphical models are more compatible with the multiple inputs, multiple outputs setup. Straightforwardly, in these models, the main input variables are those represented by "parentless" nodes on the graph in the case of DAGs; and the so-called "context" variables in the case of regression graphs. The main contribution is that we extended the application area of these types of graphical models, by employing nonparametric smoothing recursively along the directed flow of the graphical model. The numerical implementation is also discussed in detail. In Section 1.4, the implementation of the chosen kernel regression method is discussed, together with kernel and bandwidth selection according to the types and sample characteristics of the variables, further some other practical considerations are included. We prove the (mean-square) consistency of the chosen kernel regression implementation in Appendix B. As the selection of the parameters can be automated in the subsequent nonparametric regressions, the ICE algorithm can be the building block of an intelligent decision support system. Applications are presented in Section 1.5.

Chapter 2 is devoted to the analysis of regression discontinuity design (RDD). This is a pretest-posttest quasi-experimental design to measure the average treatment effect, applied in scenarios where randomization is not possible but the treatment is provided based on an additional (so-called running) variable. This chapter is based on my article with Roland Molontay [J3], where we apply a widely used RDD method to measure the effectiveness of two differently designed remedial mathematics courses at the Budapest University of Technology and Economics (BME). That paper is an extended version of our earlier conference paper [C2]. In these, using RDD tools in various settings, we study both the direct and longer-term effects of mathematical remediation. The RDD is described in Section 2.1 as a preliminary. A contribution of our work is that we point out some methodological issues regarding RDD that are neglected in other applied papers. As a methodological novelty, we propose a new alternative method (called RDD with jittering) to handle the problem of a discrete running variable in the RDD setting, detailed in Section 2.2.2. At the end of this chapter, in Section 2.3, some connections of RDD to graphical modeling are discussed. In Section 2.4, our case study is presented in more detail. We study the impact of mathematical remediation at BME on students' subsequent academic achievements regarding both first-year mathematics courses and longer-term effects. At BME, all first-year students take a mathematics assessment test in the first week of their studies, which serves as a basis for offering them remediation. Hence, participation in the remedial course highly depends on the results of the assessment. This procedure makes it suitable to study the statistical impact of remediation using RDD. We investigate two differently designed remedial courses (detailed in Section 2.4.1). We find that for students on the margin of passing the assessment test, participating in the general remedial course focusing on rehearsing the high school material has a small but statistically significant positive effect on further academic achievements. However, a differently designed remedial course, that provides more practice of the new concepts of the first-year calculus course, shows much better results, see Section 2.4.2. Our large-scale study is based on the data of almost 20,000 undergraduate students enrolled between 2010 and 2018. While the majority of similar research is conducted on US data and very few studies on European universities, to the best of our knowledge, this is the first paper of this nature from Central and Eastern European universities.

In Chapter 3, our attention is turned toward time series analysis. Here, after some preliminaries in Section 3.1, connections of the covariance structure and spectral densities of stationary multidimensional time series are investigated. Based on the theoretical results of another joint paper with my supervisor [J4], a novel dynamic principal component analysis algorithm (called dynPCA) is also introduced. This was further studied in [C4] and [C5]. Specifically, we work with multivariate weakly stationary discrete-time real-valued time series. Given a time series like that with absolutely summable autocovariances, we proved an asymptotic relation between the eigenvalues of the block Toeplitz matrix of the first n autocovariances and the union of spectra of the spectral density matrices at the n Fourier frequencies. For the proof, the spectral properties of block circulant matrices are used. We use spectra in two different meanings: for the eigenvalues and for the spectral representation of weakly stationary processes. The proved theorem leads to analogies between the time and frequency domain calculations, and to new techniques. In particular, a method for a low-rank approximation of the process; and consequently, a novel dynamic principal component analysis algorithm (called dynPCA). The latter algorithm, which has various potential applications, estimates the dynamic principal components of the multidimensional time series input. The results are illustrated and applications are presented in Section 3.6.

As already mentioned, each chapter closes with applications of the discussed techniques, methods, in order to show the importance of applications in multivariate statistical modeling. The ICE algorithm and the nonparametric version of RDD involve nonparametric regression, thus some preliminaries on these are included in Appendix B. Some additional material on time series analysis is also presented in Appendix C.

The three main “thesis points” of the dissertation are the following:

1. We introduce the ICE algorithm, and we prove its (mean-squared) consistency over probabilistic DAGs and regression graphs (Chapter 1). For its numerical implementation, we also prove (mean-squared) consistency of the utilized kernel regression method (Appendix B).

2. We analyze the RDD technique, then we propose and discuss a novel alternative method (called RDD with jittering) to handle the problem of a discrete running variable in the RDD setting (Chapter 2). A unique, detailed case study is also included.

3. We investigate the connections of the covariance structure and spectral densities of stationary multidimensional time series, and prove asymptotic relations between these (Chapter 3). Based on our theoretical results, the novel dynPCA algorithm is also introduced, with unique applications.

Chapter 1

Iterated Conditional Expectation algorithm (ICE) on DAGs and regression graphs

This chapter introduces the Iterated Conditional Expectation (ICE) algorithm, based on the article with my supervisor of similar title [J2]. Again, the proposed method is a regression technique with multiple outputs, where these are estimated one-by-one in an iterative fashion with nonparametric regression methods.

After some preliminaries on graphical models in Section 1.1, we describe the technique (Section 1.2) and prove its (mean-square) consistency (Section 1.3) over probabilistic DAGs (Section 1.1.2) and regression graphs (Section 1.1.7). In these two models, the main input variables are those represented by "parentless" nodes on the graph in the case of DAGs; and the "context" variables in the case of regression graphs. In Section 1.4, the implementation of the kernel smoothing is discussed together with some practical considerations. Additional material and a result on the mean-square consistency of the chosen kernel regression method is in Appendix B.

Our contribution is that we extended the application area of probabilistic DAGs and regression graph models, by introducing and theoretically justifying an algorithm, that applies nonparametric smoothing recursively along the directed flow of the graphical model.

1.1 Preliminaries on graphical models

In this section, we give an introduction to the theory of probabilistic graphical models (or simply just graphical models). The purpose of graphical models is to provide a clearer, computationally more efficient way to look at multidimensional probability spaces by representing them with simple graph structures. The aim of the graph-based representations is to give us a picture of the dependencies (*edges*) and conditional independences that hold between the variables (*vertices*). Independence does not often occur in complex systems. However, conditional independence does often arise and can lead to significant representational and computational savings. This is a generalization of independence, where two pieces of a system become independent once we observe a third piece. In the theory of graphical models, the concept of conditional independence plays an important role.

In Section 1.1.1, we introduce the notation and some of the theoretical background that is essential to understand the rest of the work. In the later sections, we describe examples of graphical models and try to point out connections between the different types. Two of the most commonly used models are the Bayesian networks and the Markov random fields. We start with Bayesian networks in Section 1.1.2, which are widely used models in the case of directed graph representations. The main properties of these were summarized in the '80s by Judea Pearl [3] and others. Then we continue with Markov random fields (sometimes Markov networks) in Section 1.1.4. These are so common among undirected representations that some use (incorrectly) the term undirected graphical model for these. The Markov random field was introduced as the general setting for the Ising model, also in the '80s.

In Section 1.1.6, chain graphs are introduced, as an example of mixed graphical models. They are mixed in the sense that the graph representation contains both directed and undirected edges. Different types of chain graphs are discussed, some of them will come up again in Chapter 1.

1.1.1 Graphical models in general

To characterize these models we will stick to some restrictions on the notations to avoid misunderstanding. In a graphical model representation of a multivariate probability space, let $G = (V, E)$ be the graph representation of the desired multivariate space. We work with **simple graphs**, meaning it has at most one edge for each vertex-pair and has no vertex linked to itself. There is one-to-one compliance between the vertices (V) and the random variables, therefore we regularly say things like probability distribution "over G" or "with respect to G". When it does not cause any confusion we use the words, vertex, node, and variable interchangeably. We represent the dependencies between the random variables with edges (E) in a way that these meet the specifications of the chosen model. The letters G, V, E will be used only for the graph representation:

- $G = (V, E)$ is the graph itself;
- $V = \{v_1, v_2, \ldots, v_n\}$ is the set of vertices/nodes (denoted with lower case letters);
- $E \subset V \times V$ is the set of edges: $(i, j) \in E$ or $i \to j$ will mean a directed edge from i to j , meanwhile $[i, j] \in E$ or $i - j$ will mean an undirected edge.

For a given graph $G = (V, E)$, the subgraph induced by a subset $A \subset V$ is as usual:

$$G(A) = (A, E(A)), \quad \text{where} \quad E(A) = \{(x, y) \in E \quad \text{or} \quad [x, y] \in E \mid x, y \in A\}.$$

We shorten the term "subgraph induced by a set of vertices" to "subgraph of vertices", which again just keeps the vertices and the edges present among them in a given graph. To every $i \in V$ belongs a random variable X_i:

- The random variables in the probability space are $X_1, X_2, \ldots, X_n$ (denoted with upper case letters). For every vertex $i \in V$, X_i is a (one-dimensional) random variable, which can be discrete/categorical or continuous real-valued.
- For any subset $A = \{a_1, a_2, \ldots, a_m\} \subset V$, $\boldsymbol{X}_A$ will denote the following random vector: $\boldsymbol{X}_A = \boldsymbol{X}_{\{a_1,a_2,\ldots a_m\}} := (X_{a_1}, X_{a_2}, \ldots, X_{a_m})$

Each X_i takes its values from the finite set $\mathcal{X}_i$ in the discrete/categorical case, or $\mathcal{X}_i$ can be $\mathbb{R}$ itself in the continuous case, thus the vector $\boldsymbol{X}_V$ (or sometimes just $\boldsymbol{X}$) will take its value(combination)s from $\mathcal{X}_V := \bigtimes_{i=1}^n \mathcal{X}_i$, and for $A \subset V$ similarly. A possible outcome is sometimes called a configuration or a state. We also have a probability measure $\mathcal{P}_V$ associated to $\boldsymbol{X}_V$, and similarly for the individual X_i's.

We have to use some probabilistic terms regarding dependencies that can be confusing. The concept of conditional independence has a crucial role, so we walk around this concept first. Let (X, Y, Z) be a random vector whose coordinates take values from (possible different) sets. In the case of discrete/categorical variables, we use the conventional short notation for the (conditional) probability of a configuration:

$$\mathrm{p}_{\mathrm{XY}}(\cdot,\cdot) = \mathbb{P}\left(X = \cdot, Y = \cdot\right) \quad \text{and} \quad \mathrm{p}_{\mathrm{X|Y}}(\cdot,\cdot) = \mathbb{P}\left(X = \cdot \mid Y = \cdot\right).$$

In the case of continuous (w.r.t Lebesgue measure) variables we use the same notation ($\mathrm{p}_{\mathrm{XY}}(\cdot,\cdot)$ and $\mathrm{p}_{\mathrm{X|Y}}(\cdot,\cdot)$) for the (conditional) density. The reason behind this is that most of the upcoming theoretical part stays the same regardless of the type of the variables, and even can be generalized to some specific mixed probability spaces. These are mixed in the sense that they contain both discrete/categorical and continuous variables.

Actually, we can confine ourselves to the following family of distributions, based on [4].

Definition 1.1. *The probability measure $\mathcal{P}_V$ is* marginally continuous *if it is absolutely continuous with respect to the product of its one-dimensional marginals $\mathcal{P}_i$, $i \in V$, that is, $\mathcal{P}_V \ll \bigtimes_{i\in V} \mathcal{P}_i$.*

It can be shown [4] that $\mathcal{P}_V$ is marginally continuous if and only if there is a collection of σ-finite measures μ_i for the one-dim. X_i's for which $\mathcal{P}_V \ll \mu := \bigtimes_{i\in V} \mu_i$. For marginally continuous distributions, the density p_{V} of $\boldsymbol{X}_V$ is defined as the Radon–Nikodym derivative w.r.t this μ. The family contains: every discrete measure over V since the counting measure on X_V can serve as its dominating measure; the multidimensional Gaussian distribution (dominated by the Lebesgue measure on X_V); and the so-called “conditional Gaussian” distributions in the mixed case. Note that nearly all multidimensional measures used in practice are marginally continuous, however, some singular Gaussian measures are not.

With this family of distributions in mind we define the followings:

Definition 1.2. *We say that X and Y are independent if $\mathrm{p}_{\mathrm{XY}} = \mathrm{p}_{\mathrm{X}}\,\mathrm{p}_{\mathrm{Y}}$ (or equivalently $\mathrm{p}_{\mathrm{X|Y}} = \mathrm{p}_{\mathrm{X}}$). Notated as $X \perp\!\!\!\perp Y$.*

Definition 1.3. *We say that X and Y are conditionally independent given Z if $\mathrm{p}_{\mathrm{XY|Z}} = \mathrm{p}_{\mathrm{X|Z}}\,\mathrm{p}_{\mathrm{Y|Z}}$ (or equivalently $\mathrm{p}_{\mathrm{X|YZ}} = \mathrm{p}_{\mathrm{X|Z}}$). Notated as $X \perp\!\!\!\perp Y \mid Z$.*

In the above definitions and later on as well, we mean by the product of functions, measures, the point-wise product of them over the product domain: $\mathrm{p}_{\mathrm{X}}\,\mathrm{p}_{\mathrm{Y}}(\cdot,\cdot) := \mathrm{p}_{\mathrm{X}}(\cdot)\,\mathrm{p}_{\mathrm{Y}}(\cdot)$, and by equality of these functions we mean point-wise or almost-all equality, depending on the domain. The definition of conditional independence is non-trivial, can be ambiguous, and has a deep measure theoretic background, for more details, see the book of Studený [4].

Another key concept in the above definitions and in multidimensional probability theory, in general, is marginalization. E.g., in the discrete case, for a subset $A \subset V$ of the examined variables, the A-marginal of a function $f_V : \mathcal{X}_V \to \mathbb{R}$, with other words its marginalization over $V \setminus A$, is the function $f_A : \mathcal{X}_A \to \mathbb{R}$ below:

$$f_A(\boldsymbol{x}_A) := \sum_{\substack{\boldsymbol{y}_V \in \mathcal{X}_V: \\ \boldsymbol{y}_A = \boldsymbol{x}_A}} f_V(\boldsymbol{y}_V) = \sum_{\boldsymbol{x}_{V\setminus A} \in \mathcal{X}_{V\setminus A}} f_X(\boldsymbol{x}_A, \boldsymbol{x}_{V\setminus A}), \quad \forall \boldsymbol{x}_A \in \mathcal{X}_A.$$

Briefly, we use the notation

$$f_A(\boldsymbol{x}_A) := \left(\sum_{V\setminus A} f_V \right) (\boldsymbol{x}_A),$$

or shortly $f_A := \sum_{V\setminus A} f_V$. Note that in the case of similar marginalization, f is not necessarily a probability mass or probability density function but something that is proportional to them, and measurable. The concept naturally extends to continuous variables through Riemann integration, and further extends to marginally continuous distributions through Lebesgue integration.

1.1.2 Bayesian network, DAG, a directed graphical model

We speak about a Bayesian network (BN) when the graph representation of our multivariate probability space is a directed acyclic graph (DAG). Sometimes, the models themselves are also called DAGs, e.g., by Lauritzen [5].

Definition 1.4. *A simple directed graph $G = (V, E)$ is called a* ***directed acyclic graph*** *if there is an ordering $\sigma : V \to \{1, 2, \ldots, n\}$ of the vertices such that for all $(i, j) \in E$, $\sigma(i) < \sigma(j)$, meaning there can not be a directed cycle in G.*

For a DAG, an ordering like that is called a topological ordering. Now let $(X_1, X_2, \ldots, X_n)$ be a random vector whose coordinates take values from (possible different) finite sets (*or real intervals*). Then with the well-known chain-rule, the joint probability (*or density*) can always be written as:

$$\mathrm{p}_{\mathrm{X_1 X_2 \ldots X_n}} = \mathrm{p}_{\mathrm{X_1}}\, \mathrm{p}_{\mathrm{X_2|X_1}}\, \mathrm{p}_{\mathrm{X_3|X_1X_2}} \cdots \mathrm{p}_{\mathrm{X_n|X_1 \ldots X_{n\text{-}1}}}\,.$$

Some terms in the above equation become simpler if we know or suppose some conditional independences between certain variables. After the possible simplifications induced by these, we can build a DAG as follows, from $i = 1$ to n:

- assign a vertex to every variable: $X_i \to v_i$,
- allow an edge $v_i \to v_j$ if there is a factor of the form $\mathrm{p}_{\mathrm{X_j|\ldots X_i \ldots}}$.

The result of this procedure is always a DAG, however, the joint distribution does not determine the graph uniquely, it depends on the ordering of the variables, and if the distribution is not strictly positive, then the simplifications are not unique either.

Definition 1.5. *If we already have a DAG G with n vertices, then we say that the distribution of $\boldsymbol{X}$, $\mathrm{p}_{\mathrm{X_1X_2\ldots X_n}}$ fulfills the* ***Directed Factorization*** *property (DF) over G if:*

$$\mathrm{p}_{\mathrm{X_1X_2\ldots X_n}} = \prod_{i=1}^{n} \mathrm{p}_{\mathrm{X_i|X_1\ldots X_{i\text{-}1}}} = \prod_{i=1}^{n} \mathrm{p}_{\mathrm{X_i|X_{Par(i)}}}, \tag{1.1}$$

where $\mathrm{Par}(i)$ *means the parents of vertex i in G:*

$$\mathrm{Par}(i) = \{j \in V \mid (j,i) \in E\},$$

and the ordering of the variables is a topological ordering.

This is the main property of the distributions induced by a DAG. Different Markov properties can be defined over DAG models, Bayesian networks, see Appendix A.

1.1.3 Causal Bayesian network

Bayesian networks are often used to represent causal relationships, but this is not always desired. For example, for a Bayesian network over three nodes A, B, C a directed edge from A to B does not require X_B to be causally dependent on X_A. The following Bayesian networks are in fact Markov-equivalent:

$$A \to B \to C \qquad \text{and} \qquad A \leftarrow B \leftarrow C.$$

They encompass the same conditional independence statements.

We call a Bayesian network causal; shortly, causal DAG or causal network if the edges also represent causal relations in Pearl's terminology. An additional property of causal DAGs is that we can impose intervention on it with the $\mathrm{do}(\cdot)$ operator of Pearl [6]: We fix a node A to be in a given state a, denoted as $\mathrm{do}(A = a)$. As a consequence, we do not study our original (pre-intervention) graph G anymore but a new (post-intervention) one $G_{\mathrm{do}(A)}$. We obtain this by removing the edges from the parents of A to A, and setting A to the *caused value a*. By the graph operation, the represented probability distribution may fundamentally change, and different inferences can be made, as we have a new probabilistic model, and

$$\mathbb{P}_G(X_B \mid X_A = a) \neq \mathbb{P}_{G_{\mathrm{do}(A)}}(X_B \mid \mathrm{do}(X_A = a)).$$

1.1.4 Markov random field, an undirected graphical model

If the random variables in the probability space satisfy one of the *undirected* Markov properties with respect to some undirected graph, then it is called a Markov random field (MRF). A BN is always a DAG, whereas an MRF is undirected and possibly cyclical. There are different types of Markov properties that can be satisfied. These are detailed in Appendix A.

1.1.5 Covariance graph, another undirected model

Cox & Wermuth [7, 8] define covariance graphs to represent marginal dependencies instead of conditional ones. The graphs contain undirected **dashed** edges. Pairwise, Local, and Global

Markov properties of these models were defined by Kauermann [9].

Absence of an edge between nodes A and B means that the variables X_A and X_B are marginally independent. In the case of multivariate normal distribution, absence of an edge equivalently means a zero in the corresponding position in the covariance matrix, hence the name.

A covariance graph and an MRF with the same edge set (but with dashed and regular edges, respectively) are Markov-equivalent if they build up exclusively of disjoint complete subgraphs [8].

The covariance graphs are sometimes represented with bidirected edges and called bidirected graph models. This notation is also popular because covariance graphs are a special case of Acyclic Directed Mixed Graphs [10], that contain directed and bidirected edges too. This mixed model arises from a natural extension of the d-separation (Section A.1.1), called m-separation. Nevertheless, the covariance graph and the bidirected graph models are equivalent.

1.1.6 Chain graph, a class of mixed graphical models

Chain graphs (CGs) are a bigger family of mixed graphical models. They are mixed in the sense that they contain both directed and undirected edges. They are useful when we have both symmetric and asymmetric relationships in a system.

They contain edges in the following way. If we keep only the undirected edges, the graph falls apart into connected chain components, whereas if we keep only the directed ones, we have a DAG. In other words, they do not contain any semi-directed (partially directed) cycles. A chain component C is a maximal subset of nodes such that there is a path between every pair of nodes in C containing only a type of undirected edge.

Here, the joint distribution of $X_1, \ldots X_n$ can be factorized along the chain components $C_1, \ldots, C_\ell$ ($\ell < n$) as

$$\mathrm{p}_{\mathrm{C}_\ell} \prod_{k=1}^{\ell-1} \mathrm{p}_{\mathrm{C}_{\ell\text{-k}} \,|\, \mathrm{C}_{>\ell\text{-k}}}, \tag{1.2}$$

and it further reduces along the DAG to:

$$\mathrm{p}_{\mathrm{C}_\ell} \prod_{k=1}^{\ell-1} \mathrm{p}_{\mathrm{C}_{\ell\text{-k}} \,|\, \mathrm{C}_{\mathrm{Par}\,(\ell\text{-k})}}, \tag{1.3}$$

where $C_{\mathrm{Par}\,(k)} = \cup_{i \in C_k} \mathrm{Par}\,(i)$, the union of parent variables of variables in component C_k. This formula is in terms of the joint and conditional probability mass/density functions of the chain components (over a marginally continuous distribution); the numbering of the components is a given (reversed) topological ordering of the compatible DAG over them.

For DAGs, the intuitive meaning is that the parent nodes are the cause of child nodes, but the intuitive meaning for CGs can be more difficult to grasp. The interpretation of the edges gives rise to four types of CGs, by combining two different interpretations of directed edges with two different interpretations of undirected edges, as described by Drton [11].

Other models can contain both types of undirected edges, e.g., ancestral graphs (AG), marginal AMP CGs (MAMP CGs, a superclass of AMP CGs, and MVR CGs), and regression CGs contain dashed (or bidirected), regular undirected and directed edges too but with some additional restrictions on what structures these can take.

1.1.7 Regression graph, a twist on chain graphs

During my Master's and Ph.D. studies, we mainly investigated the just mentioned regression graphs (or sometimes regression CGs). They were introduced by Wermuth and Sadeghi [12]. A regression graph is a refined chain graph that gives rise to a collection of so-called traceable regressions along the directions of its edges, see [13]. They distinguish two types of chain components: a type for **response** variables (that we are interested in), and another for **context** variables (that are given in the context of the experiment). The components are numbered so (conventionally) that the last ones (with the highest indices) correspond to the context variables. The three types of edges are used the following way:

- Context variables of the same chain component are connected with undirected edges based on the concentration graph on them. This is a distinction from MVR CGs.
- From the context variables directed edges point to variables in the lower indexed components, containing primary, secondary, etc. responses. Directed edges also may go between response variables of different chain components, based on some hierarchy of them.
- Between the joint response variables of the same connected component, dashed (undirected, but sometimes denoted as bidirected) edges are present, which indicate dependences on conditional covariance base. These joint response variables are also called to be on equal standing, and they usually correspond to similar aspects of an experiment, without direct causality between them.

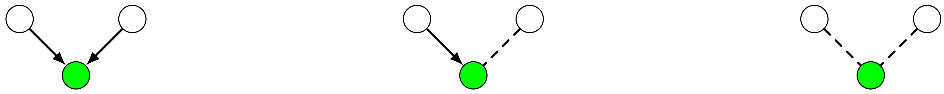
Similarly to DAG-based models, for every directed edge $j \to i$, conventionally, the relation $i < j$ holds. This so-called topological ordering is referred to as "j is the parent of i". So the "youngest" vertex has label 1, and the "older" a vertex, the larger its label is (we can think of labels as age).

In our convention (following Wermuth) for representation, the arrows point from right ("past") to left ("future"), see [J1] and Figures 1.4, 1.7 and 1.9 for examples. Let $C_1, C_2, \dots, C_\ell$ denote the chain components as in the previous section. Here C_ℓ contains the context variables, and the chain components are also indexed so that the lower index components contain "younger" vertices. With the shorthand $C_{>m} := C_{m+1} \cup \dots \cup C_\ell$, we can make the following conditional independence statements that uniquely define the missing edge positions:

- for $i < j$ such that $X_i \in C_m$ and $X_j \in C_{>m}$ are in different chain components, there is no $j \to i$ arrow if and only if $X_i \perp\!\!\!\perp X_j \mid C_{>m} \setminus \{X_j\}$;
- for $X_i, X_j \in C_\ell$ (context component), there is no solid line between them if and only if $X_i \perp\!\!\!\perp X_j \mid C_\ell \setminus \{X_i, X_j\}$;

- for X_i and X_j in the same C_m (non-context component, $m < \ell$), there is no dashed line between them if and only if $X_i \perp\!\!\!\perp X_j \mid C_{>m}$.

In Theorem 1 of [12], the authors state that two regression graphs are Markov equivalent if and only if they have the same skeleton (the graph resulting from making every edge undirected) and the same set of collision Vs, irrespective of the type of edges, see Figure 1.1. In their Theorem 2, they prove that a regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a DAG with the same skeleton if and only if it does not contain any chordless collision path in four nodes. These so-called *forbidden quadruple*s are shown in Figure 1.2.



(a) (b) (c)

Figure 1.1: Possible collision Vs of non-context nodes in a regression graph

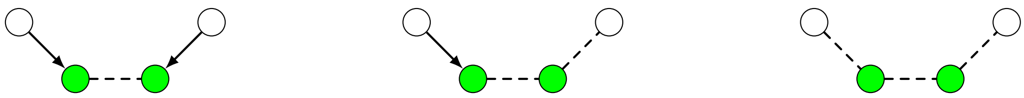


Figure 1.2: Chordless collision paths on four nodes

1.2 Description of the ICE algorithm

The main idea is the following. In an ordering of the vertices compatible with the DAG or the regression graph, instead of linear, linearized, or logistic regression, we take conditional expectation in a nonparametric way, like the ACE (Alternating Conditional Expectation) algorithm of Breiman and Fiedman [14]. The ACE is applicable to an additive model that finds a unit variance measurable function of the dependent variable and those of the independent ones, the sum of which approximates best the former in L^2 norm. In the convergence proof of the ACE, the emphasis is on the optimal function of the single regressor, when predicting the current phase (additive function) of all other variables. On the contrary, we find the measurable function of all predictors recursively, and the finite iteration ends up in predicting the target(s).Here we need not alternate, we just take "directed" conditional expectations consecutively, with the nonparametric regression concepts of [15]. Therefore, we call our recursive algorithm ICE (Iterated Conditional Expectation). In the consistency proofs we compare the conditional expectation to the smoother.

In comparison with the former similar results in the literature, our algorithm combines the theory of probabilistic DAGs and regression graphs with the nonparametric conditional expectation concept. However, we have no infinite iteration (e.g. ACE) and maximization steps (e.g. EM algorithm) as we do not assume a parametric likelihood, but work in a nonparametric setup.

The basis for the construction is a training sample, expert knowledge, and the statistical validation of the edges, see [12]. The far not trivial task of structure learning is not considered in this chapter. Here the focus is on the prediction of multiple responses, given the graph. The soundness of the ICE algorithm is also supported by the proof of its consistency, meaning that with large training samples we get near-optimal estimates (in L^2 sense) for the test samples.

The mean-square consistency of the individual smoothers, based on local averaging, is guaranteed under the conditions of the C. J. Stone's theorem, see [15] and Theorem B.1 of Appendix B. For example, k-nearest neighbor (k-NN), partitioning, and kernel regression estimates are mean-square consistent under particular conditions (see again [15]). Actually, we mainly use the product kernel regression approach of [16] which adapts well to different types of variables. It is asymptotically normal, and in turn, mean-square consistent too, as we prove it in Lemma B.1 of Appendix B. For comparison, we include some results obtained with other smoothers too, in our applications.

1.2.1 ICE on a DAG

Let the joint distribution of $X_1, \dots, X_d$ be (in this topological ordering) Markov compatible with a DAG. Say, at least the values of the variables $X_{k+1}, \dots, X_d$ are known; these include the ones without parents in the DAG. Then we first predict X_k, then X_{k-1}, $\dots$, finally X_1, based on an n-element complete training sample $x_1^{(i)}, \dots, x_d^{(i)}$, $i = 1, \dots, n$. We also introduce some notation: $\boldsymbol{X}_{>j} := \{X_{j+1}, \dots, X_d\}$, $\boldsymbol{x}_{>j} := \{x_{j+1}, \dots, x_d\}$ and $\boldsymbol{X}_{\mathrm{Par}(j)}$ denotes the set of the parents of X_j, from where directed edges point to it.

For $j = k, k-1, \dots, 1$ we define

$$P_j(\boldsymbol{x}_{>j}) := \mathbb{E}\left(X_j \mid \boldsymbol{X}_{>j} = \boldsymbol{x}_{>j}\right) = \mathbb{E}\left(X_j \mid \boldsymbol{X}_{\mathrm{Par}(j)} = \boldsymbol{x}_{\mathrm{Par}(j)}\right) \tag{1.4}$$

as the exact solution of the regression task; it is a projection and a measurable function of its arguments. The last equality follows by the Markovity over the DAG.

If our data is from a multivariate Gaussian distribution, then the above conditional expectations are linear functions of the variables in the condition, and are obtainable by linear regression, where the coefficients are estimated from the training data. Otherwise, in the case of unknown distribution, we take the conditional expectation in a nonparametric way, by the smoothing procedures discussed in [14, 15]. Therefore, we define

$$S_j^{(n)}\left(\boldsymbol{x}_{>j}\right) := \sum_{i=1}^{n} x_j^{(i)} W_j^{(n,i)}\left(\boldsymbol{x}_{>j}\right) = \sum_{i=1}^{n} x_j^{(i)} W_j^{(n,i)}\left(\boldsymbol{x}_{\mathrm{Par}(j)}\right) \tag{1.5}$$

as a smoother, based on local averaging and mapping into the convex hull of the sample space for X_j. By Markovity, instead of the set $\boldsymbol{x}_{>j}$, the parent set $\boldsymbol{x}_{\mathrm{Par}(j)}$ is considered.

Here the non-negative weights $W_j^{(n,i)}(\boldsymbol{x}_{>j})$ are calculated from the n-element training sample, and are usually normalized such that

$$\sum_{i=1}^{n} W_j^{(n,i)}\left(\boldsymbol{x}_{>j}\right) = 1.$$

The smoother $S_j^{(n)}$ itself depends on the segment $x_j^{(i)}, x_{j+1}^{(i)}, \dots, x_d^{(i)}$ for $i = 1, \dots, n$ of the training sample, and is also a measurable function of its arguments. The smoothers we implemented are mainly Nadaraya–Watson type kernel estimators, where

$$W_j^{(n,i)}\left(\boldsymbol{x}_{>j}\right) = \frac{K\left(\boldsymbol{x}_{>j}, \boldsymbol{x}_{>j}^{(i)}\right)}{\sum_{i'=1}^{n} K\left(\boldsymbol{x}_{>j}, \boldsymbol{x}_{>j}^{(i')}\right)}.$$

For possible kernels K, see the product kernel approach and other local averaging estimators, discussed in Section 1.4 and Appendix B.

Then the iterated conditional expectations

$$P_j\left(P_{j+1}\left(P_{j+2}\left(\ldots\left(P_k\left(\boldsymbol{X}_{>k}\right),\ldots\right),\boldsymbol{X}_{>k}\right),\boldsymbol{X}_{>k}\right),\boldsymbol{X}_{>k}\right)$$

are imitated with the iterated smoothings

$$S_j^{(n)}\left(S_{j+1}^{(n)}\left(S_{j+2}^{(n)}\left(\ldots\left(S_k^{(n)}\left(\boldsymbol{X}_{>k}\right),\ldots\right),\boldsymbol{X}_{>k}\right),\boldsymbol{X}_{>k}\right),\boldsymbol{X}_{>k}\right).$$

In Section 1.3, it is proved that the mean-square difference between the iterated smoothings and the iterated conditional expectations tends to 0 as $n \to \infty$, for $j = k-1, \ldots, 1$. Therefore, the successive estimates are close to the optimal (in L^2 sense) that means the consistency of the algorithm. For a detailed explanation, see Section 1.3.1.

1.2.2 ICE on a regression graph

Here the joint density of $X_1, \ldots X_d$ can be factorized along the chain components $C_1, \ldots, C_\ell$, see Equation (1.2) , thus the conditional expectations are taken from chain component to chain component, starting from the last one, C_ℓ. We postulate again that at least the context variables, included in C_ℓ, are known.

In the general step, for $k = \ell-1, \ell-2, \ldots, 2, 1$, let

$$\boldsymbol{P}_k\left(\boldsymbol{x}_{C_{>k}}\right) := \mathbb{E}\left(\boldsymbol{X}_{C_k} \mid \boldsymbol{X}_{C_{>k}} = \boldsymbol{x}_{C_{>k}}\right) \tag{1.6}$$

denote the exact solution of the regression task. The result is a vector of components

$$P_{k,j}\left(\boldsymbol{x}_{C_{>k}}\right) := \mathbb{E}\left(X_j \mid \boldsymbol{X}_{C_{>k}} = \boldsymbol{x}_{C_{>k}}\right) = \mathbb{E}\left(X_j|\boldsymbol{X}_{\mathrm{Par}(j)} = \boldsymbol{x}_{\mathrm{Par}(j)}\right), \quad X_j \in C_k. \tag{1.7}$$

This is an instance of parallel regression, where we utilized that all the parents of $X_j \in C_k$ are in $C_{>k}$. We also used that the conditional expectation of a random vector on another one is the vector of components that are the conditional expectations of its components on the same conditioning set. By the Markov properties of the regression graph, even if we condition on the same set of random variables, the individual components may have different parent sets to which the conditioning sets are limited.

The smoothers are also defined componentwise:

$$S_{k,j}^{(n)}\left(\boldsymbol{x}_{C_{>k}}\right) := \sum_{i=1}^{n} x_j^{(i)} W_{k,j}^{(n,i)}\left(\boldsymbol{x}_{C_{>k}}\right) = \sum_{i=1}^{n} x_j^{(i)} W_{k,j}^{(n,i)}\left(\boldsymbol{x}_{\mathrm{Par}(j)}\right), \quad X_j \in C_k. \tag{1.8}$$

This is an approximate solution using a smoother, based on local averaging. The smoother $\boldsymbol{S}_k^{(n)}(\boldsymbol{x}_{C_{>k}})$ consists of coordinates $S_{k,j}^{(n)}(\boldsymbol{x}_{C_{>k}})$ for $X_j \in C_k$.

Then, by the forthcoming Section 1.3.2, our estimates are close to the optimal if the mean-square difference between the iterated component-wise conditional expectations

$$\boldsymbol{P}_k\left(\boldsymbol{P}_{k+1}\left(\boldsymbol{P}_{k+2}\left(\ldots\left(\boldsymbol{P}_{\ell-1}\left(\boldsymbol{X}_{C_\ell}\right),\ldots\right),\boldsymbol{X}_{C_\ell}\right),\boldsymbol{X}_{C_\ell}\right),\boldsymbol{X}_{C_\ell}\right)$$

and the analogous smoothers

$$\boldsymbol{S}_k^{(n)}\left(\boldsymbol{S}_{k+1}^{(n)}\left(\boldsymbol{S}_{k+2}^{(n)}\left(\ldots\left(\boldsymbol{S}_{\ell-1}^{(n)}\left(\boldsymbol{X}_{C_\ell}\right),\ldots\right),\boldsymbol{X}_{C_\ell}\right),\boldsymbol{X}_{C_\ell}\right),\boldsymbol{X}_{C_\ell}\right)$$

tends to 0 as $n \to \infty$, for $k = \ell-1, \ell-2, \ldots, 2, 1$.

1.3 Consistency of the ICE algorithm

1.3.1 Consistency along a DAG

When the target variable is absolutely continuous

Now assume that $X_1, \ldots, X_d$ (in the topological DAG ordering) have finite variances and obey an absolutely continuous joint distribution. The preparation for the proof of the mean-square consistency of the ICE algorithm follows. Since

$$\mathbb{E}\left[S_k^{(n)}(\boldsymbol{X}_{>k}) - X_k\right]^2 = \mathbb{E}\left[S_k^{(n)}(\boldsymbol{X}_{>k}) - P_k(\boldsymbol{X}_{>k})\right]^2 + \mathbb{E}\left[P_k(\boldsymbol{X}_{>k}) - X_k\right]^2, \tag{1.9}$$

and the second term on the right depends only on the nature of the underlying distribution (the larger the coefficient of determination, the smaller this regression error is), it is the first term on the right that can be directly decreased by the sampling and smoothing procedure. (Note that the coefficient of determination is closely related to the maximal correlation of Rényi [17], and to the multiple correlation in the Gaussian case.) Therefore, the above objective function on the left of Equation (1.9) is close to the optimal value if the first term on the right is close to 0.

In the spirit of [15], we call our nonparametric estimate mean-square consistent if

$$\mathbb{E}\int_{\mathbb{R}^{d-k}}\left[S_k^{(n)}(\boldsymbol{x}_{>k}) - P_k(\boldsymbol{x}_{>k})\right]^2 \mu(d\boldsymbol{x}_{>k}) \to 0, \quad n \to \infty, \tag{1.10}$$

where μ denotes the joint distribution of the variables behind it. Here both the n-element sample for $X_k, X_{k+1}, \ldots, X_d$, used by the smoother, and the values of the variables $X_{k+1}, \ldots, X_d$ of a new-coming observation are random, this is why we separated these two concepts when taking the expectation. Briefly,

$$\mathbb{E}\left[S_k^{(n)}(\boldsymbol{X}_{>k}) - P_k(\boldsymbol{X}_{>k})\right]^2 \to 0, \quad n \to \infty,$$

where $\mathbb{E}$ now takes the expectation with respect to (w.r.t.) both of the above concepts of randomness.

First, we prove consistency in the first step of the iteration (see Lemma 1.1), then we prove consistency of the whole ICE on DAG process by induction (see Theorem 1.1). Going backward, to predict X_{k-1}, we have to prove the mean-square consistency of the estimate

$$S_{k-1}^{(n)}\left(S_k^{(n)}\left(\boldsymbol{X}_{>k}\right), \boldsymbol{X}_{>k}\right).$$

Here, because of the orthogonality properties of the conditional expectation operator (as a projection), we get that

$$\begin{aligned}
&\mathbb{E}\left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - X_{k-1}\right]^2 \\
&= \mathbb{E}\left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k}))\right]^2 \\
&+ \mathbb{E}\left[P_{k-1}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(P_k(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k}))\right]^2 \\
&+ \mathbb{E}\left[P_{k-1}(P_k(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(X_k, \boldsymbol{X}_{>k}))\right]^2 \\
&+ \mathbb{E}\left[P_{k-1}(X_k, \boldsymbol{X}_{>k})) - X_{k-1}\right]^2.
\end{aligned}$$

The last two terms are the regression errors that are cumulated in the consecutive steps. Therefore, it suffices to estimate the first two terms, the sum of which also estimates the mean-square difference

$$\begin{aligned}
&\mathbb{E}\left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(P_k(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k}))\right]^2 \\
&\leq 2\ \mathbb{E}\left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k}))\right]^2 \\
&+ 2\ \mathbb{E}\left[P_{k-1}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(P_k(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k}))\right]^2
\end{aligned} \tag{1.11}$$

between the iterated conditional expectation and the iterated smoothing. We will use this form in our forthcoming Lemma 1.1 and Theorem 1.1.

Lemma 1.1. *With the notation of Equations* (1.4) *and* (1.5)*, let* $P_k(\boldsymbol{x}_{>k})$ *and* $S_k^{(n)}(\boldsymbol{x}_{>k})$ *be the exact solution of the regression task and a smoother (based on local averaging and mapping into the convex hull of the sample space of* X_k*), respectively. Assume that the smoother gives a mean-square consistent estimate for the conditional expectation, i.e.,*

$$\mathbb{E}\left[S_k^{(n)}(\boldsymbol{X}_{>k}) - P_k(\boldsymbol{X}_{>k})\right]^2 \to 0, \quad n \to \infty.$$

Then the one-step-ahead smoothing is also mean-square consistent, i.e.,

$$\mathbb{E}\left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k})) - P_{k-1}(P_k(\boldsymbol{X}_{>k}), \boldsymbol{X}_{>k}))\right]^2 \to 0, \quad n \to \infty.$$

Proof. For the first term on the right of Equation (1.11),

$$\mathbb{E}\int_{\mathbb{R}^{d-k}} \left[S_{k-1}^{(n)}(S_k^{(n)}(\boldsymbol{x}_{>k}), \boldsymbol{x}_{>k}) - P_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}), \boldsymbol{x}_{>k})\right]^2 \mu(d\boldsymbol{x}_{>k}) \to 0$$

as the $\mathbb{R}^{d-k}$-dimensional integral is the integral of a non-negative function under some restrictions, where for the unrestricted integral

$$\mathbb{E}\int_{\mathbb{R}^{d-k+1}}\left[S_{k-1}^{(n)}(\boldsymbol{x}_{>k-1})-P_{k-1}(\boldsymbol{x}_{>k-1})\right]^2\mu(d\boldsymbol{x}_{>k-1})\to 0,\quad n\to\infty$$

holds by the assumed mean-square consistency. It is important that $S_k^{(n)}(\boldsymbol{x}_{>k})$ is within the convex hull of the sample space for X_k.

