6-2-2021 11:00 AM

Making Sense of Noisy Data: Theory and Applications

Lingzhi Chen, *The University of Western Ontario*

Supervisor: Zitikis, Ricardas, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

Abstract

This thesis introduces a novel and interpretable index of increase which is mathematically defined based on the distance between a given function and a set of non-increasing functions. Unlike the widely used traditional statistical methods for analyzing relationships between variables, the index does not rely on assumptions such as linearity, normality, and monotonicity, which may not be satisfied. Hence, it has the flexibility to be applied directly on pairs of data points to measure and compare non-linear, asymmetric, and non-monotonic relationships between two variables.

We begin with a review of the literature and background knowledge in Chapter 2.

In Chapter 3, we propose a distance-based index of increase, describe its properties in detail, and show its benefits through applying it to an educational dataset. In this way, we see the interpretability of the index of increase and how it can be applied. We also propose several modifications for different scenarios, such as subgroup analysis. Lastly, we provide a step-by-step implementation guideline for non-statistical researchers or practitioners.

In Chapter 4, we investigate two extensions of the index of increase, which quantify the interchangeability between variables. We discuss the usage of them in the context of developing curricula, accompanied with extensive graphical and numerical illustrations.

In Chapter 5, we introduce and explore an empirical index of increase that works in both deterministic and random environments, thus allowing to assess monotonicity of functions that are prone to random measurement-errors. We prove consistency of the empirical index and show how its rate of convergence is influenced by deterministic and random parts of the data. In particular, the obtained results suggest a frequency at which observations should be taken in order to reach any pre-specified level of estimation precision. We illustrate the index using data arising from purely deterministic and error-contaminated functions, which may or may not be monotonic.

Finally, in Chapter 6, we summarize our main results and give an outline of potential future works.

Keywords: index of increase; monotonicity; consistent estimator; distance-based measure; curriculum development; student performance evaluation

Summary for lay audience

Traditional statistical methods for analyzing relationships between variables often rely on assumptions such as linearity, normality, and monotonicity, which may not be satisfied. For example, this is the case when analyzing curves depicting sales versus prices, exports versus economic growth – they are hardly monotonic, let alone linear. Thus, the use of traditional statistical tools becomes problematic. Furthermore, random noise or random measurement errors frequently contaminate data, and thus true relationships are blurred, thus leading to misrepresentations of results.

In this thesis, we explore an index of increase and its estimator that works in both deterministic and random environments, thus enabling the assessment of monotonicity of functions that might be exposed to random noise. The index and its estimator allow us to quantify non-linear, asymmetric, and non-monotonic relationships between variables. We shall illustrate theoretical results using data arising from deterministic and error-contaminated functions, which may or may not be monotonic. We also apply the index of increase with proper modifications in educational datasets to illuminate the use cases and potential extensions. Finally, we summarize the contributions of this thesis and outline the potential future works.

Co-authorship statement

This thesis consists of materials based on three jointly authored research papers. The first and second papers were co-authored with my supervisor Dr. Ričardas Zitikis. The first paper titled "Measuring and comparing student performance: a new technique for assessing directional associations" has been published in the journal *Education Sciences*. The second paper titled "Quantifying and analyzing nonlinear relationships with a fresh look at a classical dataset of student scores" has been published in the journal *Quality & Quantity*. The third paper titled "Estimating the index of increase via balancing deterministic and random Data" was co-authored with Dr. Ričardas Zitikis, Dr. Youri Davydov, and Dr. Nadezhda Gribkova. It has been published in the journal *Mathematical Methods of Statistics*. All co-authors have contributed equally to the design of the studies presented in my co-authored papers, the execution (both mathematical and computational) of the studies, and finally, to writing up the findings and conclusions.

I certify that I am the lead author for all these three papers. I would like to sincerely thank Dr. Zitikis for his critical supervision on all of my research as well as his unwavering support. I would also like to thank all the co-authors for their valuable contributions.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Ričardas Zitikis, for his unwavering support and guidance throughout my time at the University of Western Ontario. As an academic supervisor, his patience, diligence, and erudition have helped me become a professional and a better researcher. He is also a life mentor for me who always encourages me to challenge difficulties and explore unknown areas. He provided chances for me to participate in statistics-related research projects in partnership with several industrial organizations. These projects have enriched my understanding of the application of statistical modelling to real-life data analysis problem. These invaluable experiences have broaden my horizons and expanded my future career.

I am grateful to Dr. Youri Davydov, and Dr. Nadezhda Gribkova for their contribution to the grouped estimator of the index of increase. The enlightening discussion with them has advanced our understanding on the index of increase. I gratefully acknowledge the insightful comments and suggestions provided by anonymous reviews during the paper publication process. I would also like to sincerely thank my thesis committee: Dr. Marcos Escobar-Anel, Dr. Edward Furman, Dr. Jiandong Ren, and Dr. Zheng Zhang. It is my honour to work with them to complete my thesis.

I would like to thank the research association, MITACS, for providing funding opportunities for me to complete research projects with industrial partners. Specifically, I would like to thank my direct managers during my MITACS projects, Mr. Gary Bogdani from Unilever Canada and Ms. Jenny Zhang from Sun Life Financial, and all other colleagues.

I am deeply thankful to my parents for unconditionally supporting me without complaint. Without them, I would not be able to pursue the doctoral degree. I am sincerely thankful to all the colleagues and friends in the Statistical and Actuarial Sciences department, especially my academic brothers-in-arms Junhe Chen, Ang Li, Yifan Li, and Yang Miao. Last but not least, I would like to thank my partner Mihwa Seong, whose love and encouragement make life wonderful.

This thesis is dedicated to my parents for their endless love and support.

Contents

List of figures

List of tables

Chapter 1

Introduction

Several statistical tools provide ways to capture associations between variables, particularly monotonic associations. Many of these measures are primarily applications and extensions of simple linear regression and Pearson correlation coefficient, which tell the extent of the relationship by the measures' values and indicate monotonic relationships by their signs. In the following section, we provide several examples to illustrate where measuring relationships between variables matters, and how those traditional statistical tools play on stage, as well as their main limitations. Then, we introduce a novel method, the index of increase, to overcome the limitations of traditional statistical methods when quantifying non-linear, asymmetric, and non-monotonic relationship between variables.

1.1 Motivation

One popular method of measuring relationships between variables is linear regression. Linear regression is able to depict asymmetric relationships between variables. It also has solid statistical background to provide more information such as confidence intervals of estimators. Yet, linear regression has two major limitations. One is that linear regression is unable to accurately reflect relationships between variables in the case of non-linear relationships. The other is that the results are biased if the data do not meet the normality assumption. Inaccurate or biased measurements of the relationships between variables may

have important practical consequences. We provide two examples which are extensions of linear regression in finance and economics to demonstrate the limitations in detail.

In finance, the Capital Asset Pricing Model (CAPM) proposed by Sharpe (1964) and Lintner (1965) is one of the most famous models in portfolio management used to make decisions about adding assets to a well-diversified portfolio. For CAPM's discussions and extensions, we refer to Fama and French (2004, 2006); Daniel et al. (2001); Eisenbeiss et al. (2007); Zhang (2017); Gonçalves et al. (2020), and references therein. The standard CAPM has the following structure:

$$\mathbb{E}(\mathrm{R}_i) = \mathrm{R}_f + \beta_i(\mathbb{E}(\mathrm{R}_m) - \mathrm{R}_f),$$

where $\mathbb{E}(\mathrm{R}_i)$ is the expected return on the capital asset i, R_f is the risk-free rate of interest such as interest arising from government bonds, and $\mathbb{E}(\mathrm{R}_m)$ is the expected return of the market. CAPM provides a quantity β_i to measure the i^{th} asset sensitivity of the expected excess asset returns to the market premium $\mathbb{E}(\mathrm{R}_m) - \mathrm{R}_f$, also known as systematic risk. The sign of β_i indicates whether the risk is increased or diminished from portfolio. In statistical terms, β is the slope of a simple linear regression

$$Y = \alpha + \beta X + \epsilon,$$

where X is the market return, Y is the individual asset return, and ϵ is the error term, which is independent of X. Thus, β is usually defined as:

$$\beta = \frac{Cov(X, Y)}{Var(X)}.$$

Suppose we have daily individual stock returns $\{y_i\}_{i=1}^n$ and daily market return $\{x_i\}_{i=1}^n$. Then the estimate of β is the estimate of the slope in simple linear regression:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

From the nature of simple linear regression model, the constant slope β characterizes the linear relationship between the independent variable X and the response variable Y. Particularly, it measures the monotonic relationship between X and Y. It is obviously a

directional association. When we change the positions of X and Y, β changes accordingly, which makes sense in the CAPM case as the market performance and individual asset performance are not interchangeable. Another important assumption of linear regression is the normality assumption, which assumes the asset returns to be normally distributed in the CAPM case. Practically, the asset returns seldom follow the normal distribution. Instead, fat-tailed behaviour and skewness are common, which violate the model normality assumption, thus leading to the biased estimates of the risk measure, β.

In economics, there is another commonly used concept called price elasticity of demand, which measures the sensitivity, or elasticity, of the quantity demanded of a good or service to changes in its price, all else held equal. More precisely, it is the percentage of the change in unit sales relative to one percent of the change in the price. Price elasticity plays an important role in optimizing product pricing and maximizing the revenues. For the applications of price elasticity, we refer to Havranek et al. (2012); Tucker et al. (2018); Perera and Tan (2019); Corrigan et al. (2021). The formula for the price elasticity of demand is given as follow:

$$e_p = \frac{\mathrm{d}Q/Q}{\mathrm{d}P/P} = \frac{\mathrm{d}log(Q)}{\mathrm{d}log(P)} = \frac{\mathrm{d}log(Q)}{\mathrm{d}P} \times P, \tag{1.1.1}$$

where P is the unit price of the demanded good and Q is the quantity of the demanded good. By the law of demand in economy, e_p is usually negative as people are less likely to buy a product when product price increases. From Eq. (1.1.1), we know that if we want to estimate e_p, we can build the following equation:

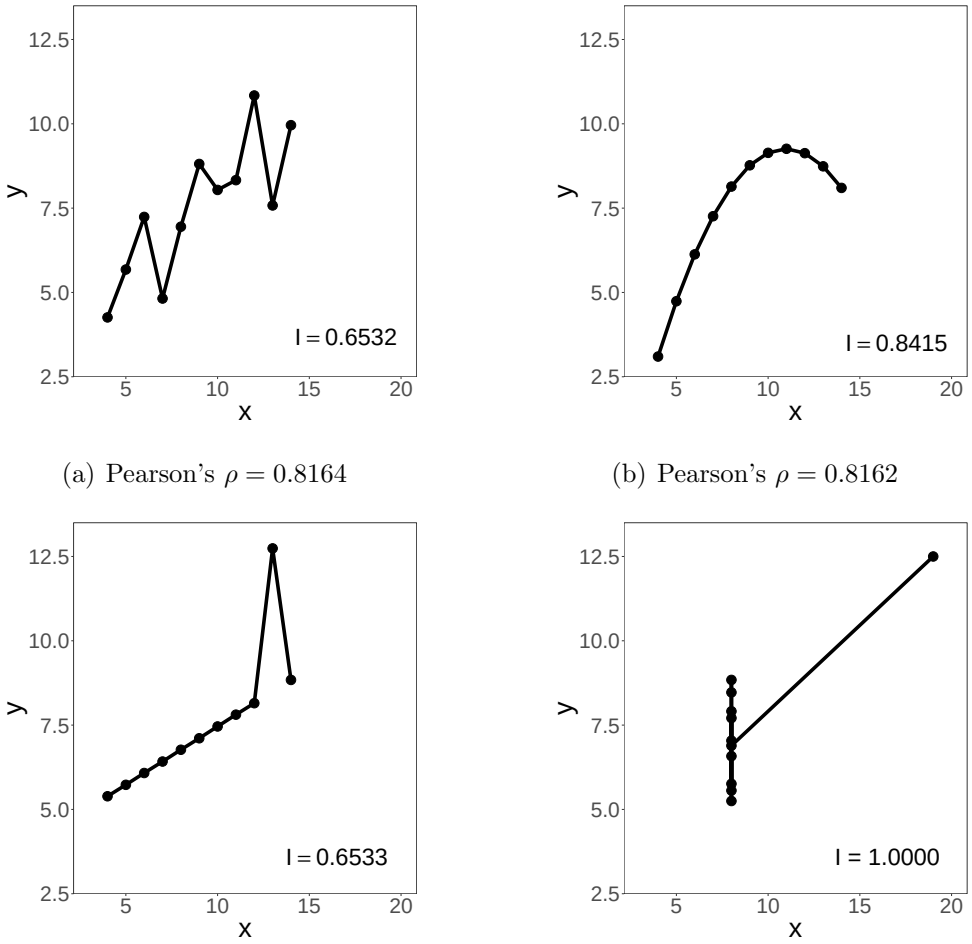
$$log(Q) = \alpha + \beta \times log(P) + \epsilon \quad or \quad log(Q) = \alpha + \beta \times P + \epsilon, \tag{1.1.2}$$

which is nothing but a simple linear regression with price P or its function as the predictor and function of unit sales as the response variable. Hence, the price elasticity of demand can be derived based on Eq. (1.1.1) and (1.1.2). In this example, we can actually describe the non-linear relationship between unit sales and its price by using a linear model through taking the logarithm transformation to variables. However, certain transformations of variables as well as probabilistic assumptions are needed when using linear regression model, which makes it inflexible in some cases when the underlying function is unknown.

Therefore, these examples illustrate the restricted uses of linear regression owing to its assumptions of data.

Next, we introduce some cases where researchers use other traditional statistical techniques when quantifying and analyzing correlations between variables. Measuring and comparing student performance are of interest for educators and psychologists. Naturally, particular attention has been paid to the design of experimental studies and careful analyses of observational data. As a rule, the Pearson correlation coefficient and its extension, the intraclass correlation coefficient (ICC), and the Spearman correlation coefficient have been extensively used in educational and psychometric literature for describing and analyzing associations in bivariate and multivariate data. We also note that the ICC and the Spearman correlation coefficient can solve some deficiencies exhibited by the Pearson coefficient. For enlightening discussions and references on the ICC, we refer to, e.g., Looney (2000), Hedges and Hedberg (2007), Zhou et al. (2011), and on the Spearman coefficient, to Gauthier (2001), Puth et al. (2015), and references therein. It should be also noted that the two coefficients are closely related to the Pearson correlation coefficient. Namely, the ICC is the Pearson correlation coefficient but with the pooled mean as the centering constant and the pooled variance as the normalizing constant. The Spearman correlation coefficient is the Pearson correlation coefficient of the ranks of the underlying random variables. Consequently, both the Spearman correlation coefficient and the ICC are symmetric with respect to the random variables, and we find this feature unnatural in the context of the research presented in later chapters.

There are other issues with the use of these coefficients when assessing trends, as we can clearly see from Figure 1.1.1. Namely, all the four panels have very different trends but virtually identical Pearson correlation coefficients. We have produced these trends by connecting the classical Anscombe (1973) bivariate data using straight lines. Each panel of Figure 1.1.1 is also supplemented with the corresponding value of the index of increase, denoted generically by I, which we shall introduce formally later in this thesis, once the necessary preliminary work has been done. At the moment, we only note some of the properties of the index that convey its idea and main features. Namely, the index is:



(a) Pearson's $\rho = 0.8164$

(b) Pearson's $\rho = 0.8162$

(c) Pearson's $\rho = 0.8163$

(d) Pearson's $\rho = 0.8165$

Figure 1.1.1: Trends and indices of increase arising from the classical Anscombe (1973) quartet.

- normalized, that is, its range is the unit interval $[0, 1]$;
- take the value 1 when the trend is increasing;
- take a value in the interval $[1/2, 1]$ when, loosely speaking, the trend is increasing more than decreasing;
- take a value in the interval $[0, 1/2]$ when, loosely speaking, the trend is decreasing more than increasing;
- take the value 0 when the trend is decreasing;

- is symmetric only when the explanatory and response variables are interchangeable;
- has a clear geometric interpretation.

Keeping these features in mind, we can now familiarize ourselves with the numerical values of the index of increase reported in the four panels of Figure 1.1.1. Rigorous definitions and modifications of the index of increase will be introduced in following chapters, based on the context.

1.2 Organization of this thesis

In this thesis, we deliberately limit our research scope to the analysis of associations of bivariate variables, which means that relationships between the variables can be expressed via curves or scatterplots. In particular, we are interested in quantifying monotonicity between two variables. We focus on proposing and developing a distance-based index of increase and its practical modifications that are intuitive and interpretable to researchers in various areas of application. We have organized this thesis as follows.

In Chapter 2, we begin with a literature review of relevant measures that quantify associations between variables, especially monotonicity of functions. Then, we provide background knowledge that is either closely related to the proposed index of increase or used throughout this thesis, such as non-parametric curve fitting. In Chapter 3, we formally introduce the novel index of increase, its properties, and practical modifications. We illustrate its benefits through analyzing and comparing student performance between genders using an education dataset. We also provide a step-by-step implementation guide for non-statistical researchers. In Chapter 4, we apply the index of increase in the context of curriculum development. In addition, we propose two extended measures based on the index of increase, which quantify the interchangeability between subjects. Discussions of advantages of the index and its extensions are based on an education dataset. In Chapter 5, we introduce and explore an empirical estimator of the index of increase that works in both deterministic and random environments, thus allowing to assess monotonicity of functions that are prone to random measurement-errors. We prove consistency of the index and

show how its rate of convergence is influenced by deterministic and random parts of the data. In particular, the obtained results suggest a frequency at which observations should be taken in order to reach any pre-specified level of estimation precision. We illustrate the index using data arising from purely deterministic and error-contaminated functions, which may or may not be monotonic. In Chapter 6, we make general comments related to our contributions through this thesis and state some possible future studies.

Chapter 2

Literature review and background knowledge

2.1 Literature review: development of measuring and quantifying monotonicity

Monotonicity is a simple but essential property stating that if one variable increases (or decreases), the related output also increases (or decreases). Studying monotonicity between two variables arises naturally from daily life problems. Are higher product prices indicating higher product sales? Are higher returns of stock A suggesting higher returns of stock B? Is higher education level resulting higher wages? Besides, considering the monotonicity between the input variable and output variable has been topic of interest for data scientists as the development of monotone constraint in machine learning algorithms, such as regressions (e.g., Brezger and Steiner, 2008), XGBoost (e.g., Wang et al., 2020), and so on. As a results, measuring the monotonicity between input and output is becoming increasingly vital. In some cases, data scientists rely on domain knowledge from experts. Nevertheless, having an objective measure is more desirable. Due to efficiency and interpretability, measures such as linear regression and the Pearson correlation coefficient and its extensions are the first choices in practice, even though they may have limitations.

To address the issues of traditional statistical tools such as linear regression and the

Pearson correlation coefficient noted in Chapter 1, we propose a novel technique, the index of increase. The index of increase is designed specifically for illuminating directional associations between variables, particularly useful when the variables follow non-linear and non-monotonic relationships. In the following section, we briefly review the recent related developments in quantitative measures for monotonicity, which provided the motivations and inspirations for developing the index of increase. Then, we describe the use of the index of increase and its properties as well as our theoretical contributions.

2.1.1 Importance to interpret statistical measures that capture dependency between variables

Except for the theoretical interests (i.e. assessing nonlinear, asymmetric, monotonic relationships) that we illustrated in the previous chapter, one of the crucial motivations of developing the index of increase is interpretability. In this thesis, by interpretability, we mean that we can use plain language to explain the meaning of the statistical measures and interpret the results. In any area where statistical methods are applied, people seek for the ability to interpret statistical results. For instance, from the perspective of academic researchers, they desire not only to identify whether a monotonic relationship between a pair of variables exists or not, but also want to quantify, compare, and interpret it. From practitioner's points of view, even though modelling with big data is becoming the mainstream nowadays, understanding and interpreting the potential dependency between variables are still essential before moving to a more structural modelling process. In many cases, machine learning algorithm's performance, such as prediction or classification ability, highly depends on the quality of data cleaning and feature engineering, which are the results of domain knowledge and data mining techniques. In order to better conduct these pre-modelling processes, it is crucial to choose proper and interpretable measures to understand the underlying relationships between variables.

With the development in statistical techniques, researchers focus more on developing complicated probability theories such as asymptotic properties and hypothesis testing, which deviate farther away from the interpretability of measures themselves (Reimherr and

Nicolae, 2013). It does not mean these statistical theories are not important. Yet, being able to explain the measures in plain language is important for a wider application beyond statistics and mathematics. Here, we give a few examples of such measures, or frameworks, which lack intuitive and straightforward explanations. Some of them only have explanation at the boundary values 0 and 1. Note that value 0 usually means "independence" while value 1 means "dependence".

- Distance correlation

  Székely, Rizzo, and Bakirov (2007) propose the distance correlation $\mathrm{dCor}^2(X, Y)$ as a measure of dependence of two random vectors of arbitrary dimensions. The definition of $\mathrm{dCor}^2(X, Y)$ is proved to be closely related to joint and marginal characteristic functions. Consistency and other asymptotic properties of the empirical distance correlation $\mathrm{dCor}_n^2(X, Y)$ have also been proved (cf. Székely, Rizzo, and Bakirov, 2007; Székely and Rizzo, 2009). Therefore, we can naturally identify independence through hypothesis testing. Compared to the Pearson's correlation, distance correlation can detect nonlinear and/or non-monotone dependency. Theoretically, when $\mathrm{dCor}^2(X, Y) = 0$, it implies X and Y are independent. If $\mathrm{dCor}^2(X, Y) = 1$, then there exists a vector $\mathbf{a}$, a nonzero value b, and an orthogonal matrix C such that $Y = \mathbf{a} + bXC$. However, the interpretation of values between 0 and 1 remains unclear.

  One of the usages of measures of dependence is in reducing dimension of predictors, which is important in many big data related modelling areas. Similar to feature screening process via the Pearsons correlation introduced by Fan and Lv (2008), feature screening process via distance correlation is also studied and implemented in different areas (e.g., Li, Zhong, and Zhu, 2012; Kong, Wang, and Wahba, 2015).

- Maximal information coefficient

  Reshef et al. (2011) present the maximal information coefficient (MIC) as a measure of the strength of the linear or non-linear relationship between two variables, which is based on information theory. MIC uses binning as a method to apply mutual information (cf. Cover and Thomas, 2006) on continuous variable. Given a finite set

D of ordered pairs, we create x bins on x-axis and y bins on y-axis so that a grid G is defined by integers (x, y). Then, for all grids G with x columns and y rows, we try to find the largest possible normalized mutual information $m_{x,y}$. Looping through all possible integers (x, y), MIC is defined as the maximum of all $m_{x,y}$. Since MIC is defined through mutual information, it is natural that MIC is symmetric. Also, when MIC equals to 0, it means the random variable X and Y are independent. When MIC tends to 1, certain noiseless relationship exists. Nevertheless, masked by the complex definition of MIC, the explanation of values between 0 and 1 is not straightforward. Other properties, technical details and applications of MIC can be found in Reshef et al. (2011, 2016, 2018). Similarly, MIC can also be implemented in feature screening process in data analysis (e.g., Ge et al., 2016; Sun et al., 2018; Wen et al., 2019).

- Copula and copula-related measures

  Copula is a widely-used framework for modelling dependency structure of random variables, especially in quantitative finance (e.g., Li, 2000; Low et al., 2013, 2016; Rad et al., 2016). Essentially, the copula function is the joint cumulative distribution function of $(U_1, ..., U_n)$, where U_i follows uniform distribution on interval $[0, 1]$. An important theorem proved by Sklar (1973) states that the joint distribution function can be expressed with marginal distribution functions and a copula function. For other properties of the copula function, we refer to Nelsen (2006). Moreover, several famous and popular measures of association of (X, Y) can be expressed by copula, for example, the Spearman's ρ and the Kendall's τ (cf. Schweizer and Wolff, 1981) as follows

$$\rho(X, Y) = 12 \int_0^1 \int_0^1 [C(u, v) - uv] \mathrm{d}u \, \mathrm{d}v,$$

$$\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u, v) \mathrm{d}C(u, v) - 1.$$

  Even though copula framework has solid background theories, it is still elusive and

cannot be easily explained in plain language due to its complicated setting.

2.1.2 Index of monotonicity inspired by the Gini index

Several researchers have worked on areas that assess and quantify monotonicity of functions. As we noted previously, the index of increase in this thesis has a clear geometric interpretation and it is a distance-based measure. Hence, we briefly introduce recent developments of distance-based methods that measure monotonicity.

Davydov and Zitikis (2005) propose an index of monotonicity to quantify monotonicity of a function, which is inspired by econometric concepts, the Lorenz curve and the Gini index. In economics, the Lorenz curve (cf. Lorenz, 1905) is a graphical representation of the distribution of income or wealth. It shows the proportion of the overall income or wealth assumed by the bottom $x\%$ of population. The Lorenz curve starts from $(0, 0)$ (the origin of coordinate) to $(1, 1)$. The curve is always monotonically increasing and never exceeds the diagonal which is drawn under the assumption of income or wealth equality. Mathematically, let X be a random variable representing the income with cdf $F(x)$ and a finite mean $\mu_F = \mathrm{E}[X]$. The actual population Lorenz function is

$$\mathrm{L}_F(p) = \frac{1}{\mu_F} \int_0^p F^{-1}(t)\,\mathrm{d}t.$$

When we assume income equality, we have the egalitarian Lorenz function $\mathrm{L}_E(p) = p$, $0 \le p \le 1$. The Gini index (cf., e.g., Gini, 1914, 1921) is used to measure social inequality in population, which is widely applied in economics (e.g., Yitzhaki and Schechtman, 2013; Inoue et al., 2015; Liao, 2006; Greselin and Zitikis, 2018). The Gini index, denoted by G_F, has the format

$$\mathrm{G}_F = 2 \int_0^1 (\mathrm{L}_E(p) - \mathrm{L}_F(p))\,\mathrm{d}p.$$

From a mathematical point of view, the Gini index quantifies the area between the Lorenz curve and the diagonal. Namely, the Gini index quantifies the gap between real social income distribution and the social income distribution under equality assumption, which makes the Gini index have a clear geometric interpretation.

The logic behind the Gini index is to set a reference assuming a statement is true (i.e., the egalitarian Lorenz function), then compare and quantify the deviation of the actual observation and the reference. Davydov and Zitikis (2005) adapt the similar logic and build the following Gini-type index of monotonicity.

Definition 2.1.1. *(Davydov and Zitikis, 2005) Let $f : [0,1] \to [0,\infty)$ be a function integrable over the interval $[0,1]$ with respect to the Lebesgue measure λ. The Gini-type index of monotonicity of function f is defined in the following:*

$$\mathrm{I}_f = \int_0^1 (F_f(t) - C_f(t))\,\mathrm{d}t = \int_0^1 \int_0^t (f(u) - G_f^{-1}(u))\,\mathrm{d}u\,\mathrm{d}t, \tag{2.1.1}$$

where $F_f(t) = \int_0^t f(u)\,\mathrm{d}u$, $C_f(t) = \int_0^t G_f^{-1}(u)\,\mathrm{d}u$, $G_f^{-1}(u) = \inf\{x \in \mathbf{R} : G_f(x) \geq u\}$, and $G_f(x) = \lambda\{s \in [0,1] : f(s) \leq x\}$.

Note 2.1.1. *To better connect the concepts to a statistical background, Davydov and Zitikis (2005) outline that we can interpret f as a random variable on the probability space $([0,1], \mathcal{B}_{[0,1]}, \lambda)$. Then, the terms introduced in Definition 2.1.1 can be interpreted as follows. $G_f(x)$ is the distribution function of f. $G_f^{-1}(u)$ is the quantile function of f, which is also called generalized inverse of the function G_f. In mathematical literatures, $G_f^{-1}(u)$ is also called a monotone (non-decreasing) rearrangement of f, which means $G_f^{-1}(u)$ is a non-decreasing measurable function such that its distribution function equals $G_f(x)$. In addition, C_f is called the convex rearrangement of the distribution function F_f.*

Note 2.1.2. *As noted by Davydov and Zitikis (2005), when a function is convex, then its convex rearrangement coincides with the function. Thus, the area between the original function and its convex rearrangement can be thought as a measure of convexity of the function. Also, if a function is non-decreasing, then its integral over the intervals $[0,t], 0 \leq t \leq 1$ is convex.*

In view of Note 2.1.2, Definition 2.1.1 shows that the index of monotonicity Davydov and Zitikis (2005) proposed adopted the general idea of the Gini index. It can measure the lack of monotonicity of a function f because it quantifies the distance between the probability distribution F_f generated by given function f and its convex rearrangement C_f.

Namely, if f is non-decreasing, F_f and C_f coincide which lead to $\mathrm{I}_f = 0$. Otherwise, $\mathrm{I}_f > 0$. Moreover, Davydov and Zitikis (2005) provided the general form of estimation of I_f and explored its convergence rate as well as the asymptotic distribution, which contributes to the theoretical results significantly.

Based on the aforementioned index of monotonicity, Qoyyimi and Zitikis (2014) further proposed a L_1-based index of non-decreasingness

$$\mathcal{I}_f = \int_0^1 |f(t) - G_f^{-1}(t)| \, \mathrm{d}t, \tag{2.1.2}$$

where f and G_f^{-1} are defined the same as the ones in Definition 2.1.1. The geometric interpretation of Eq. (2.1.2) is the L_1 distance between function f and its non-decreasing rearrangement G_f^{-1}. Theories and numerical procedures for calculating the above two indices are also provided by Qoyyimi and Zitikis (2014). In the following chapters, when we need to provide the numerical calculation for our proposed index of increase, we will also embrace the general idea in their numerical procedure, namely, discretization. Qoyyimi and Zitikis (2015) also calculated the Gini-type I_f in an educational dataset by fitting curves to pairs of data points. However, there are two main concerns that we notice, which give us reasons in favour of another distance-based measure. Firstly, the aforementioned indices are not normalized, which gives us a difficulty to interpret the numerical results given that we usually use 0 and 1 as the boundary. Secondly, the calculation of the indices highly relies on the estimation of function f. Namely, it is based on the assumption that we can choose proper curve fitting methods and the data are well fitted. In order to resolve these concerns, we discover another candidate that measures the monotonic relationship (or lack of it) between two variables.

2.1.3 Index of increase motivated from actuarial concept

Another distance-based candidate to measure monotonicity or lack of monotonicity of a function is inspired by the weighted premium calculation principle in insurance context. It also comes naturally from an optimization problem. Davydov and Zitikis (2017) proposed the index of lack of increase (LOI) as the distance between a given function and the set that contains all the non-decreasing functions.

Definition 2.1.2. *(Davydov and Zitikis, 2017) Let $\mathcal{F}$ denote the set of all absolutely continuous functions f on the interval $[0,1]$ such that $f(0)=0$. Denote the total variation of $f \in \mathcal{F}$ on the interval $[0,1]$ by $\|f\|$, that is, $\|f\| = \int_0^1 |f'|\,\mathrm{d}\lambda$. Furthermore, let $\mathcal{F}^+$ denote the set of all $f \in \mathcal{F}$ that are non-decreasing. For any $g \in \mathcal{F}$, we define its* LOI *as*

$$\mathrm{LOI}(g) = \inf\left\{\int_0^1 |g' - f'|\,\mathrm{d}\lambda : f \in \mathcal{F}^+\right\} = \inf_{f\in\mathcal{F}^+} \|g - f\|. \tag{2.1.3}$$

Note 2.1.3. *A function h_0 defined on an interval $[a,b]$ can always be 'standardized' into a function h defined on the interval $[0,1]$ such that $h(0)=0$ by considering $h(t) = h_0(a+(b-a)t) - h_0(a)$. That is why we can introduce Definition 2.1.2 for functions on the unit interval without loss of generality.*

Theorem 2.1.1. *(Davydov and Zitikis, 2017) The infimum in Definition 2.1.2 is attained at a function $f_1 \in \mathcal{F}^+$ such that $f_1' = (g')^+$, and thus*

$$\mathrm{LOI}(g) = \int_0^1 (g')^-\,\mathrm{d}\lambda. \tag{2.1.4}$$

Theorem 2.1.1 gives a simplified but computable representation of LOI. Direct proof of Theorem 2.1.1 that was not provided in Davydov and Zitikis (2017) will be given in Chapter 5. The LOI index has potential to be applied in different areas due to its intuitive and computable definition. For example, in material science, Kirk et al. (2021) use the LOI index as a new cost function to extend the computational design methodology for planning compositionally graded alloys.