For the second term on the right of Equation (1.11), we use that P_{k-1} is an $L^2(\mu)$ operator, and therefore, it can be approximated (with any small precision) with a continuous function of compact support, which is also uniformly continuous (see Theorem A.1. of [15]). So to any $\varepsilon>0$ there is a $\tilde{P}_{k-1}$ which is uniformly continuous and

$$\int_{\mathbb{R}^{d-k+1}}\left[P_{k-1}(\boldsymbol{x}_{>k-1})-\tilde{P}_{k-1}(\boldsymbol{x}_{>k-1})\right]^2\mu(d\boldsymbol{x}_{>k-1})\le\varepsilon. \tag{1.12}$$

With this,

$$\begin{aligned}
&\left[P_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))-P_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})\right]^2\\
&\le 3\left[\tilde{P}_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))-\tilde{P}_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})\right]^2\\
&+3\left[P_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))-\tilde{P}_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))\right]^2\\
&+3\left[P_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})-\tilde{P}_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})\right]^2\\
&\le 3K^2\left[(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})-(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})\right]^2\\
&+3\left[P_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))-\tilde{P}_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))\right]^2\\
&+3\left[P_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})-\tilde{P}_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})\right]^2\\
&=3K^2\left[S_k^{(n)}(\boldsymbol{x}_{>k})-P_k(\boldsymbol{x}_{>k})\right]^2\\
&+3\left[P_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))-\tilde{P}_{k-1}(S_k^{(n)}(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k}))\right]^2\\
&+3\left[P_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})-\tilde{P}_{k-1}(P_k(\boldsymbol{x}_{>k}),\boldsymbol{x}_{>k})\right]^2,
\end{aligned} \tag{1.13}$$

where K is the constant coming from the uniform continuity of $\tilde{P}_{k-1}$. As

$$\mathbb{E}\left[S_k^{(n)}(\boldsymbol{X}_{>k})-P_k(\boldsymbol{X}_{>k})\right]^2\to 0,\quad n\to\infty,$$

by the consistency assumption,

$$\begin{aligned}
&\mathbb{E}\left[\tilde{P}_{k-1}(S_k^{(n)}(\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}))-\tilde{P}_{k-1}(P_k(\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\right]^2\\
&\le K^2\mathbb{E}\left[S_k^{(n)}(\boldsymbol{X}_{>k})-P_k(\boldsymbol{X}_{>k})\right]^2\to 0,\quad n\to\infty.
\end{aligned}$$

Finally, the integral (over $\mathbb{R}^{d-k+1}$) of the last two terms of Equation (1.13), apart from the constant 3, is bounded from above with ε, due to Equation (1.12). Taking the expectation of this 2ε error term w.r.t. the sampling procedure will result in an error $2D\varepsilon$, where D is a positive constant (valid for all n) coming from the fact that the sample space is bounded. □

Now we are able to formulate the general theorem about the mean-square consistency of the ICE on a DAG.

Theorem 1.1. *Let $X_1, \dots, X_d$ obey an absolutely continuous joint distribution with finite variances that is Markov compatible with a DAG (in this ordering). They are indexed such that $\boldsymbol{X}_{>k}$ are the variables without parents in the DAG and the others are to be predicted based on an n-element sample. With the notation of Equations* (1.4) *and* (1.5)*, for $j = k, k-1, \dots, 1$, let $P_j(\boldsymbol{x}_{>j})$ and $S_j^{(n)}(\boldsymbol{x}_{>j})$ be the exact solution of the regression task and a smoother (based on local averaging and mapping into the convex hull of the sample space of X_j), respectively. Assume that the individual smoothers give mean-square consistent estimates for the conditional expectations, i.e.,*

$$\mathbb{E}\left[S_j^{(n)}(\boldsymbol{X}_{>j}) - P_j(\boldsymbol{X}_{>j})\right]^2 \to 0, \quad n \to \infty, \quad j = k, k-1, \dots, 1.$$

Then the iterated smoothings are also mean-square consistent, i.e.,

$$\begin{aligned}\mathbb{E}\Big[&S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\\ &- P_j(P_{j+1}(P_{j+2}(\dots(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2 \to 0, \quad n \to \infty,\end{aligned}$$

for $j = k, k-1, \dots, 1$.

Note that the parameters of our mainly kernel-based smoothers and of the product kernels will be adapted to the data; other than kernel-based local averaging estimators are also considered in Section 1.5. Assumptions, under which the individual smoothers give a mean-square consistent estimate, are discussed in Section 1.4 and Appendix B. Also, by Markovity, instead of the set $\boldsymbol{X}_{>j}$, the parent set $\boldsymbol{X}_{\mathrm{Par}(j)}$ is considered during the algorithm.

Proof. Using Lemma 1.1 for the one-step-ahead prediction ($j = k$), with induction we proceed backward. Assume that we have proved the consistency until some $j \in \{k, k-1, \dots, 2\}$, i.e., we have proved that

$$\begin{aligned}\mathbb{E}\Big[&S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\\ &- P_j(P_{j+1}(P_{j+2}(\dots(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2 \to 0, \quad n \to \infty.\end{aligned} \tag{1.14}$$

Now we will prove the same for $j-1$. Indeed,

$$\begin{aligned}&\mathbb{E}\Big[S_{j-1}^{(n)}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\\ &\qquad - P_{j-1}(P_j(P_{j+1}(\dots,(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2\\ &\le 2\,\mathbb{E}\Big[S_{j-1}^{(n)}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\\ &\qquad - P_{j-1}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2\\ &\quad +2\,\mathbb{E}\Big[P_{j-1}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\\ &\qquad - P_{j-1}(P_j(P_{j+1}(\dots,(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2.\end{aligned} \tag{1.15}$$

The first term on the right of the inequality tends to 0 (as $n \to \infty$) by the mean-square consistency assumption. For the second term, we use that P_{j-1} is an $L^2(\mu)$ operator, and therefore, it can be approximated (with any small precision ε) with a continuous function $\tilde{P}_{j-1}$ of compact support, which is also uniformly continuous (with constant K). So, the same argument as in the proof of Lemma 1.1 yields

$$\begin{aligned}
2\,\mathbb{E}\Big[&\tilde{P}_{j-1}(S_j^{(n)}(S_{j+1}^{(n)}(\dots,(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}) \\
&- \tilde{P}_{j-1}(P_j(P_{j+1}(\dots,(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2 \\
\le K^2\,\mathbb{E}\Big[&S_j^{(n)}(S_{j+1}^{(n)}(S_{j+2}^{(n)}(\dots(S_k^{(n)}(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}) \\
&- P_j(P_{j+1}(P_{j+2}(\dots(P_k(\boldsymbol{X}_{>k}),\dots),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k}),\boldsymbol{X}_{>k})\Big]^2,
\end{aligned}$$

that tends to 0 ($n \to \infty$) by the induction hypothesis Equation (1.14). The integral (over $\mathbb{R}^{d-j+1}$) of the two terms, containing the differences between the functions of P_{j-1} and $\tilde{P}_{j-1}$ at the same arguments, is bounded from above with ε, due to an estimate analogous to Equation (1.12). Taking the expectation of this 2ε error term w.r.t. the sampling procedure will again result in an error $2D\varepsilon$ that can be arbitrarily small. This finishes the proof. □

It is important that the target variable has finite variance and it is absolutely continuous. Intermediate responses should be continuous, but discrete ordered variables can work quite well if they have a sufficiently large sample space. If the target variable is discrete, categorical, then in that step, we use the method to be introduced in the forthcoming section.

When the target variable is categorical

When X_k is categorical, taking on c different values, then the prediction is based on the Bayes rule via mode maximization, see [15]:

$$P_k(\boldsymbol{x}_{>k}) := \underset{1\le i\le c}{\operatorname{argmax}}\ \mathbb{P}\left(X_k = i \mid \boldsymbol{X}_{>k} = \boldsymbol{x}_{>k}\right).$$

With introducing the binary variables $\mathbb{I}_{\{X_k=i\}}$ ($i = 1,\dots,c$), the posterior probabilities are the conditional expectations, i.e., regression functions:

$$P_{k,i}(\boldsymbol{x}_{>k}) := \mathbb{P}\left(X_k = i \mid \boldsymbol{X}_{>k} = \boldsymbol{x}_{>k}\right) = \mathbb{E}\left(\mathbb{I}_{\{X_k=i\}} \mid \boldsymbol{X}_{>k} = \boldsymbol{x}_{>k}\right).$$

Given the training data for $\mathbb{I}_{\{X_k=i\}}, \boldsymbol{X}_{>k}$, we construct the (smoothing) estimate $S_{k,i}^{(n)}(\boldsymbol{x}_{>k})$ for each $i = 1,\dots,c$, and the plug-in estimate

$$S_k^{(n)}(\boldsymbol{x}_{>k}) := \underset{1\le i\le c}{\operatorname{argmax}}\ S_{k,i}^{(n)}(\boldsymbol{x}_{>k}).$$

Now we show that the error of the plug-in estimate is close to the optimal error of $P_k(\boldsymbol{x}_{>k})$:

$$
\begin{aligned}
0 &\le \mathbb{P}(S_k^{(n)}(\boldsymbol{X}_{>k}) \ne X_k) - \mathbb{P}(P_k(\boldsymbol{X}_{>k}) \ne X_k) \\
&= \mathbb{P}(P_k(\boldsymbol{X}_{>k}) = X_k) - \mathbb{P}(S_k^{(n)}(\boldsymbol{X}_{>k}) = X_k) \\
&= \sum_{i=1}^{c} \left[\mathbb{P}(P_k(\boldsymbol{X}_{>k}) = X_k \mid X_k = i) - \mathbb{P}(S_k^{(n)}(\boldsymbol{X}_{>k}) = X_k \mid X_k = i) \right] \mathbb{P}(X_k = i) \\
&\le \sum_{i=1}^{c} \int |P_{k,i}(\boldsymbol{x}_{>k}) - S_{k,i}^{(n)}(\boldsymbol{x}_{>k})| \, \mu(d\boldsymbol{x}_{>k}) \\
&\le \left[\sum_{i=1}^{c} \int |S_{k,i}^{(n)}(\boldsymbol{x}_{>k}) - P_{k,i}(\boldsymbol{x}_{>k})|^2 \, \mu(d\boldsymbol{x}_{>k}) \right]^{1/2} .
\end{aligned}
$$

This is true for any data in the smoothing within the n-element sample space for $X_k, X_{k+1}, \dots, X_d$. Then taking the expectation w.r.t. the sample,

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{P}(S_k^{(n)}(\boldsymbol{X}_{>k}) \ne X_k) - \mathbb{P}(P_k(\boldsymbol{X}_{>k}) \ne X_k) \right] \\
&\le D \left[\sum_{i=1}^{c} \int \left[S_{k,i}^{(n)}(\boldsymbol{x}_{>k}) - P_{k,i}(\boldsymbol{x}_{>k}) \right]^2 \mu(d\boldsymbol{x}_{>k}) \right]^{1/2} \to 0,
\end{aligned}
$$

as $n \to \infty$, by the mean-square consistency assumption, where D is a positive constant (independent of n) resulting from the boundedness of the sample space. The same holds for $j \in \{k-1, \dots, 2\}$. We also refer to the Bayes risk consistency in discrimination, see [18].

1.3.2 Consistency along a regression graph

Now the mean-square consistency is formulated in terms of the mean-square consistency of the vectors that in turn reduces to the mean-square consistency of their coordinates. Since

$$
\begin{aligned}
&\mathbb{E}\left\| \boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{X}_{C_{\ell-1}} \right\|^2 \\
&= \mathbb{E}\left\| \boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}) \right\|^2 + \mathbb{E}\left\| \boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{X}_{C_{\ell-1}} \right\|^2 \\
&= \sum_{j:\, X_j \in C_{\ell-1}} \mathbb{E}\left| S_{\ell-1,j}^{(n)}(\boldsymbol{X}_{C_\ell}) - P_{\ell-1,j}(\boldsymbol{X}_{C_\ell}) \right|^2 + \mathbb{E}\left\| \boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}) - \boldsymbol{X}_{C_{\ell-1}} \right\|^2 ,
\end{aligned}
$$

where the last term depends only on the nature of the underlying distribution, it is the first term that can be directly decreased by the sampling and smoothing procedure. It will be proved that the finitely many terms after the summation tend to 0 as $n \to \infty$, whenever the mean-square consistency of the individual smoothers is guaranteed. In this way, the following consistency theorem is stated under similar conditions as Theorem 1.1.

Theorem 1.2. *Let $X_1, \dots, X_d$ have an absolutely continuous joint distribution with finite variances that is Markov compatible with a regression graph. The variables are organized into chain components $C_1, \dots, C_\ell$ so that their joint distribution obeys the factorization of Equation* (1.2)*; further, they are indexed such that the last chain component C_ℓ contains the context variables and the others are to be predicted backward, from component to component,* *based on an n-element sample.*

With the vector extensions of Equations (1.7) *and* (1.8) *of the conditional expectations and the smoothers, and assuming that the individual smoothers component-wise give mean-square consistent estimates for the conditional expectations, the iterated smoothings are also mean-square consistent, i.e.,*

$$\begin{aligned}\mathbb{E}\Big\|&\boldsymbol{S}_k^{(n)}(\boldsymbol{S}_{k+1}^{(n)}(\boldsymbol{S}_{k+2}^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\\ &-\boldsymbol{P}_k(\boldsymbol{P}_{k+1}(\boldsymbol{P}_{k+2}(\dots(\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big\|^2\to 0,\quad n\to\infty,\end{aligned}$$

for $k=\ell-1,\ell-2,\dots,2,1$.

Proof. In the first step,

$$\mathbb{E}\left\|\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell})-\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell})\right\|^2=\sum_{j:\,X_j\in C_{\ell-1}}\mathbb{E}\left[S_{\ell-1,j}^{(n)}(\boldsymbol{X}_{C_\ell})-P_{\ell-1,j}(\boldsymbol{X}_{C_\ell})\right]^2,$$

where the finitely many terms after the summation tend to 0 (as $n\to\infty$), whenever the mean-square consistency of the individual smoothers is guaranteed.

Analogously to the considerations of Lemma 1.1 and Theorem 1.1, by induction, we proceed backward. Assume that we have proved the consistency until some k, and now we prove it for $k-1$:

$$\begin{aligned}&\mathbb{E}\Big\|\boldsymbol{S}_{k-1}^{(n)}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\\ &\qquad-\boldsymbol{P}_{k-1}(\boldsymbol{P}_k(\dots(\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big\|^2\\ &\le 2\,\mathbb{E}\Big\|\boldsymbol{S}_{k-1}^{(n)}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\\ &\qquad-\boldsymbol{P}_{k-1}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big\|^2\\ &+2\,\mathbb{E}\Big\|\boldsymbol{P}_{k-1}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\\ &\qquad-\boldsymbol{P}_{k-1}(\boldsymbol{P}_k(\dots(\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big\|^2.\end{aligned}$$

The first term (apart from the multiplier 2) on the right of the inequality is

$$\begin{aligned}&\mathbb{E}\Big\|\boldsymbol{S}_{k-1}^{(n)}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\\ &\qquad-\boldsymbol{P}_{k-1}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big\|^2\\ &=\sum_{j:\,X_j\in C_{k-1}}\mathbb{E}\Big[S_{k-1,j}^{(n)}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\\ &\qquad-P_{k-1,j}(\boldsymbol{S}_k^{(n)}(\dots(\boldsymbol{S}_{\ell-1}^{(n)}(\boldsymbol{X}_{C_\ell}),\dots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big]^2,\end{aligned}$$

where the individual terms after the summation tend to 0 (as $n\to\infty$) by the mean-square consistency of the individual smoothers.

For the second term we use that for all j, such that $X_j\in C_{k-1}$, $P_{k-1,j}$ is an $L^2(\mu)$ operator, and therefore, it can be approximated (with any small precision ε_j) with a continuous function

$\tilde{P}_{k-1,j}$ of compact support, that is also uniformly continuous (with constant K_j). With it,

$$\begin{aligned}
&\mathbb{E}\Big[\tilde{P}^{(n)}_{k-1,j}(\boldsymbol{S}^{(n)}_k(\ldots(\boldsymbol{S}^{(n)}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\ldots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell}) \\
&\qquad - \tilde{P}_{k-1,j}(\boldsymbol{P}_k(\ldots(\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\ldots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big]^2 \\
&\le K_j^2\,\mathbb{E}\Big\|\boldsymbol{S}^{(n)}_k(\ldots(\boldsymbol{S}^{(n)}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\ldots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell}) \\
&\qquad - \boldsymbol{P}_k(\ldots(\boldsymbol{P}_{\ell-1}(\boldsymbol{X}_{C_\ell}),\ldots),\boldsymbol{X}_{C_\ell}),\boldsymbol{X}_{C_\ell})\Big\|^2,
\end{aligned}$$

that tends to 0 ($n \to \infty$) by the induction hypothesis. The integral over $\mathbb{R}^{|C_k|+\cdots+|C_\ell|}$ of the two terms, containing the differences between the functions of $P_{k-1,j}$ and $\tilde{P}_{k-1,j}$ at the same arguments, is bounded from above with ε_j, due to an estimate analogous to Equation (1.12). Taking the expectation of this ε_j error term w.r.t. the sampling procedure will again result in an error $D\varepsilon_j$ that can be arbitrarily small. Here D is a positive constant (independent of n), coming from the fact that the sample space is bounded. Applying this for all j, such that $X_j \in C_{k-1}$, the sum of these finitely many terms will also tend to 0. This finishes the proof. □

1.3.3 On the strong consistency of ICE

Previously, we investigated the mean-square consistency of our algorithm. The necessary condition was that the individual smoothers are mean-square consistent along with the iteration. The question arises: replacing the mean-square consistency assumption, and leaving other conditions intact, is the strong consistency of the individual smoothers sufficient condition to prove the strong consistency of ICE? The strong consistency with the notation of Equation (1.10) means that

$$\mathbb{P}\left(\int_{\mathbb{R}^{d-k}}\left[S^{(n)}_k(\boldsymbol{x}_{>k}) - P_k(\boldsymbol{x}_{>k})\right]^2 \mu(d\boldsymbol{x}_{>k}) \to 0\right) = 1, \quad n \to \infty.$$

We conjecture the following.

The proof of Lemma 1.1, assuming strong consistency, is analog up to Equation (1.13). Then the expectation w.r.t. the sampling is replaced by the probability w.r.t. the same. Of the remaining three terms, the first one goes to zero with probability one as we assumed the strong consistency. The two remaining members can be similarly estimated from above with $(3)\varepsilon$, which is a constant independent of the sampling, so they do not affect the probability of tending to zero.

The proof of Theorem 1.1 requires replacing the induction hypothesis with a strong consistency assumption. Then the estimation is done analogously to Lemma 1.1. Here, there would be four terms, as the second term of Equation (1.15) falls into three terms. The strong consistency assumed in the theorem holds for the first member, the strong consistency assumed in the induction hypothesis holds for the second member, and the remaining two members can be estimated in the same way from above with $(2)\varepsilon$, which is a constant independent of the sampling, so they do not affect the probability of tending to zero.

The extension of Theorem 1.2 is completely analogous to that of Theorem 1.1, but with the addition of more ε constants independent of the sampling. The number of these is proportional to the size of the chain component in the sequence. Thus, as long as we assume only finite chain components, strong consistency holds too if the individual smoothers are also like that.

We did not go beyond the mean-square consistency in [J2], the conjectures above may require further investigation.

1.4 The numerical algorithm

1.4.1 Implementation of the smoothings

Note that in Theorems 1.1 and 1.2 we utilize that the smoothing is mean-square consistent in every step. Again, there are different smoothers, for example, k-NN, partitioning, and kernel regression estimation. These are all local averaging techniques of Equation (B.1), each with different conditions regarding their mean-square consistency (see e.g., [15]), usually derived from Theorem B.1. In the realization of our iterated method, we mainly use kernel-based smoothing, more precisely the Nadaraya–Watson (NW) kernel regression estimation [19, 20]. Some theoretical background on kernel regression is provided in Section B.1.

We have implemented our method based on the Python package of **Statsmodels** [21]. This package uses a product kernel setup in the multivariate case, see Equation (B.3), thus works with a vector $\boldsymbol{h}$ of bandwidths for the kernel regression estimation. The source for this is mainly the works of Racine and Li [16, 22], who have refined the product kernel approach with the usage of mixed (continuous and discrete) regressors and a proper cross-validated choice of bandwidths. For the theoretical background behind the selection of the kernels and bandwidths, see [16] and Section B.1.

To compare the product kernel approach with other methods, we include results obtained by using a full bandwidth matrix $\boldsymbol{H}$ (see Section B.1), by the local linear version of the product kernel approach, and by the k-NN regression (see Section B.2).

Kernel selection of Racine and Li

In Equation (B.3), each univariate kernel and the corresponding bandwidth parameter are chosen according to the regressor variables. This implementation needs the type of the regressors as an input. They can be of three types: (a) continuous, (b) ordered (discrete), (c) unordered (discrete, categorical). Based on this, the three default univariate kernels are as follows:

(a) Gaussian kernel in the continuous case:

$$K_h^{(G)}\left(x, x^{(i)}\right) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{\left(x - x^{(i)}\right)^2}{2h^2}},$$

(b) Wang–Ryzin kernel in the ordered case:

$$K_h^{(WR)}\left(x, x^{(i)}\right) = (1-h) \times \mathbb{I}_{\{x = x^{(i)}\}} + \frac{1}{2}(1-h) h^{|x - x^{(i)}|} \times \mathbb{I}_{\{x \neq x^{(i)}\}},$$

(c) Aitchison–Aitken kernel in the unordered case if the categorical variable X admits c different values:

$$K_h^{(AA)}\left(x, x^{(i)}\right) = (1-h) \times \mathbb{I}_{\{x=x^{(i)}\}} + \frac{h}{c-1} \times \mathbb{I}_{\{x \neq x^{(i)}\}}.$$

In this way, the multivariate product kernel becomes

$$K_{\boldsymbol{h}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \prod_{j=1}^{r_c} K_{h_{n,j}}^{(G)}\left(x_j, x_j^{(i)}\right) \prod_{k=1}^{r_o} K_{h_{n,k}}^{(WR)}\left(x_k, x_k^{(i)}\right) \prod_{l=1}^{r_u} K_{h_{n,l}}^{(AA)}\left(x_l, x_l^{(i)}\right),$$

where r_c, r_o, r_u are the number of continuous, discrete ordered, and discrete unordered regressors, respectively. The choice of the kernel is of little importance compared to the bandwidths, and the conditions on the kernel can be relaxed as pointed out by Racine and Li:

- the Gaussian kernel can be replaced with any compactly supported kernel function that is Hölder continuous, e.g., the Epanechnikov kernel;
- the Wang–Ryzin kernel can be replaced with the Aitchison–Aitken kernel, which further simplifies things but performs slightly worse.

Bandwidth selection

The bandwidth is chosen in every iteration step independently, with two possible validation methods or with the forthcoming rule-of-thumb. The first is the leave-one-out cross-validation, by minimizing

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - S_Y^{(n)\setminus\{i\}}(\boldsymbol{x}^{(i)})\right)^2 M(\boldsymbol{x}^{(i)}), \tag{1.16}$$

where $S_Y^{(n)\setminus\{i\}}$ is the estimator in which the ith sample entry is left out from the prediction. $M(\cdot)$ is an optional weight function that trims out boundary observations. In [16], the asymptotic normality results are proved for this choice of bandwidth.

Alternatively, one can use the minimization of the corrected AIC Hurvich criteria, see [23]. These two methods produce asymptotically the same results, [24]. The optimization is performed by the Nelder–Mead downhill simplex method in the underlying Python implementation.

Scott's rule-of-thumb [25] may also work well for the selection of the bandwidth:

$$h_{n,j} \approx 1.06 \times \hat{\sigma}_j \times n^{-\frac{1}{4+r}},$$

where $\hat{\sigma}_j$ is the sample standard deviation of X_j and r is the number of regressors in the particular step. Note that this is close to optimal for Gaussian variables, and also used as the starting point of the code in the aforementioned cross-validation optimization processes.

The aforementioned bandwidth selections are data-driven, their asymptotic behavior for the NW (local constant) and the local linear kernel regression is studied in [16, 23, 24, 26]. In [16], the authors prove that under the conditions listed in Assumptions B.1 and B.2, the estimate (with the leave-one-out cross-validated choice of bandwidths) is asymptotically normal, see Theorem B.2. Their proof uses the same $\hat{h_c}$ for all continuous, and another same $\hat{h_d}$ for all discrete regressors, also they use the Aitchison–Aitken kernel for all discrete variables. They anticipate that similar results can be proved for distinct parameters and using the Wang–Ryzen kernel for discrete ordered regressors. Asymptotic normality does not always imply mean-square consistency, but under some extra conditions we are able to prove the mean-square consistency of their estimator, see Lemma B.1.

Using a full bandwidth matrix

Again, we include results of the NW estimate with a full bandwidth matrix $\boldsymbol{H}$ as well, see Equation (B.2). We use the multivariate version of the Gaussian kernel,

$$K_{\boldsymbol{H}}^{(G)}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \frac{1}{|\boldsymbol{H}|\sqrt{2\pi}^r}\, e^{-\frac{1}{2}\left[\boldsymbol{H}^{-1}\left(\boldsymbol{x}-\boldsymbol{x}^{(i)}\right)\right]^T\left[\boldsymbol{H}^{-1}\left(\boldsymbol{x}-\boldsymbol{x}^{(i)}\right)\right]},$$

and the multivariate version of Scott's rule [27] for the selection of a full bandwidth matrix:

$$\boldsymbol{H}_n \approx 1.06 \times \hat{\Sigma}^{\frac{1}{2}} \times n^{-\frac{1}{4+r}},$$

where $\hat{\Sigma}$ is the sample covariance matrix of $\boldsymbol{X}$, and r is the number of regressors in a particular step. As for the mean-square consistency of this estimator, we have not found any results in the literature yet.

1.4.2 Practical considerations

As for the realization of our algorithm, some remarks are in order.

In the case of a discrete unordered categorical intermediate response, the kernel regression method leads to a Bayes classification method via dummy indicator variables, as seen on page 21. Then we continue with the original variable in a consecutive step on the regression graph, not with its indicators.

In the case of a discrete ordered intermediate response, the regression estimation produces a value between two possible consecutive values of the variable. One can leave these predicted values as they are, or substitute a rounded value for them to have a classification instead of regression.

In the case of long chains, our iterative algorithm may result in overtly smooth predictions, compared to a direct prediction, because the smoothing will affect not only the target but all previously predicted variables too.

1.4.3 On the equivalence of DAGs and regression graphs

In lack of a forbidden quadruple of Figure 1.2, we can find a Markov equivalent DAG to a regression graph as mentioned in Section 1.1.7. If we proceed from chain component to chain component along the directed edges, and predict the joint response variables of the same chain component in parallel, we get the same result thanks to the Markov equivalence of the original chain graph and the DAG.

When we prove the consistency of our algorithm along a DAG, the starting object can be this DAG, with the labeling of the vertices given in [12, Algorithm 1], that is applicable if there is no forbidden quadruple in the underlying chain graph. The possible directed edges converted from dashed edges are taken into consideration in the iteration, but they do not add any extra information to the prediction of joint responses along the arrows of the original regression graph. This is supported by the Cochran's theorem, and by the factorization property of these models (Section 1.1.6). Note that this DAG is not unique, but the actual topological ordering of its vertices should comply with the ordering of the chain components.

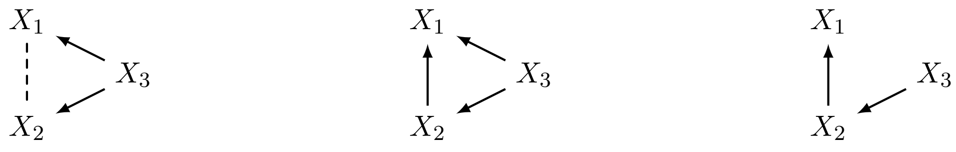
When the regression graph does not contain any forbidden quadruple, the DAG version of our algorithm, processed on the Markov equivalent DAG, gives the same results as using parallel regressions along the chain components. In the presence of a forbidden quadruple, if our joint distribution is Markov compatible with a regression graph, we proved consistency along the chain components with vector valued smoothers. By using parallel regressions, we disregard the dashed edges of the response components during the iteration. It is justified by that in regression graphs the conditioning set for a response variable does not include other responses within the same chain component.

In Figure 1.3, we illustrate the equivalence of the predictions in the two setups with three variables. Let X_1, X_2 be on equal standing, connected with a dashed edge in the same chain component, whereas their joint parent X_3 is in a chain component of their past (Figure 1.3a). So our iteration predicts X_1 only with X_3, with regression coefficient $\beta_{1|3}$. After constructing a DAG, then say, there is a directed edge $X_2 \to X_1$ (replacing the former dashed edge), as in Figure 1.3b. Along this DAG, our algorithm predicts first X_2 with X_3, then X_1 with X_3 (old parent) and with X_2 (new parent). But in this case, the direct and indirect effect of X_3 is added together as $\beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3}$, where $\beta_{1|3.2}$ is the partial regression coefficient of X_3 when regressing X_1, given also X_2 as regressor for X_1. By the Cochran's formula (see Wermuth and Sadeghi [12] and Cox and Wermuth [8]):

$$\beta_{1|3} = \beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3}, \tag{1.17}$$

so we get the same result if our variables are Gaussian, or we confine ourselves to the second moments. As the smoothings imitate the conditional expectations, the predictions would be the same if we used the Markov equivalent DAG. The above equivalence extends to several variables, Wright [28] discusses it with a more complicated notation.

From Equation (1.17) we can see that in the special case when $X_3 \to X_2 \to X_1$ form a Markov chain, i.e., $X_1 \perp\!\!\!\perp X_3 | X_2$ (see Figure 1.3c), $\beta_{1|3.2} = 0$ and $\beta_{1|2.3} = \beta_{1|2}$. Therefore, $\beta_{1|3} = \beta_{1|2}\beta_{2|3}$. If our variables have unit variances, the above product rule extends to the correlation coefficients, and so, $r_{13} = r_{12}r_{23}$. Consequently, for the regression errors we have $1 - r_{12}^2 \leq 1 - r_{13}^2$, showing that the prediction error of X_1 by X_2 is smaller than that by X_3, via X_2. During the iteration of the algorithm, there are indirect and direct effects, so the prediction errors combine in the above way. The main concepts extend to absolutely continuous distributions, other than Gaussian, if we confine ourselves to the second moments.



(a) (b) (c)

Figure 1.3: Examples for Cochran's formula

1.5 Applications of ICE

Three application examples are included. The first is based on randomly generated data; the theoretical regression function can be well controlled in this environment. The second deals with the 2014's Egypt Demographic and Health Survey (EDHS 2014) data and examines the effect of background characteristics on the ideal number of children a woman thinks manageable to have. The third considers data collected at the Budapest University of Technology and Economics (BME) to study students' academic achievements in the first two semesters, related to their background characteristics and university application scores.

In the two real-life examples, the regression graphs are built by expert knowledge and the R package **gRchain** [29] that implements the methods described in [8] for multivariate regression chain graphs. In these two examples, 90%–10% training–test data split is applied, and the results refer to the test data. The following smoothings are selected:

- the NW estimate of Racine and Li with Scott's bandwidths (NW_{sc});
- the same NW estimate with cross-validated bandwidths (NW_{cv});
- the NW estimate but with the full bandwidth matrix of Scott (NW_H);
- the local linear estimate of Racine and Li with Scott's bandwidths (LL_{sc});
- the k-NN estimate with cross-validated choices for k.

1.5.1 Randomly generated regression graph

Consider the following randomly generated example. The context variables (X_5, X_6, X_7, X_8) obey a zero-centered 4-dimensional normal distribution with covariance and concentration matrices as follows.

$$\mathbf{\Sigma} = \begin{bmatrix} 0.8 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1.2 & 0.8 & 0.4 \\ 0.4 & 0.8 & 1.2 & 0.6 \\ 0.2 & 0.4 & 0.6 & 0.8 \end{bmatrix}, \qquad \mathbf{\Sigma}^{-1} = \begin{bmatrix} -2 & -1 & 0 & 0 \\ -1 & -2 & -1 & 0 \\ 0 & -1 & -2 & -1 \\ 0 & 0 & -1 & -2 \end{bmatrix}.$$

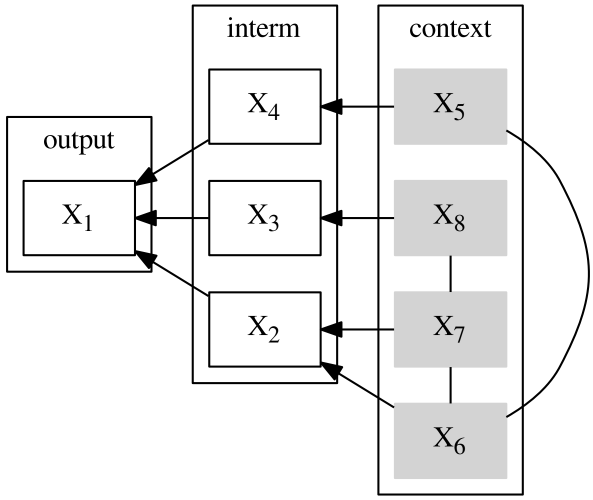


Figure 1.4: Regression graph of the generated data

The size of the sample is 1000. Response variables are generated with the equations:

$$\begin{aligned}
X_4 &:= X_5^2 + \varepsilon_4, \\
X_3 &:= X_8^3 + \varepsilon_3, \\
X_2 &:= X_6 X_7 + \varepsilon_2, \\
X_1 &:= X_2 + X_3 + X_4 + \varepsilon_1,
\end{aligned}$$

where ε_is are independent, standard normal error terms. This naturally results in the regression graph of Figure 1.4.

The context variables X_5, X_6, X_7, X_8 are considered to be known, and we predict X_1 through the intermediate responses X_2, X_3, X_4. In this artificial environment, the regression functions are known, therefore obviously, the prediction is accurate enough, see Figure 1.5.

In this example, the Gaussian kernel was used for each variable. In Figures 1.5 and 1.6, the response variable (X_1) is plotted versus the four context variables. The data points, their iterated theoretical regression, and the smoothed estimate by ICE are included; a leave-one-out cross-validation, see Equation (1.16), was applied in each step of the iteration. Error scores (R^2 and MSE) versus the true value of X_1 and the theoretical regression function, are shown in Table 1.1 with respect to the different bandwidth choices and smoothers listed in Section 1.4.1. With the relatively small sample size of 1000, the cross-validated methods were able to achieve better results than the rule-of-thumb. The scores in the first row are calculated according to

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - x_1^{\{i\}}\right]^2, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - x_1^{\{i\}}\right]^2}{\sum_{i=1}^{n}\left[\overline{x_1} - x_1^{\{i\}}\right]^2}, \tag{1.18}$$

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
X_1	4.385	3.982	4.402	3.790	4.543
$P_1(\cdot)$	0.578	0.304	0.629	0.194	1.164

(a) Mean-square errors

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
X_1	0.682	0.696	0.684	0.705	0.675
$P_1(\cdot)$	0.965	0.971	0.962	0.978	0.895

(b) R^2 scores

Table 1.1: Error scores of the generated example: versus the true value of X_1 (first rows) and the theoretical regression function (second rows)

while, in the second row,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - P_1^{\{i\}}(\cdot)\right]^2, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}\left[S_1^{\{i\}}(\cdot) - P_1^{\{i\}}(\cdot)\right]^2}{\sum_{i=1}^{n}\left[\overline{P_1(\cdot)} - P_1^{\{i\}}(\cdot)\right]^2},$$

where $S_1(\cdot)$ and $P_1(\cdot)$ are as in our Theorem 1.1, and the shorthand $(\cdot)$ for the arguments corresponds to the smoothings and conditional expectations of the previous steps, respectively. The superscript $^{\{i\}}$ indicates the ith test sample entry. These results reassure that in the continuous case and when a well-compatible regression graph is available, the estimate is accurate enough. In this simulated case, the different smoothers give similar results. The cross-validation improved the results of the NW estimate but the local linear regression, even without cross-validation, outperformed the other smoothers. Using a full bandwidth matrix had little if any impact on the results.

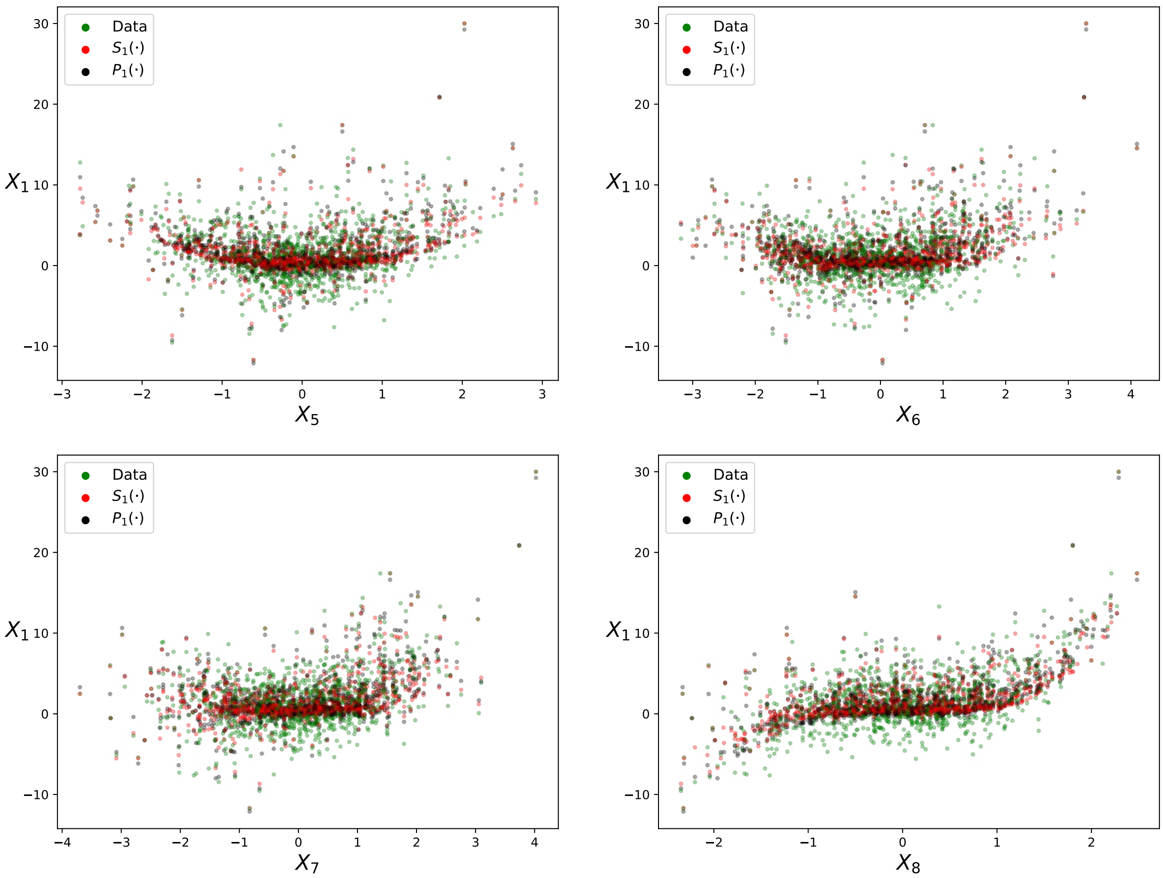


Figure 1.5: ICE on the generated example: the response variable (X_1) is plotted versus each of the four context variables (X_5, X_6, X_7, X_8); the data points are in green, the iterated theoretical regression is in black, and the smoothed estimate by ICE is in orange

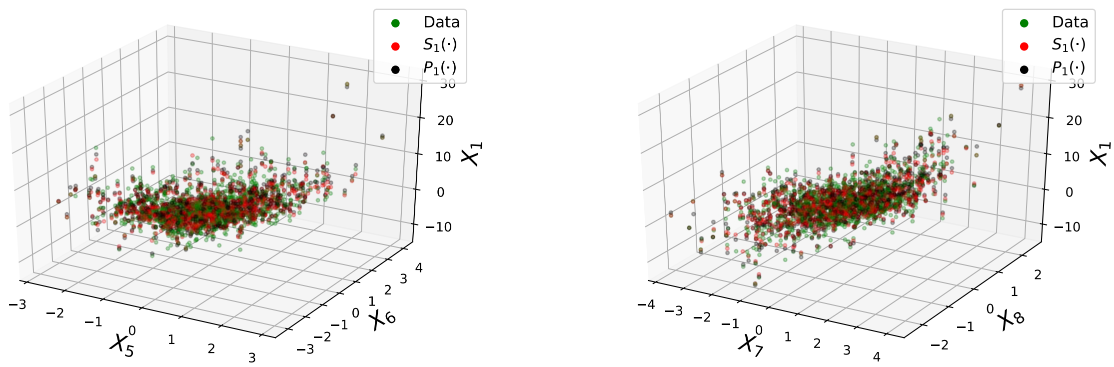


Figure 1.6: ICE on the generated example (3D): the response variable (X_1) is plotted versus the two pairs of context variables (X_5, X_6 and X_7, X_8); the data points are in green, the iterated theoretical regression is in black, and the smoothed estimate by ICE is in orange

1.5.2 The EDHS 2014 data

Based on data from the EDHS 2014, we examined the effect of background characteristics on the ideal number of children a woman thinks manageable to have. The research question is: to what extent do age and education level of married couples affect the conceivable ideal number of children, through intermediate variables. The focus is on a selected random sample of more than 17000 urban married women aged 20-49 years. Figure 1.7 shows the regression graph, built on the variables listed below. The grouping of the variables is based on the relationships between the variables suggested by the experts.