From theoretical aspect, we appreciate the beauty of the distance-based indices (2.1.1 – 2.1.4) that quantify monotonicity of functions. In the following chapters, we will adapt the general idea from Definition 2.1.2 and Theorem 2.1.1 but approach them from a different angle to form our index of increase I. We will further investigate how our index of increase is closely related to Definition 2.1.2 but easier to understand and more intuitive. That is, instead of studying the lack of increase, we switch to the lack of decrease which is defined analogously and gives bigger values when the relationship between variables is increasing more. However, when it comes to data, the underlying functions between variables are usually unknown, which poses a significant obstacle to measuring the monotonic

relationship between two variables using the indices in literatures. The index of increase we propose in this thesis can definitely overcome this obstacle, since it is not only defined in terms of given functions but also pairs of data points. We will further illustrate that these two forms of definitions coincide in a special case.

2.2 Background knowledge

2.2.1 Curve fitting

In the following chapters, we will formally introduce the index of increase which is defined based on underlying function between X and Y. In reality, it is almost impossible to know in advance the real underlying function. Generally, there are two ways for us to estimate the index of increase: 1) estimate the underlying function first, either parametric or non-parametric, then apply the definition; 2) tweak the definition to get an estimate with new formulation that only relies on pairs of data. Recall that one of the strengths of the index of increase is to capture unknown nonlinear relationship. Therefore, we briefly review a few non-parametric curve fitting methods that are widely implemented and have flexibility in estimating the underlying regression function from given data.

Generally, one-dimensional regression problem given a response variable Y and a explanatory variable X can be described as

$$Y = f(X) + \epsilon,$$

where ϵ is the noise term, and the function f is not specified. The general goal is to find an estimate of function f, such that the mean square error is minimal:

$$\min_{f \in \mathcal{F}} \mathbf{E}(Y - f(X))^2. \tag{2.2.1}$$

Proposition 2.2.1. $\min_{f\in\mathcal{F}} \mathbf{E}(Y-f(X))^2 = \mathbf{E}(Y-m(X))^2$, *where* $m(x)$ *is the conditional mean*

$$m(x) = \mathbf{E}(Y|X = x) = \int y \frac{f_{X,Y}(x,y)}{f_X(x)}\,\mathrm{d}y = \frac{\int y f_{X,Y}(x,y)\,\mathrm{d}y}{\int f_{X,Y}(x,y)\,\mathrm{d}y}, \tag{2.2.2}$$

assuming there exists a joint distribution between X and Y. That is, $m(X)$ is the optimal predictor of Y given X. As a result, the goal for solving one-dimensional regression problem in (2.2.1) turns into estimating the conditional mean $m(x)$.

The next step is to estimate the numerator and denominator in (2.2.2). In statistics, kernel density estimation is a non-parametric technique to estimate the probability density functions, either univariate or multivariate. This approach leads to the most popular non-parametric regression estimator, the Nadaraya-Watson estimator (Watson, 1964; Nadaraya, 1965), which essentially is a local constant kernel regression estimator. The Nadaraya-Watson estimate at each target point x is written as

$$\hat{m}_\lambda(x) = \frac{\sum_{i=1}^{N} K_\lambda(x, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x, x_i)}, \tag{2.2.3}$$

where $K_\lambda(\cdot)$ is the kernel function with smoothing/tuning parameter λ, which controls the width of the local neighbourhood and needs to be specified. The bigger the λ is, the smoother the fitted line is. The estimator in (2.2.3) can also be viewed as the weighted sum of response variable in a local neighbourhood of target point x, which means the underlying local approximation of the target point x is a constant. Generally, the kernel function treats the points from the left and from the right of target point x equally. These properties may lead to biased estimates when the X values are not equally spaced, especially around the boundaries.

The bias can be reduced by fitting a local polynomial regression instead of a local constant. Hence, we introduce the locally weighted polynomial regression, which is known as the LOESS method for curve fitting. For the following chapters in this dissertation, the LOESS method will be used to estimate the underlying relationships between variables, and the index of increase will be applied after. The LOESS fitting process is similar to the Nadaraya-Watson estimator. Firstly, at each point x in the data, the local neighbourhood around x and weights for points in the neighbourhood are defined based on the smoothing parameter λ and the kernel functions. Secondly, a low-degree polynomial is fitted using weighted least squares, giving more weights to points near x and less weights to points

farther away:

$$\min_{\alpha(x),\beta(x),j=1,\ldots,d} \sum_{i=1}^{N} K_\lambda(x, x_i)[y_i - \alpha(x) - \sum_{j=1}^{d} \beta_j(x) x_i^j]^2$$

with solution $\hat{f}(x) = \hat{\alpha}(x) + \sum_{j=1}^{d} \hat{\beta}_j(x) x^j$. Theoretically, the degree of polynomial is not limited to first or second degree (i.e., linear or quadratic). Nevertheless, using higher degree of local polynomial tends to overfit the data in each neighbourhood. As a result, in most cases, using linear or quadratic function will be sufficient.

In this dissertation, we implement the `loess` function in `stats` (R Core Team, 2017) in R software to generate LOESS fit. The default setting for local polynomial is with degree 2. The smoothing parameter λ that controls the size of the neighbourhood is defined as the proportion of the points, if $\lambda < 1$. And the kernel function has the form

$$K_\lambda(x, x_i) = D(\frac{|x - x_i|}{h_\lambda(x)}),$$

where $h_\lambda(x)$ is the maximum distance between target point x and the points in the neighbourhood. $D(t)$ is the tri-cube weighting function:

$$D(t) = \begin{cases} (1 - |t|^3)^3 & \text{if } |t| < 1; \\ 0 & \text{otherwise.} \end{cases}$$

2.2.2 Cross validation

In Chapter 5, we will introduce an adjusted estimator for the index of increase under the situation where the data contains deterministic trend and random errors. This adjusted estimator relies on a smoothing/grouping parameter as well. Similar to the kernel regression which requires to decide the bandwidth λ, the grouping parameter of the index estimator can be determined by a popular numerical technique, that is, cross validation.

The core of cross validation is to partition the sample data into a training set and a validation set (or test set). Then, we use the training set for fitting the statistical models, and the validation set for evaluating the performance. Through the validation set, we can have a better idea about how the statistical models can be generalized to unknown data.

We usually repeat the process several times to average out the influence of data partition. Given these motivations, considering the ways that we separate the dataset and whether we repeat the validation process or not, we have a few typical validation algorithms stated as follows:

- **Holdout method**. Using this method, we randomly separate the data into a training set used for model training and a test set used for calculating statistical measures that quantify the model performance. Thus, the validation results depend on the random splits. Mostly, the test set contains less data points than the training set. Notice that the validation process is a single run, one may get misleading results because of an "unfortunate" split. Hence, it leads to the application of the next method.

- **Repeated random subsampling method**. It repeatedly generates random splits of training and validation sets with a fixed size. For each split, similar processes in the holdout method are executed. The final measurement of performance is then averaged over the splits. In this way, the potential misleading results due to an "unfortunate" split can be reduced, even though the results still depend on the randomness. Also, since the partition is randomly decided at each split, the validation sets might be overlapped, meaning that we may not be able to fully utilize the original dataset to do the validation. To improve the cross validation, we can add more restrictions to the data partitioning process, thus moving to the next method.

- **K-fold cross validation.** Unlike the previous method, we randomly divide the original sample data into k groups with the same number of samples. For each group, we treat it as the test set and the remaining groups as the training set. Then, we conduct the model training and validation procedures. After looping through all groups, model evaluation measures are averaged across. In this way, we ensure that each data point appears in the test set once and $k-1$ times in the training set, which makes full use of all the data points. The choice of number of folds depends on the size of dataset. There is a trade-off between bias and variance when deciding the

number of folds. For example, if the number of folds is large, then we can reduce the bias of the estimate of model performance measures while increasing the variance. In practice, a common choice of fold number is $k = 10$ or $k = 5$ (James et al., 2013, pp. 184).

- **Leave-p-out cross validation.** If the number of folds in k-fold method becomes the number of samples, then the k-fold method degenerates to the leave-one-out method. Its general case is called the leave-p-out cross validation. We firstly find all the possible ways to choose p observations from original sample. Then we loop through those combinations where we use the p observations as the validation set and the remaining points as the training set.

More details can be found in James et al. (2013, Chapter 5). In this dissertation, we choose the k-fold cross validation, which is a popular and intuitive method.

2.2.3 Bootstrapping

Bootstrapping is a resampling technique used to obtain estimates of summary statistics, such as standard error and confidence interval, especially when asymptotic distribution is unknown for an estimator. In our case, even though the consistency of the estimator of the index of increase is proved in Chapter 5, the sampling distribution of the estimator is still unknown and complicated to derive. However, from the point of view of statistical inference, standard error or confidence interval of an estimator is important for completing the inference procedure, as we have no idea how good or poor the point estimate is in representing the population. That is, the point estimate contains less information and cannot reveal the uncertainty associated with it. Consequently, we need to implement the bootstrap method to estimate its confidence interval to align with the standard statistical procedures and provide the statistical inference.

Efron (1979) first developed the bootstrap method inspired by the previous jackknife method proposed by Quenouille (1949). Efron (1979) also pointed out that jackknife is a linear approximation of the bootstrap method. The main idea of jackknife method is

similar to the idea of leave-one-out cross validation. To get the jackknife estimator, given samples $X_1, X_2, ..., X_n$, the population parameter of interest T, and the statistics of interest $T_n = g(X_1, ..., X_n)$, we can calculate the leave-one-out estimators T_{-i}, for $i = 1, 2, ..., n$. Then we average all the T_{-i} to get the jackknife estimator $\frac{1}{n}\sum_{i=1}^{n} T_{-i}$. And the jackknife standard error is defined as

$$\mathrm{SE}_{\mathrm{jack}} = \left\{ \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{-i} - \frac{1}{n} \sum_{i=1}^{n} T_{-i} \right)^2 \right\}^{1/2}.$$

However, it is well-known that the jackknife method fails when we consider the sample median case. Efron (1979) provided the explanation that the naive bootstrap method can correctly estimate the variance of sample median. The main difference between jackknife and naive bootstrap is the resampling method. For naive bootstrap, we repeatedly resample n points $X_1^b, X_2^b, ..., X_n^b$ from the original sample $X_1, ..., X_n$ for B times. For each new sample, we calculate the statistic of interest $T_{n,b}, b = 1, ..., \mathrm{B}$, then $\frac{1}{B}\sum_{k=1}^{B} T_{n,b}$ as the bootstrap estimator. The standard error of the bootstrap estimator is defined as

$$\mathrm{SE}_{\mathrm{boot}} = \left\{ \frac{1}{B} \sum_{k=1}^{B} \left(T_{n,b} - \frac{1}{B} \sum_{k=1}^{B} T_{n,b} \right)^2 \right\}^{1/2}.$$

If we know that the asymptotic distribution of T_n is close to the normal distribution, it is natural to construct a normal-distribution-based confidence interval for the population parameter T. That is, the $(1-\alpha) \times 100\%$ confidence interval of T has the following form

$$T \in (T_n - z_{\alpha/2} \times \mathrm{SE}_{\mathrm{boot}},\ T_n + z_{1-\alpha/2} \times \mathrm{SE}_{\mathrm{boot}}).$$

Nevertheless, the mystery of the sampling distribution is the reason why we want to implement non-parametric resampling techniques, such as bootstrapping. We therefore are in favour of a more general format of the $(1-\alpha) \times 100\%$ confidence interval which is based on the sample quantiles of the $T_{n,1}, T_{n,2}, ..., T_{n,B}$, i.e.,

$$T \in (T^*_{\alpha/2}, T^*_{1-\alpha/2}), \tag{2.2.4}$$

where $T^*_{\alpha/2}$ is the $\alpha/2$ quantile and $T^*_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of $T_{n,1}, T_{n,2}, ..., T_{n,B}$.

Even though the naive bootstrap method can solve the problems that jackknife has, there are still cases where the naive bootstrap will fail. For example, Bickel et al. (1997) summarized a few examples where the naive bootstrap method fails, such as finding the confidence bound for an extremum, and so on. In the case of approximating the distribution of trimmed mean, the naive bootstrap also fails (Gribkova and Helmers, 2011). Other theoretical details about bootstrap methods and examples of successes and failures of bootstrap can be found in DasGupta (2008), Chapter 29. As a result, researchers argued in favour of the so-called m out of n bootstrap method to overcome the failure of the naive bootstrap (Bickel et al., 1997; Wu, 1990; Politis and Romano, 1994; Gribkova and Helmers, 2007). Hence, we decided to use the m out of n bootstrap method to produce the confidence interval of the index of increase estimator in Chapter 5. We will introduce the steps in details when we reach to this topic again in Chapter 5.

2.3 Problem statement

To show how we can apply the proposed index of increase and what insightful conclusions we can draw from it, we decided to apply our index of increase to two educational datasets in this thesis. For each dataset, we extract different information and conclusions using the index of increase. Educational data are not only meaningful but also more accessible. And the analyses are more straightforward for both educators and people who are not familiar to educational context to understand.

2.3.1 Assessing monotonicity and comparing students performance between boys and girls

Mathematics, spelling, and reading are the most important subjects for elementary education. For teachers, measuring the association between any two of these subjects can help them understand the students' studying behaviour better so that they can adjust the effort they put in a certain subject. For example, if reading scores tend to increase

more when spelling scores increase, then the teacher can put more effort into spelling class to make both subjects scores increase efficiently. This situation makes common sense because spelling is the fundamental part of reading. Also, comparing the performance between boys and girls is always an interesting topic for both educators and phycologists. For educators, if boys or girls perform significantly differently, they may adjust their way of teaching for those who fall behind to catch up. For psychologists, they can detect significant discrepancies of performance between boys and girls using the index of increase. They can also use the index of increase in other experiments where they further explore the factors that trigger different behaviours between boys and girls.

As a result, the index of increase fits the purpose of usage since it provides numbers for comparison. Also, the index of increase can be applied either directly to pairs of data points or combined with other popular curve-fitting techniques. This property gives us flexibility when we use the index. Most importantly, given the definition of index of increase, we will see it is as interpretable as those traditional statistical tools.

2.3.2 Quantifying monotonicity and interchangeability to help curriculum development

Constructing a reasonable curriculum benefits not only educators but also students. Naturally, deciding which courses should be taught first and which courses can be taught at the same time becomes an interesting research topic for educators. However, deciding which course is more basic and which course has greater influence on other courses are complex tasks without a proper measure. As we mentioned, trends for student scores between two subjects can be non-linear, non-monotonic, and asymmetric, which violates the foundations of using traditional statistical tools.

Taking advantage of a natural property that the index of increase is asymmetric, we can apply it to provide scientific and statistical support for curriculum development by analyzing students' scores in different subjects such as Mechanics, Vectors, Algebra, Analysis, and Statistics. More specifically, we can define a measure that quantifies the interchangeability between two variables by considering the absolute difference or relative

difference between $\mathrm{I}(\mathbf{x}, \mathbf{y})$ and $\mathrm{I}(\mathbf{y}, \mathbf{x})$. If the index of interchangeability between two subjects is relatively low, it means the two subjects are more interchangeable, since the difference between $\mathrm{I}(\mathbf{x}, \mathbf{y})$ and $\mathrm{I}(\mathbf{y}, \mathbf{x})$ is low. As a result, the performance of one subject is not strongly correlated to the other. In other words, they are not the "prerequisite" courses for each other. In this case, a student can construct his or her "educational portfolio" by choosing these two subjects at the same time. For schools, subjects that are more interchangeable (i.e., with a lower index of interchangeability) can be scheduled in the same semester. In contrast, if two subjects are less interchangeable, students may need to take them one after another (i.e., take the prerequisite course first).

From this aspect, we see a potential extension of the index of increase. We can not only calculate the index itself as a measure, but also develop other meaningful measures to describe and assess the relationships between variables.

2.3.3 Estimator of index of increase: balancing deterministic and random data

Educational dataset has a limited sample size. Yet, it does not stop us from thinking about using the index of increase under a large sample situation. Moreover, rather than considering deterministic data points, we would like to start considering applying the index of increase to the situations where measurement errors may exist. That is, the deterministic index of increase $\mathrm{I}(\mathbf{x}, \mathbf{y})$ will start embracing a random component since measurement errors usually appear in the form of random variable with a certain distribution. Firstly, we need to construct the estimator of the index of increase from data. Of course, we need to investigate if we can still use the same form as the deterministic case. If not, we need to create a proper estimator along with some adjustments.

Following the traditional path in statistics, we are curious what the index of increase estimator's convergence performance will be, as it is essentially a statistic. Namely, we want to know if the index of increase estimator generated from a sample with measurement errors or other noises will converge to its true value (without measurement error) when the sample size is large enough. Intuitively, our estimator should be able to smooth out the

measurement error. That is, the estimator should be capable of balancing the deterministic part and the random part of the data.

Generally, there are four types of convergence of random variables: convergence in probability, convergence in distribution, almost sure convergence, and convergence in $L_p - mean$. Convergence in probability will be our main concern because (weak) consistency is one of the most important properties of an estimator. Also, we want to know the speed of convergence which will give us a hint that how large a dataset should be to apply the index of increase estimator properly. After exploring the consistency, it is also crucial to construct the confidence interval for our estimator because we want both the point estimate and interval estimate to conclude the statistical inference.

In general, our problem becomes the following: does the novel index of increase estimator we propose have not only high interpretability but also a solid statistical background?

2.4 Notation

Lastly, we conclude the current chapter with notations in this thesis. Throughout this thesis we use the notation I when discussing the index of increase in generic terms. When the index is applied on scatterplots, we tend to use the notation $\mathrm{I}(\mathbf{x}, \mathbf{y})$ (e.g., definition (3.3.2 and 4.4.4)), where $\mathbf{x}$ and $\mathbf{y}$ are n-dimensional vectors of explanatory and response variables, respectively. When the explanatory data $\mathbf{x} = (x_1, \ldots, x_n)$ do not have ties (i.e., $x_i \neq x_j$, $1 \leq i, j \leq n$), we emphasize this fact by using the notation $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ (e.g., definition (3.3.1 and 4.4.3)). When the index is calculated from fitted to data functions, denoted by h, we use the notation $\mathrm{I}(h)$ for the corresponding index (e.g., definition (3.4.1 and 4.4.1)). A numerical approximation for $\mathrm{I}(h)$ is denoted by $\widehat{\mathrm{I}}_k(h)$ (e.g., definition (3.4.2 and 4.4.2)), with the latter approaching $\mathrm{I}(h)$ when k gets larger. In the process of analysis, we sometimes find it useful to restrict explanatory variables to certain regions, say intervals $[L, U]$, and then calculate the corresponding index values. In such instances, we denote the index by $\mathrm{I}(\mathbf{x}, \mathbf{y} \mid L, U)$ for scatterplots (e.g., definition (3.5.7 and 4.4.7)) and $\mathrm{I}(h \mid L, U)$ for fitted functions h.

Remark 2.4.1. *The notation is revealing: our dataset is in the form of scatterplots, which we sometimes analyze as they are, but sometimes truncate to certain sub-scatterplots (e.g., with explanatory variables restricted to some intervals $[L, U]$), or to which we sometimes fit continuous functions and then analyze the functions. There are several reasons for such transformation, one of them being outliers, whose ability to distort statistical analyses and thus decision making should not be underestimated. We shall illustrate this point with an example in Section 4.4.2, where we illustrate a property of the index of increase.*

Chapter 3

Measuring and comparing student performances between genders

3.1 Motivation

Measuring and comparing student performance have been topics of much interest for educators and psychologists. Are higher marks in Mathematics indicative of higher marks in other subjects such as Reading and Spelling? Alternatively, are higher marks in Reading or Spelling indicative of higher marks in Mathematics? Do boys and girls exhibit similar associations between different study subjects? These are among the many questions that have interested researchers. The literature on these topics is vast, and we shall therefore note only a few contributions; their lists of references lead to earlier results and illuminating discussions.

Given school curricula, researchers have particularly looked at student performance in Mathematics, Science, Reading, and Writing (e.g., Ma, 2001; Masci et al., 2017; Newman and Stevenson, 1990), and explored whether or not there are significant differences with respect to gender (e.g., Jovanovic and King, 1998; McCornack and McLeod, 1988; Mokros and Koff, 1978). Differences between other attributes have also been looked at, including teacher performance (e.g., Alexander et al., 2017), oral and written assessments (e.g., Huxham et al., 2012), spatial and verbal domains (e.g., Bresgi et al., 2017). As for certain

statistical techniques that are used in these area, such as interclass correlation coefficient (ICC) and Spearman's ρ, we already gave a few references in Chapter 1. We also outlined their limitations there.

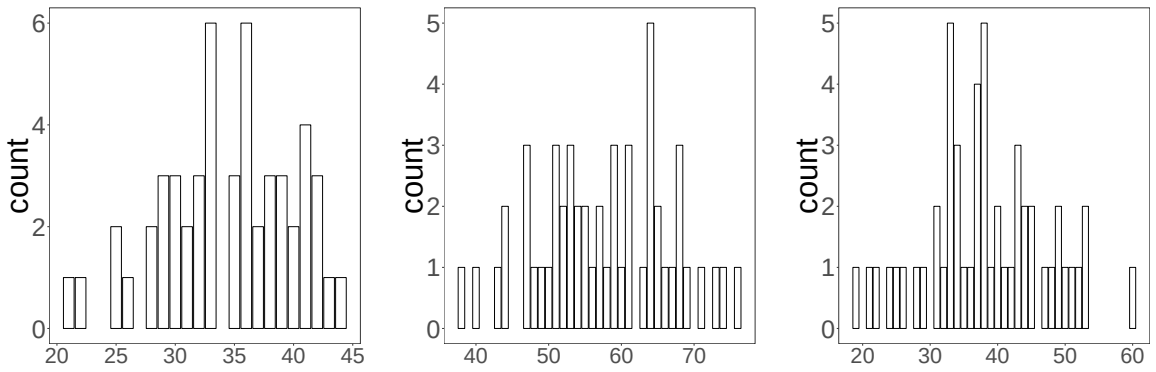
We have organized the rest of the chapter as follows. In Section 3.2 we describe a data set that we use to explore the new technique. In Section 3.3 we introduce an index for assessing directional associations in raw data, and illustrate the performance of the index. For those wishing to employ the index in conjunction with classical techniques of curve fitting, such as LOESS or other regression methods, in Section 3.4 we provide a recipe for accomplishing the task and so, for example, the proposed index can also be used as a summary index for the LOESS and other fitted curves. In Section 3.5 we demonstrate how the index of increase facilitates insights into student performance and enables comparisons between different student groups. In Section 3.6 we give a step-by-step illustration of how the index works on data. Section 3.7 concludes the chapter with a brief overview of our main contributions, and it also contains several suggestions to facilitate further research in the area. Appendix A contains illustrative computer codes and additional data-exploratory graphs.

We note at the outset that throughout this paper we deliberately restrain ourselves from engaging in interpretation and validation of the obtained numerical results, as these are subtle tasks and should be left to experts in psychology and education sciences to properly handle. We refer to Kane (1994, 2013) and references therein for an argument-based approach to validation of interpretations. Throughout this paper, we concentrate on methodological aspects of educational data analysis.

3.2 Data

We begin our considerations from the already available solid classical foundations of statistics and data analysis. To make the considerations maximally transparent, we use the easily accessible data reported in the classical text of Thorndike and Thorndike Christ (2010, pp. 24-25). A brief description of the data with relevant for our research details follow next.

The data set contains scores of 52 sixth-grade students in two classes, which we code by C1 and C2. The sizes of the two classes are the same: the enrollment in each of them is 26 students. Based on the student names, we conclude that there are 15 boys and 11 girls in class C1, and 11 boys and 15 girls in class C2; we code boys by B and girls by G. All students are examined in three subjects: Reading (R), Spelling (S), and Mathematics (M). The maximal scores for different subjects are different: 65 for Mathematics, 45 for Reading, and 80 for Spelling. The frequency plots of the three subjects are in Figure 3.2.1.



(**a**) Reading (**b**) Spelling (**c**) Mathmatics

Figure 3.2.1: Frequency plots of the scores of all students.

Our interest centers around the trends that arise when associating the scores of two study subjects. One of the classical and very common techniques employed in such studies is fitting linear regression lines which, for the sake of argument, we also do in Figure 3.2.2.

We see from Figure 3.2.2 that all the fitted lines exhibit increasing trends, with the slopes of Spelling vs. Reading, Reading vs. Spelling, Mathematics vs. Reading, and Reading vs. Mathematics being similar, with the value of the Pearson correlation coefficient r between 0.62 and 0.64. These slopes are considerably larger than those for Mathematics vs. Spelling and Spelling vs. Mathematics, whose correlation coefficient is $r = 0.15$. Of course, the coefficient is symmetric with respect to the two variables, and thus its values for, e.g., Spelling vs. Mathematics and Mathematics vs. Spelling are the same, even though the scatterplots (unless rotated) are different. More generally, we report the values of all the aforementioned correlation coefficients—Pearson, ICC, and Spearman—as

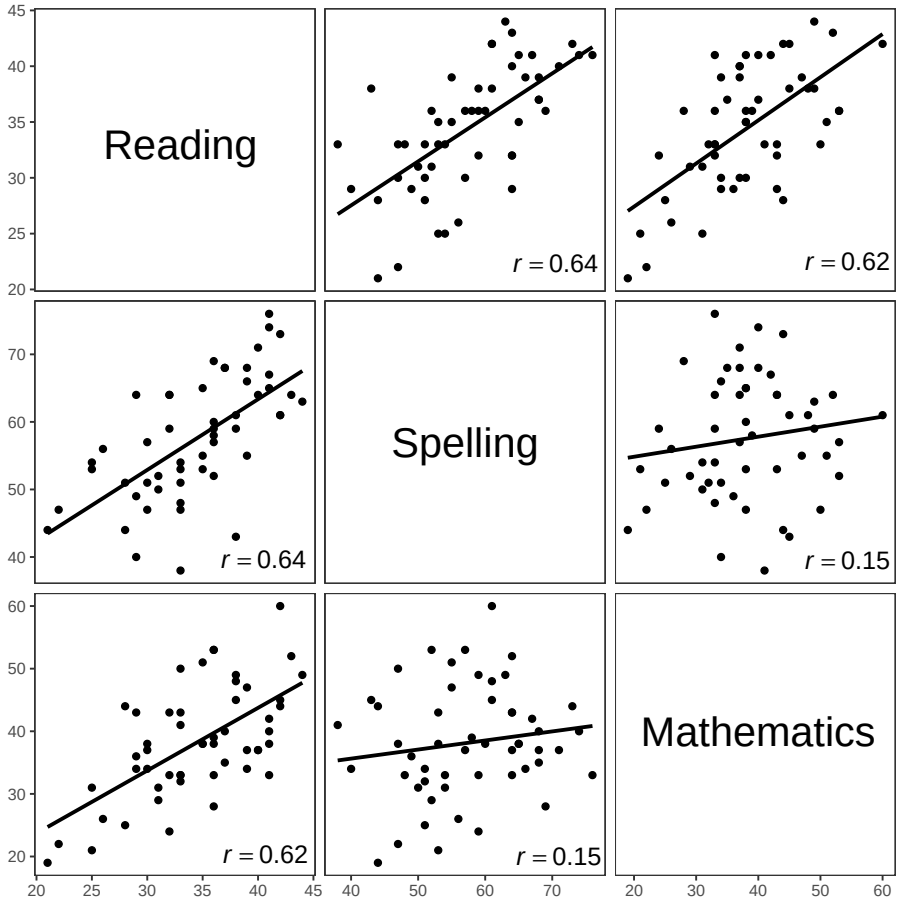


Figure 3.2.2: Scatterplot matrix of the scores of all students.

well as of the index of increase to be introduced in the next section, in Table 3.2.1.

	I	**Pearson**	**ICC**	**Spearman**
R vs. M	0.70	0.62	0.10	0.56
M vs. R	0.57	0.62	0.10	0.56
S vs. M	0.48	0.15	−0.11	0.15
M vs. S	0.53	0.15	−0.11	0.15
R vs. S	0.59	0.64	0.58	0.67
S vs. R	0.53	0.64	0.58	0.67

Table 3.2.1: The index of increase and three classical correlation coefficients.

In addition, we have visualized—in two complementing ways—the data of Thorndike

and Thorndike Christ (2010, pp. 24-25) in Figures 3.2.3 and 3.2.4.

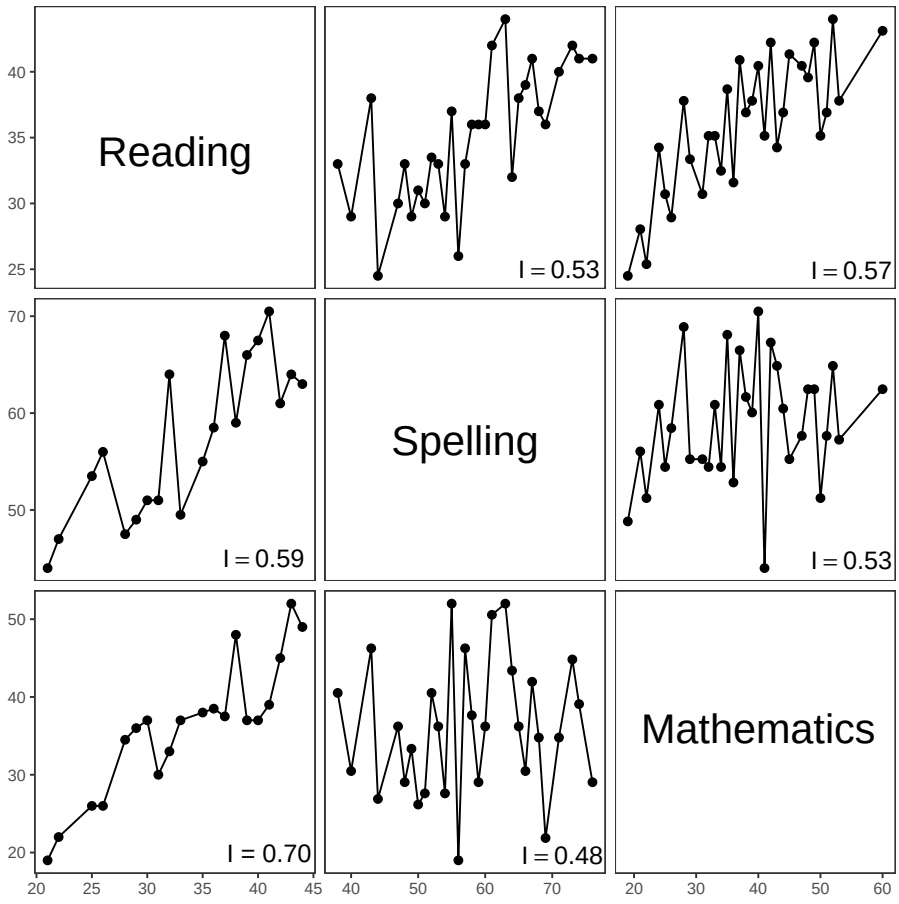


Figure 3.2.3: Scatterplot matrix of piece-wise linear fits to the scores of all students.

A close inspection of the figures suggests non-linear and especially non-monotonic relationships. In Figure 3.2.4, we have used one of the most popular regression methods, called LOESS. Specifically, we have employed the `loess` function of the R Stats Package R Core Team (2013) with the default parameter value `span` = 0.75. There are of course numerous other regression methods that we can use (e.g., Koenker et al., 2017; Young, 2017, and references therein). For example, the quantile regression method has recently been particularly popular in education literature (e.g., Haile and Nguyen, 2008; Castellano and Ho, 2013; Dehbi et al., 2015). These methods, however, are inherently smoothing methods and thus provide only general features of the trends, whereas it is the minute details that facilitate unhindered answers to questions such as those posited at the

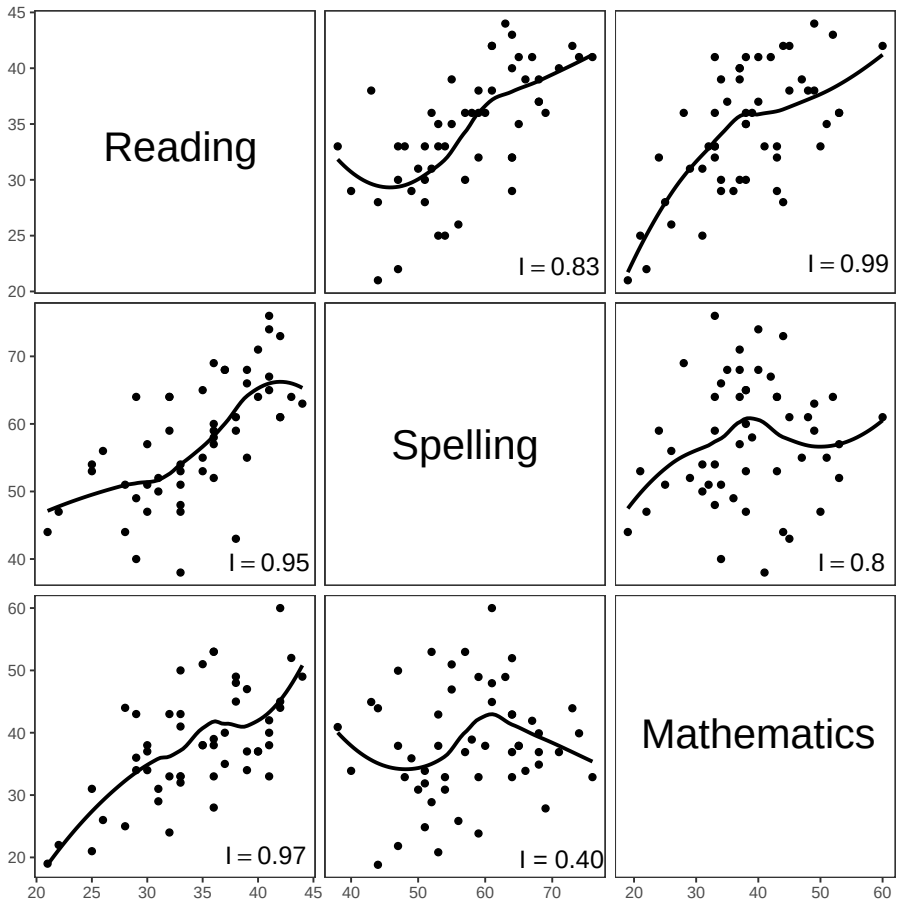


Figure 3.2.4: The fitted LOESS curves (`span` = 0.75) to the scores of all students.