For descriptive statistics of the included variables, see Table 1.2. As shown in the regression graph, Figure 1.7, the far right box includes the relevant context variables in the model:

- the Husband's and Wife's Education level in years;
- the Husband's and Wife's Age.

The next box from the right contains two intermediate variables:

- the Woman's age at the first marriage;
- Wealth index of the family (discrete, ordered, from 1 to 5).

The graph shows that some variables are considered explanatory for some variables and responses to others. Moving to the next box, the secondary responses are presented:

- Number of years the woman has been using any Contraception method;
- Number of births.

The first box on the left is the primary response variable, the ideal number of children the woman thinks to be optimal.

The choice of kernels is based on considering the age/year-related variables to be continuous, and the others to be ordered discrete. The variables were min-max scaled to the $[0, 1]$ interval before the application but were rescaled afterward for the sake of the figures. Our results show an overtly smooth regression surface after the iteration. In Figure 1.8, one can see the iterated estimates along the graph. A row of scatter plots belongs to every non-context variable. In each subplot, a non-context variable is presented versus its parent variables, one by one. These are cross-sections of the multidimensional space. Red dots correspond to the actual data points, blue and yellow dots belong to the estimates given by the ICE versus the estimated value of the parent and the actual value of it, respectively. In a subplot, the blue and yellow dots obviously coincide if the parent is a context node or a node with known values. Note that some jittering is added to the data points to have more comprehensible figures. For example, we can see that, regardless of the background characteristics, the iterated regression estimate for the ideal number of children is around 3. In our iterated regime we can calculate error scores for each non-context variable. The calculated R^2-scores and the mean-square errors, Equation (1.18), are presented in Table 1.3 for all non-context variables with respect to the different smoothings listed in Section 1.4.1. With the relatively large sample, the other kernel methods were unable to achieve much better results than the rule-of-thumb NW estimate. The k-NN provided similar results.

	count	mean	std	min	max
AgeWoman	17686	32.79	7.89	20	49
WealthIndex	17686	3.23	1.42	1	5
NumOfBornChildren	17686	2.68	1.62	0	15
AgeWomanAtFirstMar	17686	20.55	4.05	9	49
AgeHusband	17686	39.16	9.56	14	86
IdealNumOfChildren	17686	3.08	1.37	0	24
SchoolYearsWoman	17686	9.34	5.55	0	23
SchoolYearsHusband	17686	9.82	5.28	0	23
ContraceptionYears	17686	5.07	6.16	0	32

Table 1.2: Descriptive statistics of the variables in the EDHS example

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
IdealNumOfChildren	0.003	0.003	0.003	0.003	0.003
NumOfBornChildren	0.007	0.007	0.007	0.007	0.007
ContraceptionYears	0.030	0.030	0.030	0.030	0.030
WealthIndex	0.100	0.095	0.094	0.094	0.094
AgeWomanAtFirstMar	0.007	0.006	0.006	0.006	0.006

(a) Mean-square errors

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
IdealNumOfChildren	0.028	0.029	0.031	0.031	0.023
NumOfBornChildren	0.378	0.390	0.393	0.396	0.386
ContraceptionYears	0.201	0.203	0.190	0.198	0.195
WealthIndex	0.244	0.251	0.254	0.253	0.252
AgeWomanAtFirstMar	0.309	0.313	0.322	0.325	0.322

(b) R^2 scores

Table 1.3: Error scores per variable along the iteration in the EDHS example

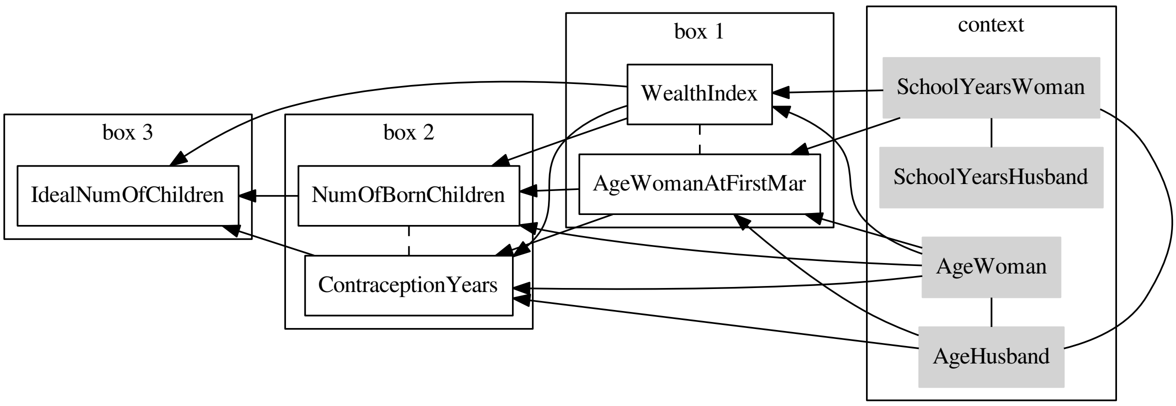


Figure 1.7: Regression graph of the EDHS example

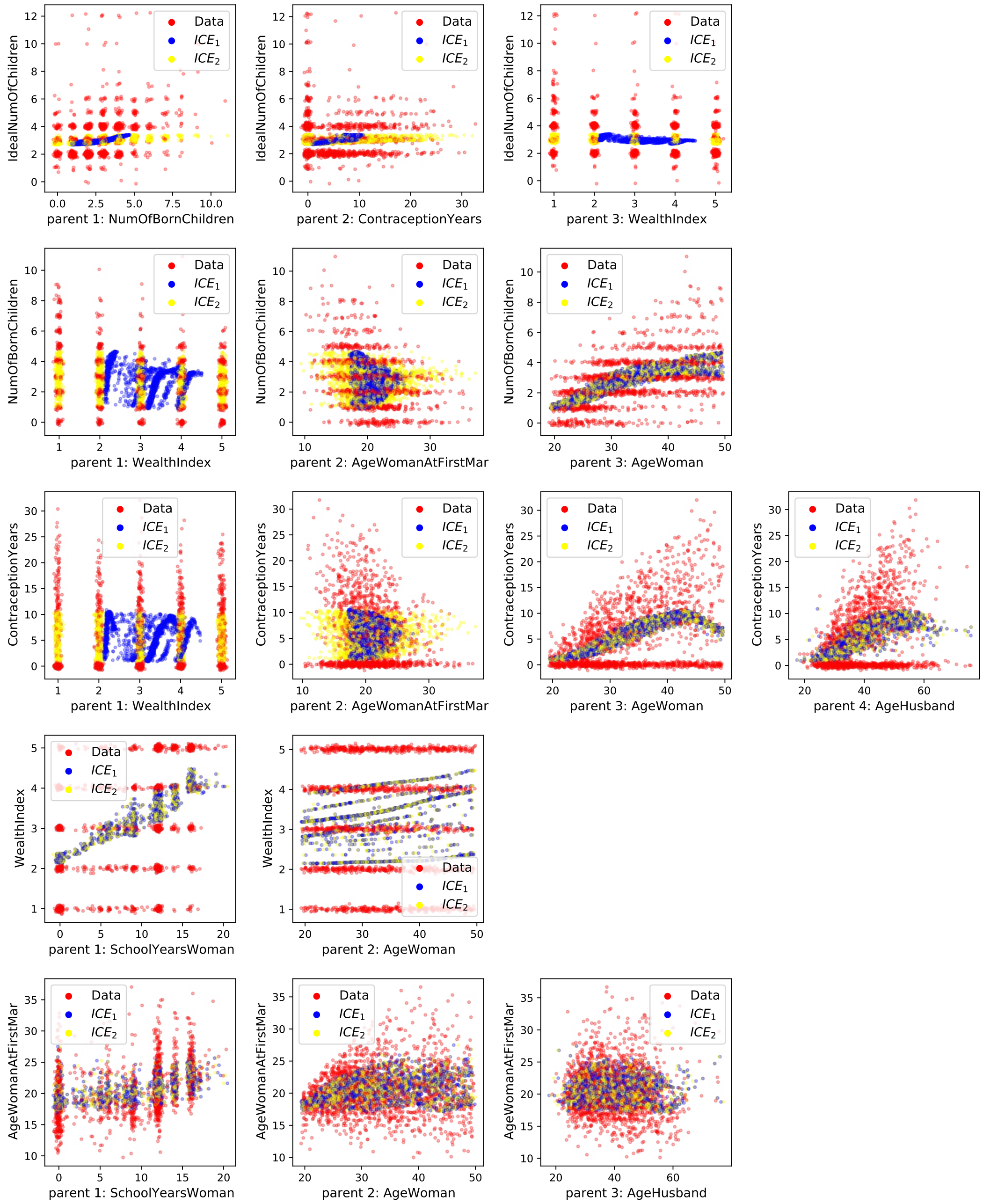


Figure 1.8: ICE on the EDHS data: to each non-context variable a row of the scatter plot belongs. In each subplot, a non-context variable is presented versus one of its parent variables. Red dots correspond to the actual data points, blue and yellow dots belong to the estimates given by the ICE versus the estimated value of the parent and the actual value of it, respectively.

1.5.3 On students academic achievements

The Central Academic Office of BME (short: KTH) registers the data of all students entering the institution. We were interested in the iterative prediction of some characteristics of the students, based on variables that are already available at their application into the university.

The involved data consists of records for over 12000 students with the limitation that it is their first time entering into our university and they wrote a so-called 'zeroth midterm' of mathematics.

In the regression graph of Figure 1.9, the boxes are formed by the chronological ordering of the included variables. The far right box includes the relevant context variables in the model, which are already available during the application process into the university:

- Matura points (secondary school exit examination);
- Study points, calculated from the grades of the core secondary school subjects;
- Extra points, earned in the application process;
- Gender and Age when starting the first semester.

For a detailed description of the Hungarian university admission system, see [30]. The next box from the right contains an intermediate variable, which is the points earned at the aforementioned "zeroth midterm". It is a mandatory entry examination written by first-year students of the BME in mathematics. It is taken in the first week of the semester by almost all entering students.

Moving to the next box, the secondary responses are:

- Grade of the first-semester maths (calculus) course (discrete, ordered, from 1 to 5);
- Credit weighted average result of the first semester without the result of the aforementioned calculus course (continuous).

The first box on the left is the primary response variable, the cumulative credit weighted average result of the first two semesters without the result of the first-semester calculus course. Some descriptive statistics of the involved variables are in Table 1.4.

The variables in this example are not all continuous, but for most of them, the sizes of the sample spaces are relatively large, so we can treat them as continuous variables. Finally, Gender is considered as binary, Age as discrete ordered (its sample space is too small), and the Grade of mathematics also as discrete ordered, every other variable is treated as continuous.

In Figure 1.10, scatter plots of the iteration steps are presented, akin to the EDHS example, with the leave-one-out cross-validated choices of bandwidths. Error scores are shown in Table 1.5 for all non-context variables and with respect to the different smoothings listed in Section 1.4.1. Again, with the relatively large sample, the other kernel methods were unable to achieve much better results than the rule-of-thumb NW estimate, and the k-NN also provided similar numbers. It seems that, with a large enough sample, the techniques aimed at improving the results – like cross-validation over rule-of-thumb, full bandwidth matrix over independent kernels, local linear over NW estimation – have only a small effect on the results when applying the ICE algorithm. However, the computational burden can be significant. The problem with k-NN is that, as the sample size grows, finding the close neighbors becomes intractable. Though implementations handle this issue with clever algorithms, the choice of k remains a challenge.

	count	mean	std	min	max
ZeroExamPoints	11876	29.50	15.66	-13	60
MathsGrade	11876	2.45	1.25	1	5
Gender	11876	0.74	0.44	0	1
Age	11876	20.19	1.02	18	43
CWA w/o Maths	11876	1.86	1.15	0	5
CumCWA w/o Maths	11876	2.10	1.28	0	6
StudyPoints	11876	174.71	14.75	87	200
MaturaPoints	11876	167.66	17.96	81	200
ExtraPoints	11876	69.20	23.46	0	100

Table 1.4: Descriptive statistics of the variables in the educational example

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
CumCWA w/o Maths	0.024	0.024	0.025	0.024	0.024
CWA w/o Maths	0.026	0.026	0.026	0.026	0.026
MathsGrade	0.072	0.072	0.074	0.070	0.071
ZeroExamPoints	0.031	0.031	0.032	0.033	0.031

(a) Mean-square errors

	NW_{sc}	NW_{cv}	NW_H	LL_{sc}	kNN
CumCWA w/o Maths	0.426	0.432	0.412	0.428	0.426
CWA w/o Maths	0.430	0.433	0.426	0.431	0.432
MathsGrade	0.368	0.356	0.326	0.347	0.333
ZeroExamPoints	0.331	0.333	0.318	0.300	0.338

(b) R^2 scores

Table 1.5: Error scores per variable along the iteration in the educational example

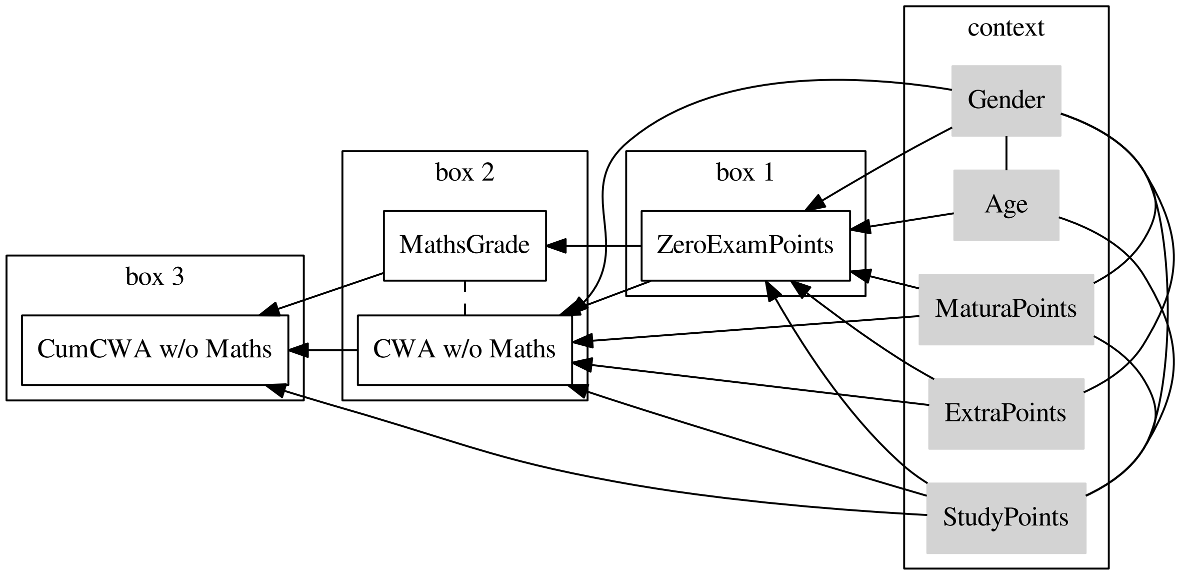


Figure 1.9: Regression graph of the educational example

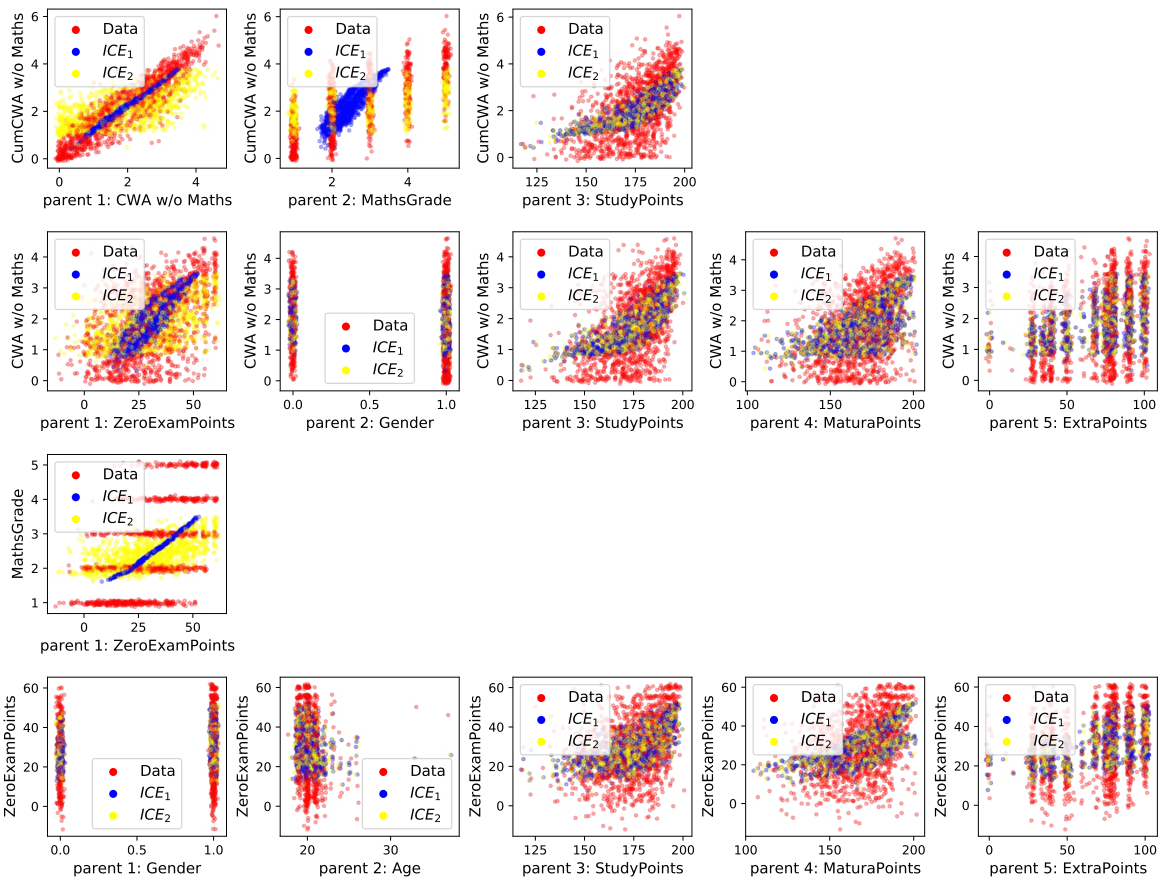


Figure 1.10: ICE on the educational data: to each non-context variable a row of the scatter plot belongs. In each subplot, a non-context variable is presented versus one of its parent variables. Red dots correspond to the actual data points, whereas, blue and yellow ones to the estimates given by the ICE versus the estimated value of the parent and the actual value of it, respectively.

Chapter 2

Regression Discontinuity Design (RDD), with Jittering

This chapter is based on my article with Roland Molontay [J3], the extended version of our earlier conference paper [C2]. In these, we apply the widely used statistical method called Regression Discontinuity Design (RDD) to measure the effectiveness of differently designed remedial mathematics courses at the Budapest University of Technology and Economics (BME). We study both the direct and longer-term effects of mathematical remediation using RDD tools in various settings. The RDD is described in Section 2.1.

As a methodological novelty, we propose an alternative method (called RDD with jittering) to handle the problem of a discrete running variable (see Section 2.2). This is detailed in Section 2.2.2. At the end of this chapter, in Section 2.3, some connections of RDD to graphical modeling are included.

In Section 2.4, our detailed case study is presented, where we studied the impact of mathematical remediation at BME on students' subsequent academic achievements regarding both first-year mathematics courses and longer-term effects. We investigate two differently designed remedial courses (detailed in Section 2.4.1).

2.1 Preliminaries on the RDD methodology

RDD is a quasi-experimental approach. Quasi-experimental is an umbrella term for statistical methods designed for different situations where randomized controlled trials are not possible. Most of these are extensively used, and important conclusions are drawn in different areas of life. In 2021, the Nobel Prize in Economic Sciences was given to David Card, Joshua D. Angrist, and Guido W. Imbens, whose methodological contributions to these methods (RDD as well as others) are inevitable.

RDD can be used to estimate the impact of a program or *treatment* if the candidates are selected for the treatment on the condition that a numeric endogenous variable, called *running variable*, exceeds (or subceeds, depending on the context) a designated threshold or *cut-point*. Over the last two decades, this approach has been used to evaluate the impact of a wide variety of programs, especially in social sciences, e.g., regarding the labor market [31, 32] and education policy [33, 34]. This procedure is also used to obtain a casual estimate of the effects of remediation on academic outcomes for students at the margin of passing the assessment test, assuming that students with scores just below and above the placement cut-point have similar abilities [35, 36, 37, 38, 39].

While RDD is intuitively appealing, as we will see, the statistical details regarding the implementation are more complicated than they might first appear. Here we provide a short overview of RDD, for a more comprehensive description we refer to Lee and Lemieux [40] and Bloom [41].In a nutshell, the method can be seen as the comparison of two regression curves at a designated point. We have an output variable Y, on which we want to estimate the effect of the treatment and the running variable X. A binary treatment variable T is assigned based on the running variable (whether it exceeds a certain cut-point, see Section 2.1.2). The basic version of this methodology was introduced by Thistlethwaite and Campbell [42], as linear regression of Y on variable X, subject to T. This idea was further developed by, e.g., using interaction terms, nonparametric regression methods, etc. The nonparametric (local linear kernel regression) approach can be favorable if individuals around the cut-point indeed have similar characteristics since the kernel method uses data close to the cut-point.

It is typically imagined that for each individual i there exists a pair of outcomes: $Y_i^{(1)}$ and $Y_i^{(0)}$ if the person was exposed to the treatment (1) and if not (0), respectively. The effect of the treatment is measured by the difference $Y_i^{(1)} - Y_i^{(0)}$. In reality, for an individual only one of the two outputs is observable, so we look at the averages of this difference over a population by taking the difference of the conditional expectations

$$\mathbb{E}\left[Y_i^{(1)} \mid X_i = x\right] - \mathbb{E}\left[Y_i^{(0)} \mid X_i = x\right]. \tag{2.1}$$

We are interested in this difference at the cut-point value c of the running variable X, thus – depending on what is observable at the two sides of the cut-point – we estimate:

$$\lim_{\varepsilon \to 0} \left\{\mathbb{E}\left[Y_i^{(1)} \mid X_i = c \pm \varepsilon\right] - \mathbb{E}\left[Y_i^{(0)} \mid X_i = c \mp \varepsilon\right]\right\}. \tag{2.2}$$

This is the average treatment effect (ATE) at the cut-point which is used to evaluate the effect. These conditional expectations are the solutions to the regression problems, which are approximated by nonparametric or parametric methods.

2.1.1 Manipulation and covariates

RDD descriptions suggest examining the density of observations of the running variable to assess for possible manipulation of the treatment status. There are tests for continuous [43] and for discrete running variables as well [44]. If there is a discontinuity in the density of the running variable at the threshold for treatment, then this may suggest that some participants were able to manipulate their treatment status, in which case RDD is not the right approach.

In RDD the inclusion of other covariates is also possible but not encouraged. Covariates usually have little impact on the effect size, but they may reduce some bias while improving the precision of the estimation [45]. Although, they may take away the explanatory property of the running variable and the treatment, which indicates manipulation at the covariates' level.

2.1.2 Sharp or Fuzzy RDD

The design is sharp if everyone above (below) the aforementioned threshold gets treatment (considered as treated), and below (above) nobody is treated. Since in reality cut-points are

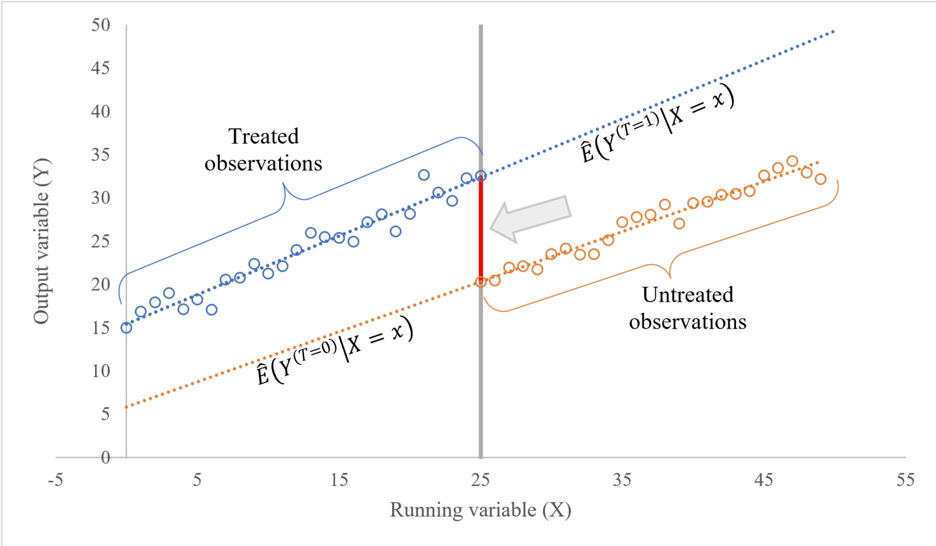


Figure 2.1: Dummy example of RDD: treatment is administered under cut-point 25, the conditional expectations can be clearly estimated with linear regression, and a visible treatment effect is indicated with the red line and arrow.

often not strictly implemented, there is another, so-called fuzzy approach. In other words, there can be overlaps between the treated and the control groups with participants having the same values of the running variable. In this case, the treatment variable is not deterministic but random, and a two-stage least squares regression is applied, where the first step is the estimation of T based on X. In the next sections, we assume the sharp scenario for simplicity.

2.1.3 Global or local method

The two most common approaches of RDD are nonparametric (local) and parametric (global), mostly polynomial, regression.

Parametric (global) method

The conventional parametric approach uses all the available observations (not just those that are near the cut-point) for the estimation, which is of the following form:

$$Y^{(0)} = \alpha_0 + f_0(X - c) + \epsilon, \tag{2.3}$$

$$Y^{(1)} = \alpha_1 + f_1(X - c) + \epsilon. \tag{2.4}$$

Instead of two estimations, a pooled version is used in practice:

$$Y = \alpha_0 + \beta_0 T + f(X - c) + \epsilon, \tag{2.5}$$

where $\beta_0 = \alpha_1 - \alpha_0$ and the function $f = f_0 - T(f_1 - f_0)$ represents the relationship between the zero-centered (c is the cut-point) running variable and the outcome. A variety of functional forms can be tested to determine which one fits the data best: $f()$ can be, e.g., linear, quadratic, or cubic; with the goal to minimize the bias. Misspecification of the parametric model can lead to false identification of the treatment effect, as in Figure 2.2. Gelman and

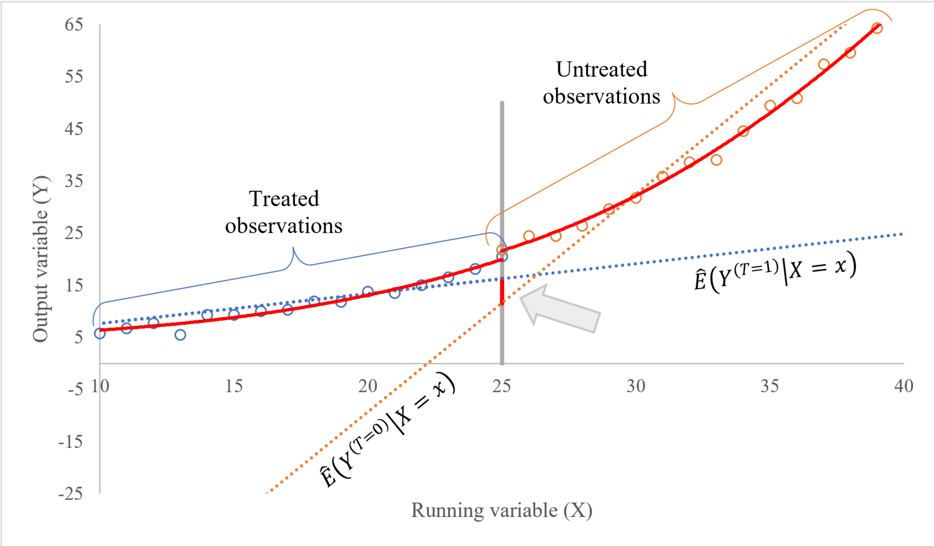


Figure 2.2: Misspecified parametric model: in this generated sample Y and X has a cubic relation without added treatment effect, yet linear regression implies a visible treatment effect, see the red line and pointing arrow.

Imbens [46] argues that higher-order polynomials should not be used in RDD analysis, and encourage the use of nonparametric models instead, up until the second order only. Examples for $f()$:

- linear: $\beta_1(X-c)$,
- linear with interaction: $\beta_1(X-c)+\beta_2 T(X-c)$,
- quadratic: $\beta_1(X-c)+\beta_2(X-c)^2$,
- cubic: $\beta_1(X-c)+\beta_2(X-c)^2+\beta_3(X-c)^3$.

Note that $f()$ should contain interaction terms such as $(X-c)T$, by this we take into consideration that the treatment may work better for some participants than for others. If the interaction is not modeled correctly or left out entirely, a false discontinuity could appear. The interaction term results in different slopes and intercepts on the two sides of the threshold.

Nonparametric (local) method

Choosing the right function $f()$ is not always clear since far away from the cut-point we can have mostly outliers with different characteristics, which would bias the estimation of the average treatment effect. In this case, we limit the analysis to observations that lie within the close neighborhood of the cut-point and use a nonparametric method.

The most common nonparametric method used in the RDD context is a local linear (kernel) regression, which is a better fit at the boundaries (note that the cut-point is a boundary) than the local constant (kernel) regression [47]. We assume that the dependence between the running variable and the chosen output is close to linear within a close neighborhood of the cut-point. This neighborhood is determined by the *bandwidth* parameter. The kernel weighted least squares minimization in the linear case uses the functional form

$$Y = \alpha_0 + \beta_0 T + \beta_1 (X - c) + \beta_2 T (X - c) + \epsilon, \tag{2.6}$$

where the estimation is carried out using data for which $c - h \leq X \leq c + h$. Here h is the bandwidth parameter and this interval is the support of the applied kernel function. Typically, either a "rectangular" kernel ($\square$) (no weighting inside the interval) or a "triangular" kernel (Δ) is used. With the notation of Appendix B and Chapter 1:

$$K_h^{(\square)}\left(x, x^{(i)}\right) = \frac{1}{2}\mathbb{I}_{\left\{\left|x - x^{(i)}\right| \leq h\right\}} \quad \text{and} \quad K_h^{(\Delta)}\left(x, x^{(i)}\right) = \left(1 - \left|x - x^{(i)}\right|\right)\mathbb{I}_{\left\{\left|x - x^{(i)}\right| \leq h\right\}},$$

The difference is that the latter puts more weight on observations closer to the cut-point, which is desirable but makes the interpretation more difficult. The triangular kernel is optimal for estimating local linear regression at the boundary [47], but the choice of the kernel typically has little impact in practice [40].

For nonparametric RDD, and kernel regression in general, the choice of an optimal bandwidth parameter is crucial. Imbens and Kalyanaraman [48] proposed a method to find the best bandwidth around the cut-point but in many cases expert judgment is also suitable. Their proposed optimal bandwidth is defined as

$$\hat{h}_{\text{opt}} = C_K \cdot \left(\frac{2 \cdot \hat{\sigma}^2(c) / \hat{f}(c)}{\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c)\right)^2 + (\hat{r}_+ + \hat{r}_-)} \right)^{1/5} \cdot N^{-1/5},$$

where $\hat{\sigma}^2(c), \hat{f}(c), \hat{m}_+^{(2)}(c), \hat{m}_-^{(2)}(c)$ are estimates for $\texttt{Var}\,(Y \mid X = c)$, the density of X at c, the right $(+)$ and left $(-)$ limits of the second derivative of $\mathbb{E}\,(Y \mid X = c)$, respectively. Also, C_K is a kernel specific constant, the $\hat{r}$ terms are regularization terms, and N is the sample size. Further discussion of the algorithm on how to calculate this optimal bandwidth would be out of scope.

2.2 The Problem of a Discrete Running Variable

In many RDD applications the running variable is discrete. Therefore, comparing outcomes for observations "just above" and "just below" the treatment threshold is not possible, and/or raises questions. There are no data in the proximity of the cut-point, so using Equation (2.2) would lead to overestimation of the true effect. For example, let us have a running variable that takes its values from the countable set $\{x_1, x_2, \ldots, x_j, \ldots, x_k\}$ with cut-point at $c = x_j$, and the treatment is given if $X \geq c$. In this case, we would actually estimate

$$\mathbb{E}\left[Y_i^{(1)} \mid X_i = c\right] - \mathbb{E}\left[Y_i^{(0)} \mid X_i = x_{j-1}\right].$$

instead of Equation (2.1) at c. It is easy to see that an overestimation happens due to extrapolation.

2.2.1 Specification error

To deal with the issue of a discrete running variable, Lee and Card [49] modify Equation (2.5) by introducing a random term s_X called *specification error* depending on the value of the running variable:

$$Y = \alpha_0 + \beta_0 T + f(X - c) + s_X + \epsilon.$$

The specification error is assumed to be random and orthogonal to X. This condition seems restrictive, but according to the original assumption this error does not even exist, i.e., it is zero. Its randomness induces a within-group correlation in the error at every level of X. In other words, the observations with the same value of the running variable are considered as members of the same "cluster".

We note that Lee and Card [49] argue that in the case of a discrete running variable, the nonparametric identification of the discontinuity is not possible. However, Bartalotti and Brummet [50] demonstrated that in some frameworks, the usage of these cluster-robust standard errors is valid even for nonparametric local polynomial estimation. We believe that our setup is also valid for a nonparametric estimation.

We also mention that Kolesár and Rothe [51] show some evidence that this clustering is generally unable to resolve bias problems in discrete RDD settings, despite its popularity. The documentation of the R package RDDapp [52] also suggests a clustering by every different value of the running variable.

2.2.2 Handling the discrete running variable with jittering

Here we propose a novel alternative method to the specification error. Instead of the orthogonal error term in the regression estimation, we add an independent noise term directly to the running variable. Adding independent noise to discrete variables is usually called jittering. It is a common solution when someone wants to have a more comprehensive figure in the case of scatter plots, or, e.g., in algorithms that require strict ordering of the data points. Our proposed method is based on the idea of Nagler [53], who theoretically justifies the jittering of discrete variables for nonparametric regression estimation. Therefore, instead of using the discrete X as a running variable, we use $\tilde{X} = X + \delta$, where δ is an independent continuous noise term. Assuming that X takes its value from an equidistant range of values, δ should be calibrated to this distance.

More precisely, assuming that the distance is 1, e.g., as is on a scale of $1, 2, 3, \ldots, 100$; the probability density function f_δ of the noise should satisfy the condition of Nagler [53], that there exists a pair of positive real numbers γ_1 and γ_2 such that

$$0 < \gamma_1 \leq 0.5 \leq \gamma_2 < 1 \quad \text{and} \quad f_\delta(x) = \begin{cases} 1 & \text{if} \quad x \in [-\gamma_1, \gamma_1], \\ 0 & \text{if} \quad x \in \mathbb{R} \setminus (-\gamma_2, \gamma_2). \end{cases} \tag{2.7}$$

The uniform distribution over the interval $[-0.5, 0.5]$ trivially satisfies the above condition with $\gamma_1 = \gamma_2 = 0.5$. A new cut-point $\tilde{c}$ is also necessary, that is the original c shifted by the above γ_2. Otherwise, measurements would end up at the wrong side of the cut-point due to the jittering. In case of uniform noise over $[-0.5, 0.5]$, the new cut-point of the RDD is at $c \pm 0.5$ depending on which side of the cut-point is strict. The treatment effect of Equation (2.1) in this jittered case becomes

$$\mathbb{E}\left[Y^{(1)} \mid \tilde{X} = \tilde{c}\right] - \mathbb{E}\left[Y^{(0)} \mid \tilde{X} = \tilde{c}\right] =$$
$$\mathbb{E}\left[Y^{(1)} \mid X + \delta = c \pm \gamma_2\right] - \mathbb{E}\left[Y^{(0)} \mid X + \delta = c \pm \gamma_2\right].$$

In the nonparametric RDD setting, the local linear estimation of Equation (2.6) with jittering becomes

$$Y = \alpha_0 + \beta_0 T + \beta_1 (X - \tilde{c} + \delta) + \beta_2 T (X - \tilde{c} + \delta) + \epsilon, \tag{2.8}$$

where the estimation is carried out using data for which $\tilde{c} - h \leq \tilde{X} \leq \tilde{c} + h$. Note that due to jittering the estimated effect size becomes random. The added perturbation is small, and can be seen as a random weighting of the sample.

This method has the benefit that the added noise term will make the running variable continuous, thus it would be easier to generalize some theoretical results regarding RDD using this set-up than using the one with specification error. Another advantage is that the proximity of the cut-point becomes populated by data points in this set-up, so the nonparametric estimation is valid again. In the case of a discrete running variable, the usage of specification error is better suited for the parametric estimation as pointed out by Lee and Card [49]. However, we believe that our proposed method is better suited for the nonparametric estimation of the treatment effect.

If we do not shift the cut-point to $\tilde{c}$, i.e. we keep the original cut-point c, then based on [53] we can conjecture the following about the jittering RDD. Denote the original treatment effect at c with $\tau(c) := \mathbb{E}[Y^{(1)}|X = c] - \mathbb{E}[Y^{(0)}|X = c]$. In the jittered case at c let it be

$$\tau^*(c) := \mathbb{E}\left[Y^{(1)} \mid X + \delta = c\right] - \mathbb{E}\left[Y^{(0)} \mid X + \delta = c\right]. \tag{2.9}$$

Given that the noise δ follows a density like in Equation (2.7), then according to [53, Proposition 1], $\tau(c) = \tau^*(c)$ for each $c \in \mathcal{X}$, where $\mathcal{X}$ is the sample space of X. If we denote the estimation of $\tau(c)$ and $\tau^*(c)$ with (local linear) kernel regression by $\hat{\tau}(c)$ and $\hat{\tau}^*(c)$, respectfully;, then according to [53, Proposition 3]:

$$\frac{\hat{\tau}^* - \tau^*}{r_n} \xrightarrow{?} W \Rightarrow \frac{\hat{\tau}^* - \tau}{r_n} \xrightarrow{?} W,$$

where r_n is a sequence that tends to 0 as the number of elements (n) tends to infinity, and W is a random variable. Here, the arrow $\xrightarrow{?}$ can denote almost sure or convergence in distribution. For a continuous running variable (the one obtained after "jittering" is such), τ^* can be estimated consistently by local linear regression, see e.g. [45], where r_n is usually a function of n and the bandwidth h_n (depending on n), for which $r_n \to 0$ is satisfied. W is usually normally distributed with parameters depending on the assumptions of the estimator. Therefore, the estimation of τ (and not just the jittered τ^*) by the jittered kernel regression estimate $\hat{\tau}^*$ is consistent at the original cut point c; the consistency results are inherited.

2.3 The Graphical Modeling of RDD

The model behind RDD is intuitively simple, yet a graphical modeling approach can help to give us a better understanding. In recent years, the causal interpretation of RDD and other quasi-experimental designs have been the subject of many studies. These represent RDD (and other methods) with causal DAGs.

Constantinou and O'Keeffe [54] take a rigorous decision-theoretic approach (using the language and calculus of conditional independences) to formally study causal effect identification and estimation in both sharp and fuzzy RDDs. In neuroscience and behavioral research, causal models are extensively applied, but quasi-experimental methods are not that popular (yet) despite their causal setup. Marinescu, Lawlor, and Kording [55] try to bridge this gap by reviewing quasi-experimental methods from a causal point of view. Steiner et al. [56] discuss and compare the identifying assumptions of quasi-experiments using causal graphs. By introducing so-called limiting graphs, it becomes clearer that RDD is a randomized experiment at the cutoff. Tadayon and Pottie [57] discuss the applications and limitations of experimental and quasi-experimental designs for educational systems and quantify them using the causal graphs' language. They also model the education system as time-varying treatments, confounders, and time-varying treatments-confounders feedback.

Causal DAGs (briefly introduced in Section 1.1.3) represent the qualitative causal knowledge or belief about the data generating process. This may contain information on the study design and the data collection procedure as well. Stronger design assumptions usually result in simpler graph representations.