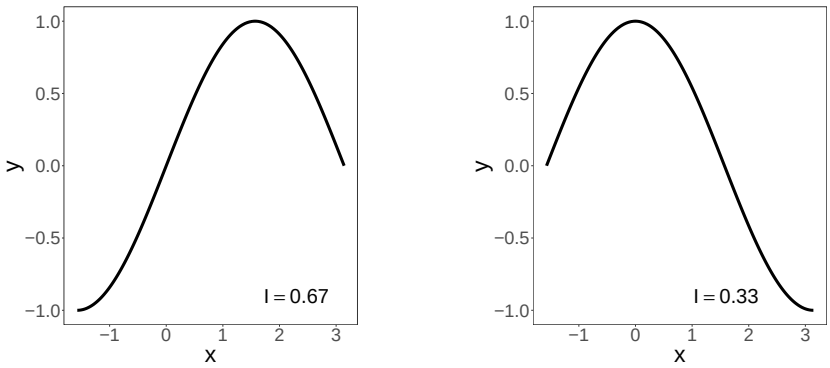
very beginning of Section 3.1.

Furthermore, when dealing with small data sets, which are common in education and psychology, we cannot reliably employ classical statistical inference techniques, such as goodness-of-fit, to assess the performance of curve fitting techniques. In this thesis, therefore, we advocate a technique that facilitates unaltered inferences from data sets of any size, $n \geq 2$. We shall discuss and illustrate the technique using the classical Thorndike and Thorndike Christ (2010, pp. 24-25) data sets, and we shall also use artificial data designed specifically for illuminating the workings of the new technique in a step-by-step manner.

3.3 Index of increase

We see from Figures 3.2.3 and 3.2.4 that trends are mostly non-monotonic. In such cases, how can we assess which of the trends are more increasing than others? We can fit linear regression lines as in Figure 3.2.2 and rank them according to their slopes or, alternatively, on the values of the Pearson correlation coefficient. However, the non-linear and especially non-monotonic trends make such techniques inadequate (e.g., Wilcox, 2001). In this section, therefore, we put forward the idea of an index of increase, whose development has been in the works for a number of years (Davydov and Zitikis, 2005, 2017; Qoyyimi and Zitikis, 2014, 2015).

To illuminate the idea, we use a very simple yet informative example. Consider two very basic trigonometric functions, $\sin(z)$ and $\cos(z)$ on the interval $[-\pi/2, \pi]$, depicted in Figure 3.3.1.



(a) $y = \sin(x)$. (b) $y = \cos(x)$.

Figure 3.3.1: Two functions and their indices of increase.

Obviously, neither of the two functions is monotonic on the interval, but their visual inspection suggests that sine must be closer to being increasing than cosine. Since one can argue that this assessment is subjective, we therefore employ the aforementioned index whose rigorous definition will be given in a moment. Using the computational algorithm presented in Section 3.4 below, the values of the index are $2/3 \approx 0.67$ for $\sin(x)$ and $1/3 \approx 0.33$ for $\cos(x)$. Keeping in mind that 3 is the normalizing constant, these values imply that sine is at the distance 2 from the set of all decreasing functions on the noted

interval, whereas cosine is at the distance 1 from the same set. In other words, this implies that sine is at the distance 1 from the set of all increasing functions on the interval, whereas cosine is at the distance 2 from the set of increasing function. Inspecting the graphs of the two functions in Figure 3.3.1, we indeed see that sine is 'twice more increasing' than cosine on the interval $[-\pi/2, \pi]$. For those willing to experiment with their own functions on various intervals, we provide a computer code in Appendix A.1.1.

The rest of the chapter is devoted to a detailed description and analysis of the index.

3.3.1 Basic idea

Suppose we possess $n \geq 2$ pairs $(x_1, y_1), \ldots (x_n, y_n)$ of data (all the indices to be introduced below can be calculated as long as we have at least two pairs), and let—for a moment—all the first coordinates (i.e., x's) be different, as well as all the second coordinates (i.e., y's) be different. Consequently, we can order all the first coordinates in the strictly increasing fashion, thus obtaining $x_{1:n} < \cdots < x_{n:n}$, called order statistics, with the corresponding second coordinates $y_{[1:n]}, \ldots, y_{[n:n]}$, called concomitants (e.g., David and Nagaraja, 2003, and references therein). Hence, instead of the original pairs, we are now dealing with the pairs $(x_{1:n}, y_{[1:n]}), \ldots (x_{n:n}, y_{[n:n]})$ ordered according to their first coordinates. We connect these pairs, viewed as two-dimensional points, using straight lines and, unlike in regression, obtain an unaltered genuine description of the trend exhibited by the first coordinates plotted against the second ones (see Figure 3.2.3). Having the plot, we can now think of a method for measuring its monotonicity, and for this purpose we use the index of increase

$$\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=2}^{n} (y_{[i:n]} - y_{[i-1:n]})_+}{\sum_{i=2}^{n} |y_{[i:n]} - y_{[i-1:n]}|}, \tag{3.3.1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, and z_+ denotes the positive part of real number z, that is, $z_+ = z$ when $z > 0$ and $z_+ = 0$ otherwise. The superscript "0" means that there are no ties among coordinates.

Obviously, the index is not symmetric, that is, $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ is not, in general, equal to $\mathrm{I}^0(\mathbf{y}, \mathbf{x})$. This is a natural and desirable feature in the context of the present chapter because student performance on different subjects is not interchangeable, and we indeed

see this clearly in the figures above. The symmetry of the Pearson, ICC and Spearman correlation coefficients is one of the features that makes their uses inappropriate in a number of applications, especially when the directionality of associations is of particular concern. In general, there is a vast literature on the subject, which goes back to at least the seminal works of C. Gini a hundred years ago (e.g., Giorgi, 1990, 1993) and, more generally, to the scientific rivalry between the British and Italian schools of statistical thought. For very recent and general discussions on the topic, we refer to Reimherr and Nicolae (2013); as well as to Furman and Zitikis (2017), and references therein, where a (non-symmetric) Gini-type correlation coefficient arises naturally and plays a pivotal role in an insurance/financial context.

To work out initial understanding of the index $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$, we first note that the numerator on the right-hand side of Equation (3.3.1) sums up all the upward movements, measured in terms of concomitants, whereas the denominator sums up the absolute values of all the upward and downward movements. Hence, the index $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ is normalized, that is, always takes values in the interval $[0, 1]$. It measures the proportion of upward movements in the piece-wise linear plot originating from the pairs $(x_1, y_1), \ldots (x_n, y_n)$. Later in this chapter (Note 3.4.2), we give an alternative interpretation of, and thus additional insight into, the index $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$, which stems from a more general and abstract consideration of Davydov and Zitikis (2017). We note in passing that the index employed by Qoyyimi and Zitikis (2015) is not normalized, and is actually difficult to meaningfully normalize, thus providing yet another argument in favour of the index that we employ in the present thesis. A cautionary note is in order.

Namely, definition (3.3.1) shows that the index $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ can be sensitive to outliers (i.e., very low and/or very high marks). The sensitivity can, however, be diminished by removing a few largest and/or smallest observations, which would mathematically mean replacing the sum $\sum_{i=2}^{n}$ in both the numerator and the denominator on the right-hand side of Equation (3.3.1) by the truncated sum $\sum_{i=2+\kappa_1}^{n-\kappa_2}$ for some integers $\kappa_1, \kappa_2 \geq 1$. This approach to dealing with outliers has successfully worked in Statistics, Actuarial Science, and Econometrics, where sums of order statistics and concomitants arise frequently (e.g.,

Brazauskas et al., 2007, 2009, and references therein). In our current educational context, the approach of truncating the sum is also natural, because exceptionally well and/or exceptionally badly performing students have to be, and usually are, dealt with on the individual basis, instead of treating them as members of the statistically representative majority. This approach of dealing with outliers is very common and, in particular, has given rise to the very prominent area called Extreme Value Theory (e.g., De Haan and Ferreira, 2006; Reiss and Thomas, 2007, and references therein) that deals with various statistial aspects associated with exceptionally large and/or small observations.

We next modify the index $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ so that its practical implementation would become feasible for all data sets, and not just for those whose all coordinates are different.

3.3.2 A practical modification

The earlier made assumption that the first and also the second coordinates of paired data are different prevents the use of the above index on many real data sets, including that of Thorndike and Thorndike Christ (2010, pp. 24-25) as we see from the frequency histograms in Figure 3.2.1. See also panel (d) in Figure 1.1.1 for another example. Hence, a natural question arises: how can piece-wise linear plots be produced when there are several concomitants corresponding to the single value of a first coordinate? To overcome the obstacle, we suggest using the median-based approach that we describe next.

Namely, given $n \geq 2$ arbitrary pairs $(x_1, y_1), \ldots (x_n, y_n)$, the order statistics of the first coordinates are $x_{1:n} \leq \cdots \leq x_{n:n}$ with the corresponding concomitants $y_{[1:n]}, \ldots, y_{[n:n]}$. Let there be m ($\leq n$) distinct values among the first coordinates, and denote them by $x_1^*, \ldots x_m^*$. For each x_i^*, there is always at least one concomitant, usually more, whose median we denote by y_i^*. Hence, we have m pairs $(x_1^*, y_1^*), \ldots (x_m^*, y_m^*)$ whose first coordinates are strictly increasing (i.e., $x_1^* < \cdots < x_m^*$) and the second coordinates are unique. Note that x_i^* actually means $x_{i:m}^*$ and y_i^* is equivalent to $y_{[i:m]}^*$. We use x_i^* and y_i^* for simplicity. We connect these m pairs, viewed as two-dimensional points, with straight lines and obtain a piece-wise linear plot (e.g., Figure 3.2.3). The values of the index $\mathrm{I} := \mathrm{I}(\mathbf{x}, \mathbf{y})$ reported in

the right-bottom corners of the panels of Figure 3.2.3 refer to the following modification

$$\mathrm{I}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=2}^{m}(y_i^* - y_{i-1}^*)_+}{\sum_{i=2}^{m}|y_i^* - y_{i-1}^*|} \tag{3.3.2}$$

of the earlier defined index of increase. Note that index (3.3.2) collapses into index (3.3.1) when all the coordinates of the original data are different, thus implying that definition (3.3.2) is a genuine extension of definition (3.3.1), and it works on all data sets.

Property 3.3.1. *The index of increase is translation and scale invariant, which means that the equation*

$$\mathrm{I}(\boldsymbol{x},\boldsymbol{y}) = \mathrm{I}(\alpha + \beta\boldsymbol{x}, \delta + \gamma\boldsymbol{y}) \tag{3.3.3}$$

holds for all real 'locations' α and δ, and all positive 'scales' $\beta > 0$ and $\gamma > 0$.

Proof. Since β is positive, the order of the coordinates of the vector $\alpha + \beta\mathbf{x}$ is the same as the order of the coordinates of the vector $\mathbf{x}$, and the relationship between the order statistics is $(\alpha + \beta x)_{i:n} = \alpha + \beta x_{i:n}$. Consequently, and also recalling that γ is positive, all the median-adjusted concomitants satisfy the relationship $(\delta + \gamma y)_i^* = \delta + \gamma y_i^*$ and so

$$\begin{aligned}
\mathrm{I}(\alpha + \beta\mathbf{x}, \delta + \gamma\mathbf{y}) &= \frac{\sum_{i=2}^{m}(\delta + \gamma y_i^* - \delta - \gamma y_{i-1}^*)_+}{\sum_{i=2}^{m}|\delta + \gamma y_i^* - \delta - \gamma y_{i-1}^*|} \\
&= \frac{\gamma\sum_{i=2}^{m}(y_i^* - y_{i-1}^*)_+}{\gamma\sum_{i=2}^{m}|y_i^* - y_{i-1}^*|} \\
&= \frac{\sum_{i=2}^{m}(y_i^* - y_{i-1}^*)_+}{\sum_{i=2}^{m}|y_i^* - y_{i-1}^*|} \\
&= \mathrm{I}(\mathbf{x},\mathbf{y}).
\end{aligned}$$

This concludes the proof of Property 3.3.1.

□

The property is handy because it allows us to unify the scales of each subject's scores, which are usually different. In our illustrative example, the Mathematics scores are from 0 to 65, the Reading scores are from 0 to 45, and those of Spelling are from 0 to 80. Due to Property 3.3.1, we can apply—without changing the value of $\mathrm{I}(\mathbf{x},\mathbf{y})$—the

linear transformation $x_i/x_{\max} \times 100\%$ on the original scores x_i, thus turning them into percentages, where $x_{\max}$ is the maximal possible score for the subject under consideration, like 65 for Mathematics.

3.3.3 Discussion

To facilitate a discussion of the values of the index $\mathrm{I}(\mathbf{x}, \mathbf{y})$ reported in the panels of Figure 3.2.3, we organize the values in the tabular form as follows:

$$\begin{aligned} \mathrm{I}(\mathrm{R}, \mathrm{M}) &= 0.70 \qquad \mathrm{I}(\mathrm{M}, \mathrm{R}) = 0.57 \\ \mathrm{I}(\mathrm{S}, \mathrm{M}) &= 0.48 \qquad \mathrm{I}(\mathrm{M}, \mathrm{S}) = 0.53 \\ \mathrm{I}(\mathrm{R}, \mathrm{S}) &= 0.59 \qquad \mathrm{I}(\mathrm{S}, \mathrm{R}) = 0.53 \end{aligned}$$

The value 0.70 for Reading vs. Mathematics is largest, thus implying the most increasing trend among the panels. The value suggests that there is high confidence that those performing well in Reading would also perform well in Mathematics. Note that the value 0.57 for Mathematics vs. Reading is considerably lower than 0.70, thus implying lower confidence that students with better scores in Mathematics would also perform better in Reading.

The index value of Spelling vs. Mathematics is the smallest (0.48), which suggests neither an increasing nor a decreasing pattern, and we indeed see this in the middle-bottom panel of Figure 3.2.3: the scores in Mathematics form a kind of noise when compared to the scores in Spelling. In other words, the curve in the panel fluctuates considerably and the proportion of upward movements is almost the same as the proportion of downward movements. On the other hand, higher scores in Mathematics seem to be slightly better predictors of higher scores in Spelling, with the corresponding index value equal to 0.53.

Finally, the values of Reading vs. Spelling and Spelling vs. Reading are fairly similar, 0.59 and 0.53 respectively, even though higher Reading scores might suggest higher Spelling scores in a slightly more pronounced way than higher Spelling scores would suggest higher Reading scores.

Note 3.3.1. On a personal note, having calculated the indices of increase for the three

subjects and then having looked at the graphs of Figure 3.2.3, the authors of this article unanimously concluded that the trends do follow the patterns suggested by the respective index values. Yet, interestingly, prior to calculating the values and just having looked at the graphs, the authors were not always in agreement as to how much and in what form a given study subject influences the other ones. In summary, even though no synthetic index can truly capture every aspect of raw data, they can nevertheless be useful in forming a consensus about the meaning of data.

3.4 The Index of increase for fitted curves

Raw data may not always be possible to present in the way we have done in Figure 3.2.3, because of a variety of reasons such as ethical and confidentiality. Indeed, nothing is masked in the figure—the raw data can be restored immediately. The fitted regression curves in Figure 3.2.4, however, mask the raw data and can thus be more readily available to the researcher to explore. Therefore, we next explain how the index of increase can be calculated when the starting point is not raw data but a smooth fitted curve, say h defined on an interval $[L, U]$, like those we see in the panels of Figure 3.2.4 (for a computer code, see Appendix A.1.1). Hence, in particular, the index of increase can be used as a summary index for the LOESS or other fitted curves. Namely, the index of increase $\mathrm{I} := \mathrm{I}(h)$ for the function h is defined by the formula

$$\mathrm{I}(h) = \frac{\int_L^U (h'(x))_+ \, \mathrm{d}x}{\int_L^U |h'(x)| \, \mathrm{d}x}, \tag{3.4.1}$$

where h' is the derivative of h. Two notes follow before we resume our main consideration.

Note 3.4.1. Definition (3.4.1) is compatible with that given by Equation (3.3.1). Indeed, let the function h be piece-wise linear with knots $d_1 < \cdots < d_n$ such that the function h is linear on each interval $[d_{i-1}, d_i]$ whose union is equal to $[L, U]$ with $L = d_1$ and $U = d_n$. The derivative $h'(x)$ in this case is replaced by $(h(d_i) - h(d_{i-1}))/(d_i - d_{i-1})$ for all $x \in [d_{i-1}, d_i]$. Index (3.4.1) for this piece-wise linear function is exactly the one on the right-hand side of Equation (3.3.1).

Note 3.4.2. It is shown by Davydov and Zitikis (2017) that the integral $\int_L^U (h'(x))_- \, \mathrm{d}x$ is a distance of the function h from the set of all non-decreasing functions, where z_- denotes the negative part of real number z, that is, $z_- = -z$ when $z < 0$ and $z_- = 0$ otherwise. Hence, the larger the integral, the farther the function h is from being non-decreasing. For this reason, the integral is called by Davydov and Zitikis (2017) the index of lack of increase (LOI), whose normalized version—always taking values between 0 and 1 – is given by the formula

$$\mathrm{LOI}(h) = \frac{\int_L^U (h'(x))_- \, \mathrm{d}x}{\int_L^U |h'(x)| \, \mathrm{d}x}.$$

Since $|z| = z_+ + z_-$ for every real number z, we have the relationship

$$\mathrm{I}(h) = 1 - \mathrm{LOI}(h),$$

which implies that the index of increase $\mathrm{I}(h)$ takes the maximal value 1 when the function h is non-decreasing everywhere; it also follows the other properties noted in Chapter 1.

In general, given any differentiable function h, calculating index (3.4.1) in closed form is time consuming. Hence, an approximation at any pre-specified precision is desirable. This can be achieved in a computationally convenient way as follows. Let d_i for $i = 1, \ldots, k$ be defined by

$$d_i = L + \frac{i-1}{k-1}(U - L),$$

where k is sufficiently large: the larger it is, the smaller the approximation error will be. (We note that the underlying index of increase can be calculated whenever we have at least two pairs of data; the k used throughout the current section is the 'tuning' parameter that governs the precision of numerical integration.) The numerator on the right-hand side of Equation (3.4.1) can be approximated as follows:

$$\begin{aligned}\int_L^U (h^{'}(x))_+ \,\mathrm{d}x &= \sum_{i=2}^{k} \int_{d_{i-1}}^{d_i} (h^{'}(x))_+ \,\mathrm{d}x \\ &\approx \sum_{i=2}^{k} \int_{d_{i-1}}^{d_i} (h^{'}(d_i))_+ \,\mathrm{d}x \\ &= \frac{U-L}{k-1} \sum_{i=2}^{k} (h^{'}(d_i))_+ \\ &\approx \frac{U-L}{k-1} \sum_{i=2}^{k} \frac{(h(d_i)-h(d_{i-1}))_+}{d_i - d_{i-1}} \\ &= \sum_{i=2}^{k} (h(d_i)-h(d_{i-1}))_+.\end{aligned}$$

Likewise, we obtain an approximation for the denominator on the right-hand side of Equation (3.4.1):

$$\int_L^U |h^{'}(x)| \,\mathrm{d}x \approx \sum_{i=2}^{k} |h(d_i)-h(d_{i-1})|.$$

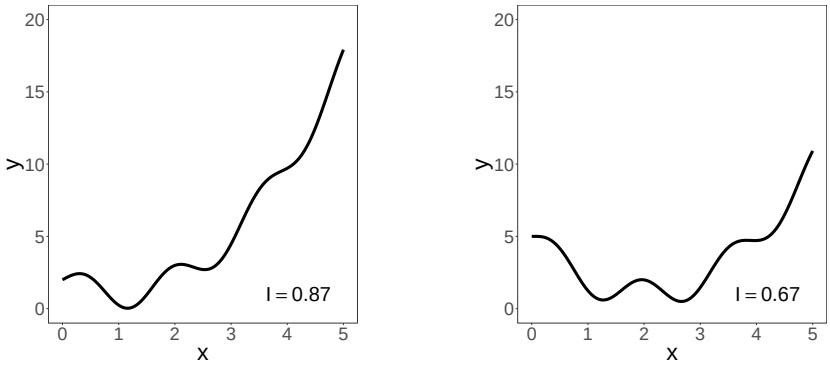
Consequently,

$$\mathrm{I}(h) \approx \widehat{\mathrm{I}}_k(h) := \frac{\sum_{i=2}^{k} (h(d_i)-h(d_{i-1}))_+}{\sum_{i=2}^{k} |h(d_i)-h(d_{i-1})|}. \tag{3.4.2}$$

This approximation turns out to be very efficient, with no time or memory related issues when aided by computers, as we shall see from the next example.

Example 3.4.1. *To illustrate the computation of the index* $\mathrm{I}(h)$ *via approximation (3.4.2), we use the functions* $h_1(x) = (x-1)^2 + 1 + \sin(4x)$ *and* $h_2(x) = (x-2)^2 + 1 + \sin(4x)$ *defined on the interval* $[0, 5]$. *We visualize them in Figure 3.4.1.*

In the figure, the corresponding values of the index of increase are reported in the bottom-right corners of the two panels. They have been calculated using approximation (3.4.2) with $L = 0$, $U = 5$, *and k=10,000. In order to check the approximation performance with respect to* k, *we have produced Table 3.4.1.*



(a) $y = h_1(x)$. (b) $y = h_2(x)$.

Figure 3.4.1: The functions h_1 and h_2 and their indices of increase.

k	h_1	h_2
10	0.9014996	0.6829716
50	0.8719495	0.6670377
500	0.8712794	0.6664844
1000	0.8712688	0.6664817
10,000	0.8712662	0.6664808
20,000	0.8712662	0.6664808
actual	0.8712662	0.6664808

Table 3.4.1: Performance of the approximation $\widehat{\mathrm{I}}_n$.

The 'actual' value in the bottom row of the table is based on Formula (3.4.1) and calculated using the `integrate` *function of the R Stats Package (R Core Team, 2013). We emphasize that the index of increase can be calculated as long as we have at least two pairs of data; the k used in the current example (and throughout Section 3.4) is the 'tuning' parameter that governs the precision of numerical integration. This concludes Example 3.4.1.*

We now come back to Figure 3.2.4, whose all panels report respective values of the index of increase, calculated using approximation (3.4.2) with $k = 10{,}000$. To check whether this choice of k is sufficiently large, we have produced Table 3.4.2.

k	M vs. R	M vs. S	R vs. M	R vs. S	S vs. M	S vs. R
10	1.0000000	0.8182878	0.9776247	1.0000000	0.3866698	0.8259335
50	0.9947776	0.8038232	0.9750816	0.9906566	0.3954920	0.8252337
500	0.9940148	0.8035203	0.9722912	0.9905645	0.3957937	0.8252272
1000	0.9939938	0.8035174	0.9722785	0.9905559	0.3957953	0.8252266
10,000	0.9939889	0.8035172	0.9722651	0.9549793	0.3957954	0.8252262
20,000	0.9939889	0.8035172	0.9722651	0.9549793	0.3957954	0.8252262

Table 3.4.2: Convergence performance of each case in Figure 3.2.4.

Note that the index values when $k = 10{,}000$ and $k = 20{,}000$ are identical with respect to the reported decimal digits, and since in the panels of Figure 3.2.4 we report only the first two decimal digits, we can safely conclude that $k = 10{,}000$ is sufficiently large for the specified precision.

We see from the two bottom rows of Table 3.4.2 that the values for Spelling vs. Reading, Reading vs. Mathematics, Reading vs. Spelling, and Mathematics vs. Reading are the largest ones, similarly to what we have seen in Figure 3.2.3. However, the values for these four cases are larger when we use LOESS. This is natural because the technique has smoothed out the minute details of the raw data and thus shows only general trends. Of course, the LOESS parameters can be adjusted to make the fitted curves exhibit more details, but getting closer to the raw data would, in practice, make confidentiality issues more acute and thus perhaps unwelcome. At any rate, when it comes to minute details and raw data, which can be of any size as long as there are at least two pairs, we have the above introduced index of increase accompanied with an efficient method of calculation.

3.5 Group comparisons

Suppose that we wish to compare two student groups, ω_1 and ω_2—which could for example be the boys and girls in the data set of Thorndike and Thorndike Christ (2010, pp. 24-25)—with respect to their monotonicity trends in Mathematics vs. Reading, Mathematics vs.

Spelling, and so on. We could set out to calculate the indices of increase for ω_1 and ω_2, and then compare the index values and make conclusions, but we have to be careful because of possibly different minimal and maximal scores for the two groups. Indeed, comparing monotonicity patterns over ranges of different length should be avoided because the wider the interval, the more fluctuations might occur. Hence, to make meaningful comparisons, we have to perform them over intervals of the same length, even though locations of the intervals can be different, due to the earlier established translation-invariance Property 3.3.1.

In the context of our illustrative example, we find it meaningful to compare monotonicities of plots over the same range of scores. Hence, in general, given two groups ω_1 and ω_2 of sizes $n(\omega_1) \geq 2$ and $n(\omega_2) \geq 2$, respectively, let the two data sets consist of the pairs

$$\big(x_i(\omega_1), y_i(\omega_1)\big), \quad i = 1, \ldots, n(\omega_1),$$

and

$$\big(x_j(\omega_2), y_j(\omega_2)\big), \quad j = 1, \ldots, n(\omega_2).$$

These data sets give rise to two piece-wise linear plots: the first one ranges from $x_{1:n(\omega_1)}(\omega_1)$ to $x_{n(\omega_1):n(\omega_1)}(\omega_1)$, and the second one from $x_{1:n(\omega_2)}(\omega_2)$ to $x_{n(\omega_2):n(\omega_2)}(\omega_2)$. The overlap of the two ranges is the interval $[L, U]$, where

$$L = \max\left\{x_{1:n(\omega_1)}, x_{1:n(\omega_2)}\right\}$$

and

$$U = \min\left\{x_{n(\omega_1):n(\omega_1)}, x_{n(\omega_2):n(\omega_2)}\right\}.$$

We identify and order the distinct first coordinates in each of the two samples and calculate the medians of the corresponding concomitants. We arrive at the following two sets of paired data:

$$\big(x_i^*(\omega_1), y_i^*(\omega_1)\big), \quad i = 1, \ldots, m(\omega_1), \tag{3.5.1}$$

and

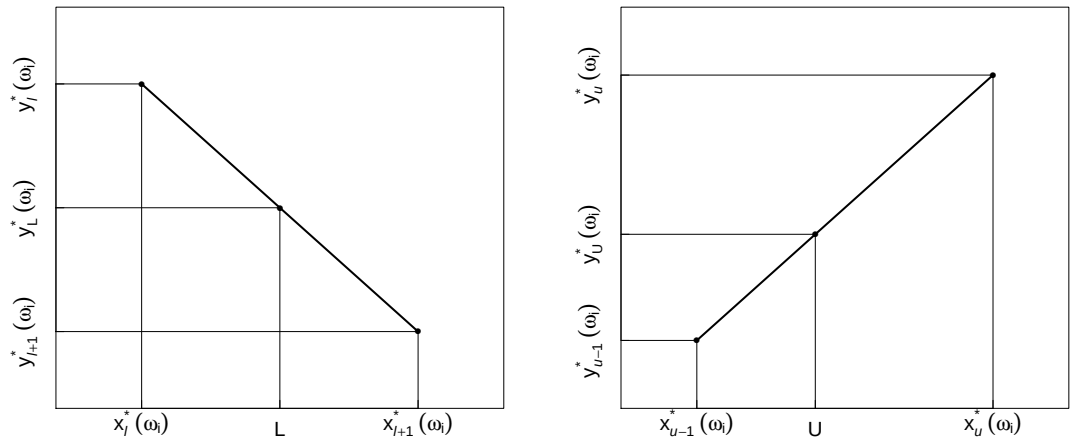
$$\big(x_j^*(\omega_2), y_j^*(\omega_2)\big), \quad j = 1, \ldots, m(\omega_2), \tag{3.5.2}$$

where $m(\omega_1)$ and $m(\omega_2)$ are the numbers of distinct x's in the original data sets. Note that here we use $x_i^*(\omega_1)$ (or $x_j^*(\omega_2)$) and $y_i^*(\omega_1)$ (or $y_j^*(\omega_2)$) instead of $x_{i:m(\omega_1)}^*(\omega_1)$ (or $x_{j:m(\omega_2)}^*(\omega_2)$) and $y_{[i:m(\omega_1)]}^*(\omega_1)$ (or $y_{[j:m(\omega_2)]}^*(\omega_2)$) in order to simplify the notations. We next modify data set (3.5.1):

Step 1: If $L = x_1^*(\omega_1)$, then we do nothing with the data set. If $L > x_1^*(\omega_1)$, then let the pairs $\big(x_l^*(\omega_1), y_l^*(\omega_1)\big)$ and $\big(x_{l+1}^*(\omega_1), y_{l+1}^*(\omega_1)\big)$ be such that $x_l^*(\omega_1)$ is the closest first coordinate in data set (3.5.1) to the left of L, and thus $x_{l+1}^*(\omega_1)$ is the closest one to the right of L. We delete all the pairs from set (3.5.1) whose first coordinates do not exceed $x_l^*(\omega_1)$, and then augment the remaining pairs with $(L, y_L^*(\omega_1))$, where

$$y_L^*(\omega_1) = y_l^*(\omega_1) + \frac{y_{l+1}^*(\omega_1) - y_l^*(\omega_1)}{x_{l+1}^*(\omega_1) - x_l^*(\omega_1)}(L - x_l^*(\omega_1)). \tag{3.5.3}$$

Formula (3.5.3) is useful for computing purposes. We have visualized the pair $(L, y_L^*(\omega_1))$ as an interpolation result in panel (a) of Figure 3.5.1.



(a) $(L, y_L^*(\omega_1))$ via Formula (3.5.3) (b) $(U, y_U^*(\omega_1))$ via Formula (3.5.4)

Figure 3.5.1: The two augmenting-pairs via the interpolation technique.

Step 2: If $U = x_{m(\omega_1)}^*(\omega_1)$, then we do nothing with the data set. If $U < x_{m(\omega_1)}^*(\omega_1)$, then let the pairs $\big(x_{u-1}^*(\omega_1), y_{u-1}^*(\omega_1)\big)$ and $\big(x_u^*(\omega_1), y_u^*(\omega_1)\big)$ be such that $x_{u-1}^*(\omega_1)$ is the

closest first coordinate in the data set to the left of U, and thus $x_u^*(\omega_1)$ is the closest one to the right of U. From the set of pairs available to us after the completion of Step 1, we delete all the pairs whose first coordinates are on or above $x_u^*(\omega_1)$, and then augment the remaining pairs with $(U, y_U^*(\omega_1))$, where

$$y_U^*(\omega_1) = y_{u-1}^*(\omega_1) + \frac{y_u^*(\omega_1) - y_{u-1}^*(\omega_1)}{x_u^*(\omega_1) - x_{u-1}^*(\omega_1)}(U - x_{u-1}^*(\omega_1)). \tag{3.5.4}$$

As an interpolation result, we have visualized the pair $(U, y_U^*(\omega_1))$ in panel (b) of Figure 3.5.1.

In the case of data set (3.5.2), we proceed in an analogous fashion:

Step 3: If $L = x_1^*(\omega_2)$, then we do nothing, but if $L > x_1^*(\omega_2)$, then we produce a new pair $(L, y_L^*(\omega_2))$ that replaces all the deleted ones on the left-hand side of data set (3.5.2).

Step 4: If $U = x_{m(\omega_1)}^*(\omega_2)$, then we do nothing with the pairs available to us after Step 3, but if $U < x_{m(\omega_1)}^*(\omega_2)$, then we produce a new pair $(U, y_U^*(\omega_2))$ that replaces the deleted ones on the right-hand side of the data set obtained after Step 3.

In summary, we have turned data sets (3.5.1) and (3.5.2) into the following ones:

$$\big(v_i(\omega_1), w_i(\omega_1)\big), \quad i = 1, \ldots, k(\omega_1), \tag{3.5.5}$$

and

$$\big(v_j(\omega_2), w_j(\omega_2)\big), \quad j = 1, \ldots, k(\omega_2), \tag{3.5.6}$$

with some $k(\omega_1)$ and $k(\omega_2)$ that do not exceed $m(\omega_1)$ and $m(\omega_2)$, respectively. Note that the smallest first coordinates $v_1(\omega_1)$ and $v_1(\omega_2)$ are equal to L, whereas the largest first coordinates $v_{k(\omega_1)}(\omega_1)$ and $v_{k(\omega_2)}(\omega_2)$ are equal to U. Hence, both data sets (3.5.5) and (3.5.6) produce piece-wise linear functions defined on the same interval $[L, U]$. The

corresponding index $\mathrm{I} := \mathrm{I}(\mathbf{x}, \mathbf{y}|L, U) := \mathrm{I}(\mathbf{v}(\omega), \mathbf{w}(\omega))$ for any $\omega \in \{\omega_1, \omega_2\}$ is calculated using the formula

$$\mathrm{I}(\mathbf{x}, \mathbf{y}|L, U) := \mathrm{I}(\mathbf{v}(\omega), \mathbf{w}(\omega)) = \frac{\sum_{i=2}^{k(\omega)} (w_i(\omega) - w_{i-1}(\omega))_+}{\sum_{i=2}^{k(\omega)} |w_i(\omega) - w_{i-1}(\omega)|}. \tag{3.5.7}$$

To illustrate, we report the index values and the comparison ranges for boys and girls in the three subjects, and in all possible combinations in Table 3.5.1.

	Span	M vs. R	M vs. S	R vs. M	R vs. S	S vs. M	S vs. R
	N/A	0.58	0.53	0.72	0.63	0.52	0.56
Boys	0.35	0.61	0.52	0.73	0.72	0.54	0.62
	0.75	0.85	0.49	1.00	0.99	0.91	0.99
	N/A	0.55	0.50	0.61	0.57	0.49	0.54
Girls	0.35	0.62	0.52	0.61	0.58	0.48	0.57
	0.75	0.94	0.57	0.94	0.94	0.37	0.61
Range		33.85–81.54	33.85–81.54	48.89–93.33	48.89–93.33	50.00–92.50	50.00–92.50

Table 3.5.1: Performance of boys and girls in the two classes combined on the three subjects as measured by the index I.

The values in the rows with `span` = N/A have been calculated based on the piece-wise linear fits. The rows with `span` values 0.35 and 0.75 are based on the LOESS curve fitting. We note in this regard that the parameter `span` controls the smoothness of the fitted curves: the larger the `span`, the smoother the resulting curve, with the default value `span` = 0.75 that we already used in Figure 3.2.4. The graphs corresponding to Table 3.5.1 are of interest, but since they are plentiful, we have relegated them to Appendix A.2 at the end of the thesis.

3.6 A step-by-step guide

To have a better understanding and appreciation of the suggested method, we next implement it in a step-by-step manner using a small artificial data set. Namely, suppose that we wish to compare the performance of two groups of students, say ω_1 and ω_2, in two

subjects, which we call x and y. Let the scores, given in percentages, be those in Table 3.6.1.

i	1	2	3	4	5	6
$x_i(\omega_1)$	100	75	75	50	75	87.5
$y_i(\omega_1)$	90	40	40	50	50	70
$x_i(\omega_2)$	100	87.5	75	37.5	75	87.5
$y_i(\omega_2)$	100	60	50	50	50	70

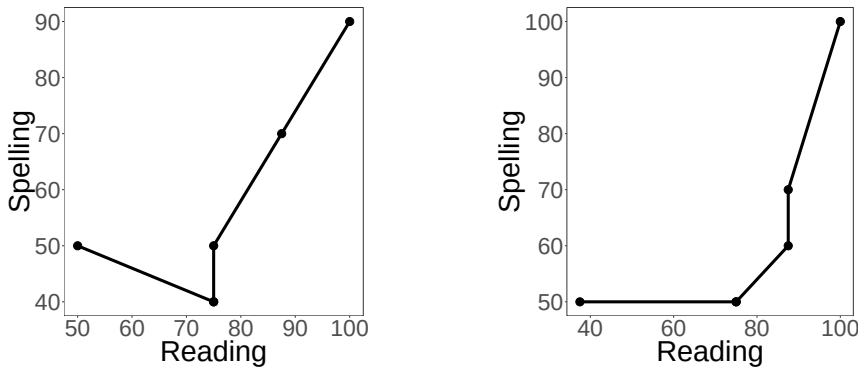
Table 3.6.1: Illustrative scores.

As the first step, we order the data according to the x's and also record their concomitants, which are the y's. The numerical outcomes are reported in Table 3.6.2.

i	1	2	3	4	5	6
$x_{i:6}(\omega_1)$	50	75	75	75	87.5	100
$y_{[i:6]}(\omega_1)$	50	40	40	50	70	90
$x_{i:6}(\omega_2)$	37.5	75	75	87.5	87.5	100
$y_{[i:6]}(\omega_2)$	50	50	50	60	70	100

Table 3.6.2: Ordered scores and their concomitants.

Using piece-wise linear plots, we have visualized them in Figure 3.6.1.



(a) ω_1 (b) ω_2

Figure 3.6.1: Piece-wise linear fits to the illustrative scores.

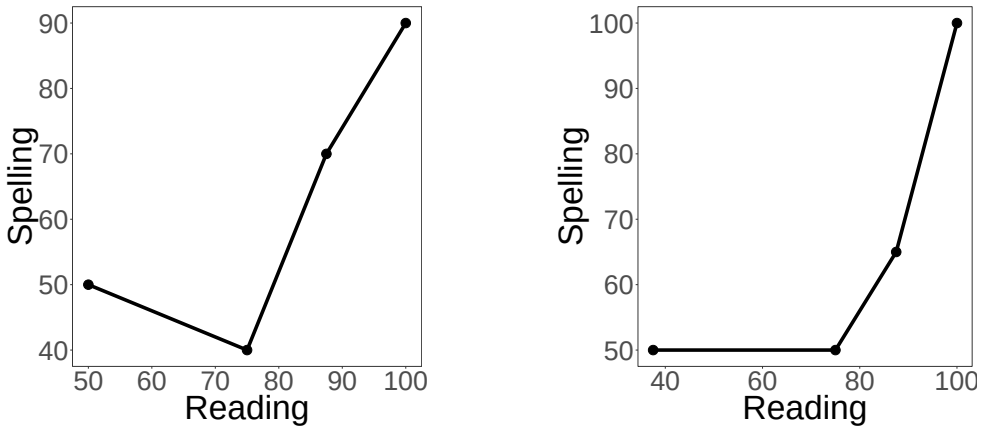
In both panels of the figure, we see only five points, whereas the sample sizes of the two groups are six: $n(\omega_1) = 6$ and $n(\omega_2) = 6$. The reason is that each of the two groups has two identical points: $(75, 40)$ in the case of ω_1, and $(75, 50)$ in the case of ω_2.

Next we apply the median-based aggregation: in the case of ω_1, the median of 40, 40, and 50 (which correspond to 75) is 40, and in the case of ω_2, the median of 50 and 50 (which correspond to 75) is 50, and that of 60 and 70 (which correspond to 87.5) is 65. (There are several definitions of median, but in this thesis we use the one implement in the R Stats Package (R Core Team, 2013): given two data points in the middle of the ranked sample, the median is the average of the two points.) We have arrived at two data sets with only $m(\omega_1) = 4$ and $m(\omega_2) = 4$ pairs, which are reported in Table 3.6.3.

i	1	2	3	4
$x_i^*(\omega_1)$	50	75	87.5	100
$y_i^*(\omega_1)$	50	40	70	90
$x_i^*(\omega_2)$	37.5	75	87.5	100
$y_i^*(\omega_2)$	50	50	65	100

Table 3.6.3: Condensed scores using the median approach.

The corresponding piece-wise linear plots are given in Figure 3.6.2.



(a) ω_1 (b) ω_2

Figure 3.6.2: Piece-wise linear fits to median adjusted data.

Note that the x-ranges of the two data sets are different: $[50, 100]$ and $[37.5, 100]$. We therefore unify them. Since the range for ω_1 is a subset of that for ω_2, we keep all the ω_1 pairs, and thus $k(\omega_1) = 4$. In the case of ω_2, the lower bound 50 is not among the x's and thus we apply the interpolation method to find the y-value corresponding to $L = 50$. Using Equation (3.5.3), we have

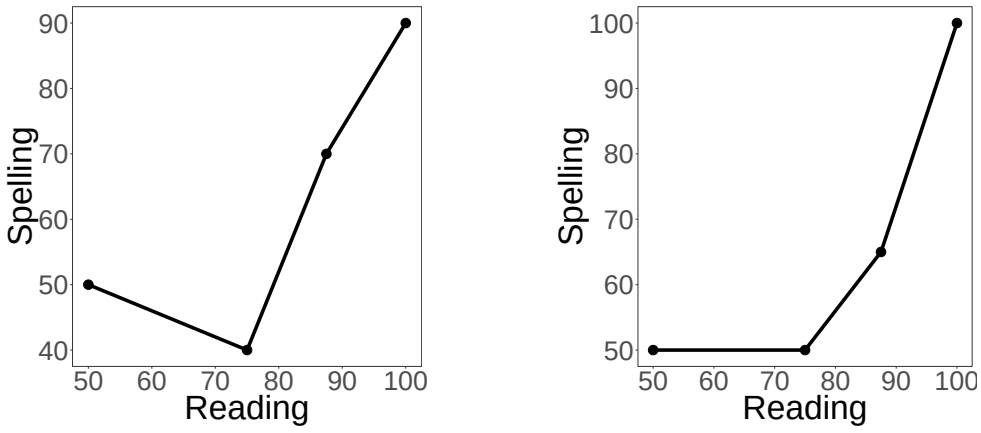
$$y_L^*(\omega_2) = 50 + \frac{50 - 50}{75 - 37.5}(50 - 37.5) = 50. \tag{3.6.1}$$

For the upper bound U, since 100 is present in the data set, nothing needs to be done. We have $k(\omega_2) = 4$ because one pair was removed and one added. The new data set is reported in Table 3.6.4.

$v_i(\omega_1)$	50	75	87.5	100
$w_i(\omega_1)$	50	40	70	90
$v_i(\omega_2)$	50	75	87.5	100
$w_i(\omega_2)$	50	50	65	100

Table 3.6.4: Ordered scores with unified ranges.

The corresponding piece-wise linear plots are depicted in Figure 3.6.3.



(a) ω_1 (b) ω_2

Figure 3.6.3: Piece-wise linear fits with unified ranges.

For the pairs in Table 3.6.4, we can now calculate the index of increase according to Equation (3.5.7). Namely, for ω_1 we have

$$\begin{aligned} \mathrm{I} &= \frac{(40-50)_+ + (70-40)_+ + (90-70)_+}{|40-50| + |70-40| + |90-70|} \\ &= \frac{30+20}{10+30+20} \\ &= \frac{5}{6}, \end{aligned} \tag{3.6.2}$$

and for ω_2 we have

$$\begin{aligned} \mathrm{I} &= \frac{(50-50)_+ + (65-50)_+ + (100-65)_+}{|50-50| + |65-50| + |100-65|} \\ &= \frac{15+35}{15+35} \\ &= 1. \end{aligned} \tag{3.6.3}$$

Both values are greater than 0.5, and thus both trends are more increasing than not, but the trend corresponding to ω_1 is less increasing than that for ω_2, which is of course obvious from Figure 3.6.3. Finally, the index for ω_2 is 1, which means that the corresponding trend is non-decreasing everywhere.

3.7 Concluding notes

In many real-life applications, trends tend to be non-monotonic and researchers wish to quantify and compare their departures from monotonicity. In this chapter, we have argued in favour of a technique that, in a well-defined and rigorous manner, has been designed specifically for making such monotonicity assessments.

We have illustrated the technique on a small artificial data set, and also on the larger data set borrowed from the classical text by Thorndike and Thorndike Christ (2010, pp. 24-25). Extensive graphical and numerical illustrations have been provided to elucidate the workings of the new technique, and we have compared the obtained results with those arising from classical statistical techniques.

To facilitate free and easy implementation of the technique in various real-life contexts, we have conducted all our explorations using the R software environment for statistical computing and graphics (R Core Team, 2013). We have implemented the new technique relying on our own R codes (see Appendix A.1 for details). For data visualization, whenever possible, we have used the R packages `ggplot2` by Wickham (2009) and `GGally` by Schloerke et al. (2017).

A large number of notes that we have incorporated in the chapter have been inspired—directly or indirectly—by the many queries, comments, and suggestions by anonymous reviewers. A few additional comments follow next. First, there can, naturally, be situations where measuring and comparing the extent of *decrease* of various (non-monotonic) patterns would be of interest. In such cases, instead of, for example, the index of increase $\mathrm{I}(\mathbf{x}, \mathbf{y})$ defined by Equation (3.3.1), we could use the index of decrease $\mathrm{D}(\mathbf{x}, \mathbf{y})$ defined by

$$\mathrm{D}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=2}^{n} (y_{[i:n]} - y_{[i-1:n]})_-}{\sum_{i=2}^{n} |y_{[i:n]} - y_{[i-1:n]}|},$$

where z_- is the negative part of z, that is, $z_- = -z$ when $z < 0$ and $z_- = 0$ otherwise. The above developed computational algorithms do not need to be redone for the index $\mathrm{D}(\mathbf{x}, \mathbf{y})$ because of the identity

$$\mathrm{D}(\mathbf{x}, \mathbf{y}) = 1 - \mathrm{I}(\mathbf{x}, \mathbf{y}),$$

which follows immediately from the equation $z_+ + z_- = |z|$ that holds for all real numbers z. As a consequence of the identity, for example, the ratio

$$\mathrm{O}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=2}^{n} (y_{[i:n]} - y_{[i-1:n]})_-}{\sum_{i=2}^{n} (y_{[i:n]} - y_{[i-1:n]})_+},$$

which can be used for measuring the extent of downward movements relative to the upward movements, can be rewritten as the "odds ratio"

$$\mathrm{O}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{D}(\mathbf{x}, \mathbf{y})}{\mathrm{I}(\mathbf{x}, \mathbf{y})} = \frac{\mathrm{D}(\mathbf{x}, \mathbf{y})}{1 - \mathrm{D}(\mathbf{x}, \mathbf{y})}$$

with $\mathrm{D}(\mathbf{x}, \mathbf{y})$ viewed as a probability, which can indeed be viewed this way because

$\mathrm{D}(\mathbf{x}, \mathbf{y})$ is the proportion of downward movements with respect to all—downward and upward—movements.

The next comment that we make is that, as noted by a reviewer, researchers may wish to assess the degree of non-exchangeability between $\mathbf{x}$ and $\mathbf{y}$. In the case of, for example, the index of increase $\mathrm{I}(\mathbf{x}, \mathbf{y})$, such an assessment can be done either in absolute terms with the help of

$$\mathrm{AI}(\mathbf{x}, \mathbf{y}) = \big|\mathrm{I}(\mathbf{x}, \mathbf{y}) - \mathrm{I}(\mathbf{y}, \mathbf{x})\big|$$

or, in relative terms, using

$$\mathrm{RI}(\mathbf{x}, \mathbf{y}) = \left|\frac{\mathrm{I}(\mathbf{x}, \mathbf{y})}{\mathrm{I}(\mathbf{y}, \mathbf{x})} - 1\right|.$$

It might be desirable to remove the absolute values from the right-hand sides of the above two definitions, thus making the newly formed two indices either positive or negative, and both of them being equal to 0 when $\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{I}(\mathbf{y}, \mathbf{x})$. We will explore these two extended measures of the degree of non-exchangeability with another education dataset in the following chapter.

Chapter 4

Quantifying directional associations and interchangeability among subjects for curriculum development

4.1 Motivation

Assessing and comparing student performance have been important and fascinating areas of educational research. Literature is abundant and covers diverse topics such as measuring differences in student performance due to differences in teacher performance (e.g., Ross, 1992; Hill et al., 2005), study subjects (e.g., Gamoran and Hannigan, 2000; Chen and Zitikis, 2017), examination formats (e.g., Agarwal et al., 2008; Heijne-Penninga et al., 2008, 2010), and gender (e.g., Leedy et al., 2003; Nguyen et al., 2005; Putwain, 2008; Wade et al., 2017).

Various methods for collecting relevant data have been employed, including observational studies and experiments, open- and closed-book examinations. Furthermore, various statistical techniques have been used, including linear and nonlinear regression, with the Pearson correlation coefficient naturally arising as a measure of relationship between variables (e.g., Krasne et al., 2006; Agarwal et al., 2008; Heijne-Penninga et al., 2008, 2010; Thorndike and Thorndike Christ, 2010).

In addition to research by professional educators, a considerable body of specialized statistical literature has utilized educational data to illustrate various methods and techniques, including distance-based and classical multivariate analyses (e.g., Groenen and Meulman, 2004), Bayesian analysis (e.g., Efron, 2012), orthogonal simple component anlaysis (e.g., Anaya-Izquierdo et al., 2011), and robust structural equation modelling with missing data and auxiliary variables (e.g., Yuan and Zhang, 2012). Furthermore, Qoyyimi and Zitikis (2014, 2015) have employed Gini-based arguments to assess the lack of relationship in multivariate educational data. Chen and Zitikis (2017) use an index of increase to quantify the amount of monotonicity in nonlinear relationships. Duzhin and Gustafsson (2018) suggest an automated procedure for analyzing educational data based on machine learning, with features such as decision making that accounts for students' prior knowledge.

As is usually the case with methods that condense raw data into a few parameters, some information inevitably gets lost in the process. The loss is sometimes acceptable, but sometimes is not. An example of the latter case would be the use of the Pearson correlation coefficient, as it gives the same value irrespective of which of the two variables under consideration is explanatory or response. Later in this chapter, we shall illuminate these issues using educational data, and will in turn put forward arguments in favour of an index of increase (Davydov and Zitikis, 2017) as a measure for quantifying the presence of monotonicity in inherently non-monotonic scatterplots. The index has recently been employed by Chen and Zitikis (2017) to revisit a dataset of Thorndike and Thorndike Christ (2010), with further theoretical insights worked out by Chen et al. (2018).

Our current research builds upon the work of Chen and Zitikis (2017), but unlike that work, we explore the rich data reported by Mardia et al. (1979, pp. 3-4). Due to the popularity of this textbook, the data have been frequently used by statisticians and others to illustrate various notions and techniques of Multivariate Analysis. Consequently, and naturally, the data are available in several computing packages, such as `MVT` (Osorio and Galea, 2015). In the current chapter we revisit the data with the aid of additional insights on the topic that have been acquired since the publication of Chen and Zitikis (2017).

We have organized the rest of this chapter as follows. In Section 4.2, we describe the dataset of Mardia et al. (1979, pp. 3-4) and give its preliminary analysis. In Section 4.3, we fit certain functions to the data and give reasons why this exercise is of interest, and sometimes even necessary. In Section 4.4, we explain basic concepts and intuition behind the index of increase, which can take on several forms depending on the nature of available data (e.g., scatterplots, fitted functions, etc.). In Section 4.5, we use the index to illuminate directional relationships between several subjects. Section 4.6 finishes the chapter with a summary of main contributions and concluding notes.

4.2 Data and an idea of measuring increase

The dataset of Mardia et al. (1979, pp. 3-4) consists of $n = 88$ examination scores in five subjects: Algebra, Analysis, Mechanics, Vectors, and Statistics. The scores are out of 100 possible in each of the five subjects, with the scores in Mechanics and Vectors coming from closed-book examinations, and the scores in Algebra, Analysis, and Statistics coming from open-book examinations. In Figure 4.2.1 we give a snapshot of the data based on commonly used scatterplots and least-squares regression lines. The response variables are noted in the rows and the explanatory ones in the columns. The corresponding values of the Pearson correlation coefficient r are inside of each of the twenty off-diagonal panels. For example, the panel with $r = 0.553$ in the top row is the scatterplot of Vectors vs Mechanics, whereas the panel with the same $r = 0.553$ one row below is the scatterplot of Mechanics vs Vectors. The slopes of the fitted lines are different because the variances of the explanatory and response variables are different. Each of the five diagonal panels has, obviously, $r = 1$.

Note that the reported r values are symmetric with respect to the two variables under consideration, although the study subjects clearly lack symmetry with respect to each other. Hence, the use of r in the current context is hardly suitable. The slope $b = rs_y/s_x$ of the linear regression line is a better choice, where s_x and s_y denote the standard deviations of the explanatory and response variables, respectively. However, the scatterplots can hardly suggest linear patterns. Hence, neither r not b seem to be particularly informative

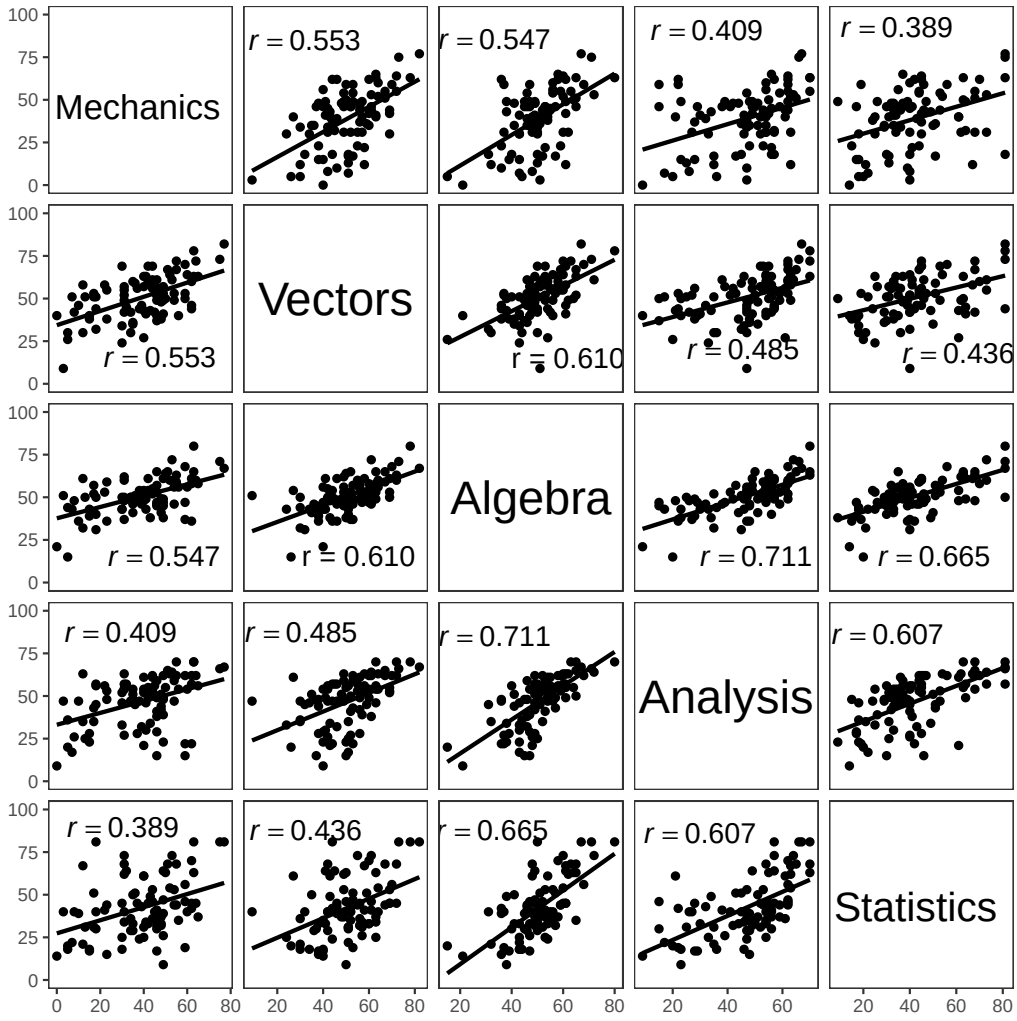


Figure 4.2.1: Least-squares regression lines fitted to the data of Mardia et al. (1979, pp. 3-4) with the corresponding values of the Pearson correlation coefficient $r = r(\mathbf{x}, \mathbf{y})$.

in the current context. Given our goal to understand and even predict how changes in the scores of one study subject are reflected in the scores of another subject, we therefore find it desirable to search for alternative ways for quantifying nonlinear relationships.

Note that although b is not perfect, it is nevertheless better than r, and this is in part due to asymmetry of b with respect to the explanatory and response variables. This feature is natural when quantifying dependence, as elucidated by Reimherr and Nicolae (2013, p. 119). If, however, symmetry is desirable for any reason, then it can be imposed by symmetrization, which can be achieved in many ways (see, e.g., Reimherr and Nicolae,

2013, p. 120). We shall briefly come back to this topic at the end of Section 4.3, noting now that the measure that we are to employ for quantifying relationships between study subjects is asymmetric, which we find natural and appropriate.

Namely, to assess how much a pattern (scatterplot, function, etc.) is increasing, we measure its distance from the set of decreasing patterns. Hence, if the pattern is decreasing, the distance is 0. By normalizing the distance, we do not allow it to exceed 1. Not going into any more mathematical details at the moment (Davydov and Zitikis, 2017), we obtain an index of increase, denoted by I, with the following features:

- it takes values only in the interval $[0, 1]$,
- vanishes when there are no segments of increase,
- takes the maximal value 1 when there are no segments of decrease,
- exceeds 0.5 when the pattern is more upward than downward,
- is smaller than 0.5 when the pattern is more downward than upward.

To illustrate the features, in Figure 4.2.2 we have depicted the dataset of Mardia et al. (1979, pp. 3-4) by connecting the consecutive data points using straight lines, which have enabled us to calculate the index of increase for each panel using a computational formula that we shall give and discuss later in this chapter. Note, for example, that Algebra vs. Vectors, Algebra vs. Analysis, Algebra vs. Statistics have the largest three values, thus implying that the corresponding patterns are most increasing among the twenty off-diagonal panels. We can interpret this by saying that students with higher scores in Algebra tend to have higher scores in Vectors ($\mathrm{I} = 0.576$), Analysis ($\mathrm{I} = 0.575$), and Statistics ($\mathrm{I} = 0.578$) than in any other study subject. This, we think, is due to Algebra being a fundamental subject for Vectors, Analysis, and Statistics. For strong arguments and evidence in favour of Algebra, we refer to Gamoran and Hannigan (2000).

Vectors vs. Algebra ($\mathrm{I} = 0.527$) and Statistics vs. Algebra ($\mathrm{I} = 0.546$) have lower indices than the three ones mentioned in the previous paragraph, and so we are less confident that better performance in Vectors and Statistics would lead to higher scores in Algebra.

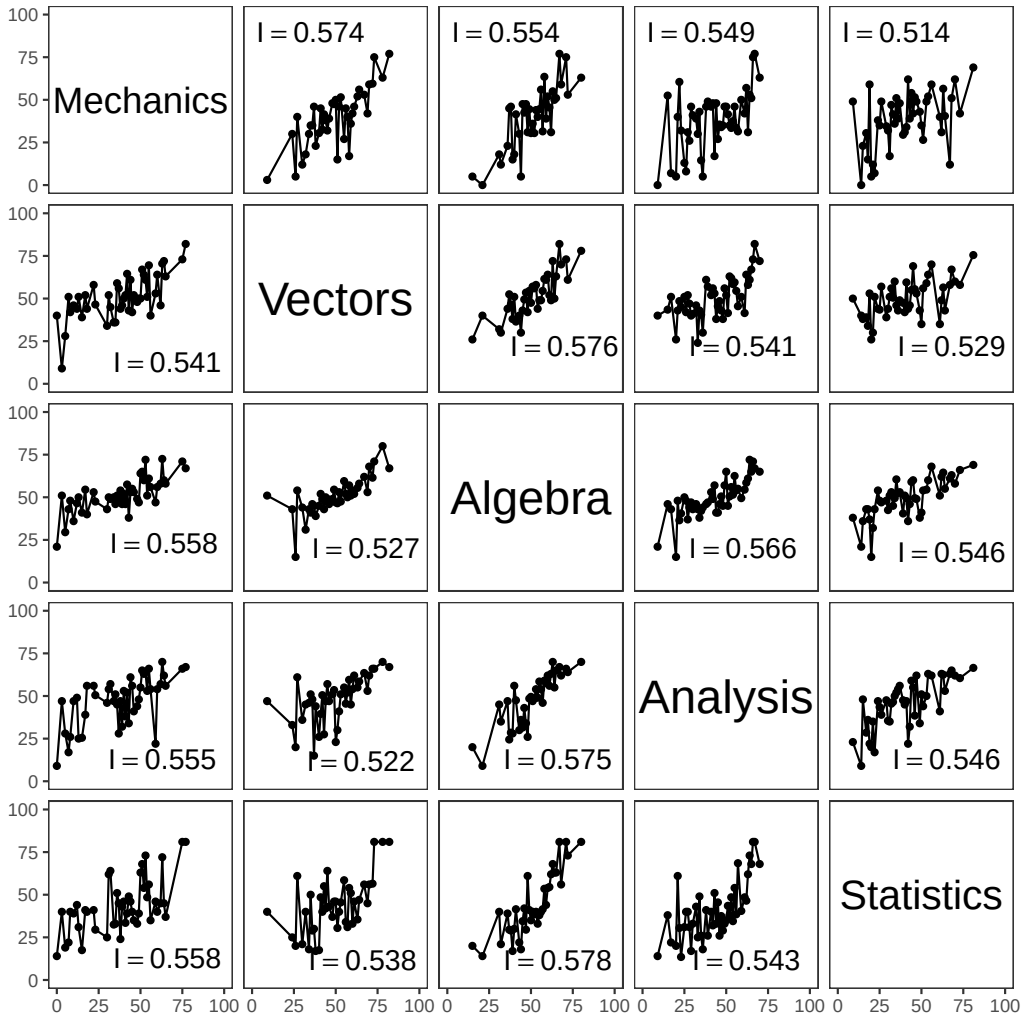


Figure 4.2.2: Piece-wise linear fits to the data of Mardia et al. (1979, pp. 3-4) with the corresponding values of the index of increase $\mathrm{I} = \mathrm{I}(\mathbf{x}, \mathbf{y})$.

Furthermore, Analysis vs. Algebra (0.566) has just a slightly lower index than the three top ones. This, we think, is due to Analysis and Algebra being fundamental subjects, and thus students possibly viewing them as equally important, or equally challenging, and thus demanding similar study efforts.

Among the twenty panels, Statistics vs. Mechanics has the lowest index ($\mathrm{I} = 0.514$), which is not far away from the boundary value 0.500 separating more increasing patterns from more decreasing ones.

Remark 4.2.1. *In the above discussion, to illustrate the mathematical concept of the*

index of increase, we treated the dataset of Mardia et al. (1979, pp. 3-4) as a "population," and not as a sample with variability. We shall do so quite often throughout the chapter, but we shall also let the reader know our thoughts on the statistical side of the subject matter (see, e.g., Remark 4.4.3, and also the second half of concluding Section 4.6).

4.3 Functions, fitted curves, and interchangeability

The index of increase can be calculated not only from (discrete) scatterplots, such as those in Figure 4.2.2, but also from continuous functions. The latter ones naturally, and sometimes inevitably, arise due to several reasons:

- The phenomena under consideration might be modelled using continuous functions, which could, for example, arise as solutions to differential equations, as is frequently the case in mathematical biology, as well as in other areas dealing with dynamical modelling.

- Continuous functions may arise due to fitting curves to scatterplots (e.g., Hastie et al., 2009; Murphy, 2012, and references therein). Such fitting might also be done by the researcher already possessing raw data but wishing to smooth out noise from the data, mitigate the influence of potential outliers, or due to some other statistical considerations.

- Fitted curves may be the only objects available to the researcher for analysis and decision making, due to reasons such as ethics and confidentiality. For example, research that involves the use of personal data, irrespective of whether the data are identifiable or de-identified, requires a research ethics board review at most institutions. Scatterplots would be among such datasets, but the fitted curves would hardly be such.