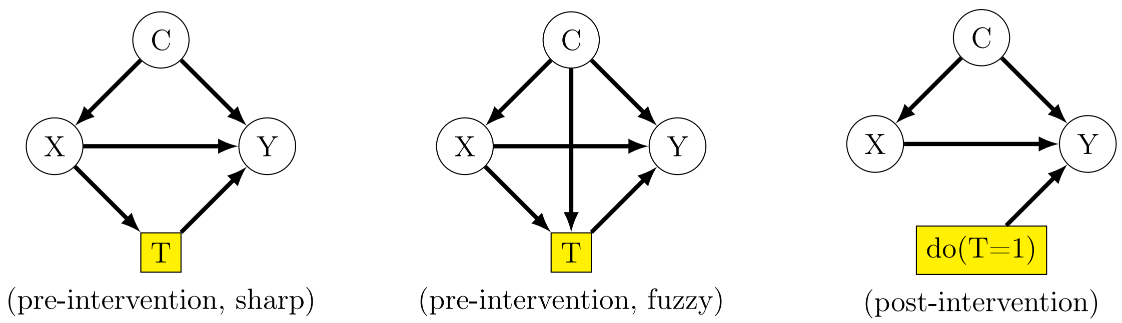


Figure 2.3: Causal representations of RDD

The previously cited works [54, 55, 56, 57] draw a similar causal model for RDD as in Figure 2.3. Here X, T, Y denote the running variable, the binary treatment variable, and the studied output variable, respectively; whereas C denotes all confounding variables (observed or unobserved) present in the study. The pre-intervention DAG represents the data generating process, determined by the model itself in this case. In the sharp scenario, T is a deterministic function of X. However, in the fuzzy scenario, the participation in the treatment is subject to some confounding variable of C too, so we have an extra edge $C \to T$. The post-intervention DAG is the pre-intervention DAG after performing the intervention in the sense of Pearl's $\mathrm{do}(\cdot)$ operator [6]: removing the parents of the nodes we are doing an intervention on.

Small differences in the representation happen due to slightly different notions, but they do not have an impact on the model.

- In [54], an extra variable (Σ) is introduced for the "regime" the model is in.
- In [55], the edge $X \to Y$ is not included. Meaning, they assume that the effect of X on Y can be explained entirely by C.
- In [56], a bi-directed dashed arrow between C and X is included, which is a slightly confusing notation for some undefined association or unknown common cause.
- In [57], C is separated into two groups of observed and unobserved confounders.

The causal graph representation alone does not say more about the functional form of the generating process. The graph stays the same for discrete running variables handled with specification error, or with jittering as well. The additional information is encoded in the so-called structural causal model (SCM) [6], which accompanies the graph. In other words, the functions that define the data generating process.

2.4 Application of RDD, a case study at the BME

A longer, detailed case study is presented. We apply RDD to measure the effectiveness of two differently designed remedial mathematics courses at the Budapest University of Technology and Economics (BME).

Our large-scale study is based on data of almost 20,000 undergraduate students enrolled between 2010 and 2018. We study both the direct and longer-term effects of remediation. We also investigate the jittering method proposed in Section 2.2.2.

2.4.1 Mathematics remediation at BME

The data for this study were collected at BME, the flagship of Hungarian engineering education. BME is a large public university with eight faculties and about 19,000 current students. Here we detail how mathematics remediation works at BME. This system is similar to the practice of several other higher education institutes all over the world, making our approach and findings generally applicable. At BME every first-year undergraduate student is required to take a mathematics assessment test at the beginning of the academic year to determine their mathematics skills. The mathematics assessment has been in effect since 2010 without any significant changes since then. The topics of the test are in line with the normal level secondary school exit examination (matura). For a detailed description and analysis of the assessment test see the work of Csákány [58]. The test consists of 15 multiple choice questions each with 5 possible answers, of which only one answer is correct. A right, a wrong, and a blank answer are worth 4, -1, and 0 points, respectively, so the total number of attainable points ranges from -15 to 60 points.

A successful assessment test is required for the completion of first-year mathematics courses. The test is considered to be successful if the student gains at least 24 points. A student who achieves less than 24 points has two options: either enroll in the remedial course called *Introductory Mathematics* (short: *IntroMath*) in the first semester as a co-requisite of the first-semester calculus course (short: *Calc1*) – this is the recommended option – or retake the assessment test during the same semester. The co-requisite nature means that students enroll simultaneously in the remedial course and the gateway course. The remedial course consists of one weekly session (135 minutes) for 9 weeks. The course design and syllabus of the remedial course differ across the university, more precisely, there are two variants of the course.

One of them, which is used by the majority of the faculties, covers the most relevant topics of secondary school mathematics with respect to further education. This variant will be referred to as *IntroMathHS* course from now on.

On the other hand, at one single faculty, namely at the Faculty of Chemical Technology and Biotechnology (short: VBK) the mathematical remediation is handled differently. Instead of reciting secondary school mathematics, here the *IntroMath* focuses on the same concepts as the mandatory first-year calculus course, but at a slower pace with more explanations and with more emphasis on high school material if necessary. Essentially, here the remedial course acts as an extra recitation course for the calculus course that can be a good supplement

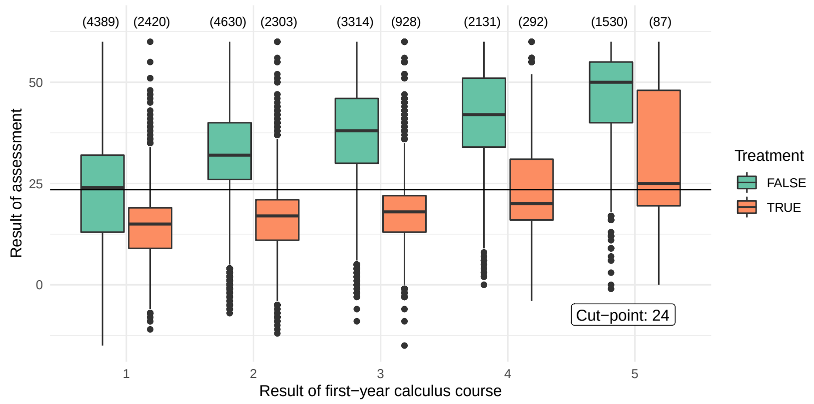


Figure 2.4: Box plot for the grade of the *Calc1* course and the points earned on the assessment test with respect to remediation (treated or not), with the sample sizes in parentheses over the boxes

for those who are less skilled in mathematics. The topics covered in this latter variant of *IntroMath* is in alignment with the syllabus of the first-semester calculus course. This other variant will be referred to as *IntroMathCalc* course from now on. Also, at the VBK, there is an advanced calculus practice course, which is recommended for students with good assessment test results. It will be referred to as *AdvCalc* course. The two courses are exclusive of one another, the accomplishment of both is not permitted.

The vast majority of students who fail the assessment test (i.e. achieve less than 24 points) enroll in the remedial course, the number of those who choose to retake the assessment test instead is negligible so we left them out from our analysis. *IntroMath* is an elective course meaning that students who managed to gain the minimum 24 points on the assessment could also enroll, which would implicate a fuzzy design scenario, see Section 2.1.2. However, some mathematically well-trained students enroll in the remedial course to obtain extra "easy" credit points. It is also confirmed by Figure 2.4 where we can see that around 50% of those with accomplished *IntroMathHS* course and grade 4 or 5 from *Calc1* were over the designed cut-point, although, it is not a big group of students. Also, around 50% of those without remediation and grade 1 (failing) from calculus should have done the *IntroMathHS*, as they were below the cut-point. They might suspend their university studies mid-semester, could not finish the remedial course successfully, or chose to retake the assessment test. They form a larger subgroup of students, as can be seen in Figure 2.4. The fuzzy design would include these two groups among others, which would possibly bias the results. We aim to study the impact of the remediation, and the inclusion of these students in the treated group would result in an overestimation, so we also left them out from our further analysis.

Therefore, we define the treated group as the students who have completed the remedial course with an assessment score less than 24 points, while the control group consists of students with more than 24 points on the assessment and without enrollment in the remedial course. Our sharp RDD scenario is as follows:

- Treated: students with <24 points on the assessment AND accomplished the remedial course

- Control: students with ≥ 24 points on the assessment AND NOT accomplished the remedial course
- Running variable: the result of the mathematical assessment
- Cut-point: 24 points (23.5 for the jittered variant)

In Table 2.1 we see the sample sizes, but note that in the studied sharp RDD scenarios, we restrict the samples for those who are in the main diagonals, so from now on the sample sizes are the sums of the elements in the main diagonal (in bold). We measure the direct and longer-term impact of remediation on the following outputs:

- accomplishment of the first-year calculus (short: *Calc1*) course (binary variable indicating whether the student completed the calculus course successfully), which is held in the same semester as the remedial course
- the grade of the first-year calculus course (ordered numeric variable from 1 to 5, where 1 is the failing grade and 5 corresponds to excellent),
- adjusted credit weighted average of grades without the result of the first-year calculus course (short: ACWA-C) of the first two semesters (continuous variable),[1]
- active enrollment in the third semester after the assessment test (binary variable).

To consider students with more similar characteristics, we restrict the analysis to those who did their assessment, remediation, and calculus course in their first semester at BME.

Treat. \ cut-p.	≥ 24	< 24	Total (Treat.)
FALSE	**12851**	3143	15994
TRUE	694	**5336**	6030
Total (cut.-p.)	13545	8479	22024

(a) *IntroMathHS* course

Treat. \ cut-p.	≥ 24	< 24	Total (Treat.)
FALSE	**827**	191	1018
TRUE	819	**793**	1612
Total (cut.-p.)	1646	984	2630

(b) *IntroMathCalc* course

Table 2.1: The number of students subject to the cut-point (24) of the assessment test and treatments (the two remediation courses). The sample sizes for the sharp RDD studies are in bold.

The setup of the remedial mathematics education at BME is quite similar to the practice of other colleges and universities all over the world, making the presented methodology and our results generally applicable. Suggesting remedial courses based on assessment test results is a common policy in several countries, such as the USA [59], Canada [60], Australia [61], Chile [62] or Italy [38]. Another important feature of the mathematical remediation at BME is the co-requisite nature of the remedial course. Co-requisite mathematics remediation is gaining popularity since it was found to be a more effective strategy than traditional pre-requisite remediation [63, 64, 65].

[1] The sample sizes for this variable are a bit smaller due to students without second-semester results.

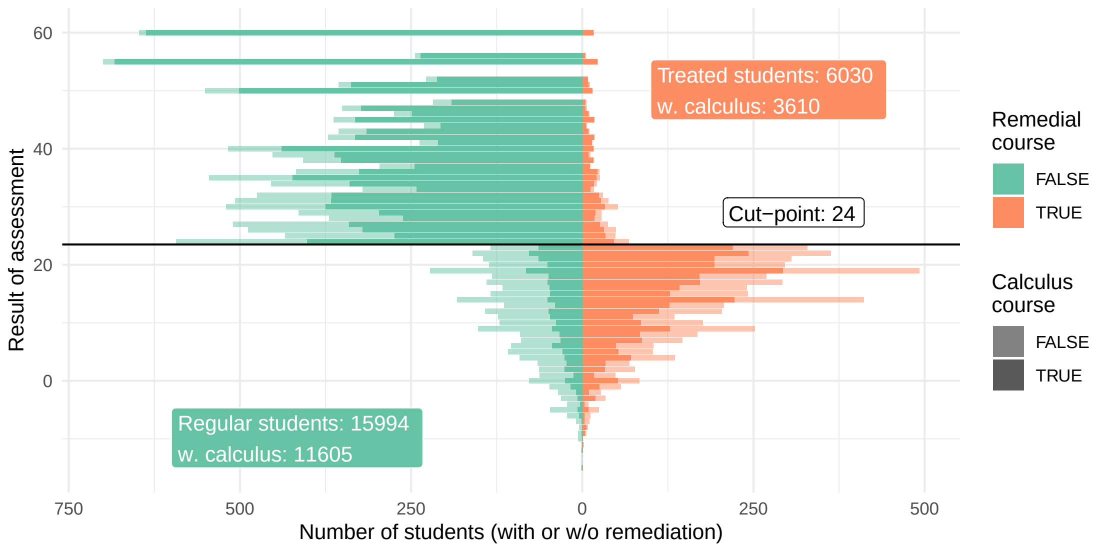


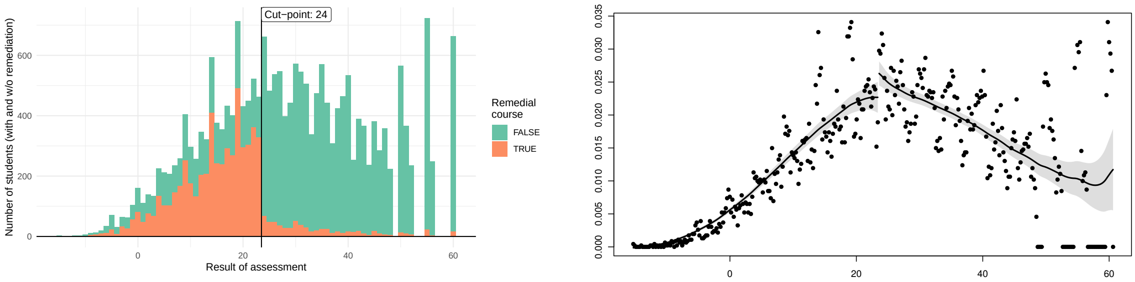
Figure 2.5: Comparative histogram of the points earned on the assessment test (vertical axis) against the number of students (horizontal axis). It is colored and counted separately by the accomplishment of the *IntroMath* course. Transparency indicates the accomplishment of the *Calc1* course in a stacked manner.

2.4.2 Results of RDD

In this section, we present the results of our RDD analysis. We aim to be comprehensive in order to be more certain about our findings. First, we demonstrate our findings with respect to the effectiveness of the *IntroMathHS* course, then remediation at the faculty VBK, the *IntroMathCalc* course is put under the microscope. Later on, we check how robust our results are using sensitivity, yearly and faculty-wise analyses, and finally covariates. We also include some results obtained by our jittered variant of the RDD, see Section 2.2.2, in order to compare it with the more traditional setups.

Let us take a step back from the RDD analysis first, and look at just the data. Figure 2.5 illustrates the relation between the result of the assessment test, the accomplishment of the *IntroMathHS*, and the *Calc1* courses. By a naive approach, one can say that the remediation is ineffective since there are many students with accomplished remediation still failing in calculus. However, the two groups are similar in the sense that around two-thirds of them were able to complete the calculus course. On the other hand, if we look a bit closer we can observe that for a fixed score on the assessment test, taking the remedial course makes it more likely to complete the calculus course successfully. However, this observation cannot be accepted as definite proof of the effectiveness of the *IntroMathHS* course since for a fixed assessment score the self-selection bias arises, meaning that it is the students' own decision whether they enroll in the remedial course or not. That is why we assess the effect with a more sophisticated quasi-experimental approach, the regression discontinuity design.

We used the RDD implementation of the R package RDDapp [52]. A visual inspection of Figure 2.6a, and the McCrary-test (see Section 2.1.1) conducted on the jittered points of the assessment test (Figure 2.6b) implicates discontinuity. It is due to the scoring system of the test (detailed in Section 2.4.1), and the manipulation at this scale is unlikely, thus we rejected



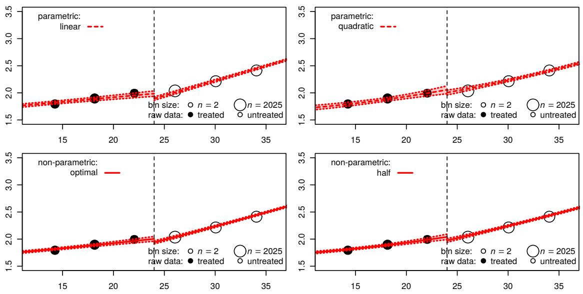
(a) Stacked histogram of the assessment test results (b) Result of the McCrary-test on the jittered results

Figure 2.6: Discontinuity of the assessment test results due to the described pointing scale

the hypothesis of possible manipulation. As of the writing of this dissertation, no implementation of the manipulation test of Frandsen [44] is available in R for discrete running variables, thus this test was left out from our study. Although the rules would implicate a fuzzy design, a restricted sharp design is more accurate in our case, as explained in Sections 2.1.2 and 2.4.1. We find the non-parametric approach more suitable (see Section 2.1.3) because at the two ends of the pointing scale there are many students with irregular performance. Nevertheless, some results are included for global linear and global quadratic regression estimations. The optimal bandwidth parameters are calculated using the method of Imbens and Kalyanaraman [48], abbreviated as the IK method. If not specified differently, we used the specification error setup of Section 2.2.1, but some outputs were compared against our proposed jittering method of Section 2.2.2 as well.

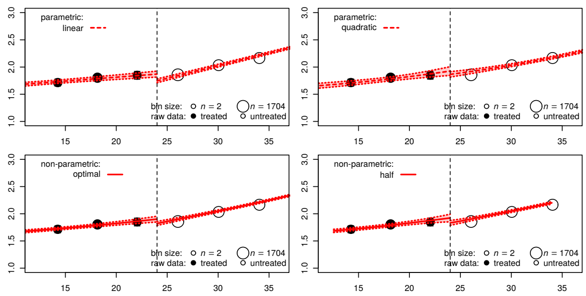
Take a look at the *IntroMathHS* course first. The graphical interpretation of Figure 2.7 suggests that there is a small but considerable effect of the course on most of the studied variables. We also assess the statistical significance of the estimates for the average treatment effect (ATE) using a Z-test. The null hypothesis of this test is that the coefficient of the treatment variable (T) in the (pooled) model – which is the average treatment effect of Equation (2.2) – is zero. At significance level of 10% we can reject the null hypothesis for all studied variables in the local linear case with the IK bandwidth selection. Therefore, we can say that the remedial course has a statistically significant effect on all the examined outputs based on students in the calculated window (i.e. at the margin of passing the assessment test). Note that the significance may be due to the huge sample size. If we look at the results of the global linear cases, they are quite different. For the two non-binary output variables (Figures 2.7a and 2.7b) a clear overestimation happens due to students at the two ends of the pointing scale. However, for the two binary output variables (Figures 2.7c and 2.7d) the effect is not observable. This inconsistency is due to using linear regression with a binary dependent variable. The smallest estimate for the first two outputs (Figures 2.7a and 2.7b) is measured in the quadratic case. It is still positive but seems statistically less significant. The most substantial effect of the *IntroMathHS* seems to be on the third-semester enrollment.

Now take a look at the *IntroMathCalc* course of the VBK. It is clear by Figure 2.8 that this course – by focusing on the *Calc1* material – shows a much larger positive effect on all studied direct and longer-term outputs. Also, the results here are statistically more significant, despite the smaller sample size. Note, that the *IntroMathCalc* attracts many students with



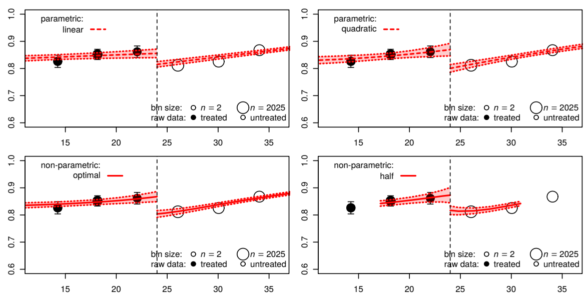
	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			37.57	18.79
Observations	18187	18187	18186	12961
Estimate	0.092	0.052	0.063	0.064
Std. Error	0.043	0.036	0.036	0.035
$\mathbb{P}(>\lvert z\rvert)$	0.034	0.148	0.076	0.063

(a) the grade of the *Calc1* course



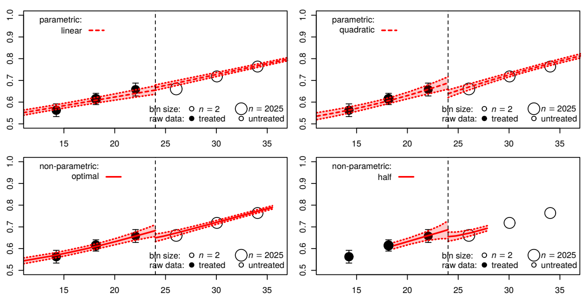
	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			24.93	12.46
Observations	16972	16972	14132	8865
Estimate	0.123	0.061	0.080	0.090
Std. Error	0.047	0.042	0.035	0.042
$\mathbb{P}(>\lvert z\rvert)$	0.008	0.147	0.023	0.033

(b) ACWA-C of the first two semesters



	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			18.07	9.03
Observations	18187	18187	12961	7460
Estimate	0.041	0.069	0.064	0.056
Std. Error	0.010	0.016	0.017	0.020
$\mathbb{P}(>\lvert z\rvert)$	0.000	0.000	0.000	0.004

(c) being enrolled in the third semester

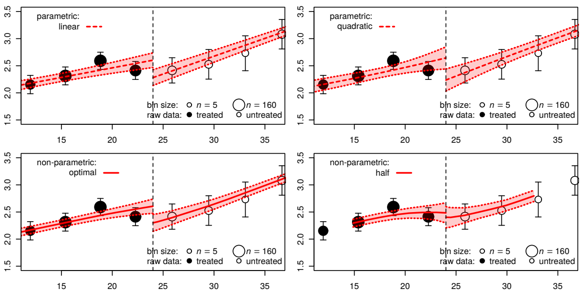


	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			14.14	7.07
Observations	18187	18187	10713	6182
Estimate	-0.011	0.050	0.033	0.032
Std. Error	0.015	0.017	0.017	0.020
$\mathbb{P}(>\lvert z\rvert)$	0.444	0.003	0.048	0.117

(d) the accomplishment of the *Calc1* course

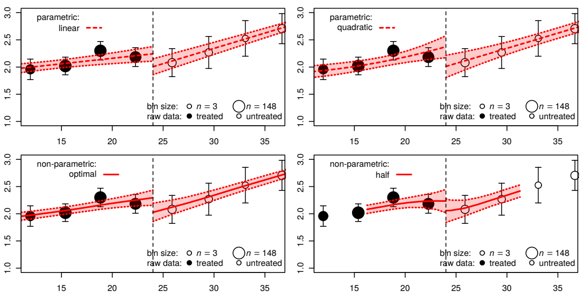
Figure 2.7: Sharp RDD results on the chosen outputs for the *IntroMathHS* course: on the left side, the usual two-sided regression plots of the RDD estimation with 95% confidence intervals, and binned data points; while on the right side, the summary statistics of the estimations.

assessment scores above the threshold, so the interest for this course, among such students, is higher than at other faculties where the *IntroMathHS* course is available. Since these students were excluded from the analysis it reduced the size of the control group in our sharp setup. However, students of the *AdvCalc* course were included in the control group for this study. These constraints resulted in the largest treated and smallest control group at the VBK among the faculties, proportionally, that may somewhat bias the results of the VBK.



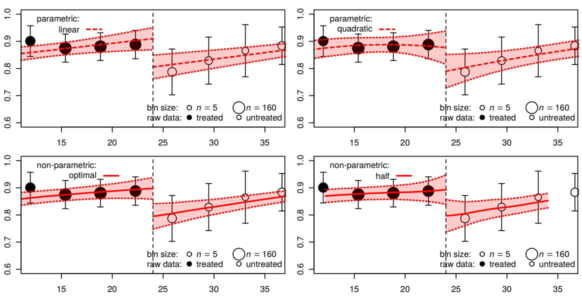
	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			22.44	11.22
Observations	1620	1620	1293	748
Estimate	0.321	0.408	0.306	0.081
Std. Error	0.112	0.170	0.141	0.163
$\mathbb{P}(>\lvert z\rvert)$	0.004	0.017	0.030	0.620

(a) the grade of the *Calc1* course



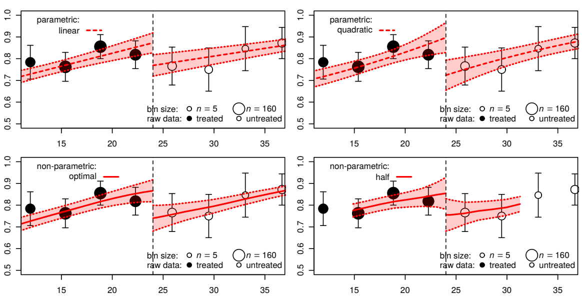
	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			19.31	9.66
Observations	1535	1535	1093	581
Estimate	0.237	0.362	0.269	0.176
Std. Error	0.106	0.138	0.116	0.148
$\mathbb{P}(>\lvert z\rvert)$	0.025	0.009	0.021	0.236

(b) ACWA-C of the first two semesters



	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			26.70	13.35
Observations	1620	1620	1408	859
Estimate	0.104	0.087	0.103	0.096
Std. Error	0.026	0.040	0.034	0.045
$\mathbb{P}(>\lvert z\rvert)$	0.000	0.029	0.003	0.033

(c) being enrolled in the third semester

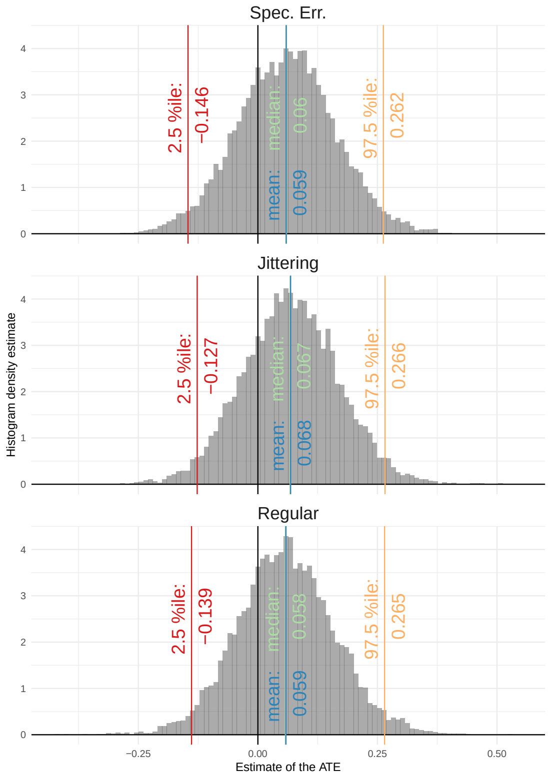


	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			20.66	10.33
Observations	1620	1620	1197	684
Estimate	0.104	0.173	0.126	0.099
Std. Error	0.036	0.052	0.042	0.046
$\mathbb{P}(>\lvert z\rvert)$	0.004	0.001	0.003	0.032

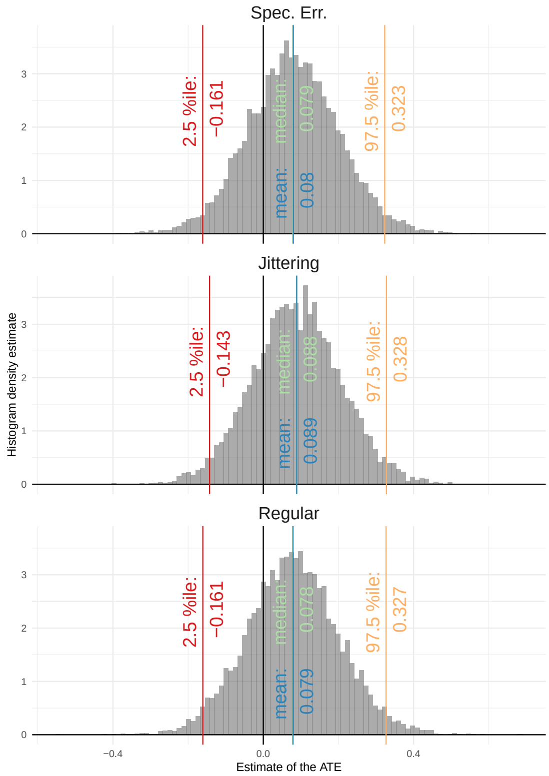
(d) the accomplishment of the *Calc1* course

Figure 2.8: Sharp RDD results on the chosen outputs for the *IntroMathCalc* course: on the left side, the usual two-sided regression plots of the RDD estimation with 95% confidence intervals, and binned data points; while on the right side, the summary statistics of the estimations.

The ATE on the grade of the *Calc1* course and the grade point average of the first two semesters (ACWA-C, see Section 2.4.1) was estimated by bootstrapping as well, to compare the results of the different non-parametric estimations (using the IK bandwidth): with specification error, with jittering, and without modifications regarding the discrete running variable. We have (re)sampled 2,000 students 10,000 times to estimate the ATE. In Figure 2.9, we can compare the bootstrapped estimates of the ATE in the three setups. The bandwidth was calculated for each subset one by one. The bootstrapped distribution of the ATE estimates is roughly the same in all three cases. Our jittered version gives a slightly larger estimate in median and mean. These results reassure us of the positive effect of the *Calc1* course, and that our jittered variant provides a good estimate of the treatment effect.



(a) The grade of the *Calc1* course



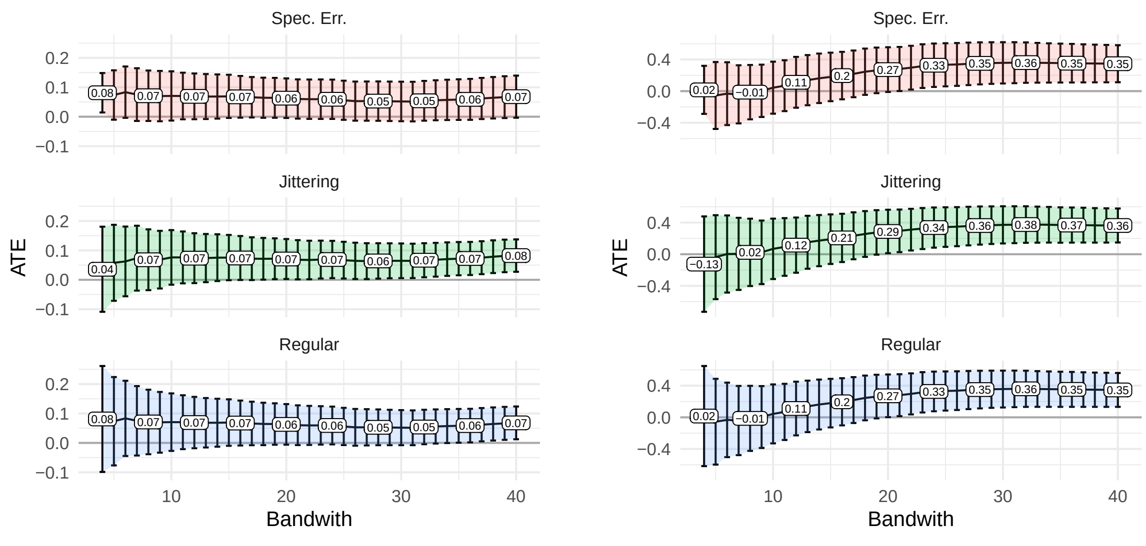
(b) ACWA-C of the first two semesters

Figure 2.9: Bootstrapping the ATE estimates on two different outputs for the *IntroMathHS* course; in order to compare the results for the three different types of discrete running variable handling.

2.4.3 Bandwidth sensitivity

Non-parametric regression estimation is usually sensitive to bandwidth selection. As stated before, in RDD setups, the method of Imbens and Kalyanaraman [48] is the quasi-standard for bandwidth selection. However, it is worth checking how the ATE changes with different bandwidth parameters. We tested the sensitivity of the effect size on the grade of the *Calc1* course, for both remedial courses. We compared the ATE estimates over different bandwidth parameters, ranging from 4 to 40, for the three different non-parametric (local linear) cases: with specification error, with jittering, and without modification. In Figure 2.10 we can see that the ATE is not sensitive to the chosen bandwidth. It is roughly the same over the range of the checked bandwidths. A small increase is observable for the largest bandwidths in the case of the *IntroMathHS* course since for large bandwidths the local estimation reduces back to the global linear case, which gives us larger estimates, as discussed before. The increase is larger for the *IntroMathCalc* course due to the lack of data in the very close proximity of the cut-point. Note that the confidence intervals look different with specification error because the standard error calculation is handled differently in this case by the RDDapp package [52].

This bandwidth sensitivity analysis confirms the stability of the local estimations, and provides another reassurance that our positive results are not just by chance. Also, the comparison of the three discrete variable handling setups shows that the jittered variant is as stable as the other two.



(a) *IntroMathHS* course (b) *IntroMathCalc* course

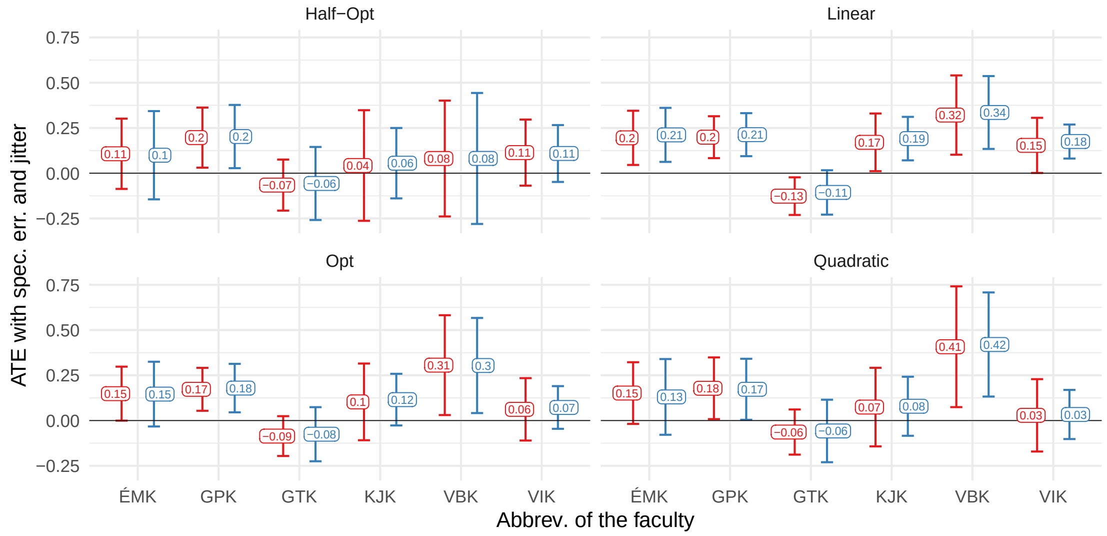
Figure 2.10: Sensitivity to different bandwidths in the non-parametric setup for the three types of discrete running variable handling: studied on the grade of the *Calc1* course as an output for both remedial courses.

2.4.4 Faculty-wise results

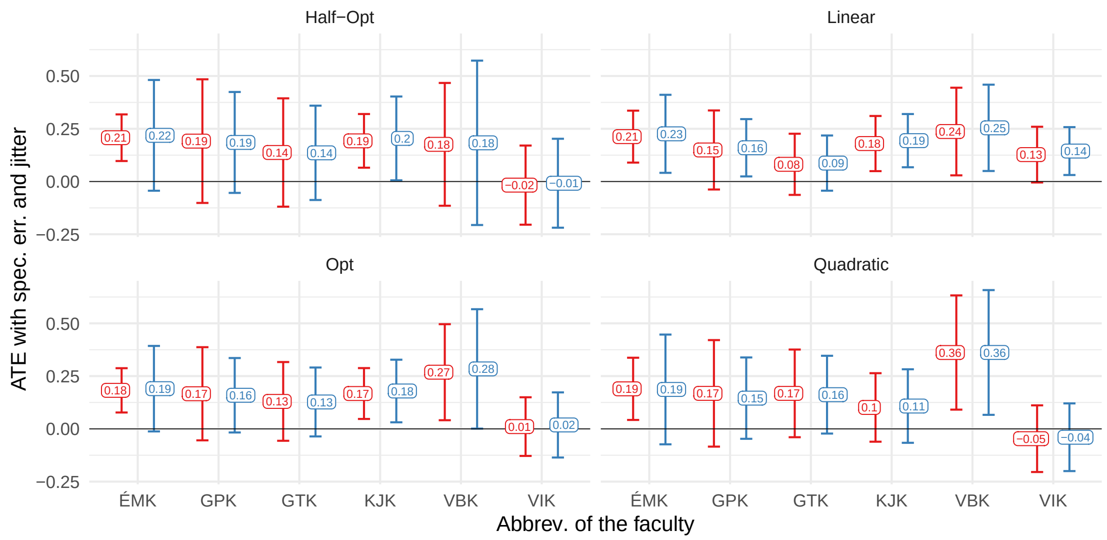
We have estimated the ATE faculty-wise as well. It is necessary since students of different faculties may react differently to the same treatment – in the case of *IntroMathHS*–, and the sample sizes are more balanced this way which makes the comparison (to VBK) more sensible. In Figures 2.11a and 2.11b, we can see the estimates with the 95% confidence intervals per faculty. The red and blue error bars belong to the specification error and the jittered variant of RDD, respectively. Faculties are denoted by their official Hungarian abbreviations. In Figure 2.11c, we can see the sample sizes of the tested sharp scenarios for each faculty.

The figures show that the results are much better at the VBK, where the *IntroMathCalc* takes place, focusing remediation on the material of the *Calc1* course. The results point out the weakness of RDD in assessing the effect of possible hidden factors in the estimation of the treatment effect: there are faculties where the effect of the *IntroMathHS* is positive and statistically significant, whereas other faculties show zero or even negative effect sizes. This will be further assessed in Section 2.4.6.

As a comparison of the specification error and the jittering variants of RDD, the estimated effect size is roughly the same in all cases. The main difference is in the sizes of the confidence intervals. Again, it is the consequence of the robust standard error calculation applied in case of specification error. The study of how standard errors and confidence intervals should be calculated in the case of jittering may be of future interest.



(a) ATE on the grade of the *Calc1* course



(b) ATE on the ACWA-C of the first two semesters

faculty / Treat.	ÉMK	GPK	GTK	KJK	**VBK**	VIK
FALSE	994	3739	1467	1383	**827**	5268
TRUE	740	860	1442	872	**793**	1422
Total (faculty)	1734	4599	2909	2255	**1620**	6690

(c) Sample sizes per faculty; the one with the *IntroMathCalc* is in bold

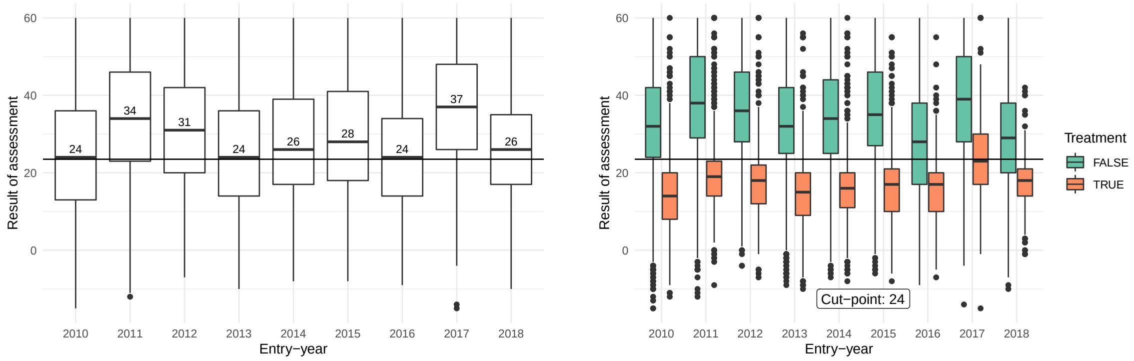
Figure 2.11: ATE estimates with 95% confidence intervals at the different faculties: on two selected outputs each with four regression model setups, where red and blue error bars indicate the spec. error and the jittered variants, respectively.

2.4.5 Yearly results

It is worth checking the ATE estimates year-by-year as well, since the studied 9 years period is relatively long and the effect may vary over the years. In Figures 2.13a and 2.13b, we investigated the ATE on the grade of the *Calc1* course and the grade point average of the first two semesters (ACWA-C), respectively, year by year, for the *IntroMathHS* course. In Figures 2.14a and 2.14b, we checked the ATE on the same outputs for the *IntroMathCalc* course as well, year by year. Under these figures, the yearly sample sizes of the tested sharp scenarios are also included. We have found that the results are varying. There can be several reasons behind this: the quality of the remediation, the ability of the students, the predictive power of the assessment test may change by year. Also, the variety may be caused entirely by some unknown differences in the teaching of the calculus courses. Another caveat is that the results of the global linear setup are not so similar to the others for the *IntroMathHS*, which indicates that it may not be the most adequate setup in our case. The results are still promising. The ATE and its 95% confidence interval are usually in the positive range, except for a couple of years. However, for the *IntroMathCalc* the yearly sample sizes are too small to draw far-reaching conclusions.

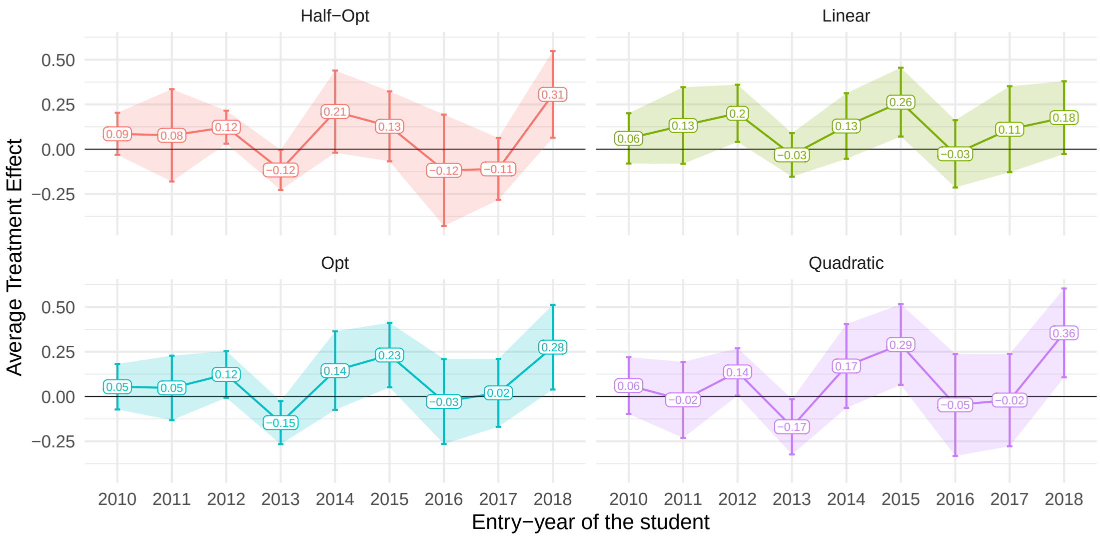
In Figure 2.12, we can see the yearly results of the assessment test. Figure 2.12a shows that the median result is usually at the cut-point or a bit over it. In 2011 and 2012 the results were much better, and even more outstanding in 2017. The latter is due to the much easier test of that year, which was a reaction to previous years' more difficult tests. Figure 2.12b shows that, except for the last three years, around 75% or more of the treated and the control group students are on the "right" side of the cut-point. It means that in those later years the sharp RDD setup is too restrictive as we left out too many students from the analysis. Interestingly, in 2017 a lot of students over the cut-point felt the need for remediation, despite the illusion created by the too easy assessment test.