Irrespective of the origins of continuous functions, calculating their indices of increase is discussed in Section 4.4.1 below. In the next subsection, for comparative and illustrative

purposes, we shall fit curves to the scatterplots of Figure 4.2.2 and also provide the values of their indices of increase calculated using a method to be described later in this chapter.

4.3.1 Fitted curves

To illustrate, we employ one of the most commonly used regression methods for fitting nonlinear relationships, which is locally estimated scatterplot smoothing, or LOESS for short. It is a non-parametric method that combines multiple regression models and k-nearest-neighbor-based meta-models. Jacoby (2000) describes the LOESS methodology in detail, including how to fit LOESS functions and perform goodness-of-fit tests, with particular attention on those cases when subject-matter knowledge suggests nonlinear relationships but little, if anything, is known about the actual underlying functional forms. This is precisely the situation we deal with in the current chapter.

There have been many uses of LOESS in educational research, and from those studies we gain valuable insights relevant to the topic of the present paper. For example, Abramo et al. (2012) use LOESS regression to explore the influence of research group's size on research productivity, with emphasis on the Italian higher-education system. Avendano et al. (2009) employ LOESS to explore the impact of educational level on changes in health outcomes among Europeans, with analyses performed separately for regions with different welfare state regimes.

Coming back to the dataset of Mardia et al. (1979, pp. 3-4) and using the R package `stats` (R Core Team, 2017), we have implemented the `loess` function with its default parameter `span` = 0.75. The resulting curves are depicted in Figure 4.3.1. We note in this regard that the parameter `span` controls smoothness: the larger the value, the smoother (i.e., less wiggly) is the fitted function. Some of the reported values of I in the panels of Figure 4.3.1 are equal to 1, thus implying that the fitted functions are increasing everywhere on their domains of definition. Interestingly, some index values are equal to 1 even when the horizontal and vertical axes are interchanged, as is, for example, for Algebra vs. Analysis and Analysis vs. Algebra. We should not, however, hastily infer from these values that Algebra and Analysis are interchangeable subjects: first, the rates at which

the two fitted functions increase are different, and second, the values of the two indices are influenced by the degree of smoothing, governed by the parameter `span`. We shall illustrate the latter feature later in the paper, when we set `span` = 0.35, in addition to the default value `span` = 0.75.

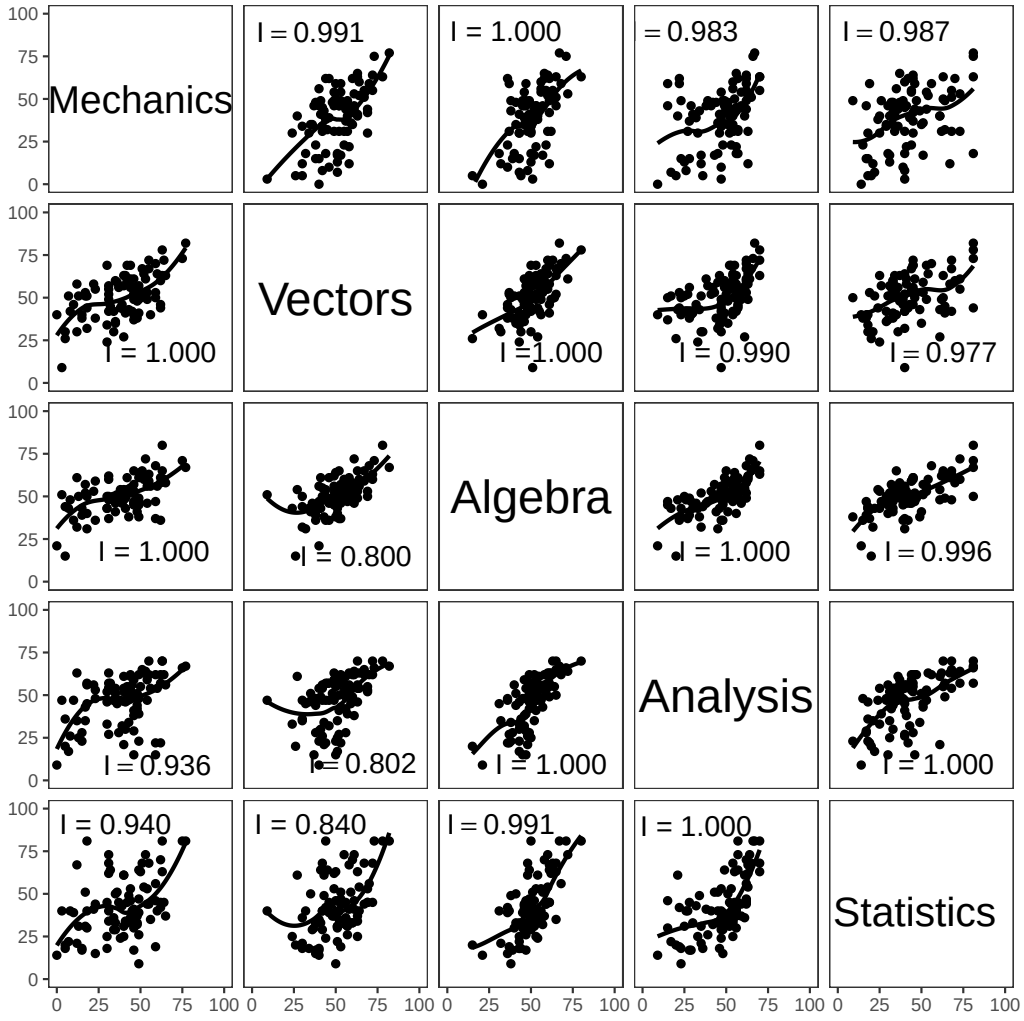


Figure 4.3.1: LOESS fitted functions $h = h_{0.75}$ to the data of Mardia et al. (1979, pp. 3-4) with the corresponding values of the index of increase $\mathrm{I} = \mathrm{I}(h_{0.75})$.

We conclude this subsection with Table 4.3.1, which summarizes our findings so far. Specifically, in the table we report the values of the Pearson correlation coefficient $r = r(\mathbf{x}, \mathbf{y})$ (Figure 4.2.1), and also those of $\mathrm{I} = \mathrm{I}(\mathbf{x}, \mathbf{y})$ for the raw data (Figure 4.2.2) and $\mathrm{I} = \mathrm{I}(h)$ for the LOESS fits under the default parameter `span` = 0.75 (Figure 4.3.1).

Student scores		Pearson		Data		$\text{LOESS}_{0.75}$	
x	**y**	r	RI%	I	RI%	I	RI%
Mechanics	Vectors	0.553	0.000	0.541	-5.749	1.000	0.908
Vectors	Mechanics	0.553	0.000	0.574	6.099	0.991	-0.900
AI%		0.000		3.300		0.900	
Vectors	Algebra	0.610	0.000	0.527	-8.507	0.800	-20.000
Algebra	Vectors	0.610	0.000	0.576	9.298	1.000	25.000
AI%		0.000		4.900		20.000	
Algebra	Analysis	0.711	0.000	0.575	1.590	1.000	0.000
Analysis	Algebra	0.711	0.000	0.566	-1.565	1.000	0.000
AI%		0.000		0.900		0.000	
Analysis	Statistics	0.607	0.000	0.543	-0.549	1.000	0.000
Statistics	Analysis	0.607	0.000	0.546	0.552	1.000	0.000
AI%		0.000		0.300		0.000	
Mechanics	Algebra	0.547	0.000	0.558	0.722	1.000	0.000
Algebra	Mechanics	0.547	0.000	0.554	-0.717	1.000	0.000
AI%		0.000		0.400		0.000	
Vectors	Analysis	0.485	0.000	0.522	-3.512	0.802	-18.990
Analysis	Vectors	0.485	0.000	0.541	3.640	0.990	23.441
AI%		0.000		1.900		18.800	
Algebra	Statistics	0.665	0.000	0.578	5.861	0.991	-0.502
Statistics	Algebra	0.665	0.000	0.546	-5.536	0.996	0.505
AI%		0.000		3.200		0.500	
Mechanics	Analysis	0.409	0.000	0.555	1.093	0.936	-4.781
Analysis	Mechanics	0.409	0.000	0.549	-1.081	0.983	5.021
AI%		0.000		0.600		4.700	
Vectors	Statistics	0.436	0.000	0.538	1.701	0.840	-14.023
Statistics	Vectors	0.436	0.000	0.529	-1.673	0.977	16.310
AI%		0.000		0.900		13.700	
Mechanics	Statistics	0.389	0.000	0.558	8.560	0.940	-4.762
Statistics	Mechanics	0.389	0.000	0.514	-7.885	0.987	5.000
AI%		0.000		4.400		4.700	

Table 4.3.1: Summary statistics for all subjects.

4.3.2 Interchangeability of study subjects

In Table 4.3.1 we have also reported the values of the *relative* index RI% := RI × 100% of interchangeability of $\mathbf{x}$ and $\mathbf{y}$, where

$$\mathrm{RI} := \mathrm{RI}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{I}(\mathbf{x}, \mathbf{y})}{\mathrm{I}(\mathbf{y}, \mathbf{x})} - 1,$$

and also the values of the *absolute* index of interchangeability AI% := AI × 100% of $\mathbf{x}$ and $\mathbf{y}$, where

$$\mathrm{AI} := \mathrm{AI}(\mathbf{x}, \mathbf{y}) = \big|\mathrm{I}(\mathbf{x}, \mathbf{y}) - \mathrm{I}(\mathbf{y}, \mathbf{x})\big|.$$

We note that the indices RI and AI, which are also mentioned in the concluding section of Chen and Zitikis (2017), are not specific to the index I. Indeed, RI and AI can be calculated for any index of interest, including the Pearson correlation coefficient $r = r(\mathbf{x}, \mathbf{y})$, but in the latter case, the values of RI and AI are always 0 due to the symmetry of r with respect to $\mathbf{x}$ and $\mathbf{y}$. The latter note highlights the unsuitability of r in the context of current research.

4.4 Index of increase

In the previous sections, we introduced the index of increase via its properties, and illustrated its performance with numerical results. The latter task required actionable formulas, adapted for the two scenarios of particular interest: scatterplots and functions. We next provide and discuss such formulas, starting with functions.

4.4.1 The index for functions

Let $h : [L, U] \to \mathbb{R}$ be a real-valued function defined on an interval $[L, U]$. For example, h could be a LOESS function fitted to a scatterplot $\{(x_i, y_i),\ i = 1, \ldots, n\}$, with $L = \min_i\{x_i\}$ and $U = \max_i\{x_i\}$ being the smallest and largest x-values, respectively.

The index of increase of h is, by definition, the normalized distance between the function h and the set of all decreasing (precisely speaking, non-increasing) functions (Davydov and Zitikis, 2017). Hence, the index is equal to 0 when the function h is decreasing,

and is equal to 1 when it is increasing (precisely speaking, non-decreasing). When h is differentiable, the formula for this distance-based index is (Davydov and Zitikis, 2017)

$$\mathrm{I}(h) = \frac{\int_L^U (h'(x))_+ \, \mathrm{d}x}{\int_L^U |h'(x)| \, \mathrm{d}x}, \tag{4.4.1}$$

where z_+ denotes the positive part of any real number z, that is, $z_+ = z$ when $z > 0$ and $z_+ = 0$ otherwise.

A practical way to calculate the index $\mathrm{I}(h)$ is via discretization. Namely, we first divide the interval $[L, U]$ into many small subintervals $[d_{i-1}, d_i]$, $2 \le i \le k$, where $d_i = L + \frac{i-1}{k-1}(U - L)$, $1 \le i \le k$. Then we calculate

$$\widehat{\mathrm{I}}_k(h) = \frac{\sum_{i=2}^k (h(d_i) - h(d_{i-1}))_+}{\sum_{i=2}^k |h(d_i) - h(d_{i-1})|}. \tag{4.4.2}$$

It has been shown (Davydov and Zitikis, 2017; Chen and Zitikis, 2017) that when k grows indefinitely, $\widehat{\mathrm{I}}_k(h)$ converges to $\mathrm{I}(h)$. Based on this fact, we can calculate $\mathrm{I}(h)$ at any desired precision by calculating $\widehat{\mathrm{I}}_k(h)$ for a sufficiently large k.

Remark 4.4.1. *The parameter k, which is not to be confused with the scatterplot size n, is chosen by the researcher, and can be as large as computing time and power permit. For example, Chen and Zitikis (2017) show that for their chosen illustrative functions, setting $k = 20,000$ is sufficient to reach the true value of $\mathrm{I}(h)$ at the precision of six decimal digits.*

4.4.2 The index for scatterplots

By their very nature, scatterplots are discrete, but even when we connect their points with straight lines, the resulting functions, though continuous, are not differentiable and thus formula (4.4.1) cannot be engaged. For this reason, Chen and Zitikis (2017) propose a modification, which resembles formula (4.4.2) of the numerical approximation $\widehat{\mathrm{I}}_k(h)$. To describe it, let $\{(x_i, y_i),\ i = 1, \ldots, n\}$ be the scatterplot under consideration. For the sake of simplicity, let all the x_i's be different, the assumption that we shall remove in Section 4.4.3 below. Hence, we can, and thus do, uniquely order the x_i's from the smallest to the largest, thus obtaining $x_{1:n} < x_{2:n} < \cdots < x_{n:n}$ that are called order statistics (e.g., David and Nagaraja, 2003).

For every $x_{i:n}$, we find the corresponding point (x_j, y_j) in the scatterplot, with j determined by the equation $x_j = x_{i:n}$. We denote the second coordinate of the point (x_j, y_j) by $y_{[i:n]}$, which is usually called the i^{th} concomitant (e.g., David and Nagaraja, 2003). The index of increase is defined by the formula (Chen and Zitikis, 2017)

$$\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=2}^{n} (y_{[i:n]} - y_{[i-1:n]})_+}{\sum_{i=2}^{n} |y_{[i:n]} - y_{[i-1:n]}|}, \tag{4.4.3}$$

with the superscript "0" reminding us that there are no ties among the x's.

To easily interpret the index $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$, we first note that the numerator in its definition (4.4.3) sums up all the upward movements $y_{[i:n]} - y_{[i-1:n]} > 0$, while the denominator sums up the absolute values of all the movements $y_{[i:n]} - y_{[i-1:n]} \in \mathbb{R}$, upward and downward. Hence, the index of increase is the proportion of upward movements among all the movements. In particular, when $\mathrm{I}^0(\mathbf{x}, \mathbf{y}) < 0.5$, the proportion of downward movements is larger than that of upward movements, and so the pattern looks more decreasing than increasing. Analogously, when $\mathrm{I}^0(\mathbf{x}, \mathbf{y}) > 0.5$, the proportion of upward movements is larger than that of downward movements, and so the pattern looks more increasing than decreasing. When $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ is near 0.5, the proportions of upward and downward movements are similar, thus suggesting that the values of the first and the last concomitants (i.e., of $y_{[1:n]}$ and $y_{[n:n]}$) must be similar. The following property establishes this observation rigorously.

Property 4.4.1. *We have* $\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = 0.5$ *if and only if* $y_{[1:n]} = y_{[n:n]}$.

This property follows from the equations $z = z_+ - z_-$ and $|z| = z_+ + z_-$, which imply the equivalence of $\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = 0.5$ and $\sum_{i=2}^{n} (y_{[i:n]} - y_{[i-1:n]}) = 0$, the latter being equivalent to $y_{[1:n]} = y_{[n:n]}$.

Remark 4.4.2. *Based on definition (4.4.3) and Property 4.4.1, we can now complete Remark 2.4.1 by providing an example (suggested by one of the reviewers of this paper) in order to show how much outliers can skew our analysis, as they usually do with any statistical analysis. Namely, suppose that the scatterplot consists of n points, with the left- and right-hand points having the same y-coordinates (i.e.,* $y_{[1:n]} = y_{[n:n]}$*). However, all the points except the right-most point have strictly increasing y-coordinates. Hence, we can say*

that the scatterplot exhibits a strictly increasing pattern, with the right-most point being an outlier. By Property 4.4.1, we have $\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = 0.5$*, but if we remove the outlier (i.e., the right-most point) and calculate the index of increase for the just obtained sub-scatterplot, we get* $\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = 1$*, because the sub-scatterplot exhibits an increasing pattern and thus the numerator and the denominator on the right-hand side of definition (4.4.3) coincide. Of course, from the strictly mathematical point of view, given the original scatterplot with no points removed, the index* $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ *does not lie by giving us the value* 0.5*, as the trend that arises from the scatterplot ends at the same height on the right-hand side as it started on the left-hand side, thus technically making the trend neither increasing nor decreasing. Yet, the statistician would likely remove the right-hand point, calculate the index value* 1*, and would disagree with the mathematician's conclusion. Both would be right in their own ways.*

Another notable property of $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ is translation and scale invariance, utilized by Chen and Zitikis (2017) in order to unify the scales of measurement of different scatterplots.

Property 4.4.2. *For all real* $\alpha, \beta \in \mathbb{R}$ *and all positive* $\gamma, \delta > 0$*, we have*

$$\mathrm{I}^0(\mathbf{x}, \mathbf{y}) = \mathrm{I}\big(\gamma(\mathbf{x} - \alpha), \delta(\mathbf{y} - \beta)\big).$$

This property is particularly useful when dealing with student performance on different subjects, when they are assessed using different score scales. Indeed, the property says that shifting and stretching (or shrinking) data do not affect the value of the index.

Remark 4.4.3. *The parameter* n*, though arbitrary, is nevertheless fixed throughout this paper. The statistical tradition of letting* n *grow indefinitely is not appropriate in the context of the present research, since uncontrollably expanding class sizes do not facilitate insights that we aim to gain in the paper; more on this topic will be in concluding Section 4.6. Nevertheless, one may naturally wish to assess the estimator's variability for a given fixed* n*, due to reasons such as testing one- or two-sample hypotheses. In such cases, we would suggest using the (exact) permutation test (e.g., Wasserman, 2004, pp. 161–164).*

4.4.3 Adjustments due to data ties

The index of increase $\mathrm{I}^0(\mathbf{x}, \mathbf{y})$ is defined under the assumption that all x's are different, but quite often this assumption is violated. Hence, we suggest the following modification (cf. Chen and Zitikis, 2017). Given any scatterplot $\{(x_i, y_i),\ i = 1, \ldots, n\}$, let $x_1^*, x_2^*, \ldots, x_m^*$ denote all the $m(\leq n)$ distinct values among $x_1, x_2, \ldots, x_n$. For each x_i^*, let $\mathcal{Y}_i$ be the set all those y's whose corresponding x's are equal to x_i^*. Each set $\mathcal{Y}_i$ has at least one element, and let y_i^* denote the median of the elements in $\mathcal{Y}_i$. This gives rise to the modified scatterplot $\{(x_i^*, y_i^*),\ i = 1, \ldots, m\}$ with distinct x's, and thus with uniquely defined order statistics $x_{1:m}^* < x_{2:m}^* < \cdots < x_{m:m}^*$ and their corresponding concomitants $y_{[1:m]}^*, y_{[2:m]}^*, \ldots, y_{[m:m]}^*$. Applying definition (4.4.3) on the just constructed modified scatterplot, we obtain the index of increase

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=2}^m (y_{[i:m]}^* - y_{[i-1:m]}^*)_+}{\sum_{i=2}^m |y_{[i:m]}^* - y_{[i-1:m]}^*|}. \tag{4.4.4}$$

The values of I that we earlier reported in Figure 4.2.2 are actually those of the just defined index $\mathrm{I}(\mathbf{x}, \mathbf{y})$, because the data of Mardia et al. (1979, pp. 3-4) contain ties among x-coordinates.

Remark 4.4.4. *Given $(x_i^*, \mathcal{Y}_i)$, instead of calculating the median of the values inside $\mathcal{Y}_i$, we may calculate their mean or some other summary statistic. The various possibilities available to the researcher depend on the data under consideration and/or the researcher's point of view.*

4.4.4 Scatterplots over a specific range

In our explorations so far, we have utilized all the scatterplot points. Hence, piecewise linear and LOESS fitted functions have been defined on the scatterplot-specific interval $[x_{1:n}, x_{n:n}]$, where $x_{1:n} = \min_i\{x_i\}$ and $x_{n:n} = \max_i\{x_i\}$. There are, however, situations (as the one we shall encounter in the next section) when we wish to assess monotonicity only on a certain subinterval $[L, U]$ of $[x_{1:n}, x_{n:n}]$. This can be desirable due to a number of reasons, such as:

- A few left- and right-hand points of the scatterplot might be outliers, and we shall

encounter such a situation in the next section; see Remark 4.5.1 therein. Hence, removing the points might be warranted. This idea of truncation in order to improve the robustness of statistical analysis has long been employed by statisticians, and in various situations. For example, to robustify the classical sample mean as an estimator of the population mean, one typically uses trimmed or winsorized means (e.g., Serfling, 1980; Jurečková et al., 2019).

- One may wish to explore the scores of only a certain portion of the entire class, such as the middle 80% of students, with 10% of under- and 10% of over-performing students treated in special ways in order to make their learning experience more fulfilling.

- When comparing several scatterplots, which we frequently do throughout this chapter, it is advisable to make their ranges comparable, since comparing monotonicity of, for example, two scatterplots with one covering the entire interval $[0, 100]$ and another only $[60, 100]$ may not lead to meaningful conclusions.

Hence, since L and U may not be the minimal and maximal x's of the scatterplot, we therefore need a modification of our previous considerations. This can be done by artificially, though quite naturally, augmenting the scatterplot with points (L, y_L^*) and (U, y_U^*) with specially constructed y-coordinates y_L^* and y_U^*, as described next. Namely, let $\{(x_i, y_i),\ i = 1, \ldots, n\}$ be the scatterplot under consideration, and let $[L, U]$ be a subinterval of $[x_{1:n}, x_{n:n}]$ of particular interest to the researcher. We convert this scatterplot into the modified one $\{(x_i^*, y_i^*),\ i = 1, \ldots, m\}$ with $m(\leq n)$ distinct x-coordinates. Among the points of the modified scatterplot, we find $(x_{l:m}^*, y_{[l:m]}^*)$ and $(x_{(l+1):m}^*, y_{[(l+1):m]}^*)$ such that $x_{l:m}^*$ is the closest x-coordinate to the left of (or equal to) L, and $x_{(l+1):m}^*$ is the closest x-coordinate to the right of (or equal to) L. To L we attach

$$y_L^* = y_{[l:m]}^* + \frac{y_{[(l+1):m]}^* - y_{[l:m]}^*}{x_{(l+1):m}^* - x_{l:m}^*}(L - x_{l:m}^*) \tag{4.4.5}$$

and arrive at the point (L, y_L^*), which we add to the modified scatterplot. Analogously we

arrive at the point (U, y^*_U) with

$$y^*_U = y^*_{[(u-1):m]} + \frac{y^*_{[u:m]} - y^*_{[(u-1):m]}}{x^*_{u:m} - x^*_{(u-1):m}}(U - x^*_{(u-1):m}), \tag{4.4.6}$$

where $(x^*_{(u-1):m}, y^*_{[(u-1):m]})$ and $(x^*_{u:m}, y^*_{[u:m]})$ are the two points in the modified scatterplot such that $x^*_{(u-1):m}$ is the closest x-coordinate to the left of (or equal to) U, and $x^*_{u:m}$ is the closest x-coordinate to the right of U. With

$$z^*_{[i:m]} = \begin{cases} y^*_L & \text{when } i = l, \\ y^*_{[i:m]} & \text{when } i = l+1, \ldots, u-1, \\ y^*_U & \text{when } i = u, \end{cases}$$

we define the (conditional on $[L, U]$) index of increase

$$\mathrm{I}(\mathbf{x}, \mathbf{y} \mid L, U) = \frac{\sum_{i=l+1}^{u} (z^*_{[i:m]} - z^*_{[i-1:m]})_+}{\sum_{i=l+1}^{u} |z^*_{[i:m]} - z^*_{[i-1:m]}|}. \tag{4.4.7}$$

Our following explorations of the dataset of Mardia et al. (1979, pp. 3-4) rely on this index.

4.5 A revisit of Mardia et al. (1979, pp. 3-4)

Based on the dataset of Mardia et al. (1979, pp. 3-4) and using the just introduced conditional index of increase, we next explore relationships between the scores from closed-book examinations (Section 4.5.1), open-book examinations (Section 4.5.2), and also general performance based on the combined scores arising from closed- and open-book examinations (Section 4.5.3). When comparing any pair of scatterplots, we do so based on only those points whose x-coordinates are in the largest common interval $[L, U]$.

Namely, let the two scatterplots be $\{(x_i, y_i),\ i = 1, \ldots, n_1\}$ and $\{(v_i, w_i),\ i = 1, \ldots, n_2\}$ with some n_1 and n_2; for every scatterplot of Mardia et al. (1979, pp. 3-4), we have $n_1 = n_2 = n = 88$. Using the median adjustment described in Section 4.4.3, the two scatterplots reduce to the modified scatterplots $\{(x^*_i, y^*_i),\ i = 1, \ldots, m_1\}$ and $\{(v^*_i, w^*_i),\ i = 1, \ldots, m_2\}$, respectively. The endpoints of their common interval $[L, U]$ are calculated by the formulas

$$L = \max\{x^*_{1:m_1}, w^*_{1:m_2}\} \quad \text{and} \quad U = \min\{x^*_{m_1:m_1}, w^*_{m_2:m_2}\}. \tag{4.5.1}$$

4.5.1 Closed-book examinations

Vectors and Mechanics are the only two subjects in the dataset of Mardia et al. (1979, pp. 3-4) that were assessed using closed-book examinations. To illuminate relationships between the scores in these subjects, in Figure 4.5.1 we have depicted Mechanics vs. Vectors as well as Vectors vs. Mechanics over their common range $[L, U] = [9, 77]$, which we obtained using formula (4.5.1). For the LOESS fits, we have used the default `span` $= 0.75$ and

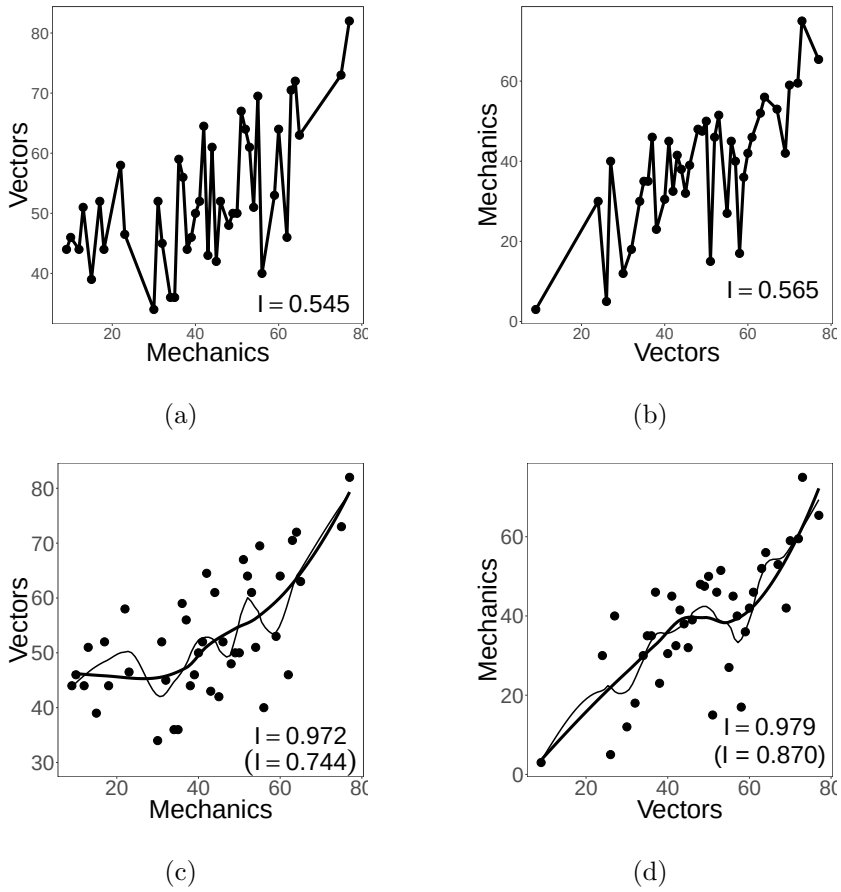


Figure 4.5.1: Piece-wise linear fits (panels (a) and (b)), and the LOESS fits (panels (c) and (d)) when the `span` is 0.75 (thicker) and 0.35 (thinner) with the index $\mathrm{I} = \mathrm{I}(h_{0.35})$ in parentheses.

also `span` $= 0.35$. The former smoothes out more fluctuations and thus reveals general patterns, which are fairly increasing, whereas `span` $= 0.35$ maintains more minute details. Table 4.5.1 summarizes the results.

Study subjects		Data		$\text{LOESS}_{0.75}$		$\text{LOESS}_{0.35}$	
x	**y**	I	RI%	I	RI%	I	RI%
$\text{Mechanics}_{[9,77]}$	Vectors	0.545	-3.540	0.972	-0.715	0.744	-14.483
$\text{Vectors}_{[9,77]}$	Mechanics	0.565	3.670	0.979	0.720	0.870	16.935
AI%		2.000		0.700		12.600	

Table 4.5.1: Closed-book examination summaries

The reported values of the index I suggest that Vectors vs. Mechanics exhibits a more increasing pattern than Mechanics vs. Vectors. This is also seen from the values of the relative index of interchangeability, RI%, which is positive for Vectors vs. Mechanics (and thus negative for Mechanics vs. Vectors) irrespective of the degree of smoothing. Hence, we conclude that students with higher scores in Vectors are more likely to get higher scores in Mechanics than the other way around, that is, when Mechanics precedes Vectors. This, we think, is due to the fact that Vectors is a fundamental subject for learning Mechanics; think of, e.g., the notion of force. To support this observation, we refer to the introductory sections of the classical textbook by Synge and Griffith (1949), who first recall basics of Vectors and only then teach Mechanics.

In view of the above, it becomes revealing why curriculum developers tend to include Mechanics modules into Mathematics classes. To illustrate the point, Kitchen et al. (1997) argue that in order to strengthen the appreciation of Mathematics, students should study Kinematics, Statics, and Dynamics, which make up parts of Mechanics and require knowledge of Vectors. Moreover, the authors argue that the use of illustrations based on Mechanics make Mathematics more relevant and thus more appreciated. Consequently, changes in Mathematics curricula have the potential of affecting Mechanics modules, which can in turn become particulary worrisome among those who teach first-year engineering students at universities (e.g., Lee et al., 2006, and references therein). The results reported in Table 4.5.1 are in good agreement with the aforementioned observations, and may therefore lend support to those in favour of encouraging students not to avoid "harder" study subjects.

We now take a look at the issue of interchangeability of Mechanics and Vectors with the aid of the absolute index of interchangeability, AI%. For the raw scatterplot, AI% is 2%, which is a relatively small number, likely due to the noise, but not to the pattern itself. We can smooth out the noise using a LOESS fit with a large `span` value. For example, the default value `span` = 0.75 smoothes out a lot of variability and makes the two fits virtually increasing: the index I values are 0.972 and 0.979, quite close to the maximum 1. By setting `span` to 0.35, the absolute index of interchangeability surges to 12.6%, which is large, and we would therefore hesitate to state that Mechanics and Vectors are interchangeable. Reiterating our earlier discussion based on RI%, and also recalling our note concerning Synge and Griffith (1949), and further arguments by Kitchen et al. (1997), we would tend to believe that viewing Vectors as an explanatory variable for Mechanics is more appropriate than the other way around.

4.5.2 Open-book examinations

Algebra, Analysis, and Statistics are the three subjects in the dataset of Mardia et al. (1979, pp. 3-4) that were assessed using open-book examinations. Hence, we have three pairs of scatterplots, whose summaries are in Table 4.5.2,

Study subjects		Data		$\text{LOESS}_{0.75}$		$\text{LOESS}_{0.35}$	
x	**y**	I	RI%	I	RI%	I	RI%
$\text{Algebra}_{[15,70]}$	Analysis	0.569	7.156	0.992	-0.800	0.879	18.623
$\text{Analysis}_{[15,70]}$	Algebra	0.531	-6.678	1.000	0.806	0.741	-15.700
AI%		3.800		0.800		13.800	
$\text{Analysis}_{[9,70]}$	Statistics	0.543	0.185	1.000	1.317	0.899	13.367
$\text{Statistics}_{[9,70]}$	Analysis	0.542	-0.184	0.987	-1.300	0.793	-11.791
AI%		0.100		1.300		10.600	
$\text{Algebra}_{[15,80]}$	Statistics	0.578	4.521	1.000	4.493	0.865	11.326
$\text{Statistics}_{[15,80]}$	Algebra	0.553	-4.325	0.957	-4.300	0.777	-10.173
AI%		2.500		4.300		8.800	

Table 4.5.2: Open-book examination summaries

with corresponding Figures B.1.1–B.1.3 relegated to Appendix B.1. Note the different intervals $[L, U]$ for each of the three pairs, and we shall therefore restrain from comparing, for example, Algebra vs. Analysis and Algebra vs. Statistics. However, we shall compare and discuss, for example, Algebra vs. Analysis with Analysis vs. Algebra.