The yearly findings may require further investigations but the slight fluctuation of the effect size over the years does not affect the overall results.

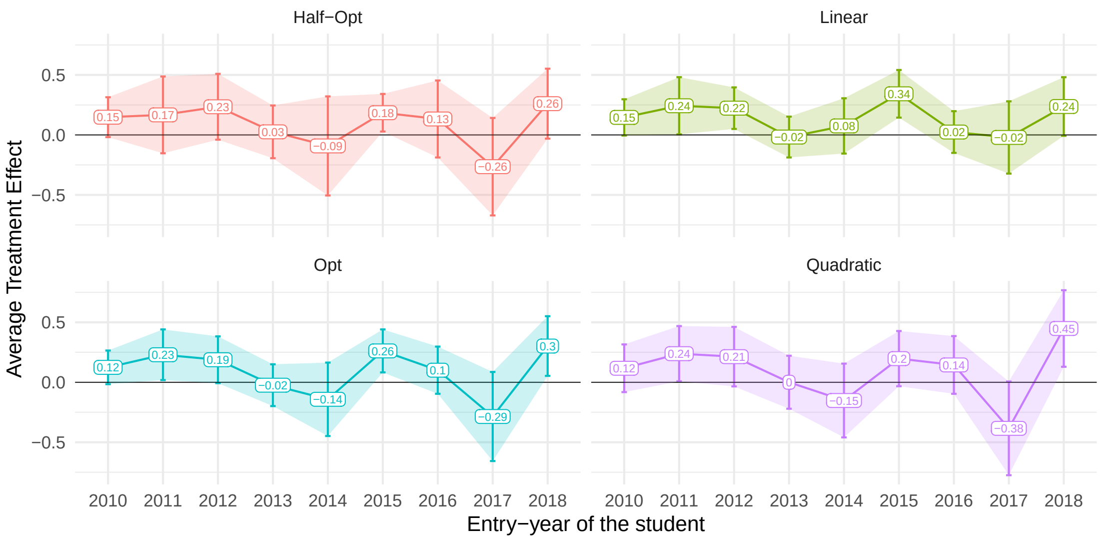


(a) No separation by the *IntroMathHS* (b) Separated by the *IntroMathHS*

Figure 2.12: Results of the assessment test over the years (without the results of the VBK)



(a) ATE on the grade of the *Calc1* course

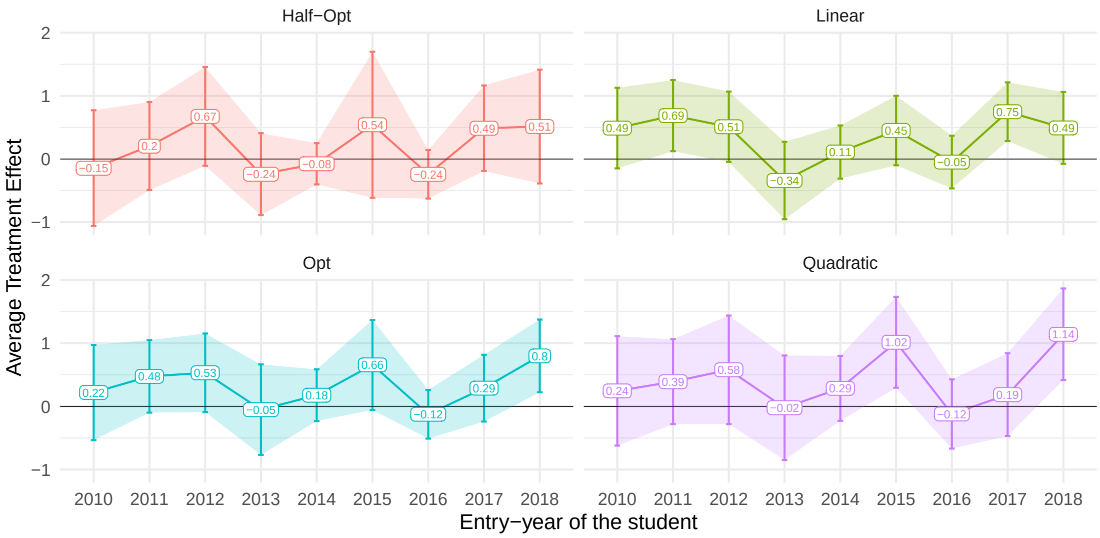


(b) ATE on the ACWA-C of the first two semesters

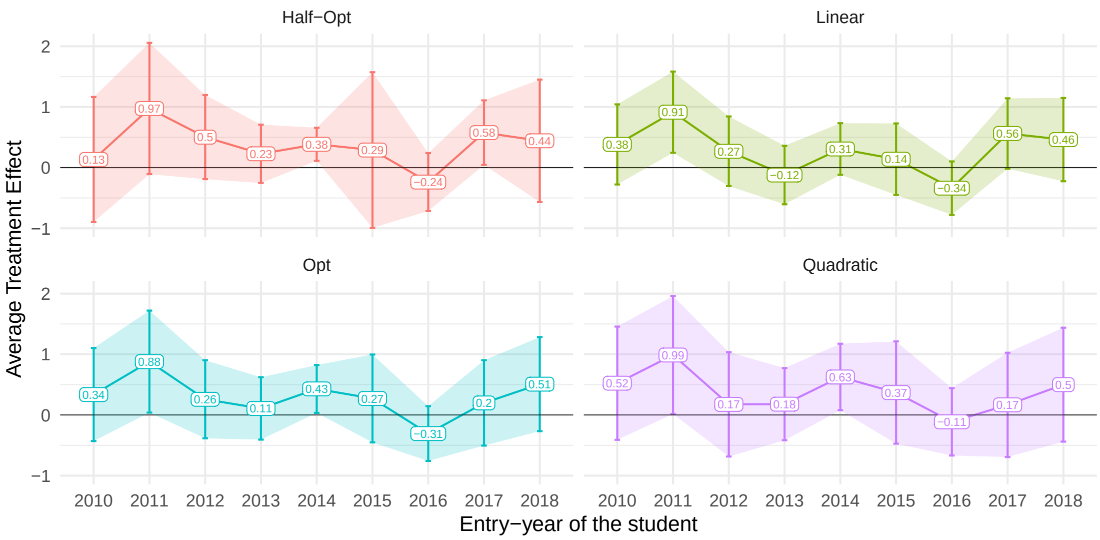
Treat. \ year	2010	2011	2012	2013	2014	2015	2016	2017	2018
FALSE	1346	1919	1597	1195	1382	1449	1169	1696	1098
TRUE	994	457	583	894	750	696	551	96	315
Total (year)	2340	2376	2180	2089	2132	2145	1720	1792	1413

(c) Sample sizes

Figure 2.13: ATE of the *IntroMathHS* course with 95% confidence intervals over the years



(a) ATE on the grade of the *Calc1* course



(b) ATE on the ACWA-C of the first two semesters

year / Treat.	2010	2011	2012	2013	2014	2015	2016	2017	2018
FALSE	81	80	96	74	101	118	78	116	83
TRUE	132	54	78	123	91	77	117	35	86
Total (year)	213	134	174	197	192	195	195	151	169

(c) Sample sizes

Figure 2.14: ATE of the *IntroMathCalc* course with 95% confidence intervals over the years

2.4.6 Covariates

After investigating the faculty-wise and yearly results, we tested how the estimation of the ATE changes if the year of enrollment and the student's faculty are included as discrete covariates in the pooled model of Equation (2.5). We did not find any literature on how discrete covariates may distort the results in an RDD setting.

	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			37.57	18.79
Observations	18187	18187	18186	12961
Estimate	0.142	0.088	0.108	0.100
Std. Error	0.043	0.033	0.034	0.032
$\mathbb{P}(>\lvert z\rvert)$	0.001	0.008	0.002	0.002

(a) the grade of the *Calc1* course

	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			24.93	12.46
Observations	16972	16972	14132	8865
Estimate	0.191	0.115	0.143	0.146
Std. Error	0.044	0.036	0.029	0.030
$\mathbb{P}(>\lvert z\rvert)$	0.000	0.001	0.000	0.000

(b) the ACWA-C of the first two semesters

Table 2.2: RDD results with covariates on two selected outputs for the *IntroMathHS* course

	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			22.44	11.22
Observations	1620	1620	1293	748
Estimate	0.465	0.524	0.448	0.290
Std. Error	0.111	0.161	0.137	0.171
$\mathbb{P}(>\lvert z\rvert)$	0.000	0.001	0.001	0.091

(a) the grade of the *Calc1* course

	Linear	Quadr.	Opt.	.5x-Opt.
Bandwidth			19.31	9.66
Observations	1535	1535	1093	581
Estimate	0.396	0.506	0.455	0.459
Std. Error	0.096	0.117	0.094	0.127
$\mathbb{P}(>\lvert z\rvert)$	0.000	0.000	0.000	0.000

(b) the ACWA-C of the first two semesters

Table 2.3: RDD results with covariates on two selected outputs for the *IntroMathCalc* course

For the *IntroMathCalc* course a third additional covariate, the accomplishment of the *AdvCalc* course (see again Section 2.4.1), was also included. The RDD (with covariates) results on the selected outputs for the *IntroMathHS* and the *IntroMathCalc* courses can be found in Tables 2.2 and 2.3, respectively. The measured effect is slightly larger in every case, and statistically significant in all the relevant cases.

We believe that including these covariates – investigated in earlier sections separately – in the model, and still measuring statistically significant positive effect for both courses is the final reassurance we needed to draw conclusions in the forthcoming section.

Chapter 3

Spectra of Multivariate Stationary Time Series, and Time Domain Relations

In this chapter, we work with multivariate weakly stationary real-valued discrete-time time series, based on articles [J4, C4] with my supervisor and various coauthors. We proved a theorem about the asymptotic relation between the eigenvalues of the block Toeplitz matrix of the first n autocovariances and the union of spectra of the spectral density matrices at the n Fourier frequencies. To achieve this, in Section 3.2, we introduce block circulant matrices and show how the existing results for their spectra are applicable in our situation. It is important that due to the duplication of the eigenvalues, eigenvectors of both real and complex coordinates are available. Both kinds of eigenvectors are intensively studied. After this preparation, we prove our main theorem in Section 3.3. The theorem leads to relations between the time and frequency domain calculations in Section 3.4, like bounds for the spectra, complex principal component transformation, and to some new techniques. In particular, a method for a low-rank approximation of the process (Section 3.4.4); and consequently, a novel dynamic principal component analysis algorithm, called dynPCA (Section 3.5), will be introduced. Applications of these are presented in Section 3.6.

3.1 Preliminaries on Multivariate Stationary Time Series

As mentioned before, we are focusing on multivariate time series. In Appendix C some preliminaries on one dimensional time series is included.

Definition 3.1. *A d-dimensional (discrete-time) time series $\{\boldsymbol{X}_t\}$, $t \in \mathbb{Z}$ is weakly stationary (or stationary in the wide sense) if it has finite expectation and finite covariance function that do not depend on time shift:*

$$\mathbb{E}\boldsymbol{X}_t = \boldsymbol{\mu}, \quad c_{jk}(h) := \operatorname{Cov}\left(X^j_{t+h}, X^k_t\right) = \operatorname{Cov}\left(X^j_h, X^k_0\right),$$

where $t, h \in \mathbb{Z}$; and $j, k = 1, \dots, d$ are indexing the elements of the covariance matrix function $\boldsymbol{C}(h)$ of $\boldsymbol{X}_t$.

Without loss of generality we assume that $\boldsymbol{\mu} = \mathbf{0}$, and then $c_{jk}(h) = \mathbb{E}\left(X^j_h \overline{X^k_0}\right)$. Further on, we study the autocovariance structure of these processes. First, *in the time domain*, let $\{\boldsymbol{X}_t\}$ be a weakly stationary d-dimensional time series, $t \in \mathbb{Z}$ and the state space is $\mathbb{R}^d$. Let $\mathbb{E}(\boldsymbol{X}_t) = \mathbf{0}$ and the sequence of the $d \times d$ autocovariance matrices $\boldsymbol{C}(0), \boldsymbol{C}(1), \boldsymbol{C}(2), \dots$

is absolutely summable (entrywise), meaning that $\sum_{h=0}^{\infty} |c_{pq}(h)| < \infty$ for $p, q = 1, \dots, d$; where $\boldsymbol{C}(-h) = \boldsymbol{C}^T(h)$ and $\boldsymbol{C}(0)$ is the usual covariance matrix of $\boldsymbol{X}_t$. Now the $nd \times nd$ covariance matrix $\mathfrak{C}_n$ of the compounded random vector $(\boldsymbol{X}_1^T, \dots, \boldsymbol{X}_n^T)^T$ is the symmetric, positive semidefinite block-Toeplitz matrix defined as

$$\mathfrak{C}_n\,(\text{block}_i, \text{block}_j) = \begin{cases} \boldsymbol{C}(j-i), & j \geq i; \\ \boldsymbol{C}^T(i-j), & i \leq j. \end{cases}$$

It can be unfolded as

$$\mathfrak{C}_n := \begin{bmatrix} \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) & \cdots & \boldsymbol{C}(n-1) \\ \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) & \cdots & \boldsymbol{C}(n-2) \\ \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \cdots & \boldsymbol{C}(n-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}^T(n-1) & \boldsymbol{C}^T(n-2) & \boldsymbol{C}^T(n-3) & \cdots & \boldsymbol{C}(0) \end{bmatrix}, \tag{3.1}$$

the (i, j) block of which is $\boldsymbol{C}(j-i)$. The symmetry comes from the fact, that the (j, i) block is $\boldsymbol{C}(i-j) = \boldsymbol{C}^T(j-i)$. At the same time *in the frequency domain*, we consider the spectral density matrix $\boldsymbol{f}$ of the process, that exists under the previous conditions. This is a self-adjoint, positive semidefinite, $d \times d$ matrix, which is, in fact, the Fourier transform of the autocovariance matrices:

$$\boldsymbol{f}(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \boldsymbol{C}(h) e^{-ih\omega}, \quad \omega \in [0, 2\pi]. \tag{3.2}$$

It has real diagonal, but usually complex off-diagonal entries. In view of $\boldsymbol{f}(-\omega) = \overline{\boldsymbol{f}(\omega)}$, it suffices to confine ourselves to the $[0, \pi]$ interval (here $\bar{\cdot}$ denotes the complex, entrywise conjugation).

In our main theorem, we state that for "large" n, the eigenvalues of $\mathfrak{C}_n$ asymptotically comprise the union of the spectra (eigenvalues) of the spectral density matrices $\boldsymbol{f}(\omega_j)$, $j = 0, 1, \dots, n-1$. The theorem also has the computational benefit that instead of finding the eigenvalues of an $nd \times nd$ matrix (that needs $\mathcal{O}(n^3 d^3)$ operations) we can find the eigenvalues of number n of $d \times d$ matrices (that means only $\mathcal{O}(nd^3)$ operations).

Under our conditions, the rank r of $\boldsymbol{f}(\omega)$ is almost surely constant on $[0, \pi]$, but it can be less than d. It is true for sliding summations (two-sided moving average processes), in particular, for regular ones (one-sided moving average processes). Processes of rational spectral density, e.g., VARMA processes, are such [2]. At the same time, the dimension of the innovation subspaces, which equals to r, can be concluded from the diagonal blocks of the block Cholesky decomposition of $\mathfrak{C}_n$, and it can also be reduced in the time domain with usual principal component (PC) analysis techniques. In parallel, in the frequency domain, an m-rank approximation with $m \leq r$ is also possible by using PC transformation with m complex PCs, where m depends on the gap of the spectrum of $\boldsymbol{f}$, see also [66].

3.2 Characterizing the Spectra in the Time Domain

To characterize the eigenvalues of the block-Toeplitz matrix $\mathfrak{C}_n$, we need the symmetric block circulant matrix $\mathfrak{C}_n^{(s)}$ that we consider for odd n, say $n = 2k+1$ here (for even n, the calculations are similar); for the definition, see [67]. It is crucial that k is the integer part of $\frac{n}{2}$, as the constructed block circulant matrix is symmetric only with this choice of k. (In the case of even n, the choice of $\frac{n}{2}$ is perfect, and it causes only minor changes in the spectrum, as will be discussed later.) $\mathfrak{C}_n^{(s)}$ is the same as $\mathfrak{C}_n$ for block indices $|j-i| \le k := \lfloor n/2 \rfloor$ and defined as

$$\mathfrak{C}_n^{(s)}\,(\text{block}_i, \text{block}_j) = \begin{cases} \boldsymbol{C}(j-i), & 0 \le j-i \le k; \\ \boldsymbol{C}(n-(j-i)), & 0 \le j-i > k; \\ \boldsymbol{C}^T(i-j), & 0 < i-j \le k; \\ \boldsymbol{C}^T(n-(i-j)), & 0 < i-j > k. \end{cases}$$

It could be defined along its off-diagonal blocks, unfolded as:

$$\mathfrak{C}_n^{(s)} := \begin{bmatrix}
\boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) & & \ddots & \boldsymbol{C}(k) & \boldsymbol{C}(k) & \ddots & \ddots & \boldsymbol{C}(2) & \boldsymbol{C}(1) \\
\boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) & & & \ddots & \boldsymbol{C}(k) & \boldsymbol{C}(k) & \ddots & \boldsymbol{C}(3) & \boldsymbol{C}(2) \\
\boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & & & & \ddots & \boldsymbol{C}(k) & \boldsymbol{C}(k) & \ddots & \boldsymbol{C}(3) \\
\ddots & \ddots & \ddots & \ddots & & & & & & \ddots & \\
\boldsymbol{C}^T(k-1) & & \ddots & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}^T(1) & \ddots & & & \boldsymbol{C}(k) & \boldsymbol{C}(k) \\
\boldsymbol{C}^T(k) & \ddots & & \ddots & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}^T(1) & \ddots & & \ddots & \boldsymbol{C}(k) \\
\boldsymbol{C}^T(k) & \boldsymbol{C}^T(k) & & & \ddots & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}^T(1) & \ddots & & \boldsymbol{C}(k-1) \\
\ddots & \ddots & \ddots & \ddots & & & & & & \ddots & \\
\boldsymbol{C}^T(3) & \ddots & \boldsymbol{C}^T(k) & \boldsymbol{C}^T(k) & \ddots & & & & \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) \\
\boldsymbol{C}^T(2) & \boldsymbol{C}^T(3) & \ddots & \boldsymbol{C}^T(k) & \boldsymbol{C}^T(k) & & & \ddots & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) \\
\boldsymbol{C}^T(1) & \boldsymbol{C}^T(2) & \ddots & \ddots & \boldsymbol{C}^T(k) & \boldsymbol{C}^T(k) & \ddots & & \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0)
\end{bmatrix}. \tag{3.3}$$

This is a symmetric block-Toeplitz matrix, like $\mathfrak{C}_n$, but $\mathfrak{C}_n^{(s)}$ is also a block circulant matrix that fits our purpose better.[1] Spectra of block circulant matrices are well characterized, but $\mathfrak{C}_n$ is not block-circulant, in general; this is why $\mathfrak{C}_n^{(s)}$ is constructed. The rows of a circulant matrix are cyclic permutations of the preceding ones; whereas, in the block circulant case, when permuting, we take transposes of the blocks if those are not symmetric themselves, see Equation (3.3). In $\mathfrak{C}_n^{(s)}$, the autocovariances of order greater than $\frac{n}{2}$ are disregarded. This can be done only on the assumption that the sequences of autocovariances are (entrywise) absolutely summable.

This choice also has benefits when the autocovariance matrices are estimated based on an n-length long time series $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with

$$\hat{\boldsymbol{C}}(h) = \hat{\boldsymbol{C}}^*(-h) = \sum_{t=1}^{n-h} \boldsymbol{X}_{t+h}\boldsymbol{X}_t^*, \quad h = 0, 1, \dots, \lfloor \frac{n}{2} \rfloor$$

where $\mathbb{E}\boldsymbol{X}_t = 0$, or otherwise the estimated mean is subtracted. Here at least $\frac{n}{2}$ sample entries

[1] For even n, there is only one $\boldsymbol{C}(K)$ in each row.

are used for each $\hat{\boldsymbol{C}}(h)$ with $h \leq \frac{n}{2}$, that makes the estimate more accurate. Note that the above estimate is asymptotically unbiased even though its ergodicity is guaranteed only if $h = o(n)$ as $n \to \infty$. Though $\frac{n}{2}$ is obviously not $o(n)$, as the proof of the main theorem will show, disregarding the last $\frac{n}{2}$ autocovariances will guarantee the convergence of the spectrum.

Eigenvalues of $\mathfrak{C}_n^{(s)}$

When $\boldsymbol{C}(h)$s are $d \times d$ matrices, by inflation techniques and applying Kronecker products, we use blocks instead of entries and the eigenvectors also follow a block structure. In [67, 68], the eigenvalues and eigenvectors of a general symmetric block circulant matrix are characterized. We apply this result in our situation, when $n = 2k+1$ is odd. In view of this, the spectrum of $\mathfrak{C}_n^{(s)}$ is the union of spectra of the matrices

$$\begin{aligned}
\boldsymbol{M}_j &= \boldsymbol{C}(0) + \sum_{h=-k}^{k} \boldsymbol{C}(h)\rho_j^h \\
&= \boldsymbol{C}(0) + \sum_{h=1}^{k} [\boldsymbol{C}(h)\rho_j^h + \boldsymbol{C}^T(h)\rho_j^{-h}] \\
&= \boldsymbol{C}(0) + \sum_{h=1}^{k} [\boldsymbol{C}(h)e^{i\omega_j h} + \boldsymbol{C}^T(h)e^{-i\omega_j h}] \\
&= \boldsymbol{C}(0) + \sum_{h=1}^{k} [(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h) + i(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h)
\end{aligned}$$

for $j = 0, 1, \ldots, n-1$; whereas, the eigenvectors are obtained by compounding the eigenvectors of these $d \times d$ matrices. Here $\rho_j := e^{i\omega_j}$ is the jth primitive (complex) nth root of 1, and $\omega_j = \frac{2\pi j}{n}$ is the jth Fourier frequency. Therefore we need the spectral decomposition of the $\boldsymbol{M}_j$ matrices for $j = 1, 2, \ldots, n-1$. Since $\boldsymbol{C}(h) + \boldsymbol{C}^T(h)$ is symmetric and $\boldsymbol{C}(h) - \boldsymbol{C}^T(h)$ is anti-symmetric with 0 diagonal, $\boldsymbol{M}_j$ is self-adjoint for each j and has real eigenvalues with corresponding orthonormal set of eigenvectors of possibly complex coordinates. Indeed, $\boldsymbol{M}_j$ may have complex entries if $j \neq 0$; actually, $\sum_{h=1}^{k}(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h)$ is the real and $\sum_{h=1}^{k}(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h)$ is the imaginary part of $\boldsymbol{M}_j$.

It is easy to see that $\boldsymbol{M}_{n-j} = \overline{\boldsymbol{M}_j}$ (entrywise conjugate), therefore, it has the same (real) eigenvalues as $\boldsymbol{M}_j$, but its (complex) eigenvectors are the (componentwise) complex conjugates of the eigenvectors of $\boldsymbol{M}_j$. We also need the following form of this matrix:

$$\begin{aligned}
\boldsymbol{M}_{n-j} &= \boldsymbol{C}(0) + \sum_{h=1}^{k} [(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h) - i(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h)] \\
&= \boldsymbol{C}(0) + \sum_{h=1}^{k} [\boldsymbol{C}(h)e^{-i\omega_j h} + \boldsymbol{C}^T(h)e^{i\omega_j h}], \quad j = 1, \ldots, n-1.
\end{aligned} \tag{3.4}$$

Summarizing, for odd $n = 2k+1$, the nd eigenvalues of $\mathfrak{C}_n^{(s)}$ are obtained as the union of the (real) eigenvalues of $\boldsymbol{M}_0$ and those of $\boldsymbol{M}_j$ ($j = 1, \ldots, k$) duplicated. Note that for even n, similar arguments hold with the difference that there the spectrum of $\mathfrak{C}_n^{(s)}$ is the union of the eigenvalues of $\boldsymbol{M}_0$ and $\boldsymbol{M}_{n-1}$, whereas the eigenvalues of $\boldsymbol{M}_1, \dots, \boldsymbol{M}_{\frac{n}{2}-1}$ are duplicated.

Eigenvectors of $\mathfrak{C}_n^{(s)}$

The eigenvectors are obtainable by compounding the d (usually complex) orthonormal eigenvectors of the $d \times d$ self-adjoint matrices $\boldsymbol{M}_0, \boldsymbol{M}_1, \dots, \boldsymbol{M}_{n-1}$ as follows. For $j = 1, \dots, k$: if $\boldsymbol{v}$ is a unit-length eigenvector of $\boldsymbol{M}_j$ with eigenvalue λ, then in [67] it is proved that the compound vector

$$(\boldsymbol{v}^T, \rho_j \boldsymbol{v}^T, \rho_j^2 \boldsymbol{v}^T, \dots, \rho_j^{n-1} \boldsymbol{v}^T)^T \in \mathbb{C}^{nd}$$

is an eigenvector of $\mathfrak{C}_n^{(s)}$ with the same eigenvalue λ. It has squared norm

$$\boldsymbol{v}^* \boldsymbol{v} (1 + \rho_j \rho_j^{-1} + \rho_j^2 \rho_j^{-2} + \dots + \rho_j^{n-1} \rho_j^{-(n-1)}) = n.$$

Therefore, the vector

$$\boldsymbol{w} = \frac{1}{\sqrt{n}} (\boldsymbol{v}^T, \rho_j \boldsymbol{v}^T, \rho_j^2 \boldsymbol{v}^T, \dots, \rho_j^{n-1} \boldsymbol{v}^T)^T \in \mathbb{C}^{nd} \tag{3.5}$$

is a unit-norm eigenvector (of complex coordinates) of $\mathfrak{C}_n^{(s)}$.

Further, if

$$\boldsymbol{z} = \frac{1}{\sqrt{n}} (\boldsymbol{t}^T, \rho_\ell \boldsymbol{t}^T, \rho_\ell^2 \boldsymbol{t}^T, \dots, \rho_\ell^{n-1} \boldsymbol{t}^T)^T \in \mathbb{C}^{nd}$$

is another unit-norm eigenvector of $\mathfrak{C}_n^{(s)}$ compounded from a unit-norm eigenvector $\boldsymbol{t}$ of another $\boldsymbol{M}_\ell$ ($\ell \neq j$), then $\boldsymbol{w}$ and $\boldsymbol{z}$ are orthogonal, irrespective whether $\boldsymbol{M}_\ell$ has the same eigenvalue λ as $\boldsymbol{M}_j$ or not. Similar construction holds starting with the eigenvectors of $\boldsymbol{M}_0$.

Here for each $j = 0, 1, \dots, n-1$, there are d pairwise orthonormal eigenvectors (potential $\boldsymbol{v}$s) of $\boldsymbol{M}_j$, and the so obtained $\boldsymbol{w}$s are also pairwise orthonormal. Assume that the eigenvectors of $\boldsymbol{M}_j$ are enumerated in nonincreasing order of its (real) eigenvalues, and the inflated $\boldsymbol{w}$s also follow this ordering, for $j = 0, 1, \dots, n-1$.

Choose a unit-norm eigenvector $\boldsymbol{v} \in \mathbb{C}^d$ of $\boldsymbol{M}_j$ with (real) eigenvalue λ. Then $\overline{\boldsymbol{v}} \in \mathbb{C}^d$ is the corresponding unit-norm eigenvector of $\boldsymbol{M}_{n-j}$ with the same eigenvalue λ. Consider the compounded $\boldsymbol{w} \in \mathbb{C}^{nd}$ and $\overline{\boldsymbol{w}} \in \mathbb{C}^{nd}$ obtained from them by Equation (3.5). We learned that they are orthonormal eigenvectors of $\mathfrak{C}_n^{(s)}$ corresponding to the eigenvalue λ with multiplicity (at least) two. From them, corresponding to this double eigenvalue λ, the new orthonormal pair of eigenvectors

$$\frac{\boldsymbol{w} + \overline{\boldsymbol{w}}}{\sqrt{2}} \quad \text{and} \quad -i \frac{\boldsymbol{w} - \overline{\boldsymbol{w}}}{\sqrt{2}} \tag{3.6}$$

is constructed, but they, in this order, occupy the original positions of $\boldsymbol{w}$ and $\overline{\boldsymbol{w}}$. They have real coordinates and unit norm. (Actually, their coordinates are the $\sqrt{2}$ multiples of the real and imaginary parts of the coordinates of $\boldsymbol{w}$.) It is in accord with the fact that a real symmetric matrix, as $\mathfrak{C}_n^{(s)}$, must have an orthogonal system of eigenvectors with real coordinates too. We do not go into details, nor discuss defective cases.

Consider $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{nd}$, the so obtained orthonormal set of eigenvectors (of real coordinates) of $\mathfrak{C}_n^{(s)}$ (in the above ordering), and denote by $\boldsymbol{U} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_{nd})$ the $nd \times nd$ (real) orthogonal matrix containing them columnwise. Let

$$\mathfrak{C}_n^{(s)} = \boldsymbol{U}\boldsymbol{\Lambda}^{(s)}\boldsymbol{U}^T \tag{3.7}$$

be the corresponding spectral decomposition. After this preparation, we are able to prove the following theorem.

3.3 The Main Theorem about the Spectra in the Time and Frequency Domain

The following theorem can be formulated based on the previous observations.

Theorem 3.1. *Let $\{\boldsymbol{X}_t\}$ be d-dimensional weakly stationary time series of real components. Denoting by $\boldsymbol{C}(h) = [c_{ij}(h)]$ the $d \times d$ autocovariance matrices ($\boldsymbol{C}(-h) = \boldsymbol{C}^T(h)$, $h \in \mathbb{Z}$) in the time domain, assume that their entries are absolutely summable, i.e., $\sum_{h=0}^{\infty} |c_{pq}(h)| < \infty$ for $p, q = 1, \dots, d$. Then, the self-adjoint, positive semidefinite spectral density matrix $\boldsymbol{f}$ exists in the frequency domain, defined by Equation* (3.2). *For odd $n = 2k+1$, consider $\boldsymbol{X}_1, \dots \boldsymbol{X}_n$ and the block Toeplitz matrix $\mathfrak{C}_n$ of Equation* (3.1)*; further, the Fourier frequencies $\omega_j = \frac{2\pi j}{n}$ for $j = 0, \dots, n-1$. Further, let $\boldsymbol{D}_n$ be the $dn \times dn$ diagonal matrix that contains the spectra* [2] *of the matrices $\boldsymbol{f}(0), \boldsymbol{f}(\omega_1), \boldsymbol{f}(\omega_2), \dots, \boldsymbol{f}(\omega_k)$, $\boldsymbol{f}(\omega_k), \dots, \boldsymbol{f}(\omega_2), \boldsymbol{f}(\omega_1)$ in its main diagonal* [3]*, i.e.,*

$$\boldsymbol{D}_n = \mathrm{diag}(\mathrm{spec}\, \boldsymbol{f}(0), \mathrm{spec}\, \boldsymbol{f}(\omega_1), \dots, \mathrm{spec}\, \boldsymbol{f}(\omega_k), \mathrm{spec}\, \boldsymbol{f}(\omega_k), \dots, \mathrm{spec}\, \boldsymbol{f}(\omega_1)).$$

Then, with the spectral decomposition of Equation (3.7),

$$\boldsymbol{U}^T \mathfrak{C}_n \boldsymbol{U} - 2\pi \boldsymbol{D}_n \to \boldsymbol{O}, \quad n \to \infty,$$

i.e., the entries of the matrix $\boldsymbol{U}^T \mathfrak{C}_n \boldsymbol{U} - 2\pi \boldsymbol{D}_n$ tend to 0 uniformly.

Proof. We saw that $\boldsymbol{U}^T \mathfrak{C}_n^{(s)} \boldsymbol{U} = \boldsymbol{\Lambda}^{(s)}$. Recall that the eigenvalues in the diagonal of $\boldsymbol{\Lambda}^{(s)}$ comprise the union of spectra of the matrices $\boldsymbol{M}_0$ and those of $\boldsymbol{M}_1, \dots, \boldsymbol{M}_{n-1}$, which are the same as the eigenvalues of $\boldsymbol{M}_0$ and those of $\boldsymbol{M}_{n-1}, \dots, \boldsymbol{M}_{n-k}$ of Equation (3.4), duplicated. However, these matrices are finite sub-sums (for $|h| \le k$) of the infinite summations

$$2\pi \boldsymbol{f}(\omega_j) = \sum_{h=-\infty}^{\infty} \boldsymbol{C}(h) e^{-ih\omega} = \boldsymbol{C}(0) + \sum_{h=1}^{\infty} [\boldsymbol{C}(h) e^{-i\omega_j h} + \boldsymbol{C}^T(h) e^{i\omega_j h}].$$

So, by the absolute summability of the autocovariances, and because the eigenvalues depend continuously on the underlying matrices, the pairwise distances between the eigenvalues of $\boldsymbol{M}_j$ and the corresponding eigenvalues of $2\pi \boldsymbol{f}(\omega_j)$ (both in nonincreasing order) tend to 0 as $n \to \infty$, for $j = 0, 1, \dots, k$. Indeed, the absolute summability of the entries of $\boldsymbol{C}(h)$s implies that the diagonal entries of the diagonal matrix $\boldsymbol{\Lambda}^{(s)} - 2\pi \boldsymbol{D}_n$ are bounded in absolute value

[2] The eigenvalues (spec $(\cdot)$) of the affected matrix are in nonincreasing order if not otherwise stated.
[3] The duplication is due to the fact that $\boldsymbol{f}(\omega_j) = \boldsymbol{f}(\omega_{n-j})$, $j = 1, \dots, k$, for real time series.

by

$$\max_{p,q\in\{1,\dots,d\}} \sum_{|h|>k} |c_{pq}(h)| \to 0, \quad n = 2k+1 \to \infty.$$

Consequently, the matrix $\boldsymbol{\Lambda}^{(s)} - 2\pi \boldsymbol{D}_n$ tends to the zero matrix entrywise uniformly as $n \to \infty$. Therefore, it remains to show that the entries of $\boldsymbol{U}^T \mathfrak{C}_n \boldsymbol{U} - \boldsymbol{U}^T \mathfrak{C}_n^{(s)} \boldsymbol{U}$ tend to 0 uniformly as $n \to \infty$.

Before doing this, some facts should be clarified.

- The pth row sum of $\boldsymbol{M}_j$ is bounded by

$$\sum_{q=1}^{d} |c_{pq}(0)| + \sum_{q=1}^{d}\sum_{h=1}^{k} |c_{pq}(h)| + \sum_{q=1}^{d}\sum_{h=1}^{k} |c_{qp}(h)| \le d c_{pp}(0) + 2dL,$$

  for $p \in \{1,\dots,d\}$ with $L = \max_{p,q\in\{1,\dots,d\}} \sum_{h=1}^{\infty} |c_{pq}(h)| > 0$, independently of n, because of the absolute summability of the entries of $\boldsymbol{C}(h)$. This is true for any $j \in \{0,1,\dots,n-1\}$. For simplicity, consider (any) one of the $\boldsymbol{M}_j$s, and denote it by $\boldsymbol{M} = [m_{pq}]_{p,q=1}^{d}$. Then

$$\|\boldsymbol{M}\|_\infty = \max_{p\in\{1,\dots,d\}} \sum_{q=1}^{d} |m_{pq}| \le d \max_{p\in\{1,\dots,d\}} c_{pp}(0) + 2dL = K.$$

  As the spectral radius of $\boldsymbol{M}$ is at most $\|\boldsymbol{M}\|_\infty$, any eigenvalue λ of $\boldsymbol{M}$ is bounded in absolute value by K (independently of n).

- Recall that $\boldsymbol{u}$ is compounded via Equations (3.5) and (3.6) from the primitive roots. Therefore, its coordinates are bounded by $\sqrt{\frac{2}{n}}$ in absolute value.

Now we are ready to show that

$$|\boldsymbol{u}_i^T \mathfrak{C}_n^{(s)} \boldsymbol{u}_j - \boldsymbol{u}_i^T \mathfrak{C}_n \boldsymbol{u}_j| \to 0, \quad n \to \infty$$

uniformly in $i,j \in \{1,\dots,nd\}$. Recall that in the $nd \times nd$ matrices $\mathfrak{C}_n^{(s)}$ and $\mathfrak{C}_n$ the (m,ℓ) blocks are the same if $|m-\ell| \le k$. Denote by $\boldsymbol{u}_{i,m}$ and $\boldsymbol{u}_{j,\ell}$ the mth and ℓth blocks of the unit-norm eigenvectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$, respectively. Then

$$
\begin{aligned}
&|\boldsymbol{u}_i^T(\mathfrak{C}_n^{(s)} - \mathfrak{C}_n)\boldsymbol{u}_j| \\
&= \left| \sum_{m=1}^{k} \sum_{\ell=1}^{m} [\boldsymbol{u}_{i,\ell}^T(\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\boldsymbol{u}_{j,n-m+\ell} + \boldsymbol{u}_{i,n-m+\ell}^T(\boldsymbol{C}(n-m) - \boldsymbol{C}(m))\boldsymbol{u}_{j,\ell}] \right| \\
&\le \left| \sqrt{\frac{2}{n}} \sum_{m=1}^{k} \mathbf{1}_d^T(\boldsymbol{C}(m) - \boldsymbol{C}(n-m)) \sum_{\ell=1}^{m} \boldsymbol{u}_{j,n-m+\ell} \right| \\
&+ \left| \sqrt{\frac{2}{n}} \sum_{m=1}^{k} \sum_{\ell=1}^{m} \boldsymbol{u}_{i,n-m+\ell}^T(\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{1}_d \right| \\
&\le 2\sqrt{\frac{2}{n}}\sqrt{\frac{2}{n}} \left| \sum_{m=1}^{k} m\mathbf{1}_d^T(\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{1}_d \right| \\
&\le \frac{4}{n}\left(\sum_{m=1}^{k} m \sum_{p=1}^{d}\sum_{q=1}^{d} |c_{pq}(m)| + \sum_{m=1}^{k} m \sum_{p=1}^{d}\sum_{q=1}^{d} |c_{pq}(n-m)| \right) \\
&\le 4d^2 \left(\max_{p,q\in\{1,\dots,d\}} \sum_{m=1}^{k} \frac{m}{n}|c_{pq}(m)| + \max_{p,q\in\{1,\dots,d\}} \sum_{m=1}^{k} \frac{m}{n}|c_{pq}(n-m)| \right) \\
&\le 4d^2 \left(\max_{p,q\in\{1,\dots,d\}} \sum_{m=1}^{k} \frac{m}{n}|c_{pq}(m)| + \max_{p,q\in\{1,\dots,d\}} \sum_{m=n-k}^{n-1} \frac{k}{n}|c_{pq}(m)| \right),
\end{aligned}
$$

where $\mathbf{1}_d \in \mathbb{R}^d$ is the vector of all 1 coordinates and so, the quadratic form $\mathbf{1}_d^T(\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{1}_d$ is the sum of the entries of $\boldsymbol{C}(m) - \boldsymbol{C}(n-m)$. In the last line, the second term converges to 0, since it is bounded by $\sum_{m=k}^{\infty} |c_{pq}(m)|$ (indeed, $\sum_{m=n-k}^{n-1} \frac{k}{n}|c_{pq}(m)| \le \sum_{m=k}^{\infty} |c_{pq}(m)|$ as $k < n-k$), and together with n, k tends to ∞ too; further, it holds uniformly for all $p, q \in \{1, \dots, d\}$. The first term for every p, q pair also tends to 0 as $n \to \infty$ by the discrete version of the dominated convergence theorem (for series), see the forthcoming Lemma 3.1. Indeed, the summand is dominated by $|c_{pq}(m)|$ and $\sum_{m=1}^{\infty} |c_{pq}(m)| < \infty$; further, $\frac{m}{n}|c_{pq}(m)| \to 0$ as $n \to \infty$, for any fixed m. Consequently, $\sum_{m=1}^{\infty} \frac{m}{n}|c_{pq}(m)|$ tends to 0, and so does $\sum_{m=1}^{k} \frac{m}{n}|c_{pq}(m)|$ as $n \to \infty$. It holds uniformly for all p, q, and also for all i, j, so the proof is complete. □

Lemma 3.1 (Dominated convergence theorem for sums, discrete version). *Consider the sum $\sum_{m=1}^{\infty} f_n(m)$ and assume that $|f_n(m)| \le g(m)$ with $\sum_{m=1}^{\infty} g(m) < \infty$. If $\lim_{n\to\infty} f_n(m) = f(m)$ exists $\forall m \in \mathbb{N}$, then*

$$\lim_{n\to\infty} \sum_{m=1}^{\infty} f_n(m) = \sum_{m=1}^{\infty} f(m).$$

The proof of Theorem 3.1 suggests that the speed of convergence depends on how rapidly the tail of the absolutely summable autocovariances tends to zero. For example, in the case of VARMA processes autocovariances decrease exponentially, and this implies the convergence at a geometric rate.

3.4 Relations and Consequences of the Main Theorem

3.4.1 Bounds for the eigenvalues of $\mathfrak{C}_n$

Analogously to the 1D statement (see [69]), for the eigenvalues of the block Toeplitz matrix $\mathfrak{C}_n$ the following lower and upper bounds are obtainable.

Proposition 3.1. *Assume that for the spectra of the spectral densities $\boldsymbol{f}$ of the d-dimensional weakly stationary process $\{\boldsymbol{X}_t\}$ of real coordinates the following hold:*

$$\lambda_{\text{inf}} := \inf_{\omega\in[0,2\pi],q\in\{1,\dots,d\}} \lambda_q(\boldsymbol{f}(\omega)) > 0,$$

$$\lambda_{\text{sup}} := \sup_{\omega\in[0,2\pi],q\in\{1,\dots,d\}} \lambda_q(\boldsymbol{f}(\omega)) < \infty.$$

(Note that under the conditions of Theorem 3.1, $\boldsymbol{f}(\omega) > 0$ and it is continuous almost everywhere on $[0, 2\pi]$, so the above conditions are readily satisfied.)

Then for the eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_{nd}$ of the block Toeplitz matrix $\mathfrak{C}_n$ the following holds:

$$2\pi\lambda_{\text{inf}} \le \lambda_1 \le \lambda_{nd} \le 2\pi\lambda_{\text{sup}}.$$

Proof. Let λ be an arbitrary eigenvalue of $\mathfrak{C}_n$ with a corresponding eigenvector $\boldsymbol{x} \in \mathcal{C}^{nd}$ (i.e., $\mathfrak{C}_n\boldsymbol{x} = \lambda\boldsymbol{x}$), and denote by $\boldsymbol{x}_j$ the jth block of it ($j = 1,\dots,n$). Take the spectral decomposition of the spectral density matrix $\boldsymbol{f}$:

$$\boldsymbol{f}(\omega) = \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \cdot \boldsymbol{u}_\ell(\boldsymbol{f}(\omega)) \cdot \boldsymbol{u}_\ell^*(\boldsymbol{f}(\omega)).$$

Then we can write that

$$\begin{aligned}
\lambda|\boldsymbol{x}|^2 &= \lambda\boldsymbol{x}^*\boldsymbol{x} = \boldsymbol{x}^*\mathfrak{C}_n\boldsymbol{x} \\
&= \boldsymbol{x}^* \cdot \int_{-\pi}^{\pi} \left[e^{-i(j-k)\omega}\boldsymbol{f}(\omega)\right]_{j,k=1}^{n} d\omega \cdot \boldsymbol{x} \\
&= \int_{-\pi}^{\pi} \sum_{j,k=1}^{n} e^{-i(j-k)\omega}\boldsymbol{x}_j^*\boldsymbol{f}(\omega)\boldsymbol{x}_k d\,\omega \\
&= \int_{-\pi}^{\pi} \sum_{j,k=1}^{n} e^{-i(j-k)\omega} \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \cdot \boldsymbol{x}_j^* \cdot \boldsymbol{u}_\ell(\boldsymbol{f}(\omega)) \cdot \boldsymbol{u}_\ell^*(\boldsymbol{f}(\omega)) \cdot \boldsymbol{x}_k\, d\omega \\
&= \int_{-\pi}^{\pi} \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \sum_{j,k=1}^{n} e^{-ij\omega}\boldsymbol{x}_j^* \cdot \boldsymbol{u}_\ell(\boldsymbol{f}(\omega)) \cdot \boldsymbol{u}_\ell^*(\boldsymbol{f}(\omega)) \cdot \boldsymbol{x}_k \cdot e^{ik\omega}\, d\omega \\
&\le \lambda_{\text{sup}} \sum_{j,k=1}^{n} \boldsymbol{x}_j^* \cdot \int_{-\pi}^{\pi} e^{-i(j-k)\omega} \sum_{\ell=1}^{d} \boldsymbol{u}_\ell(\boldsymbol{f}(\omega)) \cdot \boldsymbol{u}_\ell^*(\boldsymbol{f}(\omega))\, d\omega \cdot \boldsymbol{x}_k \\
&= 2\pi\lambda_{\text{sup}} \sum_{j=1}^{n} \boldsymbol{x}_j^*\boldsymbol{x}_j = 2\pi M|\boldsymbol{x}|^2.
\end{aligned}$$

This proves that $\lambda \le 2\pi\lambda_{\text{sup}}$ for any eigenvalue of $\mathfrak{C}_n$. The proof of the fact that $\lambda \ge 2\pi\lambda_{\text{inf}}$ is similar. □

3.4.2 Complex PC transformation

The complex PC transform of the collection of random vectors $\boldsymbol{X} = (\boldsymbol{X}_1^T, \dots \boldsymbol{X}_n^T)^T$ of real coordinates is the random vector $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \dots, \boldsymbol{Z}_n^T)^T$ of complex coordinates obtained by

$$\boldsymbol{Z} = \boldsymbol{W}^* \boldsymbol{X}.$$

Here, analogously to Equation (3.7), $\mathfrak{C}_n^{(s)}$ has the spectral decomposition

$$\mathfrak{C}_n^{(s)} = \boldsymbol{W} \boldsymbol{\Lambda}^{(s)} \boldsymbol{W}^* \tag{3.8}$$

with complex eigenvectors too: $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_{nd})$, where $\boldsymbol{w}$s are the orthonormal eigenvectors of $\mathfrak{C}_n^{(s)}$ (of complex coordinates), introduced by Equation (3.5). By Theorem 3.1, $\mathbb{E}\boldsymbol{Z}\boldsymbol{Z}^* \sim 2\pi \boldsymbol{D}_n$, so the coordinates of $\boldsymbol{Z}$ are asymptotically uncorrelated, for "large" n.

Instead of the intractable compounded transformation, we consider the blocks $\boldsymbol{Z}_j$s of $\boldsymbol{Z}$, and perform a "partial principal component transformation" (in d-dimension) on them. For this purpose, we use the eigenvectors in the columns of $\boldsymbol{W}$ (of complex coordinates) in the ordering described in Section 3.2. We utilize their block structure and also assume that they are already normalized to have a complete orthonormal system in $\mathbb{C}^{nd}$. Let $\boldsymbol{w}_{1j}, \dots, \boldsymbol{w}_{dj}$ be the columns of $\boldsymbol{W}$ corresponding to the coordinates of $\boldsymbol{Z}_j$.In view of Equation (3.5), the block $\boldsymbol{Z}_j$ can be written as

$$\boldsymbol{Z}_j = \frac{1}{\sqrt{n}} \left(\boldsymbol{V}_j^* \otimes \boldsymbol{r}^* \right) \boldsymbol{X},$$

where $\boldsymbol{r}^* = (1, \rho_j^{-1}, \rho_j^{-2}, \dots, \rho_j^{-(n-1)})$ and $\boldsymbol{V}_j$ is the $d \times d$ unitary matrix in the spectral decomposition $\boldsymbol{M}_j = \boldsymbol{V}_j \boldsymbol{\Lambda}_j \boldsymbol{V}_j^*$.

3.4.3 Relation to discrete Fourier transformation

We can relate this complex PC transformation to discrete Fourier transformation. Because of $\mathbb{E}\boldsymbol{Z}_j \boldsymbol{Z}_j^* = \boldsymbol{\Lambda}_j$, from the proof of Theorem 3.1, we have that

$$\mathbb{E}(\boldsymbol{V}_j \boldsymbol{Z}_j)(\boldsymbol{V}_j \boldsymbol{Z}_j)^* = \boldsymbol{V}_j \boldsymbol{\Lambda}_j \boldsymbol{V}_j^* = \boldsymbol{M}_j.$$

At the same time,

$$\boldsymbol{V}_j \boldsymbol{Z}_j = \frac{1}{\sqrt{n}} \boldsymbol{V}_j \left(\boldsymbol{V}_j^* \otimes \boldsymbol{r}^* \right) \boldsymbol{X} = \frac{1}{\sqrt{n}} \left(\mathbb{I}_d \otimes \boldsymbol{r}^* \right) \boldsymbol{X} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{X}_t e^{-it\omega_j}, \quad j = 1, \dots, n.$$

This is the discrete Fourier transform of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. It is in accord with the existence of the orthogonal increment process $\{\boldsymbol{Z}_\omega\}$ (see [69]) of which $\boldsymbol{V}_j \boldsymbol{Z}_j \sim \boldsymbol{Z}_{\omega_j} - \boldsymbol{Z}_{\omega_{j-1}}$ is the discrete analogue. Note that a weakly stationary time series itself can be represented (with probability one) as

$$\boldsymbol{X}_t = \int_{-\pi}^{\pi} e^{it\omega} d\boldsymbol{Z}_\omega, \quad t \in \mathbb{Z}$$

with the process $\boldsymbol{Z}_\omega$ of orthogonal increments, where $\boldsymbol{F}(\omega) = \mathbb{E}(\boldsymbol{Z}_\omega \boldsymbol{Z}_\omega^*)$, $\boldsymbol{F}(\omega) = \int_{-\pi}^{\omega} \boldsymbol{f}(\theta)\, d\theta$, $\omega \in [-\pi, \pi]$. This is sometimes called Cramér representation. (It generalizes the case of the superposition of sinusoids, where the process has a point spectrum.)

Also, $\boldsymbol{Z}_1, \dots \boldsymbol{Z}_n$ are asymptotically pairwise orthogonal akin to $\boldsymbol{V}_1\boldsymbol{Z}_1, \dots, \boldsymbol{V}_n\boldsymbol{Z}_n$. Further,

$$\mathbb{E}(\boldsymbol{V}_j\boldsymbol{Z}_j)(\boldsymbol{V}_j\boldsymbol{Z}_j)^* \sim 2\pi \boldsymbol{f}(\omega_j),$$

and it is in accord with the fact that

$$\mathbb{E}\boldsymbol{Z}_j\boldsymbol{Z}_j^* \sim 2\pi \text{ diag spec } \boldsymbol{f}(\omega_j),$$

for $j = 0, 1, \dots, n-1$ when n is "large".

3.4.4 Low-rank approximation of the process

Let $\{\boldsymbol{X}_t\}_{t=1}^n$ be the finite part of a d-dimensional process of real coordinates and constant rank $1 \le r \le d$. Its discrete Fourier transform, discussed in Section 3.4.3, is

$$\boldsymbol{T}_j := \boldsymbol{V}_j\boldsymbol{Z}_j = \frac{1}{\sqrt{n}}\sum_{t=1}^n \boldsymbol{X}_t e^{-it\omega_j}, \quad j = 0, \dots, n-1.$$

More precisely, $\boldsymbol{T}_0 = \frac{1}{\sqrt{n}}\sum_{t=1}^n \boldsymbol{X}_t$, $\boldsymbol{T}_j = \frac{1}{\sqrt{n}}\sum_{t=1}^n \boldsymbol{X}_t[\cos(t\omega_j) - i\sin(t\omega_j)]$, and $\boldsymbol{T}_{n-j} = \overline{\boldsymbol{T}}_j$, for $j = 1, \dots, k$ $(n = 2k+1)$. Therefore,

$$\boldsymbol{Z}_j = \boldsymbol{V}_j^{-1}\boldsymbol{T}_j = \boldsymbol{V}_j^*\boldsymbol{T}_j, \quad j = 0, \dots, n-1;$$

and it can easily be seen that $\boldsymbol{Z}_{n-j} = \overline{\boldsymbol{Z}}_j$ too.

To find the best m-rank approximation $(1 < m \le r)$ of the process, we project the d-dimensional vector $\boldsymbol{T}_j$ *onto the subspace* spanned by the m leading eigenvectors of $\boldsymbol{V}_j$. Important that the eigenvalues in $\boldsymbol{\Lambda}_j$ are in nonincreasing order. Let us denote the eigenvectors corresponding to the m largest eigenvalues by $\boldsymbol{v}_{j1}, \dots, \boldsymbol{v}_{jm}$, compounded into the matrix $\boldsymbol{V}_j^{(m)}$. Then

$$\widehat{\boldsymbol{T}}_j^{(m)} := \texttt{Proj}_{\,\texttt{Span}\,\{\boldsymbol{v}_{j1},\dots,\boldsymbol{v}_{jm}\}}\boldsymbol{T}_j = \left[\boldsymbol{V}_j^{(m)}\left(\boldsymbol{V}_j^{(m)}\right)^*\right]\boldsymbol{V}_j\boldsymbol{Z}_j = \sum_{\ell=1}^m \left(\boldsymbol{v}_{j\ell}^*\boldsymbol{V}_j\boldsymbol{Z}_j\right)\boldsymbol{v}_{j\ell} = \sum_{\ell=1}^m Z_{j\ell}\boldsymbol{v}_{j\ell},$$

and $\widehat{\boldsymbol{T}}_{n-j} = \overline{\widehat{\boldsymbol{T}}}_j$, for $j = 1, \dots, k$ (by the previous considerations), were $n = 2k+1$. So, for each j, the resulting vector is the linear combination of the vectors $\boldsymbol{v}_{j\ell}$s with the corresponding coordinates $Z_{j\ell}$s of $\boldsymbol{Z}_j$, $\ell = 1, \dots, m$.

Eventually, we find the m-rank approximation of $\boldsymbol{X}_t$ in the time domain by inverse Fourier transformation:

$$\widehat{\boldsymbol{X}}_t^{(m)} := \underset{j}{\text{IDFT}}\left[\widehat{\boldsymbol{T}}_j^{(m)}\right] = \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\widehat{\boldsymbol{T}}_j^{(m)} e^{it\omega_j}.$$

Apparently, the vectors $\widehat{\boldsymbol{X}}_t$, $t = 1, \dots, n$ (omitting the $^{(m)}$ notion for conciseness) all have real coordinates ($n = 2k+1$):

$$\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\widehat{\boldsymbol{T}}_j e^{it\omega_j} == \frac{1}{\sqrt{n}}\left\{\widehat{\boldsymbol{T}}_0 + \sum_{j=1}^{k}[(\widehat{\boldsymbol{T}}_j + \overline{\widehat{\boldsymbol{T}}}_j)\cos(t\omega_j) + i(\widehat{\boldsymbol{T}}_j - \overline{\widehat{\boldsymbol{T}}}_j)\sin(t\omega_j)]\right\}$$
$$= \frac{1}{\sqrt{n}}\left\{\widehat{\boldsymbol{T}}_0 + \sum_{j=1}^{k}[2\mathrm{Re}(\widehat{\boldsymbol{T}}_j)\cos(t\omega_j) + i\cdot 2i\cdot \mathrm{Im}(\widehat{\boldsymbol{T}}_j)\sin(t\omega_j)]\right\}$$
$$= \frac{1}{\sqrt{n}}\left\{\widehat{\boldsymbol{T}}_0 + 2\sum_{j=1}^{k}[\mathrm{Re}(\widehat{\boldsymbol{T}}_j)\cos(t\omega_j) - \mathrm{Im}(\widehat{\boldsymbol{T}}_j)\sin(t\omega_j)]\right\}.$$

In this way, we have a lower rank process with a spectral density of rank $m \le r$. Note that if the process is regular (e.g., it has a rational spectral density), then so is its low-rank approximation.

3.5 dynPCA: low-dimensional approximation from complex PCs

The previous *rank* reduction gives rise to a *dimension* reduction technique too. In the frequency domain, *instead of projection onto* the d-dimensional subspace spanned by the m leading eigenvectors of $\boldsymbol{V}_j$, *we map into it.* We have seen this transformation before but with the whole $\boldsymbol{V}_j$ matrix:

$$\boldsymbol{Z}_j^{(m)} = \left[\boldsymbol{V}_j^{(m)}\right]^* \boldsymbol{T}_j, \quad j = 0, \dots, n-1.$$

Or component-wise:

$$Z_{j\ell} = \boldsymbol{v}_{j\ell}^* \boldsymbol{T}_j, \quad \ell = 1, \dots m$$

defines the coordinates of an m-dimensional approximation of $\boldsymbol{T}_j$, $m \le r \le d$. $\boldsymbol{Z}_j^{(m)} = (Z_{j1}, \dots, Z_{jm})^T$. I.e., we take the first m complex PCs in each block. Note again, that the entries in the diagonal of each $\boldsymbol{\Lambda}_j$ are in nonincreasing order. To back-transform the PC process of the frequency domain into the time domain, we use inverse Fourier transformation, as we did with the rank reduction:

$$\boldsymbol{Q}_t^{(m)} = \underset{j}{\mathrm{IDFT}}\left[\boldsymbol{Z}_j^{(m)}\right] = \underset{j}{\mathrm{IDFT}}\left[\left[\boldsymbol{V}_j^{(m)}\right]^* \underset{t}{\mathrm{DFT}}(\boldsymbol{X}_t)\right] = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left[\left[\boldsymbol{V}_j^{(m)}\right]^* \boldsymbol{T}_j\right] e^{it\omega_j},$$

where $\boldsymbol{Q}_t^{(m)}$ is the resulting m-dimensional time series. This is similar to the *time series PCA in the frequency domain*[4] by Brillinger [66], and with our numerical considerations in Section 3.5.1 we call our method *dynPCA*. Component-wise the m-dimensional PC process is:

$$\boldsymbol{Q}_t^{(m)} = \frac{1}{\sqrt{n}}\left(\sum_{j=0}^{n-1} Z_{j1}e^{it\omega_j}, \dots, \sum_{j=0}^{n-1} Z_{jm}e^{it\omega_j}\right)^T. \tag{3.9}$$

[4]This was later on referred to as dynamic principal component analysis in the literature.

It can be shown that $\boldsymbol{Q}_t$, $t = 1, \ldots, n$, (again, omitting the $^{(m)}$ notion for conciseness) also have real coordinates ($n = 2k + 1$):

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\boldsymbol{Z}_j e^{it\omega_j} = &= \frac{1}{\sqrt{n}}\left\{\boldsymbol{Z}_0 + \sum_{j=1}^{k}[(\boldsymbol{Z}_j + \overline{\boldsymbol{Z}}_j)\cos(t\omega_j) + i(\boldsymbol{Z}_j - \overline{\boldsymbol{Z}}_j)\sin(t\omega_j)]\right\}\\ &= \frac{1}{\sqrt{n}}\left\{\boldsymbol{Z}_0 + \sum_{j=1}^{k}[2\mathrm{Re}(\boldsymbol{Z}_j)\cos(t\omega_j) + i\cdot 2i\cdot \mathrm{Im}(\boldsymbol{Z}_j)\sin(t\omega_j)]\right\}\\ &= \frac{1}{\sqrt{n}}\left\{\boldsymbol{Z}_0 + 2\sum_{j=1}^{k}[\mathrm{Re}(\boldsymbol{Z}_j)\cos(t\omega_j) - \mathrm{Im}(\boldsymbol{Z}_j)\sin(t\omega_j)]\right\}.\end{aligned}$$

Reconstruction from the lower dimensional space

The previous transformation is a lossy dimension reduction technique too. We can reconstruct from $\boldsymbol{Q}_t$ a time series, close to the original one, similarly to how static principal component analysis (PCA) works. The reconstructed time series is actually the low-rank approximation of Section 3.4.4:

$$\tilde{\boldsymbol{X}}_t := \underset{j}{\mathrm{IDFT}}\left[\left[\boldsymbol{V}_j^{(m)}\right]\underset{t}{\mathrm{DFT}}\left(\boldsymbol{Q}_t^{(m)}\right)\right] = \widehat{\boldsymbol{X}}_t^{(m)}.$$

The larger the average gap in the spectra between the eigenvalue processes λ_k and λ_{k+1}, the smaller the loss is. A schematic workflow of the algorithm is in Figure 3.1.

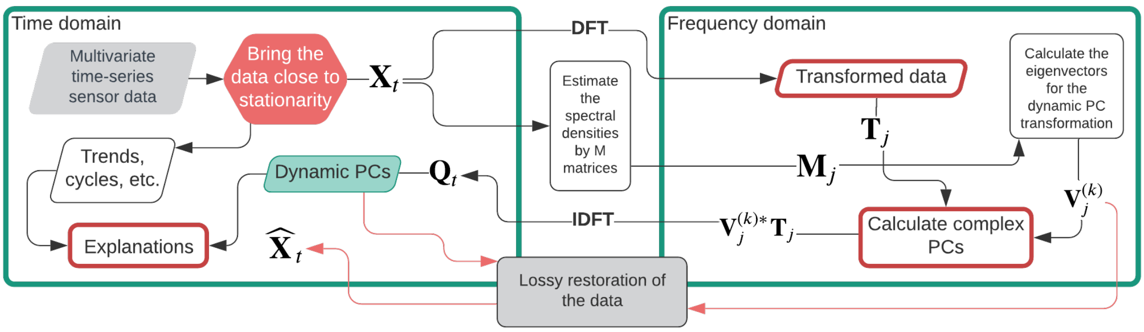


Figure 3.1: A schematic workflow of the dynPCA algorithm

Comparison with other methods

The original PCA method [70] is developed for static data sets. It does not take into account the time-shifted relationships of variables in a time series data. Our method is related to Brillinger's dynamic PCA [66], hence the name. Their technique also exploits the properties of spectral density to select quasi-orthogonal principal component processes. However, there the transformation relies more that than the covariance structure of the time series. The marker dynamic comes from the fact that the principal components here also take into account the time dependence of the data, in the sense that

- the autocovariance of the principal component processes decays and its cross-covariance is filtered out (Figure 3.2a);

- the reconstructed time series preserves the auto/crosscovariance structure of the original time series (Figure 3.2b);

- the reconstruction error(process) becomes more white noise-like, with components more orthogonal to each other (Figure 3.2c).

These properties cannot be achieved by applying static PCA to the time series, as can be seen from the red dots in the aforementioned figures.

This aspect of our algorithm is not yet studied theoretically, but papers on similar techniques (e.g., [71]) show results on this property. The algorithm of Dong and Qin [71] is also related to our method, but in their case the transition to the frequency domain is avoided and the auto/crosscovariance processes are utilized differently. They apply an infinite iteration to solve a prescribed optimization problem, while in our case the parameter estimation and transformation are are carried out in finite steps. Several extensions, improvements of Brillinger's dynamic PCA method exist, e.g., Kwon, Oh, and Lim [72] deals with missing values, Hörmann, Kidziński, and Hallin [73] marries the methodology with functional PCA, and Ku, Storer, and Georgakis [74] describes their dynamic PCA for process monitoring.

3.5.1 Numerical considerations of dynPCA

In theory, for an n-length time series, we have to calculate the $\boldsymbol{V}_j^{(k)}$ matrices (and thus the $\boldsymbol{M}_j$ matrices) at n points (Fourier frequencies). We also have to estimate the $\boldsymbol{C}(h)$ matrices ($\widehat{\boldsymbol{C}}(h)$) until $m = \lfloor n/2 \rfloor$ lags. It quickly becomes intractable as n grows.

In practice, we apply two simplifications [C4]. We estimate the covariances up until a certain point only, as they tend to "decrease"; and we calculate the $\boldsymbol{M}_j$ matrices at fewer Fourier frequencies.

$$\begin{aligned} \boldsymbol{M}_j &= \sum_{h=-k}^{k} \boldsymbol{C}(h) e^{i\omega_j h} & \rightarrow \quad \widehat{\boldsymbol{M}}_{j'} &= \sum_{h=-l}^{l} \widehat{\boldsymbol{C}}(h) e^{i\omega_{j'} h} \\ \text{for } j &= 1, \dots n; & \rightarrow \quad \text{for } j' &= 1, \dots s; \end{aligned}$$

where $s \ll n$ is the number of utilized equidistant Fourier freq.'s, and $l \ll k$ is the lag until the covariances were estimated.

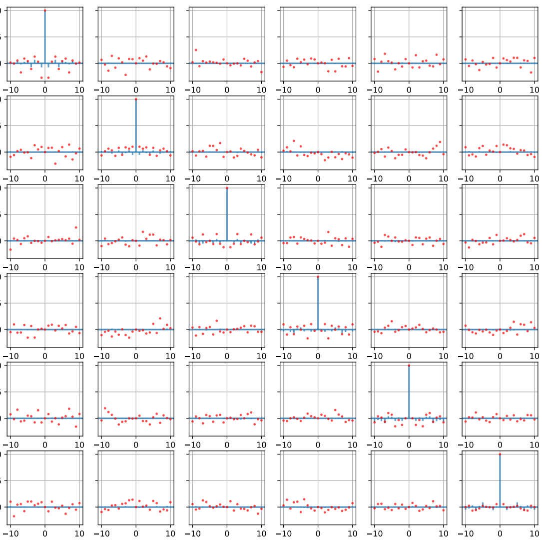
After this, the $\widehat{\boldsymbol{M}}_{j'}$s can be inter/extrapolated to the n Fourier freq.'s, supported by the continuity of the spectral density matrix process.

Dimension reduction as lossy compression

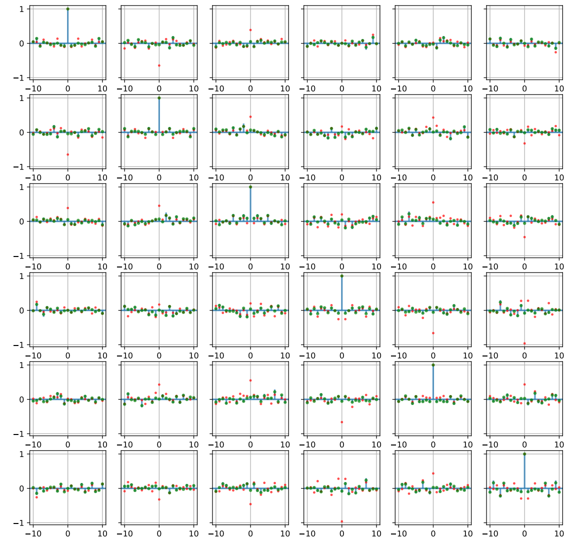
The previous simplifications reduce the original time series $\boldsymbol{X}_t$ ($n \cdot d$ floating-point numbers) to

- $\boldsymbol{Q}_t^{(m)}$: $n \cdot m$ (n floating point vectors of dim. m),

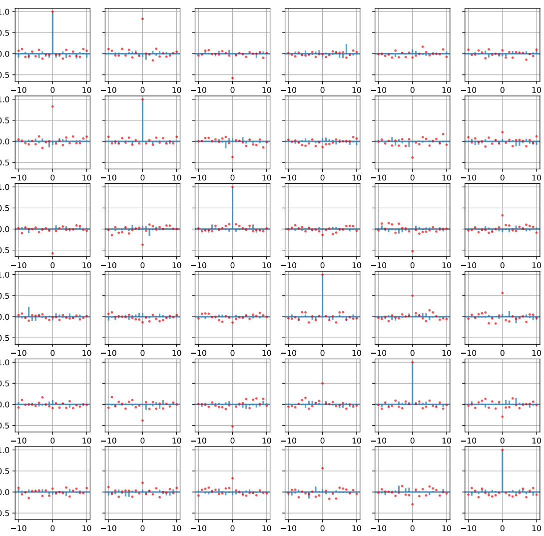
- $\boldsymbol{V}_j^{(m)}$: $s \cdot d \cdot m$ (s complex matrices, composed of m eigenvectors of dim. d).



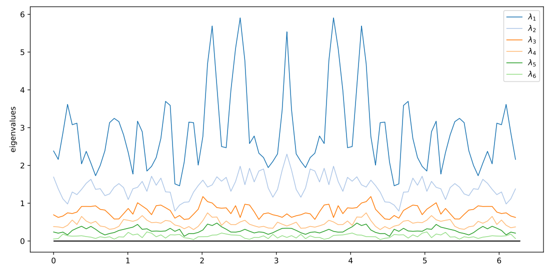
(a) Cross-correlation structure of six dynamic PCs (blue lines) against static PCs (red dots).



(b) Cross-correlation structure of the reconstructed time series from 3 dynamic PCs (blue lines), from 3 static PCs (red dots), against the original time series.



(c) Cross-correlation structure of the error sequence from 3 dynamic PCs (blue lines), from 3 static PCs (red dots).



(d) Eigenvalue processes of the estimated spectral densities (through $\boldsymbol{M}_j$ matrices) for the data at the 100 Fourier frequencies

Figure 3.2: Comparison of the outputs of the dynPCA and the static PCA, calculated for a simulated 6-dimensional VARMA(12,12) process of length 10,000. The dynamic PCs were estimated based on 100 Fourier frequencies. For each subfigure, the graph in row i, column j shows the cross- or autocorrelations (normalized by the 0th lagged (co)variance) of the ith and jth components up until ± 10 lags.

This yields a data compression ratio (DCR) (=uncompr./compr. size) of

$$\mathrm{DCR} = \left(\frac{m}{d} + \frac{2sm}{n} \right)^{-1} . \tag{3.10}$$

It could be further improved because the number of free parameters in $\boldsymbol{V}_j^{(m)}$ is $\frac{m(2d-m-1)}{2}$ instead of $d \cdot m$.

The variance explained by the PCs

Note that in the static PCA (e.g., [75]), the uncorrelated PCs explain the largest possible proportion of the total variance of the (usually correlated) components of a random vector, their variances are in decreasing order and those are the eigenvalues of the covariance matrix. Even more is true: for any m, which is at most the dimension of the random vector, the first m principal components provide the best m-rank approximation of the random vector in mean square sense. This theory also guarantees that –in the case of dynamic PCs– the "larger" the gap between the mth and $(m+1)$th eigenvalues (in nonincreasing order) of the spectral density matrix, the "smaller" the approximation error is. Unlike the static PCA, the complex PCs, i.e. $\boldsymbol{Z}_j$s in the frequency domain make rise to explain proportions of both the cross-sectional and the temporal (briefly, *vertical* and *horizontal*) variance in the complete data, as follows.

Consider the spectral decomposition of the $d \times d$ matrix $\boldsymbol{M}_j = \boldsymbol{V}_j \boldsymbol{\Lambda}_j \boldsymbol{V}_j^*$, and denote by $\lambda_j^{(1)} \geq \cdots \geq \lambda_j^{(d)}$ the eigenvalues of $\boldsymbol{M}_j$, $j = 1, \dots, k$ $(n = 2k+1)$, at frequency ω_j. The eigenvalue $\lambda_j^{(\ell)}$ is the contribution of the ℓth PC to the variance of the increment process at frequency ω_j. Let

$$\lambda_{cum}^{(\ell)} := \sum_{j=1}^{k} \lambda_j^{(\ell)}, \quad \ell = 1, \dots, m$$

be the cumulative contribution of the ℓth PC to the total variance. The relative contribution of the first m PCs to the total variance is

$$\frac{\sum_{\ell=1}^{m} \lambda_{cum}^{(\ell)}}{\sum_{\ell=1}^{d} \lambda_{cum}^{(\ell)}}. \tag{3.11}$$

Then the number m^* of the PCs retained is the smallest index for which the relative contribution (defined above) exceeds a certain percent, or else, when it increases drastically. This is the *vertical* contribution.

As for the *horizontal* contribution, we investigate the contribution of the Fourier frequencies to the total variance. For $\ell = 1, \dots, m$: let $\omega_{j^*}^{(\ell)}$ denote the argmax of the first local maximum of $\lambda_j^{(\ell)}$. Then the corresponding eigenvector of $\boldsymbol{M}_{j^*}$, i.e. the ℓth column of $\boldsymbol{V}_{j^*}$ contains the (complex) factor loadings of the ℓth PC at that prominent frequency. Based on the absolute values (which is calculated from the real and imaginary parts of these coordinates) we can explain the ℓth PC, by telling which variables (with the largest loadings) contribute to it most significantly.

3.6 Applications of dynPCA

Two different applications of the dynPCA algorithm are presented. The first one is applied to a financial dataset (as in [J4]), the second one to electrocardiography (ECG) datasets (as in [C4]). Each application focuses on different aspects of the algorithm's output, showing its versatility.

3.6.1 Application to financial data

The previously detailed low-rank approximation is illustrated on a financial dataset [76] containing stock exchange returns: Istanbul stock exchange national 100 index (ISE), S&P 500 return index (SP), stock market return index of Germany (DAX), UK (FTSE), Japan (NIKKEI), Brazil (BOVESPA), the MSCI European index (EU), and the MSCI emerging markets index (EM); ranging from Jan 5, 2009 to Feb 22, 2011. It is a $d = 8$ dimensional time series dataset that we studied up to length $n = 535$. The days for which the Turkish stock exchange was closed are excluded, and in case of missing data for the other indices, the previous day's value was imputed, see [76].

In Figure 3.3, the three largest eigenvalue processes of the estimated M_j matrices are shown in the frequency domain, the others were close to zero. Based on this, the time series is (approximately) of rank 3, thus we can apply the outlined low-rank approximation with $m = 3$. In Figure 3.4, the rank 3 approximation of the individual variables and the original data is illustrated. There are the calculated RMSE values under each subplot. The 3 leading PCs, back-transformed to the time domain, are to be found in Figure 3.5.

However, according to Equation (3.11), the first two PCs explain 90.2% of the total variance vertically, so $m^* = 2$ will be used in the sequel.

The coordinates of the first two PCs at the leading Fourier frequencies are shown in Table 3.1: the absolute values, further, the real and imaginary parts of the coordinates. As for PC1, the leading frequency is 0.458 that indicates a period $2\pi/0.458 = 13.718$ which is about a biweekly period to which ISE has the largest contribution. As for PC2, the leading frequency is 2.866 that indicates a period $2\pi/2.866 = 2.193$ which is about a 2 days period to which BOVESPA has the largest contribution.

	PC1			PC2		
max_freq_at	0.458			2.866		
	abs. val.	real part	imag. part	abs. val.	real part	imag. part
ISE	0.498	0.498	0.000	0.529	0.529	-0.000
SP	0.302	0.301	0.019	0.393	-0.383	0.086
DAX	0.372	0.370	-0.040	0.255	-0.160	0.199
FTSE	0.297	0.297	-0.013	0.190	-0.124	0.144
NIKKEI	0.389	0.347	-0.177	0.195	0.130	0.146
BOVESPA	0.316	0.316	-0.016	0.601	-0.573	0.180
EU	0.312	0.311	-0.024	0.246	-0.174	0.174
EM	0.292	0.260	-0.134	0.068	0.049	0.047

Table 3.1: The weights inside the first two principal components at the frequency (on scale $[0, \pi]$) of the maximal eigenvalue, as in Figure 3.3.

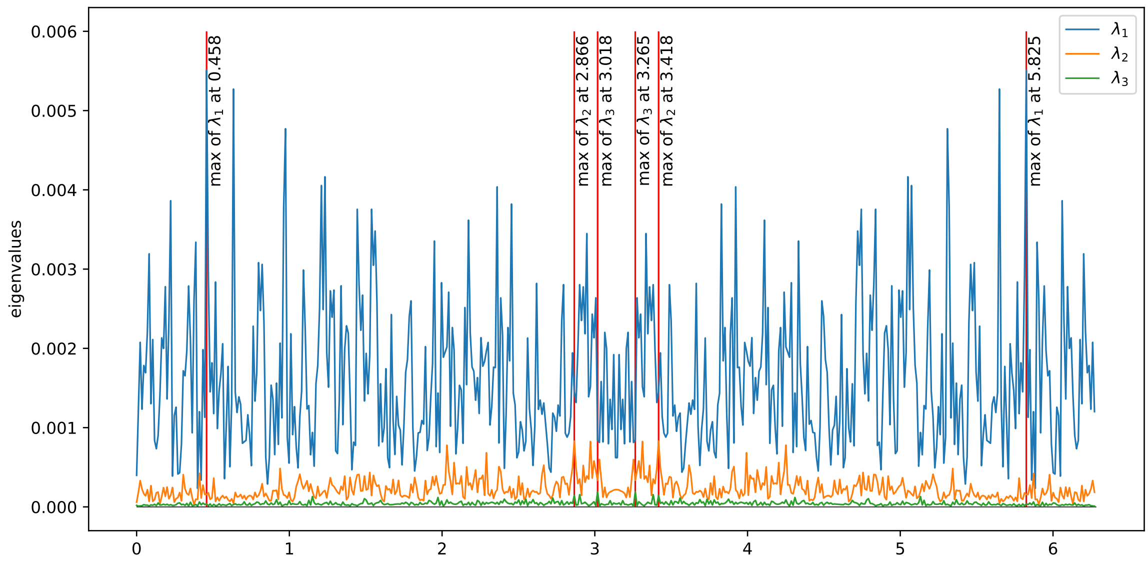


Figure 3.3: The three largest eigenvalue processes of the estimated M_j ($j = 0, \ldots, 534$) matrices over $[0, 2\pi]$ (symmetric with respect to π), ordered decreasingly.

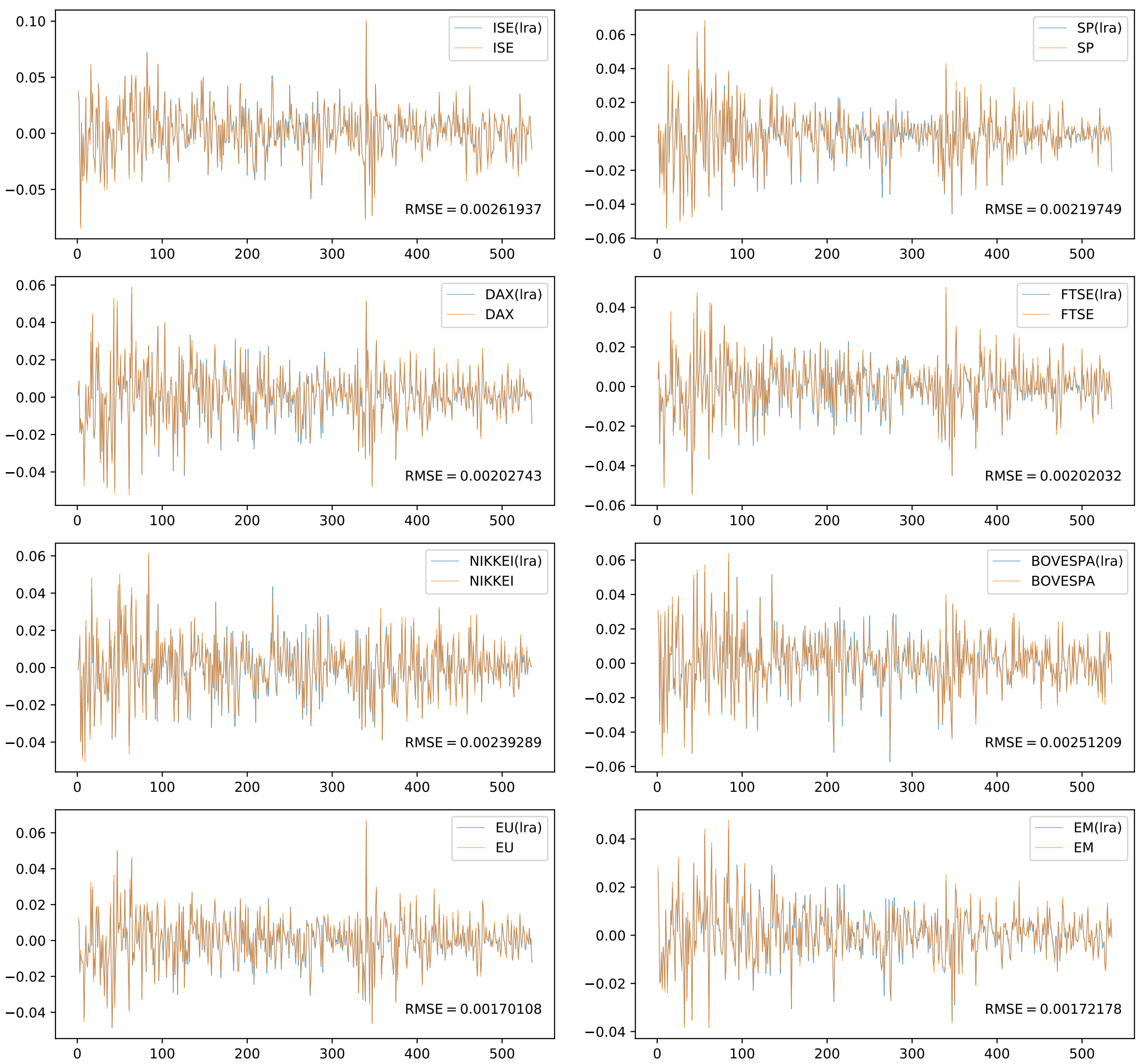


Figure 3.4: Rank 3 approximation of the variables (blue) on the original data (orange).