Algebra and Analysis provide fundamental concepts for other subjects, such as Statistics, with Algebra playing a particularly prominent role, as argued by, e.g., Gamoran and Hannigan (2000). Based on the data of Mardia et al. (1979, pp. 3-4), we reach this conclusion from the raw data (RI%=7.156) as well as from the moderate $\text{LOESS}_{0.35}$ fit (RI%=18.623). The default $\text{LOESS}_{0.75}$ fit (RI%=−0.800) gives a slight preference to Analysis over Algebra.

Remark 4.5.1. *A possible reason for this change of preference is likely due to an outlier: one student's Algebra score deviates considerably from the overall pattern of scores. Obviously, the LOESS fit under the default value* `span` $= 0.75$ *smoothes out the outlier, making* $\text{I}(\text{Analysis}, \text{Algebra})$ *equal to* 1*, whereas* $\text{I}(\text{Algebra}, \text{Analysis})$ *takes the value* 0.992*.*

The observed slight uncertainty when deciding which of the two study subjects – Algebra or Analysis – should be taught first does not seem to really matter in practice because, as far as we are aware of, Algebra and Analysis are considered fundamental subjects, focussing on different aspects of mathematics, and are thus often taught at the same time. Hence, neither of them can be easily substituted by another one: better performance in these two subjects leads to better performance in other subjects, such as Statistics, as seen from the RI values in Table 4.5.2. Note in this regard that irrespective of the degree of smoothing, the RI values for Statistics vs. Analysis and Statistics vs. Algebra are negative, and thus the empirical evidence provided by Mardia et al. (1979, pp. 3-4) suggests that Analysis and Algebra should be taught first and only then Statistics.

4.5.3 Closed-book vs. open-book examinations

In the previous two sections, we discussed subjects within closed-book examinations and also within open-book examinations. In the current section, we look at the six combinations

with one subject from closed-book examinations and another subject from open-book examinations. Table 4.5.3 summarizes our findings, with corresponding Figures B.1.4–B.1.9

Study subjects		Data		$LOESS_{0.75}$		$LOESS_{0.35}$	
x	**y**	I	RI%	I	RI%	I	RI%
$Mechanics_{[15,77]}$	Algebra	0.545	-1.089	1.000	0.000	0.756	0.265
$Algebra_{[15,77]}$	Mechanics	0.551	1.101	1.000	0.000	0.754	-0.265
AI%		0.600		0.000		0.200	
$Mechanics_{[9,70]}$	Analysis	0.516	-6.011	0.876	-12.400	0.640	-23.900
$Analysis_{[9,70]}$	Mechanics	0.549	6.395	1.000	14.155	0.841	31.406
AI%		3.300		12.400		20.100	
$Mechanics_{[9,77]}$	Statistics	0.541	7.129	0.966	15.137	0.700	24.113
$Statistics_{[9,77]}$	Mechanics	0.505	-6.654	0.839	-13.147	0.564	-19.429
AI%		3.600		12.700		13.600	
$Vectors_{[15,80]}$	Algebra	0.554	-3.819	1.000	0.000	0.964	15.865
$Algebra_{[15,80]}$	Vectors	0.576	3.971	1.000	0.000	0.832	-13.693
AI%		2.200		0.000		13.200	
$Vectors_{[9,70]}$	Analysis	0.517	-4.436	0.775	-22.111	0.560	-20.680
$Analysis_{[9,70]}$	Vectors	0.541	4.642	0.995	28.387	0.706	26.071
AI%		2.400		22.000		14.600	
$Vectors_{[9,81]}$	Statistics	0.538	1.701	0.835	-14.534	0.723	9.380
$Statistics_{[9,81]}$	Vectors	0.529	-1.673	0.977	17.006	0.661	-8.575
AI%		0.900		14.200		6.200	

Table 4.5.3: Comparison for cross category

relegated to Appendix B.1. Note from Table 4.5.3 that the values of the index of increase differ from those in Table 4.3.1. The piecewise linear and LOESS fits also differ from the corresponding ones in Figures 4.2.2 and 4.3.1, because the latter two figures are not based on unified ranges calculated by formula (4.5.1), whose notion was only introduced in Section 4.4.4.

From the RI values in Table 4.5.3, we see that irrespective of the degree of smoothing, Analysis as a study subject should precede Mechanics, which in turn should precede Statistics. Furthermore, Analysis should precede Vectors. If we do not take into account the RI values based on raw data and concentrate only on the two LOESS fits, then we conclude that both Mechanics and Vectors should precede Algebra. As to Vectors and Statistics, the two LOESS fits give somewhat conflicting suggestions, thus implying that the two subjects may not be good at determining each other's scores. This we find natural: given our teaching experience, these two subjects – on the introductory level – are hardly related to each other. We should add, however, that advanced statistics requires good knowledge of vectors, matrices, and related concepts, which can in turn be used as illuminating examples when teaching vectors and matrices.

4.6 Concluding notes

Measuring relationships and, consequently, monotonicity relationships between paired variables is an important and highly challenging problem, especially when relationships

- are inherently non-linear,
- cannot be described using closed-form formulas.

To tackle such problems, we have employed the index of increase, which is a relatively new technique that has emerged from the works of Davydov and Zitikis (2017), Chen and Zitikis (2017), and Chen et al. (2018). Since the use of computers is essential, we have thoroughly described the packages and algorithms that we have used in our computations and explorations.

By revisiting the popular dataset of Mardia et al. (1979, pp. 3-4), which is frequently used by university teachers to illustrate various classical concepts of multivariate analysis, we have enabled those familiar with the textbook and the dataset to see the need for, and benefits of, thinking outside the box. To facilitate the task, we have provided a comprehensive explanation of the index of increase, its calculation techniques under various

scenarios, and interpretations. For example, we have found the following relationships between different study subjects with respect to the timing of exposure to students:

- Vectors $\prec$ Mechanics (Section 4.5.1)
- Algebra $\prec$ Statistics (Section 4.5.2)
- Analysis $\prec$ Statistics (Section 4.5.2)
- Algebra $\perp\!\!\!\perp$ Analysis (Section 4.5.2)
- Analysis $\prec$ Mechanics $\prec$ Statistics (Section 4.5.3)
- Analysis $\prec$ Vectors (Section 4.5.3)
- Mechanics $\prec$ Algebra (Section 4.5.3)
- Vectors $\prec$ Algebra (Section 4.5.3)
- Vectors $\perp\!\!\!\perp$ Statistics (Section 4.5.3)

where $S_1 \prec S_2$ means that prior familiarity with subject S_1 is beneficial for learning subject S_2, and $S_1 \perp\!\!\!\perp S_2$ when the two subjects do not clearly exhibit $S_1 \prec S_2$ or $S_2 \prec S_1$, and can thus be taught in any order. (The sign $\perp\!\!\!\perp$ is frequently used in Statistics and Probability to indicate independence, which in the current context connotes "timing independence.")

Next, we make a few cautionary notes that we think are particularly important when dealing with problems such as those we have tackled in the present paper.

First, our interpretations and suggested decision-making are based on the data of Mardia et al. (1979, pp. 3-4), and should not be lightheartedly generalized or extended to other educational contexts. Nevertheless, as is the case with many statistical methods and techniques, they are insightful when used with care and in conjunction with subject-matter knowledge.

Second, not only the subject-matter knowledge that determines whether or not we are likely to be right (or wrong) when making decisions but also the knowledge of instructor's personality and performance are crucial. For more details and references on this topic,

and for associated consequences when teaching, e.g., Calculus and Algebra, we refer to Wade et al. (2017). The "conversation" by Taylor (2019) provides further enlightening thoughts and additional references.

Third, the classically trained statistical researcher would spontaneously ask what would happen if the sample size n (i.e., the class size in the current context) would grow indefinitely. Firstly, such situations cannot happen in the context of educational research, but if, for the sake of argument, this happens, then the answer would undoubtedly be "it would be a mess." Interestingly, in contexts outside of educational research, such as insurance and finance (e.g., Gribkova and Zitikis, 2018; Ren et al., 2019) and engineering (e.g., Gribkova and Zitikis, 2019a,b), exploring the index of increase when the sample size n grows indefinitely is meaningful and even pivotal.

It is the latter studies from which we know that the above reply "it would be a mess" is indeed the correct answer, in the sense that if each observation is a non-deterministic outcome (as is the case with student marks), then when n grows to infinity, the index of increase inevitably converges to 0.5, meaning that the underlying scatterplot grows into a chaotic pattern, with no clear upward or downward trends. In a sense, this is natural and does manifest in large-size (say, more than 200 students) introductory statistics/calculus classes, whose main purpose, roughly speaking, is not to make subtle recommendations to students such as directing them to theoretical or computational statistics/calculus studies – this is usually done in upper-year and small class-size environments – but to simply make a general assessment of student suitability to achieve a comprehensive university-level education.

Finally, a few notes concerning future work are in order. First, we reiterate that our choice of the classical dataset of Mardia et al. (1979, pp. 3-4) has been deliberate: we have aimed at contrasting classical and new techniques in a highly accessible way. But this, in turn, raises an interesting research question. Namely, with the currently rapidly developing societal need for more computer proficiency and familiarity with topics such as machine learning and artificial intelligence, are the above reached conclusions based on an old dataset still relevant today? To have well-informed answers, and we think

there is no single correct answer to this question, one would need to run observational and experimental studies, whose outcomes may depend on geographical regions, societal traditions, and so on. These are very interesting research problems, and much has already been done by educational researchers; the present thesis offers them an additional tool of analysis.

Chapter 5

Estimating the index of increase via balancing deterministic and random data

5.1 Motivation

Dynamic processes in populations are often described using functions (e.g., Bebbington et al., 2007, 2011, and references therein). They are observed in the form of data points, usually contaminated by measurement errors. We may think of these points as randomly perturbed true values of underlying functions, whose measurements are taken at certain time instances. The functions, their rates of change, and de/acceleration can be and frequently are non-monotonic. Nevertheless, it is of interest to assess and even compare the extent of their monotonicity, or lack of it. We refer to Qoyyimi (2015) for a discussion and literature review of various applications.

Several methods for assessing monotonicity have been suggested in the literature (e.g., Davydov and Zitikis, 2005, 2017; Qoyyimi and Zitikis, 2014, 2015). In particular, Davydov and Zitikis (2017) show the importance of such assessments in insurance and finance, especially when dealing with weighted insurance calculation principles (Furman and Zitikis, 2008), among which we find such prominent examples as the Esscher (Bühlmann, 1980,

1984), Kamps (1998), and Wang (1995, 1998) premiums. Furthermore, Egozcue et al. (2011) provide problems in economics where the sign of the covariance

$$\mathbf{Cov}[X, w(X)] \tag{5.1.1}$$

needs to be determined for various classes of function w. One of such examples concerns the slope of indifference curves in two-moment expected utility theory (e.g., Eichner and Wagener, 2009; Sinn, 1990; Wong, 2006, and references therein). Another problem concerns decision making (e.g., speculation, normal backwardation, contango, etc.) of competitive companies under price uncertainty (e.g., Feder et al., 1980; Hey, 1981; Meyer and Robinson, 1988, and references therein).

Lehmann (1966) has shown that if the function w is monotonic, then covariance (5.1.1) is either positive (when w is increasing) or negative (when w is decreasing). This monotonicity assumption on w, though satisfied in a number of cases of practical interest, excludes a myriad of important cases with more complex risk profiles. For example, when dealing with the aforementioned economics-based problems, the role of w is played by the derivative u' of the underlying utility function, which may not be convex or concave everywhere, as argued and illustrated by, e.g., Friedman and Savage (1948), Markowitz (1952), Kahneman and Tversky (1979), Tversky and Kahneman (1992), among others. Hence, since w might be non-monotonic, how far can this function be from being monotonic, or increasing? Furthermore, since the population risk- or utility-profile cannot be really known, the non-monotonicity of w needs to be assessed from data, and this leads us to the statistical problem of this paper.

In addition, supported by the examples of Anscombe (1973) on potential pitfalls when using the classical correlation coefficient, Chen and Zitikis (2017) argue in favour of using the index of increase, as defined by Davydov and Zitikis (2017), for assessing non-monotonicity of scatterplots. Chen and Zitikis (2017) apply this approach to analyze and compare student performance in subjects such as mathematics, reading and spelling, and illustrate their reasoning on data provided by Thorndike and Thorndike Christ (2010). One of the methods discussed by Chen and Zitikis (2017) deals with scatterplots representing finite populations, in which case large-sample estimation is not possible. The other method

involves large-sample regression techniques (Figure 5.1.1), in which case Chen and Zitikis (2017) calculate the corresponding indices of increase using a numerical approach, that gives rise to the values denoted by I and reported in the bottom-right corners of the panels of Figure 5.1.1. Though important, these methods do not allow direct large-sample non-

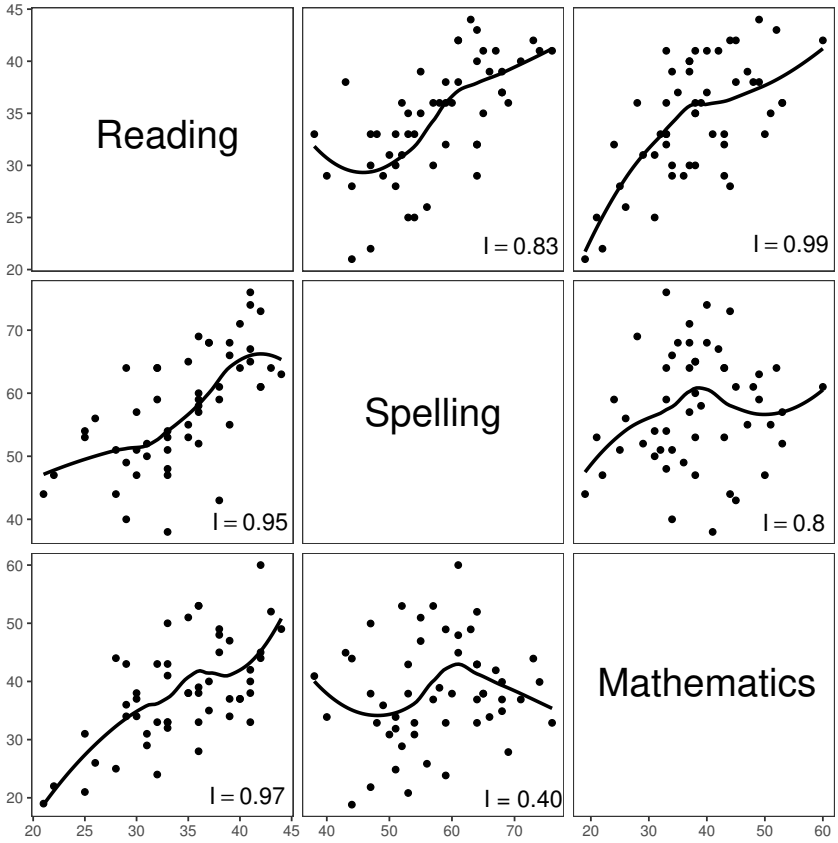


Figure 5.1.1: Regression curves fitted to the student scores reported by Thorndike and Thorndike Christ (2010), and their indices of increase.

monotonicity quantifications and thus inferences about larger populations. In this paper, therefore, we offer a statistically attractive and computationally efficient procedure for assessing data patterns that arise from non-monotonic patterns contaminated by random measurement errors.

We have organized the rest of the chapter as follows. In Section 5.2, we introduce the index and provide basic arguments leading to it. In Section 5.3, we explain why and how the index needs to be adjusted in order to become useful in situations when random

measurement errors are present. In Section 5.4, we rigorously establish consistency of the estimator and introduce relevant data-exploratory and cross-validatory techniques. Since the limiting distribution of the estimator is complex, in Section 5.5 we implement a bootstrap-based procedure for determining standard errors and, in turn, for deriving confidence intervals. Section 5.6 concludes the paper with a brief summary of our main contributions.

5.2 The index of increase

Davydov and Zitikis (2017) have introduced the index of increase

$$\mathrm{I}(h_0) = \frac{\int_a^b (h_0')_+ \mathrm{d}\lambda}{\int_a^b |h_0'| \mathrm{d}\lambda} \quad \left(:= \frac{\int_a^b (h_0'(t))_+ \mathrm{d}t}{\int_a^b |h_0'(t)| \mathrm{d}t} \right) \tag{5.2.1}$$

for any absolutely continuous (e.g., differentiable) function h_0 on interval $[a, b]$, where $(h_0')_+ := \max\{h_0', 0\}$, and ":=" denotes equality by definition. Throughout the chapter, we use λ to denote the Lebesgue measure, which helps us to write integrals compactly, as seen from the ratios above. We shall explain how the index arises later in the current section. Of course, this framework reduces to the unit interval $[0, 1]$ by considering the function $h(t) := h_0(a + (b - a)t)$ instead of h_0. Namely, we have

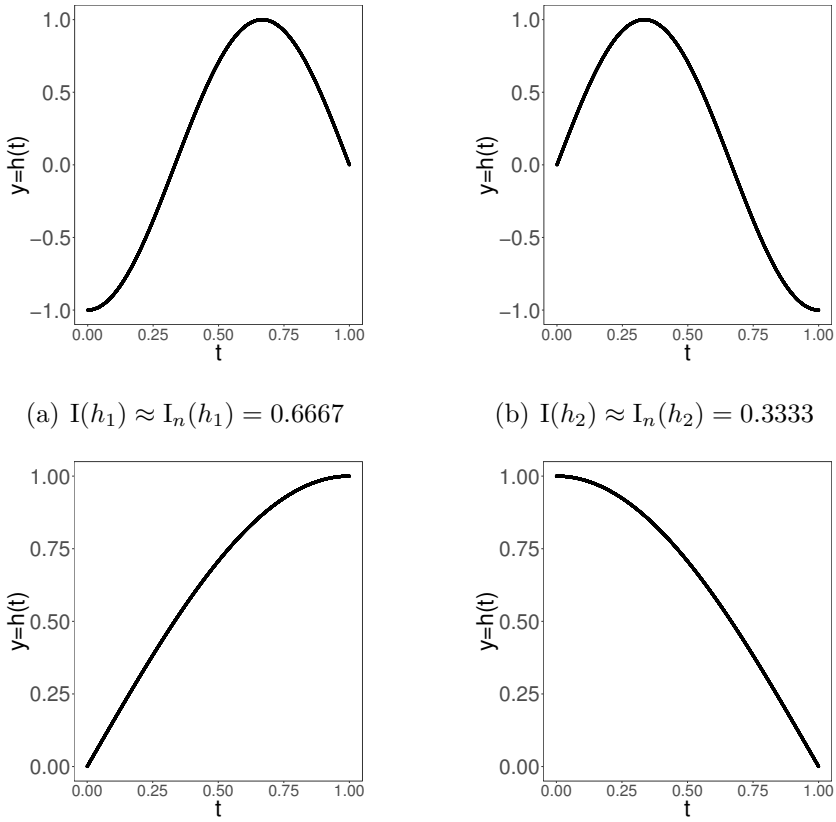
$$\mathrm{I}(h_0) = \frac{\int_0^1 (h')_+ \mathrm{d}\lambda}{\int_0^1 |h'| \mathrm{d}\lambda} =: \mathrm{I}(h). \tag{5.2.2}$$

To illustrate, in Figure 5.2.1 we have visualized the following quartet of functions

$$\begin{aligned} h_1(t) &= \sin\Big(-\frac{\pi}{2} + \frac{3\pi}{2}t\Big), \quad h_2(t) = \cos\Big(-\frac{\pi}{2} + \frac{3\pi}{2}t\Big), \\ h_3(t) &= \sin\Big(\frac{\pi}{2}t\Big), \quad h_4(t) = \cos\Big(\frac{\pi}{2}t\Big), \end{aligned} \tag{5.2.3}$$

and we have also calculated their indices of increase. Since h_3 and h_4 are monotonic functions on the interval $[0, 1]$, calculating their indices of increase using formula (5.2.2) is trivial, but the same task in the case of non-monotonic functions h_1 and h_2 requires some effort. To facilitate such calculations in a speedy fashion, and irrespective of the complexity of functions, we suggest using the numerical approximation

$$\mathrm{I}_n(h) := \frac{\sum_{i=2}^n (h(t_{i,n}) - h(t_{i-1,n}))_+}{\sum_{i=2}^n |h(t_{i,n}) - h(t_{i-1,n})|} \tag{5.2.4}$$



(a) $\mathrm{I}(h_1) \approx \mathrm{I}_n(h_1) = 0.6667$ (b) $\mathrm{I}(h_2) \approx \mathrm{I}_n(h_2) = 0.3333$

(c) $\mathrm{I}(h_3) = \mathrm{I}_n(h_3) = 1$ (d) $\mathrm{I}(h_4) = \mathrm{I}_n(h_4) = 0$

Figure 5.2.1: The functions of quartet (5.2.3) and their indices of increase

with $t_{i,n} = (i-1)/(n-1)$ for $i = 1, \ldots, n$. Intuitively, $\mathrm{I}_n(h)$ is the proportion of the upward movements of the function h with respect to all the movements, upward and downward.

Knowing the convergence rate of $\mathrm{I}_n(h)$ to $\mathrm{I}(h)$ when $n \to \infty$ is important as it allows us to set a frequency n at which the measurements of $h(t_{i,n})$ could be taken during the observation period (e.g., unit interval $[0, 1]$) so that any pre-specified estimation precision of $\mathrm{I}(h)$ would be achieved. For example, we have used $n = 10000$ to calculate the index values with the four-digit precision reported in Figure 5.2.1. We refer to Chen and Zitikis (2017) for details on computational precision.

The following proposition, which is a special case of Lemma 5.4.1 below, establishes the convergence rate based on the level of smoothness of the function h.

Proposition 5.2.1. *Let h be a differentiable function defined on the unit interval $[0, 1]$,*

and let its derivative h' be γ-Hölder continuous for some $\gamma \in (0,1]$. Then, when $n \to \infty$, we have

$$\sum_{j=2}^{n} \ell\Big(h(t_{i,n}) - h(t_{i-1,n})\Big) = \int_0^1 \ell\big(h'\big)\, d\lambda + O(n^{-\gamma}) \tag{5.2.5}$$

for any positively homogeneous and Lipschitz function ℓ (e.g., $\ell(t) = t_+$ and $\ell(t) = |t|$). Consequently,

$$\mathrm{I}_n(h) = \mathrm{I}(h) + O(n^{-\gamma}). \tag{5.2.6}$$

To explain the basic meaning of the index $\mathrm{I}(h)$, we start with an un-normalized version of it, which we denote by $\mathrm{J}(h)$. Namely, let $\mathcal{F}$ denote the set of all absolutely continuous functions f on the interval $[0,1]$ such that $f(0) = 0$. Denote the total variation of $f \in \mathcal{F}$ on the interval $[0,1]$ by $\|f\|$, that is, $\|f\| = \int_0^1 |f'| \mathrm{d}\lambda$. Furthermore, by definition, we have $(f')_+ = \max\{f', 0\}$ and $(f')_- = \max\{-f', 0\}$, and we also have the equations $f' = (f')_+ - (f')_-$ and $|f'| = (f')_+ + (f')_-$. Finally, we use $\mathcal{F}^-$ to denote the set of all the functions $f \in \mathcal{F}$ that are non-increasing. All of these are of course well-known fundamental notions of Real Analysis (e.g., Kolmogorov and Fomin, 1970; Dunford and Schwartz, 1988; Natanson, 2016).

For any function $h \in \mathcal{F}$, we define its (un-normalized) index of increase $\mathrm{J}(h)$ as the distance between h and the set $\mathcal{F}^-$, that is,

$$\mathrm{J}(h) = \inf_{f \in \mathcal{F}^-} \|h - f\|. \tag{5.2.7}$$

Obviously, if h is non-increasing, then $\mathrm{J}(h) = 0$, and the larger the value of $\mathrm{J}(h)$, the farther the function h is from being non-increasing on the interval $[0,1]$. Determining the index $\mathrm{J}(h)$ using its definition (5.2.7) is not, however, a straightforward task, and to facilitate it, we next establish a very convenient integral representation of $\mathrm{J}(h)$.

Theorem 5.2.1. *(Davydov and Zitikis, 2017) The infimum in definition (5.2.7) is attained at any function $f_1 \in \mathcal{F}^-$ such that $f_1' = -(h')_-$, and thus*

$$\mathrm{J}(h) = \int_0^1 (h')_+ \mathrm{d}\lambda. \tag{5.2.8}$$

A direct proof of this theorem was not provided by Davydov and Zitikis (2017), who refer to a more general and abstract result. Nevertheless, a short and enlightening proof exists, and we present it next.

Proof of Theorem 5.2.1. We start with the note that the bound $\mathrm{J}(h) \leq \|h - f\|$ holds for every function $f \in \mathcal{F}^-$, and in particular for the function f_1 specified in the formulation of the theorem. Hence,

$$\begin{aligned} \mathrm{J}(h) &\leq \int_0^1 |h' - f_1'| \mathrm{d}\lambda \\ &= \int_0^1 |h' + (h')_-| \mathrm{d}\lambda \\ &= \int_0^1 (h')_+ \mathrm{d}\lambda. \end{aligned} \tag{5.2.9}$$

It now remains to show the opposite bound. Let T^+ be the set of all $t \in [0,1]$ such that $h'(t) > 0$, and let T^- be the complement of the set T^+, which consists of all those $t \in [0,1]$ for which $h'(t) \leq 0$. Then

$$\begin{aligned} \mathrm{J}(h) &= \inf_{f \in \mathcal{F}^-} \left(\int_{T^+} |h' - f'| \mathrm{d}\lambda + \int_{T^-} |h' - f'| \mathrm{d}\lambda \right) \\ &\geq \inf_{f \in \mathcal{F}^-} \int_{T^+} |h' - f'| \mathrm{d}\lambda \\ &= \inf_{f \in \mathcal{F}^-} \left(\int_{T^+} h' \mathrm{d}\lambda + \int_{T^+} |f'| \mathrm{d}\lambda \right) \\ &= \int_0^1 (h')_+ \mathrm{d}\lambda, \end{aligned} \tag{5.2.10}$$

where the last equation holds when $f'(t) = 0$ for all $t \in T^+$, that is, when $f' = -(h')_-$. Bounds (5.2.9) and (5.2.10) establish equation (5.2.8), thus finishing the proof of Theorem 5.2.1. □

The index $\mathrm{J}(h)$ never exceeds $\|h\|$, and so the normalized version of $\mathrm{J}(h)$ is

$$\mathrm{I}(h) := \mathrm{J}(h)/\|h\|,$$

which is exactly the index of increase given by equation (5.2.2). In summary, the index of increase $\mathrm{I}(h)$ is the normalized distance of the function h from the set $\mathcal{F}^-$ of all non-increasing functions on the interval $[0,1]$: we have $\mathrm{I}(h) = 0$ when the function h is non-increasing, and $\mathrm{I}(h) = 1$ when the function is non-decreasing. The closer the index $\mathrm{I}(h)$ is to 1, the more (we say) the function h is increasing, and the closer it is to 0, the less (we say) the function h is increasing or, equivalently, the more it is decreasing.

5.3 Practical issues and their resolution

Measurements are usually taken with errors, whose natural model is some distribution (e.g., normal) with mean 0 and finite variance σ^2. In other words, the numerical index $\mathrm{I}_n(h)$ turns into the random index of increase

$$\mathrm{I}_n(h,\varepsilon) := \frac{\sum_{i=2}^{n}(Y_{i,n} - Y_{i-1,n})_+}{\sum_{i=2}^{n}|Y_{i,n} - Y_{i-1,n}|}, \tag{5.3.1}$$

where, for $i = 1, \ldots, n$,

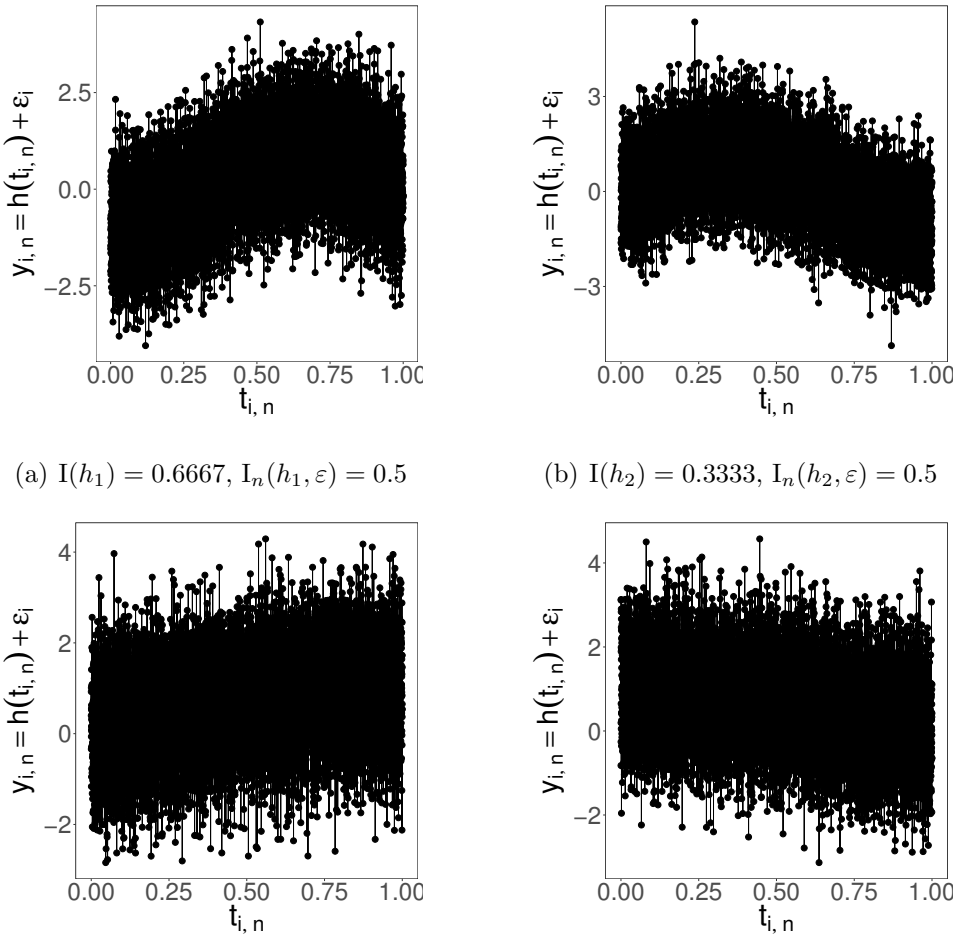
$$Y_{i,n} = h(t_{i,n}) + \varepsilon_i. \tag{5.3.2}$$

Right at the outset, however, serious issues arise. To illustrate them in a speedy and transparent manner, we put aside mathematics such as in Davydov and Zitikis (2004, 2007) and, instead, simulate $n = 10000$ standard normal errors ε_i, thus obtaining four sequences $Y_{i,n}$ corresponding to the functions of quartet (5.2.1). Then we calculate the corresponding indices of increase using formula (5.3.1). All of the obtained values of $\mathrm{I}_n(h)$ are virtually equal to 1/2 (see Figure 5.3.1). Clearly, there is something amiss.

It is not, however, hard to understand the situation: when all ε_i's are zero, the definition of the integral as the limit of the Riemann sums works as intended, but when the ε_i's are not zero, they accumulate so much when n gets larger that the deterministic part (i.e., the Riemann sum) gets hardly, if at all, visible (compare Figures 5.2.1 and 5.3.1). In summary, we are facing two extremes:

- If the model is purely deterministic in the sense that there are no measurement errors, which we can understandably argue to be outside the realm of practice, then the more frequently we observe the function h, the more precisely we can estimate its index of increase.

- If, however, there are measurement errors, as they usually are in practice, then the more frequently we observe the function, the less precisely we can estimate its index of increase, because the accumulated measurement errors obscure the deterministic part.



(a) $\mathrm{I}(h_1) = 0.6667$, $\mathrm{I}_n(h_1, \varepsilon) = 0.5$

(b) $\mathrm{I}(h_2) = 0.3333$, $\mathrm{I}_n(h_2, \varepsilon) = 0.5$

(c) $\mathrm{I}(h_3) = 1$, $\mathrm{I}_n(h_3, \varepsilon) = 0.5$

(d) $\mathrm{I}(h_4) = 0$, $\mathrm{I}_n(h_4, \varepsilon) = 0.5$

Figure 5.3.1: The indices of increase and their numerical estimators for quartet (5.2.3) with added random errors.