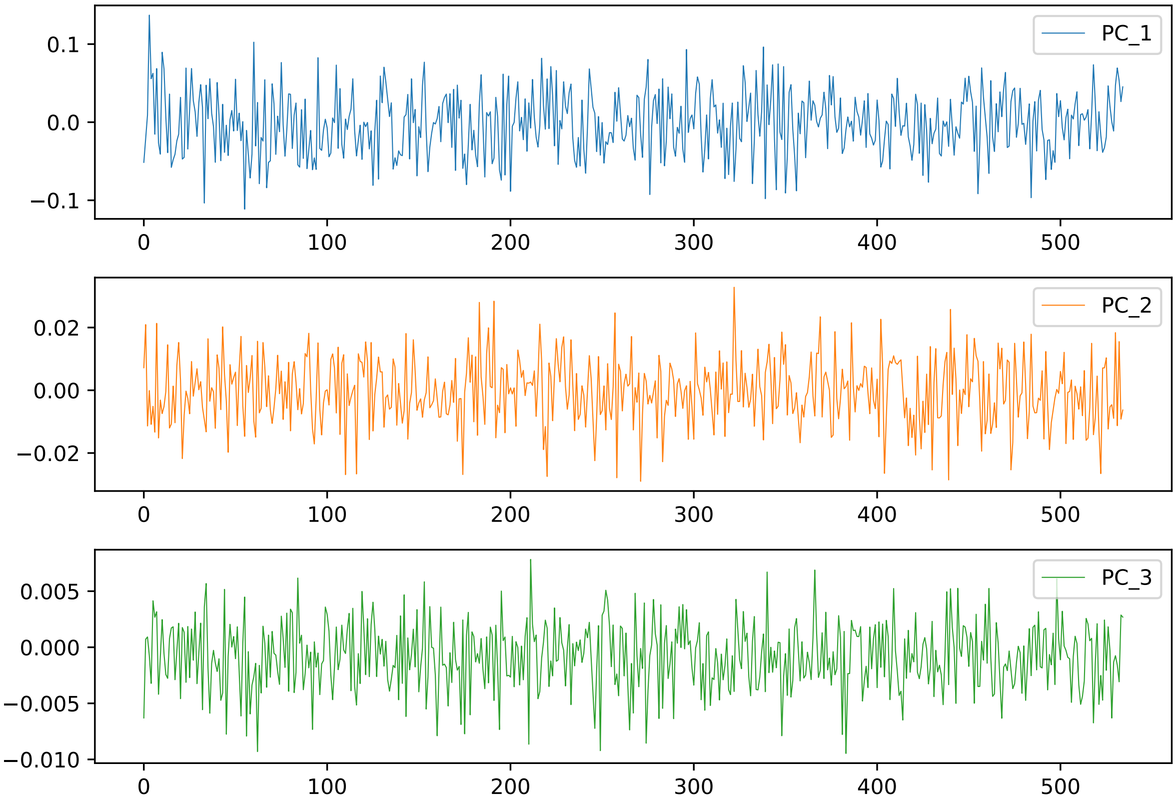


Figure 3.5: The 3 leading PCs of the stock exchange data in the time domain, colored according to the eigenvalues in Figure 3.3.

3.6.2 Application to ECG signals

We investigated two different applications to electrocardiogram (ECG) signals on two different datasets from the PhysioNet database [77], based on our conference paper [C4].

- Classification on the MIT-BIH dataset [78, 79] (2-lead ECG recordings) after dynamic principal component transformation of the signal.
- Lossy compression of the St. Petersburg INCART dataset [80] (12-lead ECG recordings).

The parameters s and l of the dynPCA algorithm (see Section 3.5.1) were both set to $\sqrt{n}$, where n is the length of one ECG recording. A schematic workflow of the two applications is in Figure 3.6.

Classification

The MIT-BIH dataset contains 2-channel (V5 and V2) ambulatory ECG recordings from 48 subjects. Computer-readable reference annotations for each beat are included in the database, that can be used for the classification of the beats.

We applied the following preprocessing steps:

- non beat-related annotation codes were ignored;
- the beats were automatically segmented by the WQRS algorithm (from J to J points) [81];

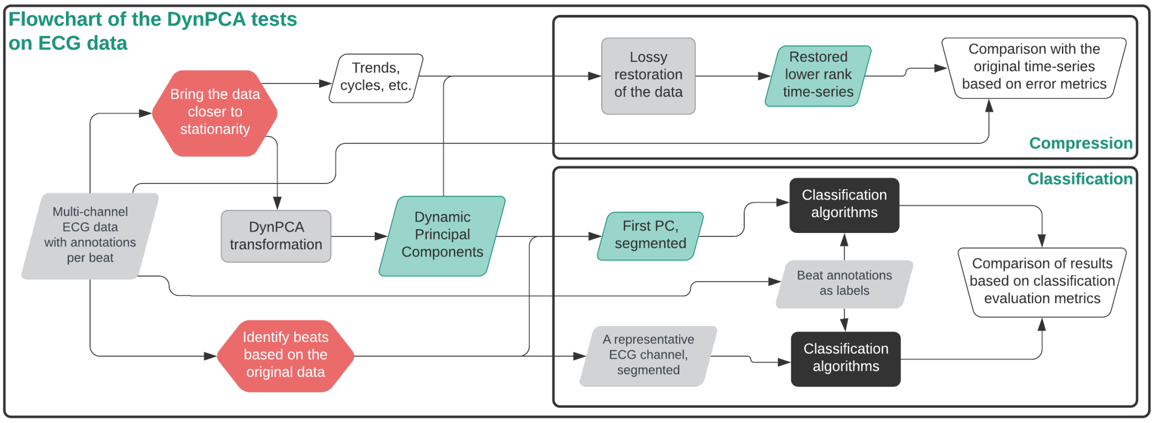


Figure 3.6: A schematic workflow for ECG signals

- the recognized beats were paired with the corresponding annotations;
- the beats were resampled (linearly) to the same length;
- annotation codes were mapped to the 5-scale AAMI classification of beats, see Table 3.2.

After these steps, we have approximately 100,000 equal-length beats with annotations.

AAMI class	**MIT-BIH heart beat types**
Normal beat (N)	Normal beat (N/.) Left bundle branch block beat (L) Right bundle branch block beat (R) Atrial escape beat (e) Nodal (junctional) escape beat (j)
Supraventricular ectopic beat (S)	Atrial premature beat (A) Aberrated atrial premature beat (a) Nodal (junctional) premature beat (J) Supraventricular premature beat (S)
Ventricular ectopic beat (V)	Premature ventricular contraction (V) Ventricular escape beat (E)
Fusion beat (F)	Fusion of ventricular and normal beat (F)
Unknown beat (Q)	Paced beat (/) Fusion of paced and normal beat (f) Unclassified beat (U/Q)

Table 3.2: AAMI classification of beats

The goal is a classification based on the first leading dynamic PC and, for comparison, on the original beat signals (of the 1st channel) in multiple ways. The main features are the equal length beats. Additional features (AF) include the estimated length of the QRS complex and the original beat-length before the resampling. Both of these are additional outputs of the WQRS algorithm. The output label is the previously mentioned AAMI labeling (5 classes: N, S, V, F, Q).

	g.detr. +PC1(512) +AF	g.detr. +PC1(102) +AF	g.detr. +PC1(102)	l.detr. +PC1(512) +AF	l.detr. +PC1(102) +AF	l.detr. +PC1(102)
J48 Tree	96.522 %	96.552 %	94.411 %	96.341 %	96.500 %	94.155 %
K-NN	97.943 %	**98.097** %	97.724 %		**98.086** %	97.800 %
REP Tree	97.428 %	97.471 %	94.729 %	96.918 %	96.498 %	94.109 %
Random Tree		93.683 %	92.340 %		93.861 %	92.251 %
Random Forest		97.503 %	96.597 %		97.584 %	96.597 %
Bagging		97.471 %	96.422 %		97.352 %	95.861 %
AdaBoost		85.209 %	82.767 %		85.209 %	82.771 %

Table 3.3: Accuracy of classification on the MIT-BIH database

Six differently preprocessed inputs were tested: beat length of 512 or 102; AF included or not; beat-by-beat (local) detrending (l.detr.) or global detrending (g.detr.) of the signal.[5] We tested different classification algorithms with their implementation of WEKA [82], after a 2 to 1 train-test split of the data. The accuracy of the algorithms on the test set with different feature sets is included in Table 3.3. For comparison, recent results on MIT-BIH are included in Table 3.5.

Compression

	1PC	2PC	3PC	4PC	5PC	6PC	7PC	8PC	9PC	10PC	11PC	12PC
RMSE (PRD)	0.165 (0.08)	0.0614 (0.03)	0.0315 (0.02)	0.0235 (0.01)	0.0202 (0.01)	0.0188 (0.01)	0.0182 (0.01)	0.0179 (0.01)	0.0179 (0.01)	0.0179 (0.01)	0.0175 (0.01)	0.0131 (0.01)
MAE	0.2317 (0.06)	0.1538 (0.04)	0.1193 (0.03)	0.1073 (0.03)	0.1012 (0.04)	0.0979 (0.04)	0.0961 (0.04)	0.0952 (0.04)	0.0952 (0.04)	0.0951 (0.04)	0.0942 (0.04)	0.0805 (0.04)
R^2 ($\tilde{\text{E}}$xp.Var.)	0.835 (0.08)	0.9386 (0.03)	0.9685 (0.02)	0.9765 (0.01)	0.9798 (0.01)	0.9812 (0.01)	0.9818 (0.01)	0.9821 (0.01)	0.9821 (0.01)	0.9821 (0.01)	0.9825 (0.01)	0.9869 (0.01)
Exp.Var.I	0.6938 (0.17)	0.9002 (0.08)	0.9704 (0.02)	0.9786 (0.01)	0.98 (0.02)	0.9803 (0.02)	0.9804 (0.02)	0.9805 (0.02)	0.9805 (0.02)	0.9805 (0.02)	0.9811 (0.01)	0.9852 (0.01)
Exp.Var.II	0.9134 (0.07)	0.965 (0.02)	0.9752 (0.02)	0.9793 (0.02)	0.9804 (0.02)	0.9808 (0.02)	0.9809 (0.02)	0.981 (0.02)	0.981 (0.02)	0.981 (0.02)	0.9813 (0.02)	0.987 (0.01)
Exp.Var.III	0.8639 (0.1)	0.9603 (0.03)	0.9766 (0.02)	0.9791 (0.02)	0.98 (0.02)	0.9803 (0.02)	0.9805 (0.02)	0.9805 (0.02)	0.9805 (0.02)	0.9806 (0.02)	0.9811 (0.02)	0.9874 (0.01)
Exp.Var.AVR	0.896 (0.07)	0.9533 (0.03)	0.9736 (0.02)	0.9794 (0.02)	0.9807 (0.02)	0.981 (0.02)	0.9811 (0.02)	0.9812 (0.02)	0.9812 (0.02)	0.9812 (0.02)	0.9815 (0.02)	0.9863 (0.01)
Exp.Var.AVL	0.7631 (0.15)	0.9366 (0.04)	0.9746 (0.02)	0.9787 (0.02)	0.9796 (0.02)	0.9799 (0.02)	0.9801 (0.02)	0.9801 (0.02)	0.9802 (0.02)	0.9802 (0.02)	0.9808 (0.02)	0.9864 (0.01)
Exp.Var.AVF	0.8983 (0.09)	0.9661 (0.02)	0.9762 (0.02)	0.9792 (0.02)	0.9802 (0.02)	0.9806 (0.02)	0.9807 (0.02)	0.9807 (0.02)	0.9808 (0.02)	0.9808 (0.02)	0.9812 (0.02)	0.9874 (0.01)
Exp.Var.V1	0.8215 (0.16)	0.9299 (0.05)	0.959 (0.02)	0.9715 (0.02)	0.9797 (0.02)	0.9824 (0.02)	0.9836 (0.02)	0.9839 (0.02)	0.9839 (0.02)	0.9839 (0.02)	0.984 (0.02)	0.9873 (0.01)
Exp.Var.V2	0.7962 (0.16)	0.9209 (0.06)	0.9643 (0.02)	0.9743 (0.01)	0.9784 (0.01)	0.9804 (0.01)	0.9815 (0.01)	0.982 (0.01)	0.982 (0.01)	0.9821 (0.01)	0.9825 (0.01)	0.9872 (0.01)
Exp.Var.V3	0.7821 (0.18)	0.9258 (0.06)	0.9659 (0.02)	0.9746 (0.01)	0.9783 (0.01)	0.9808 (0.01)	0.9821 (0.01)	0.9826 (0.01)	0.9826 (0.01)	0.9826 (0.01)	0.9829 (0.01)	0.9875 (0.01)
Exp.Var.V4	0.8585 (0.15)	0.9335 (0.09)	0.9645 (0.02)	0.9739 (0.02)	0.98 (0.01)	0.9825 (0.01)	0.9839 (0.01)	0.9843 (0.01)	0.9843 (0.01)	0.9843 (0.01)	0.9843 (0.01)	0.9873 (0.01)
Exp.Var.V5	0.8843 (0.12)	0.945 (0.05)	0.965 (0.02)	0.9747 (0.02)	0.9787 (0.02)	0.9813 (0.02)	0.9827 (0.02)	0.9834 (0.02)	0.9834 (0.02)	0.9835 (0.02)	0.9837 (0.02)	0.9862 (0.01)
Exp.Var.V6	0.8488 (0.16)	0.9261 (0.09)	0.9568 (0.04)	0.9752 (0.02)	0.9813 (0.02)	0.9836 (0.01)	0.9846 (0.02)	0.9851 (0.02)	0.9851 (0.02)	0.9851 (0.02)	0.9854 (0.01)	0.9874 (0.01)
DCR	11.59	5.8	3.86	2.9	2.32	1.93	1.66	1.45	1.29	1.16	1.05	0.97

Table 3.4: Metrics of the reconstructed time series' against the originals: mean (and SD) of the metrics over the 75 subjects of the St. Petersburg INCART dataset.

The St. Petersburg INCART dataset contains 12-lead ECG recordings from 75 subjects. Each recording was detrended beat-by-beat (R-R points), and standard scaling was applied as preprocessing. After this, the dynPCA transformation was applied, from which the time series was reconstructed. Different metrics are calculated against the reconstructed lower-rank process: Root Mean Square Error (RMSE), which is equivalent to Percentage Root Mean Square Difference (PRD) due to the scaling; Mean Absolute Error (MAE); R^2 score; the explained variance (Exp.Var.) for each channel; and the DCR, by Equation (3.10).

Table 3.4 contains the mean (and standard deviation) of the metrics over the 75 subjects. Columns denote that the reconstruction was initiated from only the leading dynamic principal component ('1PC'), or the first two ('2PC'), or three ('3PC'), etc.[6]

[5] Some combinations are omitted due to computational constraints.

[6] Reconstruction from all 12 dynamic PCs is sort of a reference, as it would be lossless without the numerical simplifications on a "nice enough" time series.

The numbers by themselves look good but algorithms fine-tuned for ECG data work better, see Table 3.6, note that we used all 12 channels. Our algorithm is a more general tool for multivariate time series data. Note as well that signal compression techniques usually compress the data along the time axis, whereas our dynPCA method compresses the data along the feature axis. Therefore our algorithm may work better as a *complement* to existing compression methods, instead of as an *alternative* one.

Literature	Features	Classifier	No. of classes	Accuracy
Übeyli, 2009 [83]	Lyapunov expon.'s + Wavelet transf. (WT) coef.'s	ANN classifier	4	93.9%
Korürek and Doğan, 2010 [84]	Temporal features	PSO and RBFNN	6	96.3%
de Chazal and Reilly, 2006 [85]	Morphology + Heartbeat intervals	Linear discriminant	5	85.9%
Jiang and Kong, 2007 [86]	Hermite transf. coef.'s + Time intervals	Block based NN	5	96.6%
Das and Ari, 2014 [87]	ST + WT + Temporal features	MLP-NN	5	97.5%
Li et al., 2014 [88]	WPE + RR	RF	5	94.6%
Martis et al., 2013 [89]	Cumulant + PCA	NN	5	94.5%
Acharya et al., 2017 [90]	End-to-end	CNN	5	94.0%
Zubair, Kim, and Yoon, 2016 [91]	End-to-end	CNN	5	92.7%
Atal and Singh, 2020 [92]	Gabor Filter + WT	BaROA-DCNN	5	93.2%
Wu et al., 2021 [93]	WT	CNN	5	97.2%
Alarsan and Younes, 2019 [94]	WT + Temporal features	RF	4	98.0%

Table 3.5: Accuracy of classification by other techniques on the MIT-BIH dataset

Literature	No. of channels	Max DCR	PRD
Martini, Polpetta, and Banelli, 2009 [95]	8	13.5	$\sim$6.60
Tan, Fang, and Wang, 2010 [96]	6	6.5	$<$5.00
Isa, Jatmiko, and Arymurthy, 2014 [97]	8	14.0	4.58

Table 3.6: Compression of the INCART dataset by other techniques

Summary

This dissertation gives an insight into some different topics of multivariate statistical modeling where –with my co-authors– we have had original contributions, over the years. There are open research questions one could follow up on, regarding these topics. First, in Chapter 1, after the general introduction, the ICE (Iterated Conditional Expectation) algorithm was introduced, operating on special graphical models. Its expansion to multivariate time series is straightforward in practice but it can be a theoretical challenge. The jittering technique can be applied here too to handle discrete variables in the model. Further investigation of the algorithm's strong consistency is another possible direction. Chapter 2 was devoted to the RDD (Regression Discontinuity Design). The jittering, in this case, was an adjustment to the model, an alternative way to handle discrete running variables. Its theoretical justification or its application in the fuzzy case are possible future research directions. Our detailed case study was also included, which also had some open questions. Last, in Chapter 3) we turn toward multidimensional time series analysis. The connections of the covariance structure and spectral densities of stationary multidimensional time series were investigated, and a novel dynamic principal component analysis algorithm (dynPCA) was also introduced. The investigation of the described numerical simplifications of dynPCA is a possible future task.

Appendix A

Graphical Models

A.1 Markov properties of DAGs, BNs

Next to the DF property, different Markov properties can be defined over DAG models.

A.1.1 The d-separation and the directed global Markov property

Bayesian networks encode the conditional independence properties of the probability space they represent. We can determine if a conditional independence holds in a BN by using a graph separation criterion called **d-separation** (see [3] for more details), which stands for direction-dependent separation.

In a given DAG G, two vertices are d-separated by the vertex-subset S if there is no **active path** between them. The active paths are those which are not **blocked** by S. Note that this blocking is not identical to separation. The formal definition of active paths and blocking is complex. In the following definitions, by *a path in a DAG* we mean an undirected path after forgetting the edge directions. Sometimes these are called chains. According to [5], in a DAG, a path from a to b is blocked by S if the path contains a vertex c such that:

- either $c \in S$, and arrows do not meet (not a sink) at c;
- or $c \notin S$, and arrows do meet (sink) at c, or c has no descendants in S.

Descendants of a vertex are those vertices that are reachable from the given vertex on a directed path. An alternative to the above definition is given in [13]. In a DAG, a path from a to b is active given S if for every vertex c on the path:

- $c \in S \cup \mathrm{Par}(S)$ if arrows meet (sink) at c;
- $c \in V \setminus S \setminus \{a, b\}$ if it is a transition vertex.

For the terms sink and transition see Figure A.1.

In [98], it is shown that for every DAG G there exists (it can be constructed) a probability distribution $\mathrm{p_V}$ which embodies all the conditional independences displayed in G. Also there is no stronger criterion than d-separation to check the conditional independences of a BN according to [5]. It also gives rise to a different Markov property of BNs.

Definition A.1. *For the distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ the* ***Directed Global Markov*** *property (DGM) holds if the statement,*

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_S \quad \Leftrightarrow \quad A \text{ and } B \text{ are d-separated by } S,$$

is fulfilled for every (A, B, S) disjoint partitioning of the vertices.

The following algorithm gives a nice graphical interpretation for the previous concepts.

Bayes-ball algorithm

To see if $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_C$ in a BN for some disjoint subsets $A, B, C \subset V$, we need to check if every variable in A is d-separated from every vertex in B by the vertices of C. In other words, we have to check $X_a \perp\!\!\!\perp X_b \mid \boldsymbol{X}_C$ for all $a \in A$ and $b \in B$.

Visually, if we condition on the vertices of C, we shade them in (see the blue nodes on Figure A.2). This is the starting step. Since we have to check paths, we have to consider how possible triples of vertices may occur. For this see Figure A.1, the names are borrowed from [99].

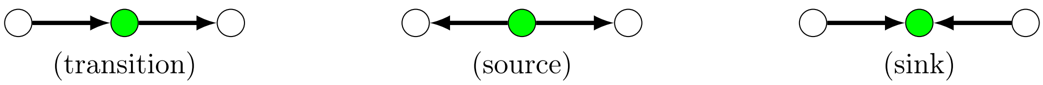


Figure A.1: The three types of triples

The Bayes-ball algorithm works as follows: we shade all nodes in C, place balls at each node in A (or B), let them "bounce around" according to the ten rules of Figure A.2, and then check if any of the balls reach any of the nodes in B (or A). If there is such a ball, we reject the conditional independence, otherwise accept it.

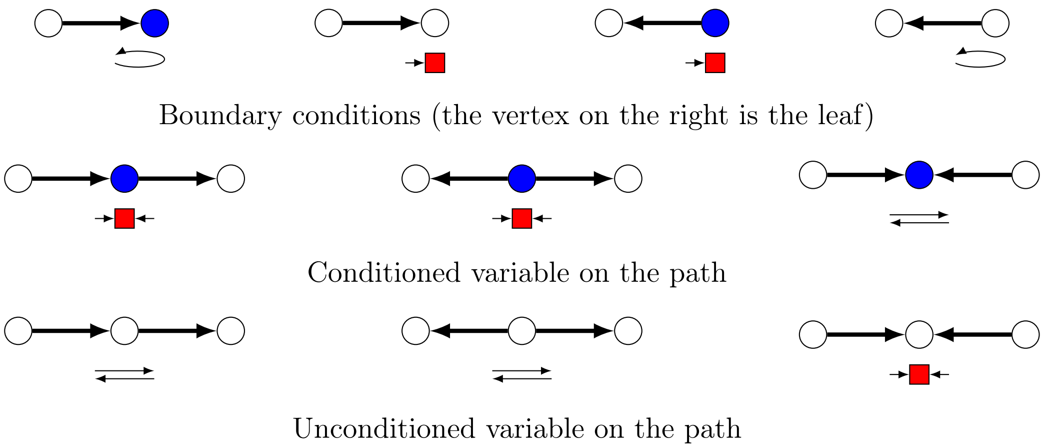


Figure A.2: The 10 rules of Bayes-ball

Note that the balls in this setup can travel opposite to the edge direction. The edge directions only indicate which rule a ball should follow during its journey. In Figure A.2, the boundary conditions show what happens when a ball reaches a leaf: it can bounce back (turn arrow) or get stuck there (red rectangle). The other conditions show what happens when a ball tries to pass through a vertex. The red rectangles mean that a ball gets stuck, can not go further, and the two arrows mean that a ball can pass through.

A.1.2 Other Markov properties of DAGs, BNs

Definition A.2. *For the distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ the **Directed Local Markov** property (DLM) holds if*

$$X_i \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{Ant}(i)} | \boldsymbol{X}_{\mathrm{Par}(i)} \quad \textit{holds} \quad \forall\, i = 1, \ldots, n, \tag{A.1}$$

where $\mathrm{Ant}(i)$ *denotes the set of **anteriors** of i, the set of its non-descendants except its parents:*

$$\mathrm{Ant}(i) = \{i-1, \ldots, 1\} \setminus \mathrm{Par}(i).$$

Nodes that are not connected represent variables that are conditionally independent of each other given the parents of the higher numbered vertex. I.e., X_i (future) and $\boldsymbol{X}_{\mathrm{Ant}(i)}$ (past) are independent conditioned on $X_{\mathrm{Par}(i)}$ (present). This generalizes the fundamental property of discrete-time Markov chains, where G is just a directed path.

Next to DF, DLM, and DGM, an additional weaker property can be defined.

Definition A.3. *For the distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ the **Directed Pairwise Markov** property (DGM) holds if*

$$X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{\mathrm{Ant}(i) \cup \mathrm{Par}(i) \setminus \{j\}}$$

holds for all $i = 1, \ldots, n$ and $j \in \mathrm{Ant}(i)$ pairs.

DLM implies DPM but the converse is not true in general. However, it can be shown that if $\mathrm{p_V}$ exists (as it is a density with respect to a product measure) then DF, DLM, and DGM are equivalent. Therefore these are used interchangeably.

A.2 Markov properties of MRFs

Definition A.4. *For a joint distribution of $\boldsymbol{X}$, the **Global Markov** property (GM) with respect to an undirected graph G is fulfilled if the statement,*

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_S \quad \Leftrightarrow \quad A \textit{ and } B \textit{ are separated by } S,$$

holds for every (A, B, S) disjoint partitioning of the vertices.

Definition A.5. *For a joint distribution of $\boldsymbol{X}$, the **Local Markov** property (LM) with respect to an undirected graph G holds if*

$$X_i \perp\!\!\!\perp X_{V \setminus \mathrm{Cl}(i)} \mid X_{\mathrm{Adj}(i)}, \qquad \forall\, i \in V, \tag{A.2}$$

where $\mathrm{Cl}(i) := \mathrm{Adj}(i) \cup \{i\}$.

Definition A.6. *For a joint distribution of $\boldsymbol{X}$, the **Pairwise Markov** property (PM) with respect to an undirected graph G holds if*

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}, \qquad \forall\, i \neq j\ (\in V). \tag{A.3}$$

It is easy to see that: $(GM) \Rightarrow (LM) \Rightarrow (PM)$. Therefore, the Global Markov property is stronger than the Local Markov property, which in turn is stronger than the Pairwise one.

These properties are hard to check by definition, but for positive probability distributions we have a sufficient and necessary condition given by Hammersly and Clifford. For this we need an extra property.

Definition A.7. *The probability distribution* $\mathrm{p_V}$ *of* $\boldsymbol{X}$ *is a* ***Gibbs distribution with respect to an undirected graph*** G *if* $\mathrm{p_V}$ *factorizes over the cliques (maximal complete subgraphs) of* G*, more precisely:*

$$\mathrm{p_{X_1 \ldots X_n}} = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C, \quad \textit{where:} \tag{A.4}$$

- $\mathcal{C} = \{C \subset V \ : \ C \textit{ is a clique of } G\}$,
- ψ_C *is a positive (real-valued) function for all* $C \in \mathcal{C}$ *over the possible states of the variables in* C,
- $Z = \sum_{X_1} \sum_{X_2} \cdots \sum_{X_n} \prod_{C \in \mathcal{C}} \psi_C$ *is a normalizing constant, which is a marginalization of the above product over all possible states (configurations) of the* n *variables.*

In some literature, e.g. [5], this property is called **undirected factorization** property. The connection with DF (Definition 1.5) can be drawn through the process of moralization [5, 1] of the directed graph.

In graph theory, cliques are often the complete subsets of a graph. However, in statistics (and that is the convention we will follow) the cliques are usually the maximal complete subsets. We will use the above notation $\mathcal{C}$ for the set of cliques over a graph from now on.

The ψ_C functions are sometimes referred to as factor potentials (e.g., if the domain is not a clique) or clique potentials. Sometimes they are defined as they already include Z. Note that another conflicting terminology is in use: the word potential is often applied to the logarithm of ψ_C. It happens because in statistical mechanics, $\log(\psi_C)$ has a direct interpretation as the potential energy of a configuration over C. The normalizing constant is sometimes called the partition function.

It is also worth noting that these functions can be multiplied together or split up in different ways. Therefore, without loss of generality, the above definition can be restated to include every complete subset, not just the maximal ones.

Theorem A.1 (Hammersly–Clifford theorem [100])**.** *A strictly positive marginally continuous distribution with respect to an undirected graph* G *satisfies the PM property if and only if it is a Gibbs distribution with respect to this* G.

The relationship between Markov random fields and Gibbs distributions was already discussed by R. Dobrushin and F. Spitzer in the context of statistical mechanics. The theorem is named after J. Hammersley and P. Clifford, who proved the equivalence in 1971 in an unpublished paper [100]. Their proof is rather complex, and simpler proofs were given in 1973-4, independently by G. Grimmett, C. J. Preston, S. Sherman, and J. Besag.

At this point we have that under the condition of Theorem A.1:

$$(GM) \Rightarrow (LM) \Rightarrow (PM) \Leftrightarrow \text{Gibbs w.r.t } G.$$

However, it can be shown that any Gibbs distribution with respect to a graph G fulfills the GM property. Therefore, under the condition of Theorem A.1 the four conditions become equivalent:

$$(GM) \Leftrightarrow (LM) \Leftrightarrow (PM) \Leftrightarrow \text{Gibbs w.r.t } G.$$

In general, the potentials do not have a probabilistic interpretation, but we can say something like "values (configurations) with higher potential are more probable".

If we do not have a graph representation yet, but we know that our joint distribution factorizes over some subsets of the variables:

$$\mathrm{p_V} := \mathrm{p_{X_1 \dots X_n}} = \frac{1}{Z} \prod_{A \subset \{X_1, \dots X_n\}} \psi_A, \tag{A.5}$$

then we can easily create a graph representation G by assigning each variable of the probability space to a vertex (as usual) and to every available ψ_A we place a clique over the vertices assigned to the variables in A. By this method $\mathrm{p_V}$ will be trivially a Gibbs distribution over the resulting graph G.

A.3 Types of chain graph models

For a far more detailed comparison of the different types, see [101, 102].

type I, LWF-CG

- They were introduced by Lauritzen and Wermuth [103] in 1989, and by Frydenberg [104] in 1990. This is the first CG interpretation chronologically.
- They have chain components in which the nodes are connected to each other by (regular) undirected edges on concentration base.
- Each parent variable (from which directed edges point to a variable of another component) affects all variables in the component of the child variable, i.e., the information travels through the children.

type II, AMP-CG

- They were introduced by Andersson, Madigan, and Perlman [105] in 1996.
- This type has a different separation criterion (than LWF) that more closely resembles the one of DAGs, as it preserves some component-wise recursive characteristics of them.
- They also have chain components in which the nodes are connected to each other by undirected edges (on concentration base), but a node depends only on its parents and not on the parents of the whole component.
- The internal structure of a chain component cannot be factorized further into smaller distributions.

type III

- They are not actively used but they exist theoretically.
- They are described by Drton as “following a Markov property that is dual to the type II property” [11]

type IV, MVR-CG

- They were introduced by Cox and Wermuth [7] in 1993, with more details in [8].
- They are equivalent to the acyclic directed mixed graphs without semi-directed cycles presented by Richardson [10].
- They combine DAGs and bidirected (covariance) graphs, as they have chain components in which the nodes are connected to each other by dashed (or bidirected) edges on covariance base.
- It is a superclass of DAGs and BGs instead of DAGs and UGs, as it is the case for LWF and AMP CGs.
- Each conditional probability distribution $\mathrm{p}_{\mathrm{C_k}\,|\,\mathrm{C_{Par\,(k)}}}$ can be factorized into smaller parameters, using Möbius parameters, as detailed by [11].
- As in AMP CGs, each child node depends only on its parents and not on the parents of the whole component.
- A typical use case is when there are hidden variables, i.e., unobserved variables that are the parents of at least two observed variables in the system.

It is noteworthy that for LWF, AMP, and MVR CGs, no interpretation contains another [102], and no interpretation is generally better than any other.

Appendix B

Nonparametric Regression

B.1 Background material on kernel regression

In this section we work with the usual (Y: target, $\boldsymbol{X}$: regressors) setup to keep the notation simple, with an n-element training sample $y^{(i)}, \boldsymbol{x}^{(i)}$, $i = 1, \dots, n$. Other notations are analogous with Chapter 1. The number or regressors will be denoted by r, out of which the number of continuous ones is p, and $\boldsymbol{x} = (\boldsymbol{x}_c, \boldsymbol{x}_d)$ decomposes into continuous and discrete coordinates.

The supervised learning procedure of the kernel regression (see [19, 20]) is historically preceded by the unsupervised learning procedure of kernel density estimation. It is a local averaging technique of Equation (1.5), which reads here:

$$S_Y^{(n)}(\boldsymbol{x}) := \sum_{i=1}^{n} y^{(i)} W^{(n,i)}(\boldsymbol{x}). \tag{B.1}$$

The mean-square consistency of such methods is usually formulated in terms of the following theorem, from which specific conditions can be derived for the different types of local averaging techniques, see e.g., [15].

Theorem B.1 (C. J. Stone [15, 106]). *Assume that the following conditions are satisfied for any distribution of $\boldsymbol{X}$:*

1. *$\exists\ c$ such that for every non-negative function $g(\cdot)$ with $\mathbb{E}[g(\boldsymbol{X})] < \infty$ and for any n:*

$$\mathbb{E}\left[\sum_{i=1}^{n} |W^{(n,i)}(\boldsymbol{X})| \cdot g(\boldsymbol{X}_i)\right] \leq c \cdot \mathbb{E}[g(\boldsymbol{X})],$$

2.

$$\exists\ D \geq 1 \quad \textit{such that} \quad \mathbb{P}\left(\sum_{i=1}^{n} |W^{(n,i)}(\boldsymbol{X})| \leq D\right) = 1,$$

3.

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} |W^{(n,i)}(\boldsymbol{X})| \cdot \mathbb{I}_{\{\|\boldsymbol{X}_i - \boldsymbol{X}\| > a\}}\right] = 0, \quad \forall\ a > 0,$$

4.

$$\sum_{i=1}^{n} |W^{(n,i)}(\boldsymbol{X})| \to 1 \quad \textit{in probability, as } n \to \infty,$$

5. *and asymptotically all weights become small:*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} |W^{(n,i)}(\boldsymbol{X})|^2\right] = 0.$$

Then for all distribution of $(\boldsymbol{X}, Y)$ *with* $\mathbb{E}\left[Y^2\right] < \infty$ *it holds that*

$$\lim_{n\to\infty} \mathbb{E}\left[\int_{\mathbb{R}^d} |S_Y^{(n)}(\boldsymbol{x}) - P_Y(\boldsymbol{x})|^2 \mu(\mathrm{d}\boldsymbol{x})\right] = 0.$$

The Nadaraya–Watson (NW) kernel estimate of the regression around an appropriate $\boldsymbol{x}$ is

$$S_Y^{(n)}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} K_{\boldsymbol{H}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) y^{(i)}}{\sum_{j=1}^{n} K_{\boldsymbol{H}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(j)}\right)}, \tag{B.2}$$

where $\boldsymbol{H}_n = \{h_{i,j}^{(n)}\}_{i=1,j=1}^{r,r}$ is the bandwidth matrix, which has to be positive definite, and

$$K_{\boldsymbol{H}}\left(\boldsymbol{x}, \tilde{\boldsymbol{x}}\right) := |\boldsymbol{H}|^{-1} \kappa\left(\boldsymbol{H}^{-1} |\boldsymbol{x} - \tilde{\boldsymbol{x}}|\right),$$

where κ is a kernel function and $|\cdot|$ refers to coordinatewise absolute value. It is a local averaging estimate of Equation (B.1) with weights

$$W^{(n,i)}(\boldsymbol{x}) = \frac{K_{\boldsymbol{H}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right)}{\sum_{j=1}^{n} K_{\boldsymbol{H}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(j)}\right)}.$$

Choosing the right bandwidth matrix $\boldsymbol{H}$ is as crucial in the multivariate setup as the selection of the only bandwidth parameter h in the univariate case (only one regressor). However, it is a computationally demanding matrix optimization task, therefore practitioners often choose diagonal bandwidth matrices $\boldsymbol{H} = \mathrm{diag}(\boldsymbol{h})$, which reduces the complexity of the task significantly. Considering a full bandwidth matrix gives more flexibility, but notably increases the number of bandwidth parameters that need to be chosen. Off-diagonal entries of $\boldsymbol{H}$ correspond to some dependency between the regressors, but it is an understudied topic in the literature regarding regression problems. Note that in our graphical modeling framework we have to perform several regression estimations, the number of which is $\#\{\text{dimensions}\} - \#\{\text{context variables}\}$, one for each intermediate node; this further increases the computation time. It would make sense to use full bandwidth matrices in our case because we have some knowledge (through edges between parents of a given node) about the dependencies between the regressors in every step. However, we wanted to decrease the computation time and make the process more automatic, therefore we also used diagonal bandwidth matrices by default.

Mean-square consistency of the NW estimator

For the mean-square consistency of the NW estimator, there are several results. In the case of a single deterministic bandwidth parameter h_n, when $\boldsymbol{H} = diag(h_n, \ldots, h_n)$, and is deterministic in the sense that only depends on the sample size, the following result holds, see [15]. The NW estimator of this case is mean-square consistent if:

- K_h is a bounded non-negative function on $\mathbb{R}^r$ with compact support; further, $K_h \geq \beta \mathbb{I}_B$ with some $\beta > 0$ and some closed sphere B centered at the origin with positive radius ($\mathbb{I}$ is the indicator function);
- $h_n \to 0$ and $nh_n^r \to \infty$ as $n \to \infty$.

This type of kernel is called *boxed* kernel in [15]. In particular, a bounded kernel with compact support, such that it is bounded away from zero at the origin, satisfies this condition. For example, the Epanechnikov kernel is such, but the Gaussian is not. However, we can take truncated Gaussian kernels, the support of which contains the sample entries.

This result can be extended (see [15]) to product kernels (multivariate kernels defined as the product of univariate kernels) with a vector of bandwidth parameters $\boldsymbol{h}_n = \operatorname{diag}(h_{n,1}, \ldots, h_{n,r})$. In this way, the multivariate kernel is the product of univariate kernels:

$$K_{\boldsymbol{h}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) = \prod_{j=1}^{r} K_{h_{n,j}}\left(x_j, x_j^{(i)}\right). \tag{B.3}$$

The NW estimator constructed from a product kernel is mean-square consistent if:

- each univariate kernel in the product is *boxed* kernel in the previous sense;
- $h_{n,j} \to 0$ for $j = 1, \ldots, r$ and $n \prod_{j=1}^{r} h_{n,j} \to \infty$.

From the second condition $nh_{n,j} = \infty$ follows for $j = 1, \ldots, r$, but not vice versa.

These consistency results are universal in the sense that they do not depend on the underlying distributions. Practitioners often use the same kernel functions for discrete and continuous regressors, which is a questionable modeling choice. There are kernel functions designed especially for discrete regressors (see the Section 1.4.1), but their consistency in mean-square sense has not been extensively studied so far.

Product kernel estimator of Racine and Li

Product kernel estimation was further refined for mixed regressors by Racine and Li [16]. Here, we list their conditions (with additional remarks) they used to prove their consistency results. Again, the number of continuous regressors will be denoted by p. For the sake of their proof, the same h_c, h_d bandwidths were chosen for all continuous and for all discrete regressors, respectively.

Assumption B.1 ([16])**.** *Conditions on the regression estimation.*

(i) (h_d, h_c) *lie in a shrinking set* $\mathcal{H}_d^{(n)} \times \mathcal{H}_c^{(n)}$, *where*

$$\mathcal{H}_d^{(n)} = \left[0, \min\left(1, \frac{C_0}{\log n}\right)\right] \quad \textit{and} \quad \mathcal{H}_c^{(n)} = \left[\frac{n^{\delta - \frac{1}{p}}}{C_1}, \frac{C_1}{n^{\delta}}\right],$$

for some $C_0, C_1, \delta > 0$*;*

- *Note that the first part of the condition is* $h_d \to 0$ *and for small* δ *the second part of the condition is virtually identical to* $h_c \to 0$ *and* $nh_c^p \to \infty$.

(ii) *the univariate kernel function* $K_c(\cdot)$ *used for the continuous variables is non-negative, symmetric, bounded,* $\int K_c(v)v^4 \, \mathrm{d}v < \infty$*, m-times differentiable and* $\int |K_c^{(s)}(v)v^s| \, \mathrm{d}v < \infty$ *for all* $s = 1, \ldots, m$*, where* $m > \max\{2 + 4/p, 1 + p/2\}$*;*

- *Note that the Gaussian kernel satisfies these conditions; further, this condition can be replaced with a compactly supported kernel function that is Hölder continuous, e.g., the Epanechnikov kernel.*

(iii) *the weighting function* $M(\cdot)$ *(optionally used in the cross-validation, see Equation* (1.16)*) is bounded and supported on a compact set with nonempty interior for all realization of the discrete variables;*

(iv) *the joint density of the variables (*$f(\cdot)$ *from now on) is bounded from below on the support of* $M(\cdot)$*.*

- *Note that conditions* (iii) *and* (iv) *are only needed to have a uniform convergence rate,* $M(\cdot)$ *can be omitted from the cross-validation.*

Assumption B.2 ([16])**.** *Conditions on the underlying distribution.*

(i) $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ *are i.i.d. and* $Y - P_Y(\boldsymbol{X})$ *has finite fourth moment;*

(ii) $\sigma(\cdot, \boldsymbol{x}_d)$, $P_Y(\cdot, \boldsymbol{x}_d)$, $f(\cdot, \boldsymbol{x}_d)$ *are all twice differentiable, and their partial derivatives (up to the second order) are all bounded by some functions with finite fourth moment for all possible* $\boldsymbol{x}_d$*, where*

$$\sigma^2(\boldsymbol{x}) := \mathbb{E}\left[(Y - P_Y(\boldsymbol{X}))^2 \mid \boldsymbol{X} = \boldsymbol{x}\right],$$

and $f(\cdot)$ *is the joint density of the variables;*

(iii) *lengthy technical condition, only needed to rule out the case when the regression function is independent of the continuous regressors.*

Theorem B.2 ([16])**.** *Under the Assumptions B.1 and B.2, the product kernel regression estimate (with the leave-one-out cross-validated bandwidth choice of Racine and Li, see Equation* (1.16)*), is asymptotically normal:*

$$\sqrt{n\hat{h_c}^p}\left(S_Y(\boldsymbol{x}) - P_Y(\boldsymbol{x}) - \hat{B}(\hat{h_c}, \hat{h_d})\right) \to \mathcal{N}\left(0, \hat{\Omega}(\boldsymbol{x})\right),$$

where

$$\begin{aligned}
\hat{\Omega}(\boldsymbol{x}) &= \frac{\hat{\sigma}^2(\boldsymbol{x})}{\hat{f}(\boldsymbol{x})} \int \left(\prod_{i=1}^{p} K_c(v_i) \right)^2 \, \mathrm{d}\boldsymbol{v}, \\
\hat{B}(\hat{h_c}, \hat{h_d}) &= \hat{h_c}^2 \Phi_1(\boldsymbol{x}) \int K_c(v) v^2 \, \mathrm{d}v + \hat{h_d} \Phi_2(\boldsymbol{x}), \\
\Phi_1(\boldsymbol{x}) &= \frac{\nabla \hat{f}(\boldsymbol{x})^T \nabla S_Y(\boldsymbol{x})}{\hat{f}(\boldsymbol{x})} + \frac{\mathrm{tr}(\nabla^2 S_Y(\boldsymbol{x}))}{2}, \\
\Phi_2(\boldsymbol{x}) &= \sum_{\tilde{\boldsymbol{x}}_d : d(\boldsymbol{x}_d, \tilde{\boldsymbol{x}}_d) = 1} \left[S_Y(\boldsymbol{x}_c, \tilde{\boldsymbol{x}}_d) - S_Y(\boldsymbol{x}_c, \boldsymbol{x}_d) \right] \frac{\hat{f}(\boldsymbol{x}_c, \tilde{\boldsymbol{x}}_d)}{\hat{f}(\boldsymbol{x}_c, \boldsymbol{x}_d)}, \\
d(\boldsymbol{x}_d, \tilde{\boldsymbol{x}}_d) &= \sum_{i=1}^{k} \mathbb{I}_{\{x_{d,i} \neq \tilde{x}_{d,i}\}}, \\
\hat{\sigma}^2(\boldsymbol{x}) &= \frac{1}{n \hat{f}(\boldsymbol{x})} \sum_{i=1}^{n} (Y_i - S_Y(\boldsymbol{X}_i))^2 \, K_{\boldsymbol{h}_n} \left(\boldsymbol{x}, \boldsymbol{x}^{(i)} \right)
\end{aligned}$$

and

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\boldsymbol{h}_n} \left(\boldsymbol{x}, \boldsymbol{x}^{(i)} \right).$$

Using the same assumptions with some additional ones we prove the mean square consistency of the product kernel estimator of Racine and Li.

Lemma B.1. *Under Assumptions B.1 and B.2, further assuming that*

$$\limsup_{n \to \infty} \mathbb{E} \left| \sqrt{n \hat{h_c}^p} \left(S_Y(\boldsymbol{x}) - P_Y(\boldsymbol{x}) - \hat{B}(\hat{h_c}, \hat{h_d}) \right) \right|^3 < \infty, \tag{B.4}$$

$$\mathbb{E} \left[\Phi_1(\boldsymbol{X})^2 \right] < \infty \quad \text{and} \quad \mathbb{E} \left[\Phi_2(\boldsymbol{X})^2 \right] < \infty, \tag{B.5}$$

the estimate is mean-square consistent.

Proof. We need that

$$\mathbb{E} \left[(S_Y(\boldsymbol{X}) - P_Y(\boldsymbol{X}))^2 \right] = \tilde{\mathbb{E}} \left[\int (S_Y(\boldsymbol{x}) - P_Y(\boldsymbol{x}))^2 \mu(\mathrm{d}\boldsymbol{x}) \right] \to 0, \text{ as } n \to \infty,$$

where $\tilde{\mathbb{E}}$ refers to the expectation over the distribution of the sample. We use the $\hat{B}()$ term of Theorem B.2:

$$\begin{aligned}
&\mathbb{E} \left[(S_Y(\boldsymbol{X}) - P_Y(\boldsymbol{X}))^2 \right] \\
&= \mathbb{E} \left[(S_Y(\boldsymbol{X}) - P_Y(\boldsymbol{X}) - \hat{B}(\hat{h_c}, \hat{h_d}) + \hat{B}(\hat{h_c}, \hat{h_d}))^2 \right] \\
&\leq 2 \, \mathbb{E} \left[(S_Y(\boldsymbol{X}) - P_Y(\boldsymbol{X}) - \hat{B}(\hat{h_c}, \hat{h_d}))^2 \right] + 2 \, \mathbb{E} \left[(\hat{B}(\hat{h_c}, \hat{h_d}))^2 \right].
\end{aligned}$$

For the first term:

$$S_Y(\boldsymbol{x}) - P_Y(\boldsymbol{x}) - \hat{B}\left(\hat{h_c}, \hat{h_d}\right) \asymp \mathcal{N}\left(0, \frac{\hat{\Omega}(\boldsymbol{x})}{\sqrt{n\hat{h_c}^{\,p}}}\right),$$

asymptotically as $n \to \infty$, by Theorem B.2. Further,

$$\lim_{n\to\infty} \int \mathbb{E}\left[(S_Y(\boldsymbol{x}) - P_Y(\boldsymbol{x}) - \hat{B}(\hat{h_c}, \hat{h_d}))^2\right] \mu(\mathrm{d}\boldsymbol{x}) = \lim_{n\to\infty} \int \frac{\hat{\Omega}(\boldsymbol{x})}{\sqrt{n\hat{h_c}^{\,p}}} \mu(\mathrm{d}\boldsymbol{x}) = 0,$$

because Equation (B.4) ensures the convergence of moments (see Example 2.21 of [107]), and $\hat{\Omega}(\boldsymbol{x})$ is essentially bounded:

$$\hat{\Omega}(\boldsymbol{x}) = \frac{\hat{\sigma}^2(\boldsymbol{x})}{\hat{f}(\boldsymbol{x})} C_K \leq \frac{D^2}{\hat{f}(\boldsymbol{x})} C_K \leq \frac{D^2}{\varepsilon} C_K,$$

where $D = Y_{max} - Y_{min}$ is the range of Y in the sample, $\varepsilon = \inf\{\hat{f}(\boldsymbol{x}) | \hat{f}(\boldsymbol{x}) > 0\}$ and C_K is the finite integral in the definition of $\hat{\Omega}(\boldsymbol{x})$. (Note that $\hat{f}(\boldsymbol{x}) = 0$ only for those $\boldsymbol{x}$s that are "too far away" from the sample, resulting in undefined $S_Y(\boldsymbol{x})$ and $\hat{\Omega}(\boldsymbol{x})$. The probability of this event goes to zero as $n \to \infty$, so the bound will hold almost surely.) Therefore, we can use the dominated convergence theorem and that $nh_c^p \to \infty$ by Assumption B.1(i) to get the zero limit. For the second term:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\hat{B}(\hat{h_c}, \hat{h_d})^2\right] &= \mathbb{E}\left[\lim_{n\to\infty} \hat{B}(\hat{h_c}, \hat{h_d})^2\right] \\
&= \mathbb{E}\left[\lim_{n\to\infty} \left(\hat{h_c}^2 \Phi_1(\boldsymbol{X}) \int K_c(v) v^2 \ \mathrm{d}v + \hat{h_d} \Phi_2(\boldsymbol{X})\right)^2\right] = 0,
\end{aligned}$$

where $\hat{h_c} \to 0$, $\hat{h_d} \to 0$ and $\int K_c(v) v^2 \ \mathrm{d}v < \infty$, by Assumption B.1. The limit and the expectation can be interchanged by the dominated convergence theorem:

$$\left|\hat{h_c}^2 \Phi_1(\boldsymbol{X}) \int K_c(v) v^2 \ \mathrm{d}v + \hat{h_d} \Phi_2(\boldsymbol{X})\right| < C_1 \Phi_1(\boldsymbol{X}) \int K_c(v) v^2 \ \mathrm{d}v + C_0 \Phi_2(\boldsymbol{X}),$$

where C_0, C_1 are from Assumption B.1, and by Equation (B.5) the second moment of the right hand side is finite. □

B.2 Background material on other smoothers

Local linear regression

The NW estimate is also called local constant estimate, since it is actually the minimizer of

$$S_Y^{(n)}(\boldsymbol{x}) = \operatorname*{argmin}_{\alpha(\boldsymbol{x})} \sum_{i=1}^{n} K_{\boldsymbol{H}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right) \left[y^{(i)} - \alpha(\boldsymbol{x})\right]^2 = \hat{\alpha}(\boldsymbol{x}),$$

where $\hat{\alpha}(\boldsymbol{x})$ is the optimal constant estimate around $\boldsymbol{x}$. Likewise, local linear and polynomial estimates can be constructed by substituting a polynomial (with coefficients depending on $\boldsymbol{x}$) for $\alpha(\boldsymbol{x})$. In the linear case:

$$\left\{\hat{\alpha}(\boldsymbol{x}), \hat{\beta}(\boldsymbol{x})\right\} = \underset{\alpha(\boldsymbol{x}),\beta(\boldsymbol{x})}{\operatorname{argmin}} \sum_{i=1}^{n} K_{\boldsymbol{H}_n}\left(\boldsymbol{x}, \boldsymbol{x}^{(i)}\right)\left[y^{(i)} - \left(\alpha(\boldsymbol{x}) - \beta(\boldsymbol{x})^T \boldsymbol{x}^{(i)}\right)\right]^2,$$

and the estimate from this is

$$S_Y^{(n)}(\boldsymbol{x}) = \hat{\alpha}(\boldsymbol{x}) + \hat{\beta}(\boldsymbol{x})^T \boldsymbol{x}.$$

Local linear estimators usually work better in practice, for example, they do not suffer from bias at the boundary of the regressors as the NW estimator does. The local linear version of Racine and Li's kernel estimate is discussed in detail in [24]. The mean-square consistency in the case of continuous regressors is proved in [108], albeit under restrictive conditions on the coefficients of the polynomials.

The k-nearest neighbor regression

This so-called k-NN regression is another type of a local averaging regression estimate, see Equation (B.1). While fixing an $\boldsymbol{x} \in \mathbb{R}^r$, reorder the training data according to the increasing values of $\|\boldsymbol{X}_i - \boldsymbol{x}\|$:

$$\left\{(\boldsymbol{X}_{1,n}(\boldsymbol{x}), Y_{1,n}(\boldsymbol{x})), (\boldsymbol{X}_{2,n}(\boldsymbol{x}), Y_{2,n}(\boldsymbol{x})), \dots (\boldsymbol{X}_{n,n}(\boldsymbol{x}), Y_{n,n}(\boldsymbol{x}))\right\}.$$

Then

$$S_Y^{(n)}(\boldsymbol{x}) := \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{i,n}(\boldsymbol{x}),$$

which is the k_n-NN regression estimate of Y.

In the case of continuous regressors, the following theorem can be stated on the mean-square consistency.

Theorem B.3 (Theorem 6.1 of [15])**.** *If* $\lim_{n\to\infty} k_n = \infty$ *and* $\lim_{n\to\infty} \frac{k_n}{n} = 0$*, then the* k_n*-NN estimate is mean-square consistent for all distribution of* $(\boldsymbol{X}, Y)$ *such that* $\mathbb{E}\left[Y^2\right] < \infty$ *and for each* $\boldsymbol{x} \in \mathbb{R}^r$*, the random variable* $\|\boldsymbol{X} - \boldsymbol{x}\|^2$ *is absolutely continuous.*

The last condition is the no-tie condition, which always can be satisfied by the inclusion of a new component Z attached to $\boldsymbol{X}$: $((\boldsymbol{X}, Z), Y)$, where Z is uniformly distributed on $[0, 1]$ and independent of $(\boldsymbol{X}, Y)$. Of course, the training sample is also supplemented with components $Z_1, Z_2, \dots, Z_n$.

Appendix C

Time Series Analysis

C.1 One-dimensional time series

Definition C.1. *A univariate (discrete-time) time series $\{X_t\}$, $t \in \mathbb{Z}$ is weakly stationary (or stationary in the wide sense) if it has finite expectation and finite covariance function that do not depend on time shift:*

$$\mathbb{E}X_t = \mu, \quad c(h) := \operatorname{Cov}(X_{t+h}, X_t) = \operatorname{Cov}(X_h, X_0),$$

where $t, h \in \mathbb{Z}$; and $c(h)$ is called the (auto)covariance function of X_t.

Let $\{X_t\}$ be a weakly stationary, real-valued time series ($t \in \mathbb{Z}$). Assume that $\mathbb{E}(X_t) = 0$, $c(0) = \mathbb{E}(X_t^2) > 0$, and the sequence of autocovariances $c(h)$, $(h = 1, 2, \dots)$ is absolutely summable; obviously, $c(h) = c(-h)$. Then it is known that the Toeplitz matrix $\boldsymbol{C}_n = [c(i-j)]_{i,j=1}^n$, that is the covariance matrix of the random vector $(X_1, \dots, X_n)^T$, is positive definite for all $n \in \mathbb{N}$ (the vectors are column vectors and ^{T} denotes the transposition). Under the absolute summability condition, the spectral density function f of the process exists and it is continuous. It is defined as

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} c(h) e^{-ih\omega}, \quad f(\omega) \geq 0, \quad \omega \in [-\pi, \pi].$$

We can as well use $[-\pi, \pi]$ or $[0, 2\pi]$ for a complete period of f, and in view of $f(\omega) = f(-\omega) = f(2\pi - \omega)$ for a real-valued time series, it suffices to confine ourselves to the $[0, \pi]$ interval. In [69], an asymptotic relation between the eigenvalues of $\boldsymbol{C}_n$ and the values $f(\omega_j)$ is proved as $n \to \infty$, where $\omega_j = j\frac{2\pi}{n}$ is the jth Fourier frequency, $j = 0, 1, \dots, n-1$.

Theorem C.1 (Prop. 4.5.2. of [69])**.** *Let $\{X_t\}$ be a weakly stationary, real-valued time series (with zero mean) where the sequence of autocovariances $c(h)$, $(h = 1, 2, \dots)$ is absolutely summable; and let $f(\omega)$ be its spectral density function. Then let D_n be the $n \times n$ matrix,*

$$D_n = \begin{cases} \operatorname{diag}\left\{f(0), f(\omega_1), f(\omega_1), \dots, f\left(\omega_{\lfloor n/2 \rfloor}\right), f\left(\omega_{\lfloor n/2 \rfloor}\right)\right\} & \textit{if } n \textit{ is odd} \\ \operatorname{diag}\left\{f(0), f(\omega_1), f(\omega_1), \dots, f\left(\omega_{(n-2)/2}\right), f\left(\omega_{(n-2)/2}\right), f\left(\omega_{n/2}\right)\right\} & \textit{if } n \textit{ is even.} \end{cases}$$

Further let $\boldsymbol{C}_n^s$ be the real symmetric circulant matrix[1]

$$\boldsymbol{C}_n^{(s)} = \begin{bmatrix} c(0) & c(1) & c(2) & \cdots & c(2) & c(1) \\ c(1) & c(0) & c(1) & \cdots & c(3) & c(2) \\ c(2) & c(1) & c(0) & \cdots & c(4) & c(3) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ c(1) & c(2) & c(3) & \cdots & c(1) & c(0) \end{bmatrix}.$$

Then with the spectral decomposition $\boldsymbol{C}_n^s = P^T \Lambda P$, where P contains real orthogonal eigenvectors,

$$P^T \boldsymbol{C}_n P - 2\pi D_n \to \mathbf{0},$$

as $n \to \infty$ uniformly, i.e, the supremum of the above matrix' elements tends to zero.

It is well known, see e.g., [69, 67, 68], that by Kronecker products (with permutation matrices) the jth (real) eigenvalue of $\boldsymbol{C}_n^{(s)}$ when $n = 2k+1$ is

$$\sum_{h=-k}^{k} c(h)\rho_j^h = c(0) + 2\sum_{h=1}^{k} c(h)\cos(h\omega_j),$$

where $\rho_j = e^{i\omega_j}$ is the jth primitive (complex) nth root of 1 and again, $\omega_j = \frac{2\pi j}{n}$ is the jth Fourier frequency ($j = 0, 1, \ldots, n-1$). Further, the eigenvector corresponding to the jth eigenvalue is $(1, \rho_j, \ldots, \rho_j^{n-1})^T$; it has norm $\sqrt{n}$. After normalizing with $\frac{1}{\sqrt{n}}$, we get a complete orthonormal set of eigenvectors (of complex coordinates). From these, real ones can be constructed that Theorem C.1 uses.

In Chapter 3 we prove a similar, albeit more complicated, result in the multivariate case.

[1] If $k := \lfloor n/2 \rfloor$, then in each row, for odd n, $c(k)$ is duplicated, and for even n, there is only one $c(k)$.

List of my Publications

Journal articles

[J1] M. Bolla, F. Abdelkhalek, and M. Baranyi. “Graphical models, regression graphs, and recursive linear regression in a unified way”. In: *Acta Scientiarum Mathematicarum* 85.12 (2019), pp. 9–57. ISSN: 0001-6969. DOI: 10.14232/actasm-018-331-4.

[J2] M. Baranyi and M. Bolla. “Iterated Conditional Expectation Algorithm on DAGs and Regression Graphs”. In: *Econometrics and Statistics* 20 (2021), pp. 131–152. ISSN: 2452-3062. DOI: 10.1016/j.ecosta.2020.05.003.

[J3] M. Baranyi and R. Molontay. “Comparing the effectiveness of two remedial mathematics courses using modern regression discontinuity techniques”. In: *Interactive Learning Environments* 29.2 (2021), pp. 247–269. DOI: 10.1080/10494820.2020.1839506.

[J4] M. Bolla, T. Szabados, M. Baranyi, and F. Abdelkhalek. “Block circulant matrices and the spectra of multivariate stationary sequences”. In: *Special Matrices* 9.1 (Jan. 2021), pp. 36–51. ISSN: 2300-7451. DOI: 10.1515/spma-2020-0121.

Conference proceeding articles

[C1] M. Baranyi, K. Gál, R. Molontay, and M. Szabó. “Modeling Students’ Academic Performance Using Bayesian Networks”. In: *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*. IEEE, Nov. 2019, pp. 42–49. DOI: 10.1109/ICETA48886.2019.9040067.

[C2] M. Baranyi and R. Molontay. “Effect of Mathematics Remediation on Academic Achievement – A Regression Discontinuity Approach”. In: *2019 International Symposium on Educational Technology (ISET)*. IEEE, July 2019, pp. 29–33. DOI: 10.1109/ISET.2019.00016.

[C3] M. Baranyi, M. Nagy, and R. Molontay. “Interpretable Deep Learning for University Dropout Prediction”. In: *Proceedings of the 21st Annual Conference on Information Technology Education*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 13–19. ISBN: 9781450370455. DOI: 10.1145/3368308.3415382.

[C4] M. Baranyi, M. Bolla, and Gy. Kocsisné Szilágyi. “A Novel Dynamic Principal Component Analysis Method, Applied to ECG Signals”. In: *2021 55th Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, CA, USA: IEEE, Oct. 2021, pp. 265–269. ISBN: 978-1-66545-828-3. DOI: 10.1109/IEEECONF53345.2021.9723158.

[C5] M. Bolla, M. Baranyi, and F. Abdelkhalek. “Dimension Reduction of High Frequency and High Dimensional Data in Time and Space”. In: *HU-MATHS-IN Success Stories of Mathematical Short-Term Projects for Industry in 2017-2021*. Ed. by Z. Horváth. Györ, Hungary: Széchenyi István University, 2021, pp. 60–66. ISBN: 978-615-5837-99-9.

Bibliography

[1] M. Baranyi. "Graphical Models and Some Related Algorithms". MSc Thesis. Budapest: Budapest University of Technology and Economics, 2018.

[2] M. Bolla and T. Szabados. *Multidimensional Stationary Time Series: Dimension Reduction and Prediction*. Boca Raton: Chapman and Hall/CRC, Apr. 2021. ISBN: 978-1-00-310729-3. DOI: 10.1201/9781003107293.

[3] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[4] M. Studený. *Probabilistic Conditional Independence Structures*. Information Science and Statistics. London: Springer, 2005. ISBN: 978-1-85233-891-6.

[5] S. L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford, New York: Oxford University Press, May 1996. ISBN: 978-0-19-852219-5.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000. ISBN: 978-0-521-77362-1.

[7] D. R. Cox and N. Wermuth. "Linear Dependencies Represented by Chain Graphs". In: *Statistical Science* 8.3 (Aug. 1993), pp. 204–218. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1177010887.

[8] D. R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation*. Vol. 67. CRC Press, 1996. ISBN: 978-0-367-40137-5.

[9] G. Kauermann. "On a Dualization of Graphical Gaussian Models". In: *Scandinavian Journal of Statistics* 23.1 (1996), pp. 105–116. ISSN: 0303-6898.

[10] T. Richardson. "Markov Properties for Acyclic Directed Mixed Graphs". In: *Scandinavian Journal of Statistics* 30.1 (2003), pp. 145–157. ISSN: 1467-9469. DOI: 10.1111/1467-9469.00323.

[11] M. Drton. "Discrete Chain Graph Models". In: *Bernoulli* 15.3 (2009), pp. 736–753.

[12] N. Wermuth and K. Sadeghi. "Sequences of regressions and their independences". In: *TEST* 21 (2012), pp. 215–279.

[13] N. Wermuth. "Traceable regressions". In: *International Statistical Review* 80.3 (2012), pp. 415–438.

[14] L. Breiman and J. H. Friedman. "Estimating optimal transformations for multiple regression and correlation". In: *Journal of the American Statistical Association* 80 (1985), pp. 580–619.

[15] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002, pp. I–XVI, 1–647. ISBN: 978-0-387-95441-7.

[16] J. Racine and Q. Li. "Nonparametric estimation of regression functions with both categorical and continuous data". In: *Journal of Econometrics* 119.1 (2004), pp. 99–130.

[17] A. Rényi. "On measures of dependence". In: *Acta Mathematica Academiae Scientiarum Hungarica* 10.3 (Sept. 1959), pp. 441–451. ISSN: 1588-2632. DOI: 10.1007/BF02024507.

[18] L. P. Devroye and T. J. Wagner. "Distribution-Free Consistency Results in Nonparametric Discrimination and Regression Function Estimation". In: *The Annals of Statistics* 8.2 (1980), pp. 231–239. ISSN: 0090-5364.

[19] É. A. Nadaraya. "On Non-Parametric Estimates of Density Functions and Regression Curves". In: *Theory of Probability & Its Applications* 10.1 (Jan. 1965), pp. 186–190. ISSN: 0040-585X, 1095-7219.

[20] G. S. Watson. "Smooth Regression Analysis". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 26.4 (1964), pp. 359–372. ISSN: 0581-572X.

[21] S. Seabold and J. Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman. 2010, pp. 57–61.

[22] J. S. Racine. "Nonparametric Econometrics: A Primer". In: *Foundations and Trends® in Econometrics* 3.1 (2008), pp. 1–88. ISSN: 1551-3076.

[23] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai. "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2 (1998), pp. 271–293. ISSN: 13697412.

[24] Q. Li and J. Racine. "CROSS-VALIDATED LOCAL LINEAR NONPARAMETRIC REGRESSION". In: *Statistica Sinica* 14.2 (2004), pp. 485–512. ISSN: 10170405, 19968507.

[25] D. W. Scott, ed. *Multivariate Density Estimation*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., Aug. 1992. ISBN: 978-0-470-31684-9.

[26] W. Hardle, P. Hall, and J. S. Marron. "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" In: *Journal of the American Statistical Association* 83.401 (1988), pp. 86–95. ISSN: 0162-1459. DOI: 10.2307/2288922.

[27] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and semiparametric models*. Springer Science & Business Media, 2004.

[28] S. Wright. "The Method of Path Coefficients". In: *The Annals of Mathematical Statistics* 5.3 (Sept. 1934), pp. 161–215. DOI: 10.1214/aoms/1177732676.

[29] T. Rusch, M. Wurzer, and R. Hatzinger. "Chain Graph Models in R: Implementing the Cox-Wermuth Procedure". Presented at the Psychoco International Workshop on Psychometric Computing, Zürich. Zürich, Schweiz, Feb. 2013.

[30] M. Nagy and R. Molontay. "Predicting Dropout in Higher Education Based on Secondary School Performance". In: *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*. IEEE, 2018, pp. 000389–000394.

[31] D. A. Black, J. Galdo, and J. A. Smith. "Evaluating the Worker Profiling and Reemployment Services System Using a Regression Discontinuity Approach". In: *American Economic Review* 97.2 (2007), pp. 104–107. DOI: 10.1257/aer.97.2.104.

[32] J. Hahn, P. Todd, and W. V. der Klaauw. *Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design*. Working Paper 7131. Cambridge, MA: National Bureau of Economic Research, May 1999. DOI: 10.3386/w7131.

[33] J. D. Angrist and V. Lavy. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". In: *The Quarterly Journal of Economics* 114.2 (1999), pp. 533–575.

[34] P. McEwan and J. S. Shapiro. "The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates Using Exact Birth Dates". In: *Journal of Human Resources* 43.1 (2008).

[35] J. C. Calcagno and B. T. Long. *Evaluating the Impact of Remedial Education in Florida Community Colleges: A Quasi-Experimental Regression Discontinuity Design*. NCPR Brief. Columbia University: National Center for Postsecondary Research, Aug. 2009.

[36] G. Sonnert and P. M. Sadler. "The impact of taking a college pre-calculus course on students' college calculus performance". In: *International Journal of Mathematical Education in Science and Technology* 45.8 (Nov. 2014), pp. 1188–1207.

[37] B. A. Jacob and L. Lefgren. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis". In: *The Review of Economics and Statistics* 86.1 (Feb. 2004), pp. 226–244.

[38] M. De Paola and V. Scoppa. "The effectiveness of remedial courses in Italy: a fuzzy regression discontinuity design". In: *Journal of Population Economics* 27.2 (2014), pp. 365–386.

[39] M. Leake and S. A. Lesik. "Do remedial English programs impact first-year success in college? An illustration of the regression-discontinuity design". In: *International Journal of Research & Method in Education* 30.1 (2007), pp. 89–99.

[40] D. S. Lee and T. Lemieux. "Regression Discontinuity Designs in Economics". In: *Journal of Economic Literature* 48.2 (2010), pp. 281–355. ISSN: 0022-0515.

[41] H. S. Bloom. "Modern regression discontinuity analysis". In: *Journal of Research on Educational Effectiveness* 5.1 (2012), pp. 43–82.

[42] D. L. Thistlethwaite and D. T. Campbell. "Regression-discontinuity analysis: An alternative to the ex post facto experiment". In: *Journal of Educational Psychology* 51.6 (1960), pp. 309–317. ISSN: 1939-2176(Electronic),0022-0663(Print). DOI: 10.1037/h0044319.

[43] J. McCrary. "Manipulation of the running variable in the regression discontinuity design: A density test". In: *Journal of Econometrics* 142.2 (2008), pp. 698–714.

[44] B. R. Frandsen. “Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design when the Running Variable is Discrete”. In: *Advances in Econometrics*. Ed. by M. D. Cattaneo and J. C. Escanciano. Vol. 38. Emerald Pub., 2017, pp. 281–315. ISBN: 978-1-78714-390-6.

[45] G. W. Imbens and T. Lemieux. “Regression Discontinuity Designs: A Guide to Practice”. In: *Journal of Econometrics* 142.2 (Feb. 2008), pp. 615–635. ISSN: 03044076. DOI: 10.1016/j.jeconom.2007.05.001.

[46] A. Gelman and G. W. Imbens. “Why high-order polynomials should not be used in regression discontinuity designs”. In: *Journal of Business & Economic Statistics* 37.3 (2019), pp. 447–456.

[47] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1996. ISBN: 9780412983214.

[48] G. W. Imbens and K. Kalyanaraman. “Optimal Bandwidth Choice for the Regression Discontinuity Estimator”. In: *The Review of Economic Studies* 79.3 (July 2012), pp. 933–959. ISSN: 0034-6527. DOI: 10.1093/restud/rdr043.

[49] D. S. Lee and D. Card. “Regression discontinuity inference with specification error”. In: *Journal of Econometrics* 142.2 (2008), pp. 655–674.

[50] O. Bartalotti and Q. Brummet. “Regression Discontinuity Designs with Clustered Data”. In: *Advances in Econometrics*. Vol. 38. Emerald Pub. Ltd, 2017. Chap. 10, pp. 383–420.

[51] M. Kolesár and C. Rothe. “Inference in Regression Discontinuity Designs with a Discrete Running Variable”. In: *American Economic Review* 108.8 (Aug. 2018), pp. 2277–2304.

[52] Z. Jin, W. Liao, and F. Thoemmes. *RDDapp: Regression Discontinuity Design Application*. R package version 1.1.0. 2018.

[53] T. Nagler. “A generic approach to nonparametric function estimation with mixed data”. In: *Statistics & Probability Letters* 137 (2018), pp. 326–330. ISSN: 0167-7152. DOI: 10.1016/j.spl.2018.02.040.

[54] P. Constantinou and A. G. O’Keeffe. “Regression Discontinuity Designs: A Decision Theoretic Approach”. In: *arXiv:1601.00439 [stat]* (Jan. 2016). arXiv: 1601.00439 [stat].

[55] I. E. Marinescu, P. N. Lawlor, and K. P. Kording. “Quasi-Experimental Causality in Neuroscience and Behavioural Research”. In: *Nature Human Behaviour* 2.12 (Dec. 2018), pp. 891–898. ISSN: 2397-3374. DOI: 10.1038/s41562-018-0466-5.

[56] P. M. Steiner, Y. Kim, C. E. Hall, and D. Su. “Graphical Models for Quasi-experimental Designs”. In: *Sociological Methods & Research* 46.2 (Mar. 2017), pp. 155–188. ISSN: 0049-1241. DOI: 10.1177/0049124115582272.

[57] M. Tadayon and G. Pottie. “Causal Inference in Educational Systems: A Graphical Modeling Approach”. In: *arXiv:2108.00654 [cs, stat]* (Aug. 2021). arXiv: 2108.00654 [cs, stat].

[58] A. Csákány. “Results of mathematics ‘test zero’ at Budapest University of Technology and Economics in 2010”. In: *Pollack Periodica* 7.Supplement 1 (2012), pp. 43–53.

[59] J. E. Hoyt and C. T. Sorensen. “High school preparation, placement testing, and college remediation”. In: *Journal of Developmental Education* 25.2 (2001), p. 26.

[60] C. Brown. “Placement testing and remedial mathematics for post-secondary students: Prescription for success”. In: *Memorial University of Newfoundland, Winter* (1999).

[61] L. J. Rylands and D. Shearman. “Mathematics learning support and engagement in first year engineering”. In: *International Journal of Mathematical Education in Science and Technology* 49.8 (2018), pp. 1133–1147.

[62] J. I. Venegas-Muggli, K. A. Muñoz-Gajardo, and M. J. González-Clares. “The Impact of Counseling and Mathematics Remedial Programs on the Academic Achievement of Higher Education Students in Chile”. In: *Journal of College Student Development* 60.4 (2019), pp. 472–488.

[63] U. Kashyap and S. Mathew. “Corequisite model: An effective strategy for remediation in freshmen level mathematics courses”. In: *Journal of STEM Education: Innovations and Research* 18.2 (2017).

[64] A. W. Logue, D. Douglas, and M. Watanabe-Rose. “Corequisite mathematics remediation: Results over time and in different contexts”. In: *Educational Evaluation and Policy Analysis* 41.3 (2019), pp. 294–315.

[65] A. McAnally. “Impact of Co-Requisite Mathematics Remediation on Students’ Perceptions of Mathematics”. PhD thesis. University of Oklahoma, 2019.

[66] D. R. Brillinger. *Time Series: Data Analysis and Theory*. Classics in Applied Mathematics 36. SIAM, 2001.

[67] G. J. Tee. “Eigenvectors of block circulant and alternating circulant matrices”. In: *New Zealand Journal of Mathematics* 36.8 (2007), pp. 195–211.

[68] B. Friedman. “Eigenvalues of composite matrices”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 57.1 (1961), pp. 37–49.

[69] P. J. Brockwell, R. A. Davis, and S. E. Fienberg. *Time series: Theory and methods*. Springer Science & Business Media, 1991.

[70] *Principal Component Analysis*. Springer Series in Statistics. New York: Springer-Verlag, 2002. ISBN: 978-0-387-95442-4. DOI: 10.1007/b98835.

[71] Y. Dong and S. J. Qin. “A Novel Dynamic PCA Algorithm for Dynamic Data Modeling and Process Monitoring”. In: *Journal of Process Control*. Big Data: Data Science for Process Control and Operations 67 (July 2018), pp. 1–11. ISSN: 0959-1524. DOI: 10.1016/j.jprocont.2017.05.002.

[72] J. Kwon, H.-S. Oh, and Y. Lim. "Dynamic Principal Component Analysis with Missing Values". In: *Journal of Applied Statistics* 47.11 (Aug. 2020), pp. 1957–1969. ISSN: 0266-4763, 1360-0532. DOI: 10.1080/02664763.2019.1699910.

[73] S. Hörmann, Ł. Kidziński, and M. Hallin. "Dynamic Functional Principal Components". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.2 (2015), pp. 319–348. ISSN: 1467-9868. DOI: 10.1111/rssb.12076.

[74] W. Ku, R. H. Storer, and C. Georgakis. "Disturbance Detection and Isolation by Dynamic Principal Component Analysis". In: *Chemometrics and Intelligent Laboratory Systems*. InCINC '94 Selected Papers from the First International Chemometrics Internet Conference 30.1 (Nov. 1995), pp. 179–196. ISSN: 0169-7439. DOI: 10.1016/0169-7439(95)00076-3.

[75] C. R. Rao. *Linear statistical inference and its applications*. Vol. 2. Wiley, 1973.

[76] O. Akbilgic, H. Bozdogan, and M. E. Balaban. "A Novel Hybrid RBF Neural Networks Model as a Forecaster". en. In: *Statistics and Computing* 24.3 (May 2014), pp. 365–375. ISSN: 1573-1375. DOI: 10.1007/s11222-013-9375-7.

[77] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. "PhysioBank, PhysioToolkit, and PhysioNet". In: *Circulation* 101.23 (June 2000), e215–e220. DOI: 10.1161/01.CIR.101.23.e215.

[78] G. B. Moody and R. G. Mark. "The Impact of the MIT-BIH Arrhythmia Database". In: *IEEE Eng. Med. Biol. Mag.* 20.3 (2001), pp. 45–50. ISSN: 07395175. DOI: 10.1109/51.932724.

[79] G. B. Moody and R. G. Mark. "MIT-BIH Arrhythmia Database". English. Version 1.0.0. In: *PhysioNet* (2005). DOI: 10.13026/C2F305.

[80] E. Yakushenko. "St Petersburg INCART 12-lead Arrhythmia Database". English. Version 1.0.0. In: *PhysioNet* (2008). DOI: 10.13026/C2V88N.

[81] W. Zong, G. Moody, and D. Jiang. "A Robust Open-Source Algorithm to Detect Onset and Duration of QRS Complexes". In: *Computers in Cardiology, 2003*. Thessaloniki Chalkidiki, GR: IEEE, 2003, pp. 737–740. ISBN: 978-0-7803-8170-4. DOI: 10.1109/CIC.2003.1291261.

[82] E. Frank, M. A. Hall, and C. J. Pal. "Appendix B: The WEKA Workbench". In: *Data Mining: Practical Machine Learning Tools and Techniques*. Ed. by I. H. Witten. 4th ed. Amsterdam, NL: Elsevier, 2017. ISBN: 978-0-12-804291-5.

[83] E. D. Übeyli. "Statistics over Features of ECG Signals". In: *Expert Syst. Appl.* 36.5 (July 2009), pp. 8758–8767. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2008.11.015.

[84] M. Korürek and B. Doğan. "ECG Beat Classification Using Particle Swarm Optimization and Radial Basis Function Neural Network". In: *Expert Syst. Appl.* 37.12 (Dec. 2010), pp. 7563–7569. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2010.04.087.

[85] P. de Chazal and R. B. Reilly. "A Patient-Adapting Heartbeat Classifier Using ECG Morphology and Heartbeat Interval Features". In: *IEEE Trans. Biomed. Eng.* 53.12 (Dec. 2006), pp. 2535–2543. ISSN: 1558-2531. DOI: 10.1109/TBME.2006.883802.

[86] W. Jiang and S. G. Kong. "Block-Based Neural Networks for Personalized ECG Signal Classification". In: *IEEE Trans. Neural Networks* 18.6 (Nov. 2007), pp. 1750–1761. ISSN: 1941-0093. DOI: 10.1109/TNN.2007.900239.

[87] M. K. Das and S. Ari. "ECG Beats Classification Using Mixture of Features". In: *International Scholarly Research Notices* 2014 (Sept. 2014), pp. 1–12. ISSN: 2356-7872. DOI: 10.1155/2014/178436.

[88] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen. "Medical Image Classification with Convolutional Neural Network". In: *2014 13th Int. Conf. on Control Automation Robotics & Vision (ICARCV)*. Singapore: IEEE, Dec. 2014, pp. 844–848. ISBN: 978-1-4799-5199-4. DOI: 10.1109/ICARCV.2014.7064414.

[89] R. J. Martis, U. R. Acharya, C. M. Lim, K. M. Mandana, A. K. Ray, and C. Chakraborty. "Application Of Higher Order Cumulant Features For Cardiac Health Diagnosis Using ECG Signals". In: *Int. J. Neural Syst.* 23.04 (Aug. 2013), p. 1350014. ISSN: 0129-0657, 1793-6462. DOI: 10.1142/S0129065713500147.

[90] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan. "A Deep Convolutional Neural Network Model to Classify Heartbeats". In: *Comput. Biol. Med.* 89 (Oct. 2017), pp. 389–396. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2017.08.022.

[91] M. Zubair, J. Kim, and C. Yoon. "An Automated ECG Beat Classification System Using Convolutional Neural Networks". In: *2016 6th Int. Conf. on IT Convergence and Security (ICITCS)*. Prague, CZ: IEEE, Sept. 2016, pp. 1–5. ISBN: 978-1-5090-3765-0. DOI: 10.1109/ICITCS.2016.7740310.

[92] D. K. Atal and M. Singh. "Arrhythmia Classification with ECG Signals Based on the Optimization-Enabled Deep Convolutional Neural Network". In: *Comput. Methods Programs Biomed.* 196 (Nov. 2020), p. 105607. ISSN: 01692607. DOI: 10.1016/j.cmpb.2020.105607.

[93] M. Wu, Y. Lu, W. Yang, and S. Y. Wong. "A Study on Arrhythmia via ECG Signal Classification Using the Convolutional Neural Network". In: *Front. Comput. Neurosci.* 14 (2021), p. 106. ISSN: 1662-5188. DOI: 10.3389/fncom.2020.564015.

[94] F. I. Alarsan and M. Younes. "Analysis and Classification of Heart Diseases Using Heartbeat Features and Machine Learning Algorithms". In: *J. Big Data* 6.1 (Aug. 2019), p. 81. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0244-x.

[95] M. G. Martini, A. Polpetta, and P. Banelli. "Context-Aware Multi-Lead ECG Compression Based on Standard Image Codecs". In: *2009 3rd Int. Conf. on Pervasive Computing Technologies for Healthcare*. IEEE, 2009, pp. 1–7.

[96] Q. Tan, B. Fang, and P. Wang. "Improved Simultaneous Matching Pursuit for Multi-Lead ECG Data Compression". In: *2010 Int. Conf. on Measuring Technology and Mechatronics Automation*. Vol. 2. IEEE, 2010, pp. 438–441.

[97] S. M. Isa, W. Jatmiko, and A. M. Arymurthy. "3D SPIHT for Multi-Lead ECG Compression". In: *2014 Int. Conf. on Robotics and Automation (ICRA)*. IEEE, May 2014, pp. 488–493. DOI: 10.1109/ICRA.2014.6906900.

[98] D. Geiger and J. Pearl. "On the Logic of Causal Models". In: *Uncertainty in Artificial Intelligence*. Ed. by R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer. Vol. 9. Machine Intelligence and Pattern Recognition. North-Holland, 1990, pp. 3–14.

[99] N. Wermuth. "Graphical Markov models, unifying results and their interpretation". In: *Wiley Stats*. Ed. by N. e. a. Balakrishnan. Wiley, 2015.

[100] J. M. Hammersley and P. E. Clifford. "Markov random fields on finite graphs and lattices". In: Unpublished manuscript (1971).

[101] D. Sonntag. *A Study of Chain Graph Interpretations*. Linköping University Electronic Press, Apr. 2014. ISBN: 978-91-7519-377-9. DOI: 10.3384/lic.diva-105024.

[102] D. Sonntag and J. M. Peña. "Chain Graph Interpretations and Their Relations Revisited". In: *International Journal of Approximate Reasoning*. Special Issue of the Twelfth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2013) 58 (Mar. 2015), pp. 39–56. ISSN: 0888-613X. DOI: 10.1016/j.ijar.2014.12.001.

[103] S. L. Lauritzen and N. Wermuth. "Graphical Models for Associations between Variables, Some of Which Are Qualitative and Some Quantitative". In: *The Annals of Statistics* 17.1 (Mar. 1989), pp. 31–57. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176347003.

[104] M. Frydenberg. "The Chain Graph Markov Property". In: *Scandinavian Journal of Statistics* 17.4 (1990), pp. 333–353.

[105] S. A. Andersson, D. Madigan, and M. D. Perlman. "An Alternative Markov Property for Chain Graphs". In: *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*. UAI'96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 1996, pp. 40–48. ISBN: 978-1-55860-412-4.

[106] C. J. Stone. "Consistent Nonparametric Regression". In: *The Annals of Statistics* 5.4 (July 1977), pp. 595–620. DOI: 10.1214/aos/1176343886.

[107] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

[108] M. Kohler. "Universal consistency of local polynomial kernel regression estimates". In: *Annals of the Institute of Statistical Mathematics* 54.4 (2002), pp. 879–899.