Neither of the two extremes can be of much interest, or use, for reasons either practical or computational. The purpose of this chapter is to offer a way out of this difficulty by showing how to strike a good balance between determinism and randomness inherent in the problem.

We next present an intuitive consideration that will guide our subsequent mathematical considerations, and it will also hint at potential applications of this research. Namely, suppose that the unit interval $[0, 1]$ represents an one-day observation period, and let an observation be taken (e.g., by a measuring equipment) every second. Hence, in total, we have $n = 86400$ observations $Y_{i,n}$ of the (unknown) function h, and they are prone

to measurement errors ε_i as in expression (5.3.2). For the sake of argument, let ε_i's be i.i.d. standard normal. If we calculate the index $\mathrm{I}_n(h, \varepsilon)$ based on these data, we already know the problem: $\mathrm{I}_n(h, \varepsilon)$ tends to $1/2$ when $n \to \infty$. To diminish the influence of these errors, we average the observed values:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} Y_{i,n} &= \frac{1}{n}\sum_{i=1}^{n} h(t_{i,n}) + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \\ &\overset{d}{\approx} \int_0^1 h \mathrm{d}\lambda + \frac{1}{\sqrt{n}}\varepsilon_0, \end{aligned}$$

where $\overset{d}{\approx}$ means 'approximately in distribution,' and ε_0 follows the standard normal distribution. However, in the process of averaging out the errors, we have inevitably also averaged the deterministic part and arrived at the mean value $\int_0^1 h \mathrm{d}\lambda$ of the function h. This value has very little to do with the index $\mathrm{I}(h)$, which fundamentally relies on the derivative h'. In short, we have clearly over-averaged the observations $Y_{i,n}$: having maximally reduced the influence of measurement errors, we have obscured the function h so much that the estimation of $\mathrm{I}(h)$ has become impossible. Clearly, we need to adopt a more tempered approach.

Hence, we group the observations into only $M < n$ groups $G_{j,n}$, $j = 1, \ldots, M$, whose cardinalities $N := \#(G_{j,n})$ we assume to be the same for all $j = 1, \ldots, M$. It is convenient to re-parametrize these choices using parameter $\alpha \in (0, 1)$, which turns M and N into

$$M = \lfloor n^{\alpha} \rfloor \quad \text{and} \quad N = \lfloor n^{1-\alpha} \rfloor.$$

This re-parametrization is not artificial. It is, in a way, connected to smoothing histograms and estimating regression functions, and in particular to bandwidth selection in these research areas. We shall elaborate on this topic more in the next section. At the moment, we only note that the aforementioned connection plays a pivotal role in obtaining practically useful and sound estimates of the parameter α.

To gain additional intuition on the grouping parameter α, we come back for a moment to our numerical example with the one-day observation period, which is comprised of $n = 86400$ observations, one per second. Suppose that we decide to average the sixty observations within each minute. Thus, we have $N = 60$ and in this way produce $M = 1440$

new data points, which we denote by $\widetilde{Y}_{j,n}$. Since $NM = n$, we have $\alpha = 1 - \log(N)/\log(n)$ and thus $\alpha = 0.6398$. If, however, instead of averaging minute-worth data we decide to average, for example, hour-worth data, then we have $N = 3600$ (=group cardinality), $M = 24$ (=number of groups), and thus $\alpha = 0.2796$.

Continuing our general discussion, we average the original observations $Y_{i,n}$, $i = 1 \ldots, n$, falling into each group $G_{j,n}$ and in this way obtain M group-averages

$$\widetilde{Y}_{j,n} := \frac{1}{N} \sum_{i \in G_{j,n}} Y_{i,n}, \quad j = 1, \ldots, M.$$

Based on these averages, we modify the earlier introduced index $\mathrm{I}_n(h, \varepsilon)$ as follows:

$$\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon) := \frac{\sum_{j=2}^{M} (\widetilde{Y}_{j,n} - \widetilde{Y}_{j-1,n})_+}{\sum_{j=2}^{M} |\widetilde{Y}_{j,n} - \widetilde{Y}_{j-1,n}|}. \tag{5.3.3}$$

The problem that we now face is to find, if exist, those values of $\alpha \in (0, 1)$ that make the index $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ converge to $\mathrm{I}(h)$ when $n \to \infty$. This is the topic of the next section.

5.4 Consistency

The following theorem establishes consistency of the estimator $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ and, in particular, specifies the range of possible α values.

Theorem 5.4.1. *Let h be a differentiable function defined on the unit interval $[0, 1]$, and let its derivative h' be γ-Hölder continuous for some $\gamma \in (0, 1]$. If $\alpha \in (0, 1/3)$, then $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ is a consistent estimator of $\mathrm{I}(h)$, that is, when $n \to \infty$, we have*

$$\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon) \xrightarrow{\mathbf{P}} \mathrm{I}(h). \tag{5.4.1}$$

The rate of convergence is of the order

$$O_{\mathbf{P}}(1) n^{-\min\{\delta(\alpha), \rho(\alpha)\}} \tag{5.4.2}$$

with $\delta(\alpha) = \alpha\gamma$ arising from the deterministic part of the problem, that is, associated with the function h, and $\rho(\alpha) = (1 - 3\alpha)/2$ arising from the random part, that is, associated with the measurement errors ε_i's.

We next discuss the choice of α from the theoretical and practical perspectives, which do not coincide due to a number of reasons, such as the fact that theory is concerned with asymptotics when $n \to \infty$, while practice deals with finite values of n, though possibly very large. Under the (practical) non-asymptotic framework, any value of $\alpha \in (0, 1]$ is, in principle, acceptable because the quantities $O_{\mathbf{P}}(1)$ and $n^{-\min\{\delta(\alpha),\rho(\alpha)\}}$ in the specification of convergence rate (5.4.2) interact, as both of them depend on h and α.

Under the (theoretical) asymptotic framework, the values $\alpha = 0$ and 1 have to be discarded immediately, as we have already noted. The remaining α's should, as Theorem 5.4.1 tells us, be further restricted to only those below $1/3$. Since we wish to chose α that results in the fastest rate of convergence, we maximize the function $\alpha \mapsto \min\{\delta(\alpha), \rho(\alpha)\}$ and get

$$\alpha_{\max} = \frac{1}{3 + 2\gamma}. \tag{5.4.3}$$

For example, if the second derivative $h''(t)$ is uniformly bounded on the interval $[0, 1]$, which is the case in all our illustrative examples, then $\gamma = 1$ and thus $\alpha_{\max} = 1/5$.

The grouping and averaging technique that we employ is closely related to smoothing in non-parametric density and regression estimation (e.g., Silverman, 1986; Härdle, 1991; Scott, 2015, and references therein). To elaborate on this connection, we recall that the number of groups is $M \approx n^{\alpha}$, whose reciprocal

$$b := 1/M \approx n^{-\alpha} \tag{5.4.4}$$

would play the role of 'bandwidth.' In non-parametric density and regression estimation, the optimal bandwidth is of the order $O(n^{-1/5})$ when $n \to \infty$, which in our case corresponds to $\alpha_{\max} = 1/5$. Hence, $\alpha = 0$ means only one bin/group and thus over-smoothing, whereas $\alpha = 1$ means as many bins/groups as there are observations, and thus under-smoothing. Of course, as we have already noted above, the values $\alpha = 0$ and $\alpha = 1$ are excluded, unless all the measurement errors vanish, in which case smoothing is not necessary and thus $\alpha = 1$ can be used, as we indeed did earlier when dealing with the numerical index $\mathrm{I}_n(h)$.

Thinking of the role of γ-Hölder continuity of h' on the problem, it is useful to look at two extreme cases: First, when $\gamma = 1$, we have $\alpha_{\max} = 1/5$ from formula (5.4.3), which

corresponds (under weak conditions) to the optimal bandwidth $O(n^{-1/5})$ in non-parametric density and regression estimation. Second, when no smoothing is applied, like in the case of the histogram density-estimator, then (under weak conditions) the optimal bandwidth is of the order $O(n^{-1/3})$, which corresponds to $\alpha_{\max} = 1/3$ when $\gamma = 0$, which essentially means boundedness but no continuity of h'.

Hence, choosing an appropriate value of the grouping parameter α is a delicate task. We next discuss two approaches: The first one is data-exploratory (visual) when we assume that we know the population and want to gain insights into what might happen in practice. The second, practice-oriented approach relies on the idea of cross-validation (e.g., Arlot and Celisse, 2010; Celisse, 2008, and references therein) and is designed to produce estimates of α based purely on data.

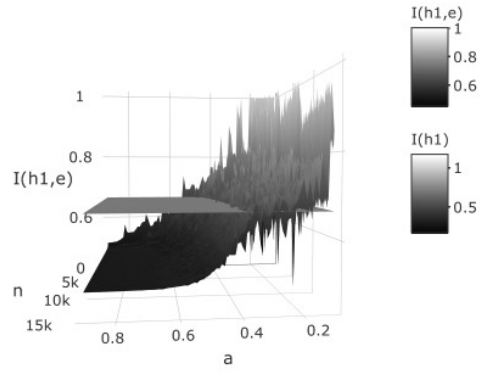
5.4.1 Data exploratory (visual) choice of α

To gain intuition on how to estimate the grouping parameter α from data, we start out with the functions in quartet (5.2.3), which we view as populations, and then we contaminate their observations with i.i.d. errors $\varepsilon_i \sim \mathcal{N}(0, 1)$ according to formula (5.3.2).

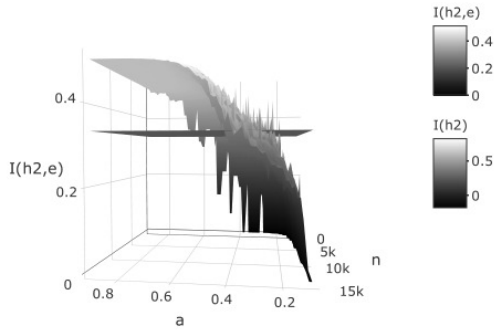
We have visualized the values of the estimator $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n and α in Figure 5.4.1, where the hyperplane in each panel is at the height of the corresponding actual index of increase $\mathrm{I}(h)$. For each panel, we visually choose a value of α which is in the intersection of the curved surface with the hyperplane, because in this case the index $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ is close to the actual index $\mathrm{I}(h)$.

Even though the chosen parameter α value, which we denote by α_{vi}, may not be optimal due to roughness of the surface, it nevertheless offers a sound choice, as we see from Figure 5.4.2 where we depict the convergence of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ to $\mathrm{I}(h)$ when n grows. In each panel, the horizontal 'reference' line is at the height of the actual index value.

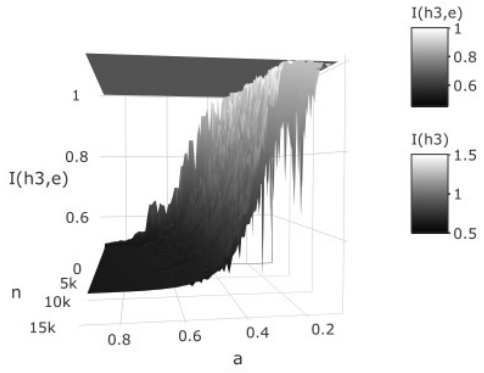
Note that in panel (a) of Figure 5.4.2, the visually obtained $\alpha_{\mathrm{vi}} = 0.35$ is slightly larger than $1/3$, but we have to say that we had decided on this value (as a good estimate) before we knew the result of Theorem 5.4.1, and thus before we knew the (theoretical) restriction $\alpha < 1/3$. Nevertheless, we have decided to leave the value $\alpha_{\mathrm{vi}} = 0.35$ as it is, without



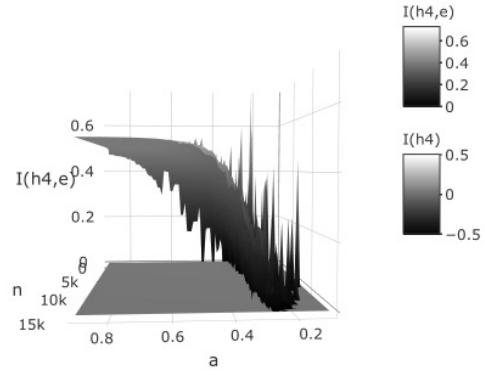
(a) The hyperplane at the height $\mathrm{I}(h_1) = 0.6667$



(b) The hyperplane at the height $\mathrm{I}(h_2) = 0.3333$



(c) The hyperplane at the height $\mathrm{I}(h_3) = 1$



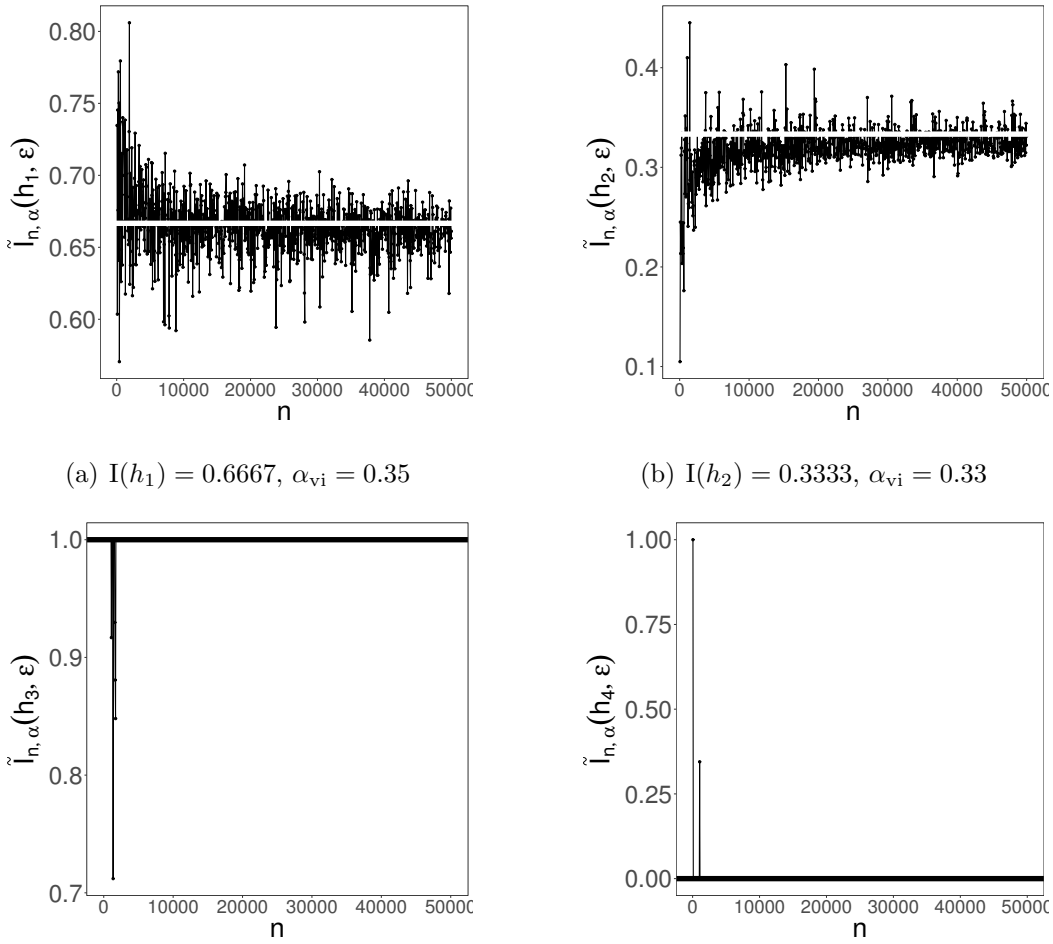
(d) The hyperplane at the height $\mathrm{I}(h_4) = 0$

Figure 5.4.1: Values of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n and α in the case of quartet (5.2.3).

tempering with our initial guess in any way. As we shall see in next Section 5.4.2, however, the purely data-driven and based on cross-validation α value is $\alpha_{\mathrm{cv}} = 0.28$, which is within the range $(0, 1/3)$ of theoretically acceptable α values.

5.4.2 Choosing α based on cross validation

As we have already elucidated, equation (5.4.4) connects our present problem with non-parametric regression-function estimation. In the latter area, researchers usually choose



(a) $\mathrm{I}(h_1) = 0.6667$, $\alpha_{\mathrm{vi}} = 0.35$

(b) $\mathrm{I}(h_2) = 0.3333$, $\alpha_{\mathrm{vi}} = 0.33$

(c) $\mathrm{I}(h_3) = 1$, $\alpha_{\mathrm{vi}} = 0.1$

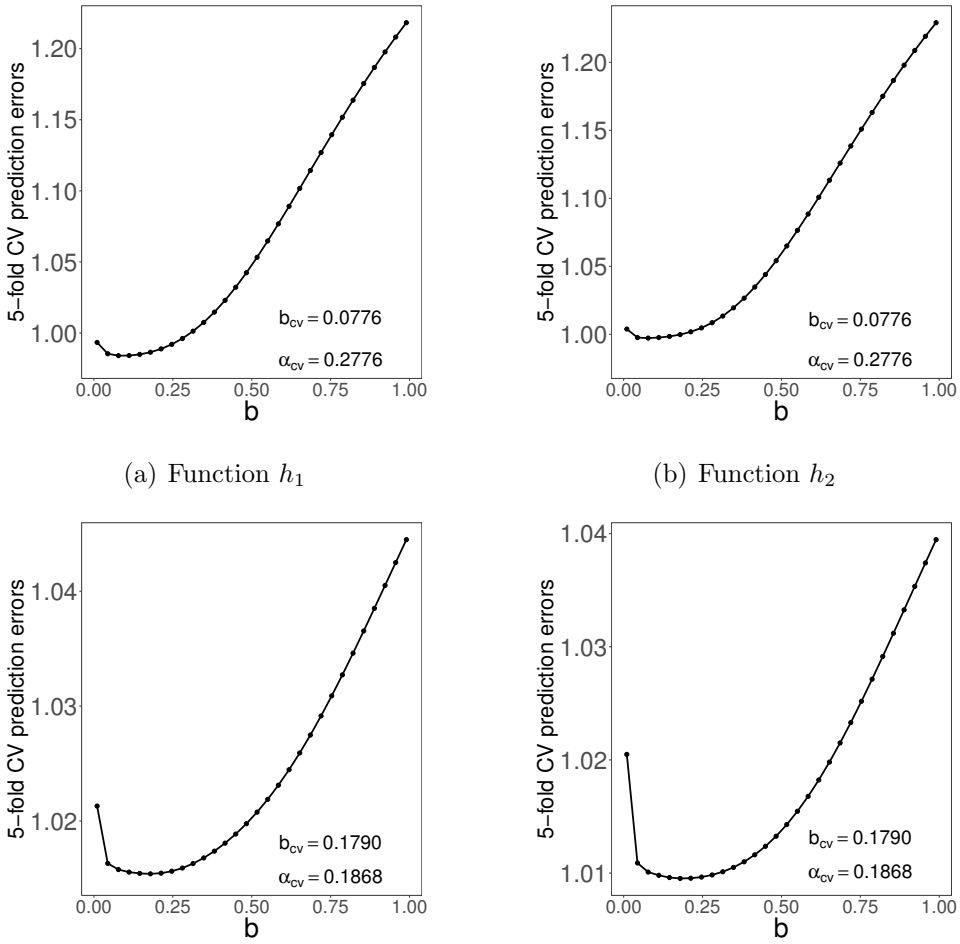
(d) $\mathrm{I}(h_4) = 0$, $\alpha_{\mathrm{vi}} = 0.1$

Figure 5.4.2: The performance of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n in the case of quartet (5.2.3) and based on visual α's.

the optimal bandwidth as the point at which cross-validation scores become minimal (e.g., Arlot and Celisse, 2010; Celisse, 2008, and references therein). We adopt this viewpoint as well. Namely, given a scatterplot, say $(t_{i,n}, Y_{i,n})$, we cross validate it (computational details and R packages will be described in a moment). Then we find the minimizing value $b = b_{\mathrm{cv}}$ and finally, according to equation (5.4.4), arrive at the 'optimal' α_{cv} via the equation

$$\alpha_{\mathrm{cv}} = \log(1/b_{\mathrm{cv}})/\log(n). \tag{5.4.5}$$

In Figure 5.4.3, we see some differences between the values of α_{vi} and α_{cv}. Nevertheless,



(a) Function h_1

(b) Function h_2

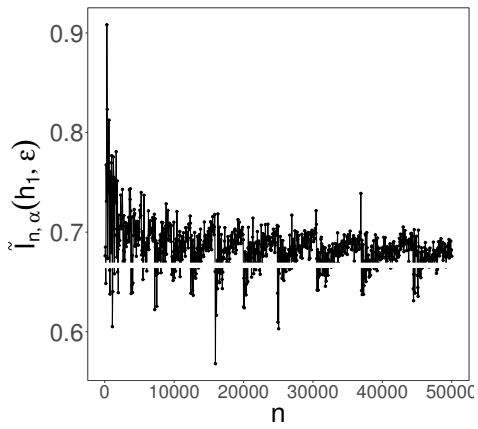
(c) Function h_3

(d) Function h_4

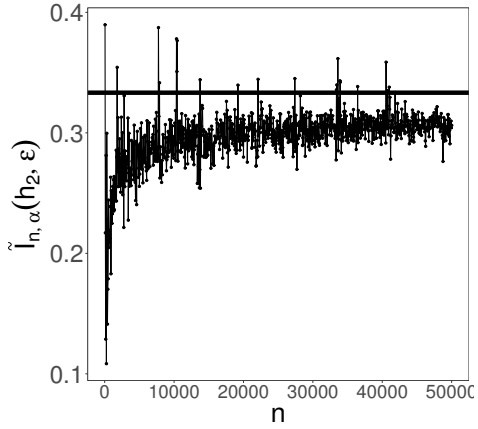
Figure 5.4.3: Cross validation, minima b_{cv}, and the grouping parameters α_{cv} for quartet (5.2.3).

we should not prejudge the situation in any way because in practice, when no hyperplanes can be produced due to unknown values of $\mathrm{I}(h)$, only the values of α_{cv} can be extracted from data. Note, however, that the four values of α_{cv} reported in the panels of Figure 5.4.3 are in compliance with the condition of Theorem 5.4.1 stipulating that α's must be in the range $(0, 1/3)$ in order to have (asymptotic) consistency.

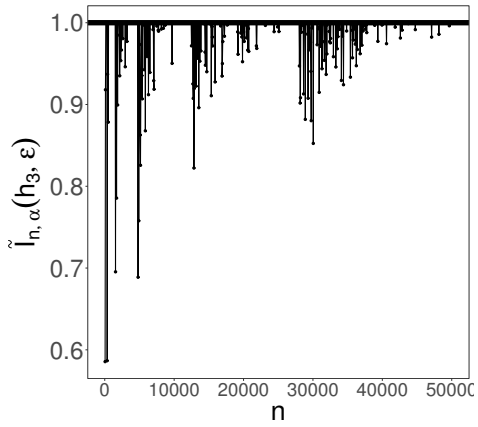
To explore how the grouped estimator $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ based on α_{cv}'s actually performs, we have produced Figure 5.4.4. Naturally, since the respective visual α_{vi}'s and cross-validatory α_{cv}'s do not coincide, the corresponding values of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ are also different. Which of them are better from the statistical point of view will become clearer only in Section 5.5,



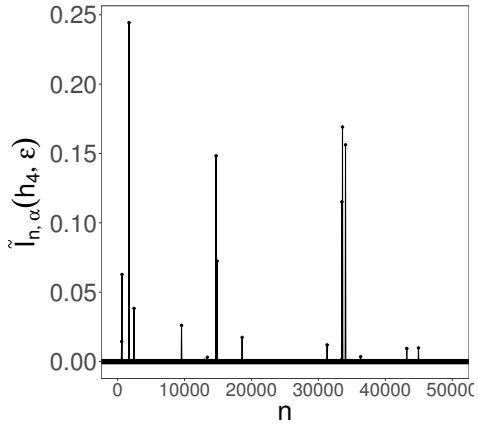
(a) $\mathrm{I}(h_1) = 0.6667$, $\alpha_{\mathrm{cv}} = 0.28$



(b) $\mathrm{I}(h_2) = 0.3333$, $\alpha_{\mathrm{cv}} = 0.28$



(c) $\mathrm{I}(h_3) = 1$, $\alpha_{\mathrm{cv}} = 0.19$



(d) $\mathrm{I}(h_4) = 0$, $\alpha_{\mathrm{cv}} = 0.19$

Figure 5.4.4: The performance of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n in the case of quartet (5.2.3) and cross validation.

where bootstrap-based standard errors and confidence intervals are derived.

We next present a detailed implementation procedure for finding cross-validatory estimates α_{cv} of the grouping parameter α. Naturally, the help of the R computing language (R Core Team, 2013) becomes indispensable, and we have used a number of R packages to accomplish the task. We also wish to acknowledge the packages `ggplot2` (Wickham, 2009) and `plotly` (Sievert et al., 2017) that we have used extensively in this chapter to draw two-dimentional plots and interactive surface plots; the latter plots have been pivotal in extracting the values α_{vi} visually.

Hence, from the purely practical computational perspective, we now utilize bandwidth

selection techniques of kernel-based regression-function estimation in order to get estimates of the grouping parameter. First, for the sake of programming efficiency, we restrict b's to the interval $(0.01, 0.99)$, and we evenly split the latter interval into bins of width $(0.99 - 0.01)/29 \approx 0.0338$, all of which can of course be refined in order to achieve, if desired, smaller computational errors. Hence, from now on, we have thirty equidistant b's, which are $b_i \approx 0.01 + (i-1)0.0338$ for $i = 1, \ldots, 30$. Next we use the common cross-validation method called repeated k-fold cross validation, and we set $k = 5$ for our purpose. The following main steps are:

1. For each function h under consideration, we generate $n = 10000$ data points based on equation (5.3.2).

2. We randomly split the given n points into k folds, denoted by $D_1, \ldots, D_k$, of roughly equal sizes.

3. For each value b_i, we use D_1 as the validation set and let other D's be training sets, which we use to fit a kernel regression model. Specifically, we use the function `ksmooth` from the R package `stats`, with the parameter `kernel` set to `normal`, which means that we use the normal kernel. Then we use the validation set D_1 to get the predicted values and calculate one prediction error, defined as the mean-square error and denoted by E_1. We repeat this step until we use up all the folds as our validation sets. Hence, we obtain k prediction errors $E_1, \ldots, E_k$. Finally, we average these k prediction errors and denote this average by $E_{b_i,1}$.

4. We repeat Step 3 for all b_i's, thus arriving at one estimated prediction error for each b_i. Hence, in total, we have $E_{b_1,1}, \ldots, E_{b_{30},1}$.

5. We repeat Steps 1–4 fifty times, for every b_i, and then take the averages of the corresponding fifty estimated prediction errors. This gives us thirty final estimates, which we denote by E_{b_i}. For example, for b_1, the final estimate E_{b_1} is the average of $E_{b_1,1}, \ldots, E_{b_1,50}$. In summary, after this step, we have $E_{b_1}, \ldots, E_{b_{30}}$ of the final estimates of the prediction error.

6. We draw the plot of the b_i's versus the corresponding estimated prediction errors. The b_i that gives the minimal prediction error is denoted by b_{cv}. Finally, we use equation (5.4.5) to get α_{cv}.

5.4.3 Proof of Theorem 5.4.1

The following lemma, whose special case is Proposition 5.2.1 formulated earlier, plays a pivotal role when proving Theorem 5.4.1.

Lemma 5.4.1. *Let h be differentiable, and let its derivative h' be γ-Hölder continuous for some $\gamma \in (0, 1]$. Furthermore, let ℓ be any positively homogeneous and Lipschitz function. Then there is a constant $c < \infty$ such that, for any set of points $s_1 := 0 < s_2 < \cdots < s_M \leq 1$,*

$$\sum_{j=2}^{M} \ell\Big(h(s_j) - h(s_{j-1})\Big) = \int_0^{s_M} \ell\big(h'\big)\, d\lambda + \theta c \sum_{j=2}^{M} |s_j - s_{j-1}|^{1+\gamma} \tag{5.4.6}$$

where θ is such that $|\theta| \leq 1$.

Proof. Since ℓ is Lipschitz and h' is γ-Hölder continuous, we have

$$\begin{aligned}
\int_0^{s_M} \ell\big(h'\big)\mathrm{d}\lambda &= \sum_{j=2}^{M} \int_{s_{j-1}}^{s_j} \ell\big(h'(s)\big) - \ell\big(h'(s_j)\big)\mathrm{d}s + \sum_{j=2}^{M} \big(s_j - s_{j-1}\big)\ell\big(h'(s_j)\big) \\
&= \theta c \sum_{j=2}^{M} \int_{s_{j-1}}^{s_j} |s - s_j|^{\gamma}\mathrm{d}s + \sum_{j=2}^{M} \big(s_j - s_{j-1}\big)\ell\big(h'(s_j)\big) \\
&= \theta c \sum_{j=2}^{M} |s_j - s_{j-1}|^{1+\gamma} + \sum_{j=2}^{M} \big(s_j - s_{j-1}\big)\ell\big(h'(s_j)\big),
\end{aligned} \tag{5.4.7}$$

where the values of $c < \infty$ and $|\theta| \leq 1$ might have changed from line to line. Next, we explore the right-most sum of equation (5.4.7), to which we add and subtract the left-hand side of equation (5.4.6). Then we use the mean-value theorem with some $\xi_j \in [s_{j-1}, s_j]$ and arrive at the equations

$$\begin{aligned}
\sum_{j=2}^{M} \big(s_j - s_{j-1}\big)\ell\big(h'(s_j)\big) &= \sum_{j=2}^{M} \ell\Big(h(s_j) - h(s_{j-1})\Big) + \sum_{j=2}^{M} (s_j - s_{j-1})\Big(\ell\big(h'(s_j)\big) - \ell\big(h'(\xi_j)\big)\Big) \\
&= \sum_{j=2}^{M} \ell\Big(h(s_j) - h(s_{j-1})\Big) + \theta c \sum_{j=2}^{M} |s_j - s_{j-1}|^{1+\gamma},
\end{aligned} \tag{5.4.8}$$

where the last equation holds because ℓ is positively homogeneous and Lipschitz, and h' is γ-Hölder continuous. Equations (5.4.7) and (5.4.8) imply equation (5.4.6) and finish the proof of Lemma 5.4.1. □

Proof of Theorem 5.4.1. We start with the equations

$$\begin{aligned}\widetilde{Y}_{j,n} &= \frac{1}{N}\sum_{i\in G_{j,n}} h(t_{i,n}) + \frac{1}{N}\sum_{i\in G_{j,n}} \varepsilon_i \\ &= \frac{1}{N}\sum_{i\in G_{j,n}} h(t_{i,n}) + \varepsilon^*_{j,n}, \end{aligned} \tag{5.4.9}$$

where

$$\varepsilon^*_{j,n} = \frac{1}{N}\sum_{i\in G_{j,n}} \varepsilon_i.$$

We next tackle the deterministic sum on the right-hand side of equation (5.4.9), and start with the equation

$$\frac{1}{N}\sum_{i\in G_{j,n}} h(t_{i,n}) = \frac{n-1}{N}\sum_{i=1}^{N} h\left(\frac{i-1}{n-1} + \frac{(j-1)N}{n-1}\right)\frac{1}{n-1}$$

because $G_{j,n} = (j-1)N + \{1,\dots,N\}$ for all $j = 1,\dots,M$. Consequently,

$$\begin{aligned}\frac{1}{N}\sum_{i\in G_{j,n}} h(t_{i,n}) &= \frac{n-1}{N}\left(\sum_{i=1}^{N} h\left(\frac{i-1}{n-1} + \frac{(j-1)N}{n-1}\right)\frac{1}{n-1} - \int_{(j-1)N/(n-1)}^{jN/(n-1)} h\mathrm{d}\lambda\right) \\ &\quad + \frac{n-1}{N}\int_{(j-1)N/(n-1)}^{jN/(n-1)} h\mathrm{d}\lambda \\ &= \frac{n-1}{N}\int_{(j-1)N/(n-1)}^{jN/(n-1)} h\mathrm{d}\lambda + O(n^{-1}), \end{aligned} \tag{5.4.10}$$

where we used the fact that h is Lipschitz. By the mean-value theorem, there is $t^*_{j,n}$ between $(j-1)N/(n-1)$ and $jN/(n-1)$ such that the right-hand side of equation (5.4.10) is equal to $h(t^*_{j,n}) + O(n^{-1})$. Consequently, with the notation

$$Y^*_{j,n} := h(t^*_{j,n}) + \varepsilon^*_{j,n},$$

we have $\widetilde{Y}_{j,n} = Y^*_{j,n} + O(n^{-1})$ and thus the increments $\widetilde{Y}_{j,n} - \widetilde{Y}_{j-1,n}$ are equal to $Y^*_{j,n} -$

$Y^*_{j-1,n} + O(n^{-1})$. This gives us the equations

$$\begin{aligned}\sum_{j=2}^{M} \ell\Big(\widetilde{Y}_{j,n} - \widetilde{Y}_{j-1,n}\Big) &= \sum_{j=2}^{M} \ell\Big(Y^*_{j,n} - Y^*_{j-1,n}\Big) + O(n^{-(1-\alpha)}) \\ &= \sum_{j=2}^{M} \ell\Big(h(t^*_{j,n}) - h(t^*_{j-1,n})\Big) + O\bigg(\sum_{j=2}^{M} |\varepsilon^*_{j,n} - \varepsilon^*_{j-1,n}|\bigg) + O(n^{-(1-\alpha)})\end{aligned} \tag{5.4.11}$$

because $|\ell(t) - \ell(s)| \le |t - s|$ for all real t and s. The random variables $\varepsilon^*_{j,n}$, $j = 1, \ldots, M$, are independent and identically distributed with the means 0 and variances σ^2/N. Hence,

$$\begin{aligned}\mathbf{E}\bigg(\sum_{j=2}^{M} |\varepsilon^*_{j,n} - \varepsilon^*_{j-1,n}|\bigg) &\le cM \max_j \sqrt{\mathbf{E}\big((\varepsilon^*_{j,n})^2\big)} \\ &\le cM \max_j \sqrt{\sigma^2/N} \\ &= O\big(n^{-(1-3\alpha)/2}\big),\end{aligned}$$

which implies

$$\sum_{j=2}^{M} |\varepsilon^*_{j,n} - \varepsilon^*_{j-1,n}| = O_{\mathbf{P}}\big(n^{-(1-3\alpha)/2}\big). \tag{5.4.12}$$

The right-hand side of equation (5.4.12) converges to 0 because $\alpha \in (0, 1/3)$. In view of equations (5.4.11) and (5.4.12), we have

$$\sum_{j=2}^{M} \ell\Big(\widetilde{Y}_{j,n} - \widetilde{Y}_{j-1,n}\Big) = \sum_{j=2}^{M} \ell\Big(h(t^*_{j,n}) - h(t^*_{j-1,n})\Big) + O_{\mathbf{P}}\big(n^{-(1-3\alpha)/2}\big) + O(n^{-(1-\alpha)}). \tag{5.4.13}$$

Furthermore, by Lemma 5.4.1 we have

$$\sum_{j=2}^{M} \ell\Big(h(t^*_{j,n}) - h(t^*_{j-1,n})\Big) = \int_0^1 \ell\big(h'\big)\mathrm{d}\lambda + O\big(n^{-\alpha\gamma}\big). \tag{5.4.14}$$

Combining equations (5.4.13) and (5.4.14), and using β to denote $\min\{\alpha\gamma, (1-3\alpha)/2\}$, we have

$$\begin{aligned}\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon) &= \frac{\int_0^1 (h')_+ \mathrm{d}\lambda + O_{\mathbf{P}}\big(n^{-\beta}\big)}{\int_0^1 |h'|\mathrm{d}\lambda + O_{\mathbf{P}}\big(n^{-\beta}\big)} \\ &\xrightarrow{\mathbf{P}} \frac{\int_0^1 (h')_+ \mathrm{d}\lambda}{\int_0^1 |h'|\mathrm{d}\lambda} = \mathrm{I}(h).\end{aligned}$$

The rate of convergence is of the order $O_{\mathbf{P}}(n^{-\beta})$. Theorem 5.4.1 is proved. □

5.5 Bootstrap-based confidence intervals

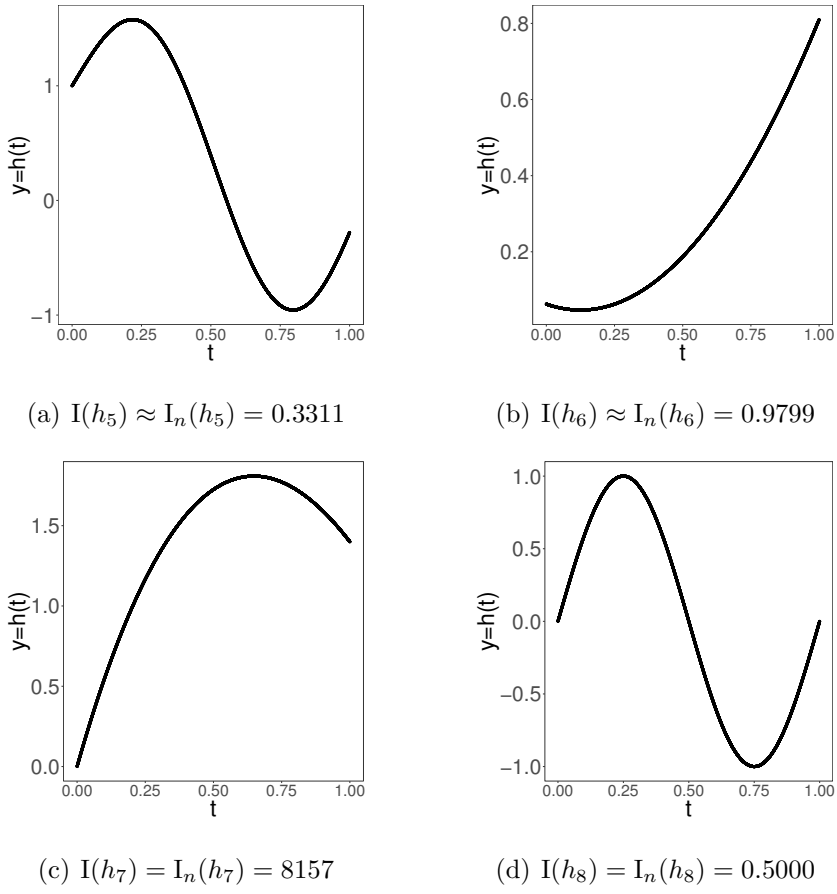
To construct confidence intervals for $\mathrm{I}(h)$ based on the estimator $\widetilde{\mathrm{I}}_{n,\alpha}(h,\varepsilon)$, we need to determine standard errors, which turns out to be a very complex task from the viewpoint of asymptotic theory. Hence, we employ bootstrap (e.g., Hall, 1992; Efron and Tibshirani, 1993; Shao and Tu, 1995; Davison and Hinkley, 1997, and reference therein). The re-sampling size m is quite often chosen to be equal to the actual sample size n, but in our case, we find it better to re-sample fewer than n observations (i.e., $m < n$) and thus follow specialized to this topic literature by Bickel et al. (1997), Bickel and Sakov (2008), Gribkova and Helmers (2007, 2011); see also references therein. Specifically, the steps that we take are:

- For a given function h, we generate $n = 10000$ values $y_1, ..., y_n$ according to the model $Y_i = h(t_{i,n}) + \varepsilon_i$, where ε_i are i.i.d. standard normal.
- We re-sample 1000 times and in this way obtain 1000 sub-samples of size m, which we choose to be $m \approx 2\sqrt{n}$ according to a rule of thumb (DasGupta, 2008, p. 478).
- We use formula (5.3.3) to calculate the grouped index of increase, thus obtaining 1000 values of it; one value for each sub-sample. We denote the empirical distribution of the obtained values by F^*.
- With Q^* denoting the (generalized) inverse of F^*, the 95% quantile-based confidence interval is $(q_{2.5\%}, q_{97.5\%})$, where $q_{2.5\%} = Q^*(0.025)$ and $q_{97.5\%} = Q^*(0.975)$.

To illustrate, we introduce a second quartet of functions of this paper, namely:

$$\begin{aligned} h_5(t) &= (t-1)^2 + \sin(6t), \quad h_6(t) = (t-0.25)^2 + \sin(0.25t), \\ h_7(t) &= t^3 - 5.6t^2 + 6t, \quad h_8(t) = \sin(2\pi t). \end{aligned} \tag{5.5.1}$$

We have visualized the functions in Figure 5.5.1. As expected, our preliminary analysis has shown that the un-groped estimators converge to 0.5 in all the four cases, but the grouped estimator $\widetilde{\mathrm{I}}_{n,\alpha}(h,\varepsilon)$ does converge under appropriate choices of the grouping (or smoothing) parameter α values. Next are summaries of our findings using two approaches: the first one is data exploratory (visual) and the second one is based on cross validation.



(a) $\mathrm{I}(h_5) \approx \mathrm{I}_n(h_5) = 0.3311$

(b) $\mathrm{I}(h_6) \approx \mathrm{I}_n(h_6) = 0.9799$

(c) $\mathrm{I}(h_7) = \mathrm{I}_n(h_7) = 8157$

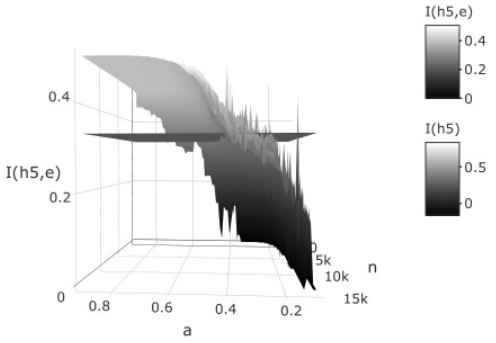
(d) $\mathrm{I}(h_8) = \mathrm{I}_n(h_8) = 0.5000$

Figure 5.5.1: Quartet (5.5.1) functions and their indices of increase

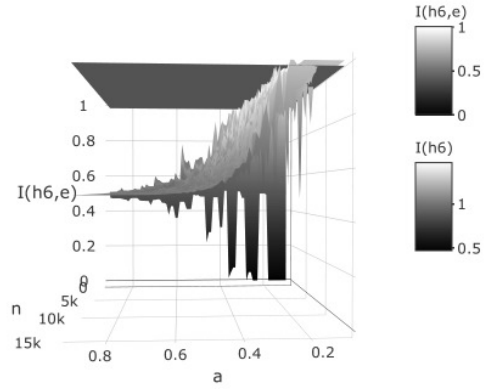
5.5.1 Data exploratory (visual) choice of α

Based on the crossings of surfaces and hyperplanes depicted in Figure 5.5.2, we choose appropriate α values, denoted by α_{vi}, for the functions of quartet (5.5.1). To check the performance of these values, we draw convergence graphs in Figure 5.5.3. Next, we use formula (5.3.3) to calculate point estimates of the actual index of increase for each of the functions in quartet (5.5.1), whose values appear in Table 5.5.1. Finally, we use bootstrap to get standard errors and confidence intervals, all of which are also reported in Table 5.5.1.

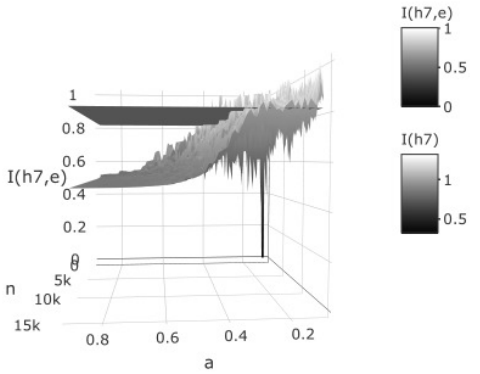
Reflecting upon the findings in Table 5.5.1, we see that the values of α_{vi} corresponding to the functions h_5 and h_8 are outside the range $(0, 1/3)$ specified by the consistency result of Theorem 5.4.1, but this of course does not invalidate anything – we are simply working



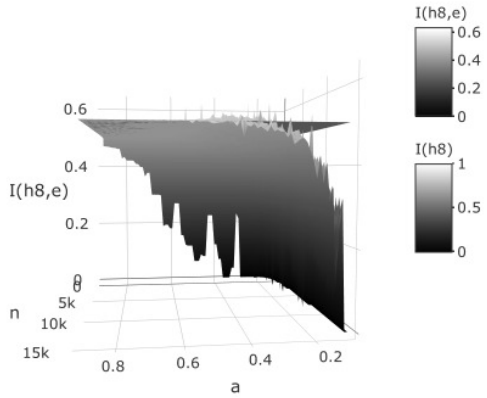
(a) The hyperplane at the height $\mathrm{I}(h_5) = 0.3311$



(b) The hyperplane at the height $\mathrm{I}(h_6) = 0.9799$



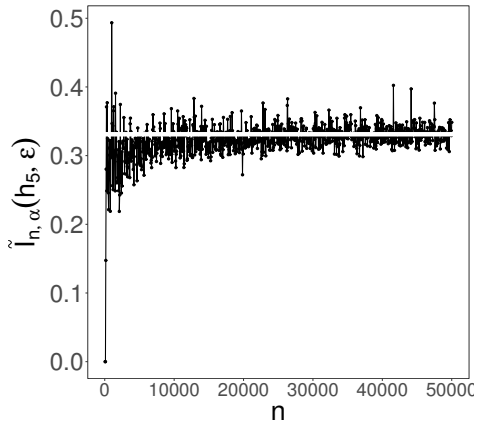
(c) The hyperplane at the height $\mathrm{I}(h_7) = 0.8157$



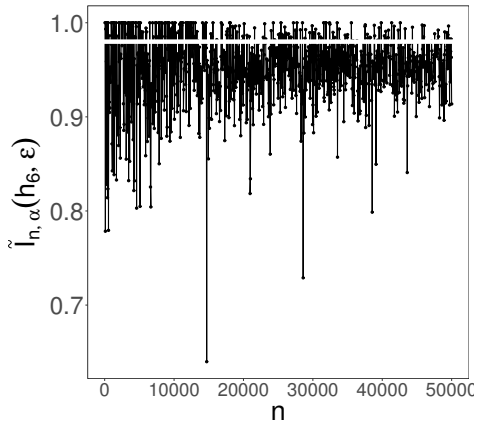
(d) The hyperplane at the height $\mathrm{I}(h_8) = 0.5000$

Figure 5.5.2: Values of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n and α in the case of quartet (5.5.1).

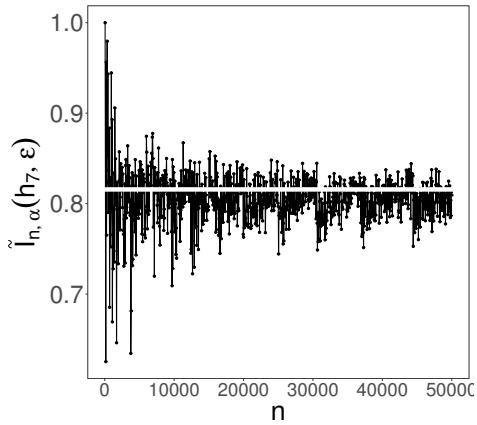
with finite sample sizes n. Naturally, we are now eager to compare all the findings reported in Table 5.5.1 with the corresponding ones obtained by cross validation, which is our next topic.



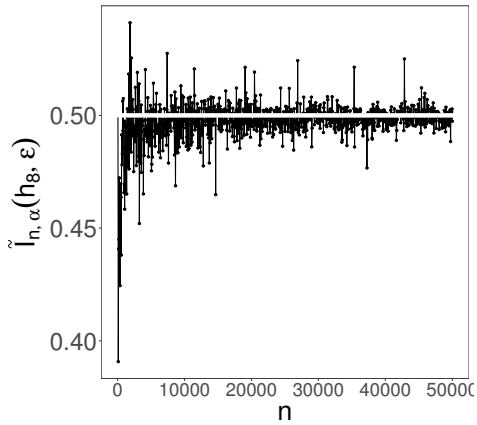
(a) $\mathrm{I}(h_5) = 0.3311$, $\alpha_{\mathrm{vi}} = 0.36$



(b) $\mathrm{I}(h_6) = 0.9799$, $\alpha_{\mathrm{vi}} = 0.25$



(c) $\mathrm{I}(h_7) = 0.8157$, $\alpha_{\mathrm{vi}} = 0.28$



(d) $\mathrm{I}(h_8) = 0.5000$, $\alpha_{\mathrm{vi}} = 0.50$

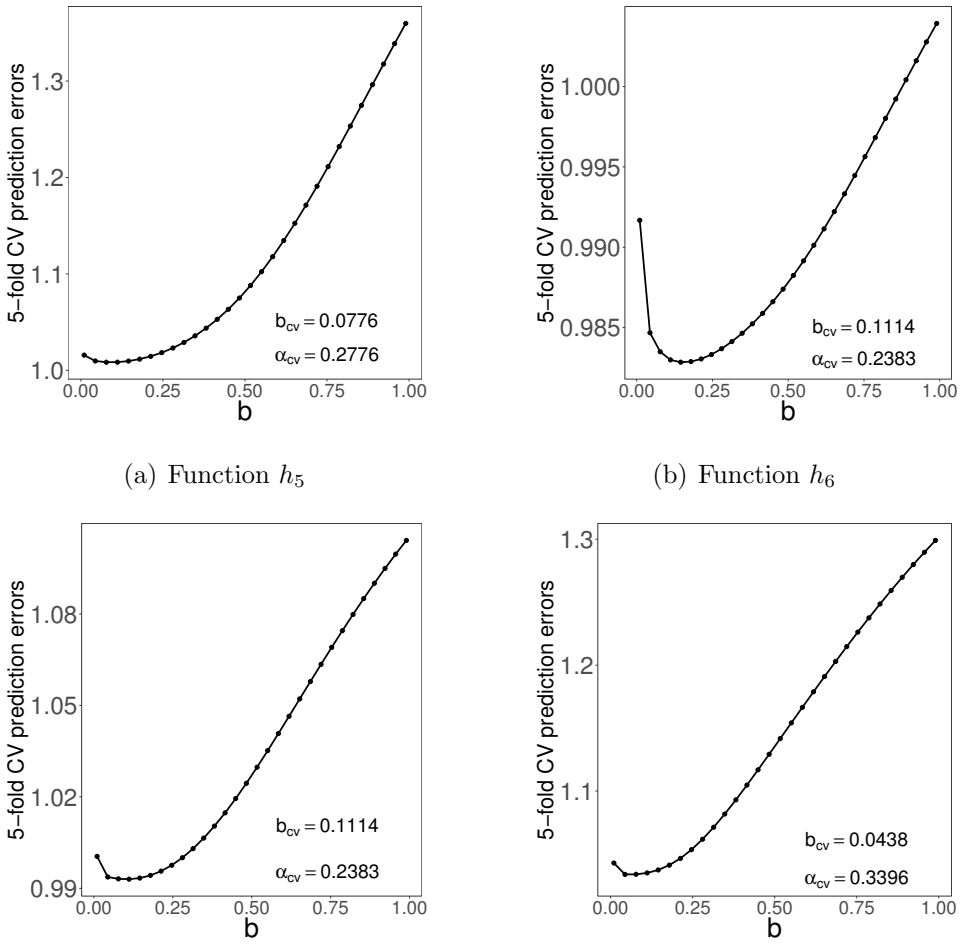
Figure 5.5.3: The performance of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n in the case of quartet (5.5.1) and based on visually assessed α's.

	h_5	h_6	h_7	h_8
True values	0.3311	0.9799	0.8157	0.5000
Point estimates	0.3274	0.9737	0.8094	0.5042
Standard deviations	0.0745	0.1103	0.1067	0.05237
Confidence intervals	(0.0372, 0.3368)	(0.6527, 1.0000)	(0.6090, 1.0000)	(0.3675, 0.5713)
Estimates α_{vi}	0.36	0.25	0.28	0.50

Table 5.5.1: Basic statistics and 95% confidence intervals for quartet (5.5.1) based on visually assessed α's.

5.5.2 Choosing α based on cross validation

We now use the cross-validation technique to get estimates α_{cv} of the grouping parameter α for all the functions of quartet (5.5.1). In Figure 5.5.4, we visualize the cross-validation



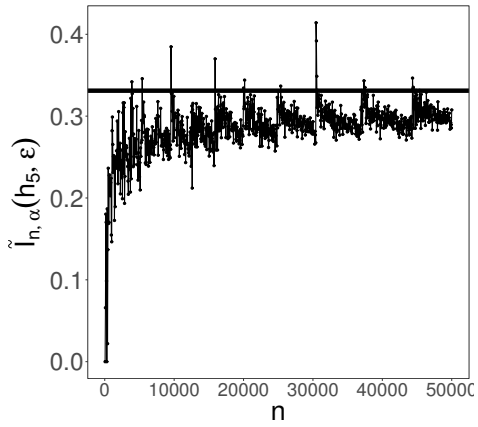
(a) Function h_5 (b) Function h_6

(c) Function h_7 (d) Function h_8

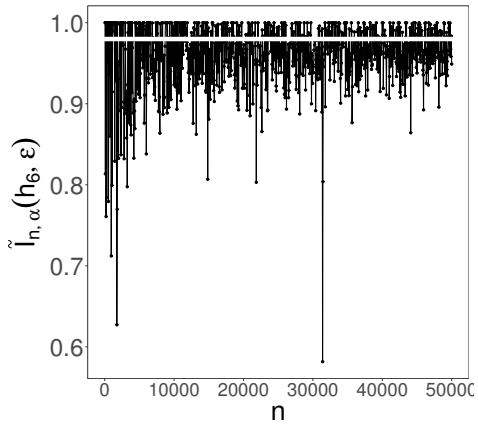
Figure 5.5.4: Cross validation, minima b_{cv}, and the grouping parameters α_{cv} for quartet (5.5.1).

scores, specify their minima b_{cv}, and also report the grouping parameters α_{cv} derived via the equation $\alpha_{\text{cv}} = \log(1/b_{\text{cv}})/\log(n)$. Based on these α_{cv} values, we explore the performance of $\widetilde{\text{I}}_{n,\alpha}(h,\varepsilon)$ using the convergence graphs depicted in Figure 5.5.5. The values of point estimates, standard errors, and confidence intervals are reported in Table 5.5.2.

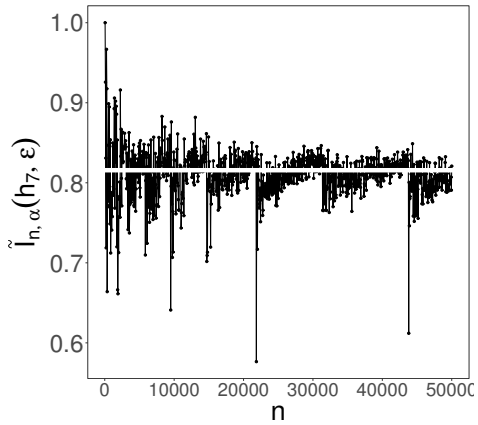
Note that the first three values of α_{cv} reported in Table 5.5.2 are inside the range $(0, 1/3)$



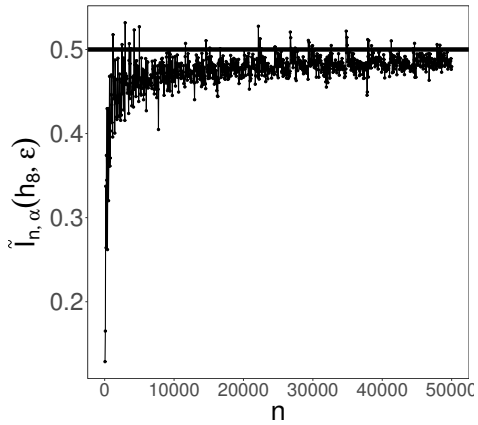
(a) $\mathrm{I}(h_5) = 0.3311$, $\alpha_{\mathrm{cv}} = 0.28$



(b) $\mathrm{I}(h_6) = 0.9799$, $\alpha_{\mathrm{cv}} = 0.24$



(c) $\mathrm{I}(h_7) = 0.8157$, $\alpha_{\mathrm{cv}} = 0.24$



(d) $\mathrm{I}(h_8) = 0.5000$, $\alpha_{\mathrm{cv}} = 0.34$

Figure 5.5.5: The performance of $\widetilde{\mathrm{I}}_{n,\alpha}(h, \varepsilon)$ with respect to n in the case of quartet (5.5.1) and cross validated α's.

	h_5	h_6	h_7	h_8
True values	0.3311	0.9799	0.8157	0.5000
Point estimates	0.2771	0.9894	0.8378	0.4703
Standard deviations	0.0797	0.1201	0.1084	0.1185
Confidence intervals	(0.0000, 0.2813)	(0.5987, 1.0000)	(0.6256, 1.0000)	(0.1803, 0.6193)
Estimates α_{cv}	0.28	0.24	0.24	0.34

Table 5.5.2: Basic statistics and 95% confidence intervals for quartet (5.5.1) based on cross validation.

specified by the consistency result of Theorem 5.4.1, whereas $\alpha_{\text{cv}} = 0.3396$ corresponding to h_8 is just slightly outside the range. Note also that the values of α_{cv} corresponding to the functions h_5 and h_8 are considerably smaller than the corresponding α_{vi}'s reported in Table 5.5.1.

The confidence intervals reported in Tables 5.5.1 and 5.5.2 comfortably cover the actual values of $\text{I}(h)$, and the widths of these confidence intervals, denoted by width_{vi} and width_{cv} respectively, are comparable for the functions h_5, h_6 and h_7. The width_{cv} of the cv-based confidence interval for the function h_8 is, however, considerably wider than the corresponding width_{vi} reported in Table 5.5.1. In summary, the relative differences $\text{width}_{\text{cv}}/\text{width}_{\text{vi}} - 1$ for the functions h_5, h_6, h_7 and h_8 are -0.0611, 0.1555, -0.0425 and 1.1541, respectively. We finish the discussion by recalling advice from Wasserman (2004): "Do not assume that, if the estimator [...] is wiggly, then cross-validation has let you down. The eye is not a good judge of risk" (Remark 20.18, page 317).

5.6 Summary and concluding notes

Davydov and Zitikis (2017) introduced an index of increase when populations are modelled with continuous functions. Chen and Zitikis (2017) explored a modification of the index when populations are discrete and presented in the form of scatterplots, and they also explored the situation when scatterplots are viewed as data sets, in which case they fitted (non-monotonic) regression functions and subsequently applied the technique by Davydov and Zitikis (2017) to assess monotonicity of the fitted functions.

In the present chapter we have extended the aforementioned technique to the case when it is not desirable, or appropriate, to view scatterplots as populations, or to use regression methods to fit curves to scatterplots. The herein proposed technique is based on grouping and averaging data, and then calculating the index of increase. Since the grouping parameter depends on both deterministic and random features of the underlying problem, we have suggested a way for grouping data so that the resulting estimator of the index of increase would be consistent. Based on this estimator, we have then suggested a construction of bootstrap-based confidence intervals for the index of increase.

The derived theoretical results have been made accessible to practitioners by detailed descriptions and analyses of various computational aspects inherent in our proposed solution of the problem.

Chapter 6

Summary and further research topics

In this dissertation, we have developed a novel technique, the index of increase, that measures non-linear, asymmetric, and non-monotonic relationships between two variables. The index of increase has been applied in educational datasets. Simulation studies and theoretical developments have also been provided to reveal the meaning of the index of increase. We summarize our results as follows.

In Chapter 3, we firstly introduced the definition of the index of increase in both discrete and continuous forms, which can be applied directly to data points and differentiable functions. Properties such as translation and scale invariance were discussed. We also provided a discretization method for any fitted curve in order to calculate the index of increase. This discretization method has been justified and illustrated mathematically. Numerical simulation results also indicate that when we discretize the target interval in a sufficiently fine way, we can achieve any desired precision. Furthermore, two practical modifications of the index of increase (discrete form), including unifying data range and ruling out ties, have been described in detail. These preparations allowed us to implement the index of increase in a dataset in classical text (Thorndike and Thorndike Christ, 2010). Numerical comparisons of students' performance between boys and girls have been provided as comprehensive explanations. This application shows that the index of increase fits the purpose for comparing student performance, since relationships between subjects are asymmetric, non-linear, and non-monotonic in most cases, which makes the index

of increase stand out from other traditional statistical tools. Moreover, convenience and interpretability of the index of increase may also attract more researchers and educators to use it.

In Chapter 4, we explored a possible and meaningful extension of the index of increase, building upon our work in Chapter 3. We developed two indices that measure the interchangeability between two variables: the relative index of interchangeability and the absolute index of interchangeability. These two indices have been developed on top of the index of increase, and so we briefly and systematically revisited the index of increase. We further emphasized the meaning of studying the index of increase in both forms (i.e., discrete and integral forms) and its practical modifications. Then, we applied the index of increase and the two indices of interchangeability to an education dataset (Mardia et al., 1979). Comprehensive discussions have been provided. We concluded that the indices of interchangeability are suitable ways to figure out the "timing independence" property of subjects as described in Section 4.6, which provides suggestions to both schools and students to construct their "education portfolios" (i.e., curricula).

In Chapter 5, we elucidated how the index of increase defined in former chapters would perform in a large sample context where data contain random component such as measurement error or other noise. We have numerically shown that if we do not modify the index of increase in this situation, the index of increase will always approach to 1/2. Next, we proposed a resolution which is an estimator of the index of increase, and it allows us to reduce the effect of random component in data based on averaging a certain number of data points. Furthermore, we proved (weak) consistency of the estimator of index of increase as well as its convergence rate. We also provided a practical and data-driven algorithm based on cross validation to choose a proper smoothing parameter for the estimator of the index of increase. Last but not least, we provided bootstrap-based confidence intervals for the estimates.

In summary, current research has expanded the area of distance-based measures that quantify non-monotonic relationships between variables in both theoretical and practical aspects. Naturally, it opens the gate to other research topics, and we would like to point

out some potential future studies:

- In Chapter 5, we provided bootstrap-based confidence intervals for the estimates of index of increase. Naturally, we would like to know the explicit asymptotic distribution of the estimator. After developing the real distribution, we can further apply standard statistical inference procedures, such as hypothesis testing, to provide more rigorous results on the performance of the index of increase in practice.

- As an extension of Chapter 3 and Chapter 4, from the practical perspective, in what other areas can we use the index of increase except for education? For cases mentioned in Chapter 1, can the index of increase be a meaningful measure? For instance, when constructing portfolio, will our index perform better than the CAPM or other advanced portfolio theory such as the Modern portfolio theory (MPT) considering to achieve high profit and low risk? Also, for the price elasticity of demand example, can we use the index of increase instead of price elasticity to determine the optimal price through maximizing the total revenue?

- We only considered 2-dimensional relationships in this thesis. That is, the index of increase can only quantify relationship between two variables. However, we can further adapt the distance-based idea for higher dimensions. Davydov et al. (2018) have developed an index of convex which is a 3-dimensional index and can quantify the convexity, or lack of it, so that we can use it to search for non-convex regions of functions. Yet, the index of convexity relies on the Hessian matrix and its eigenvalues, so the functions under consideration must have second partial derivatives. Another problem for the index of convexity is that for simple functions, it is already complicated to obtain the numerical or theoretical results. Therefore, how are we going to develop a simplified version of estimation relies on further development of the theory behind the index of convexity?. Also, how can we apply the index of convexity to data sets? Solutions to these problems still remain unclear.

- In order to popularize the index of increase and its extension, the index of convexity, it is beneficial to show its convenience when using. More specifically, providing

built-in functions or packages written in programming languages such as R or Python will be meaningful. In this way, people from different backgrounds, either researchers or practitioners, would easily implement these novel techniques.

Appendix A

Supplementary material for Chapter 3

A.1 Computer codes

The following three computer codes calculate the index of increase under various scenarios. The codes are written using the very accessible and free R software environment for statistical computing and graphics R Core Team (2013).

A.1.1 Function-based index of increase

Suppose that we wish to calculate the index of increase of a function h on the interval $[L, U]$ for some $L < U$. For this, we employ the computational algorithm described in Section 3.4. The desired precision is achieved by setting a large value of the discretization parameter n. Specifically, in the R Console, we run code:

```
indexh <- function(h,n=10000,L,U){
  temp1 <- seq(L,U,length=n)
  y <- h(temp1)
  x <- data.frame(temp1,y)
  denominator <- sum(abs(diff(x[,2])))
  temp <- c()
```

```
  for(i in 1:(length(x[,2])-1)){
    temp[i] <- ifelse(diff(x[,2])[i] > 0, diff(x[,2])[i],0)
  }
  numerator <- sum(temp)
  index <- numerator/denominator
  return(index)
}
```

As an illustration, we next run the following code in order to calculate the index of increase for the function $h(x) = \sin(x)$ on the interval $[-\pi/2, \pi]$:

```
h <- function(x){sin(x)}
indexh(h,n=10000,L=-pi/2,U=pi)
```

The result is 0.6666667, which suggests (because $0.6666667 > 0.5$) that the trend is more increasing than decreasing.

A.1.2 Index of increase for discrete Data when the are no ties

In this appendix, we provide the code pertaining to the 'basic idea' described in Section 3.3.1. In this case, all of the first coordinates (i.e., xs) are different, and all of the second coordinates (i.e., ys) are also different. (When there are ties among the x's or y's, the next appendix provides a more general and complex code, which of course also works when there are no ties.) To begin with, in the R Console, we run the following code:

```
index <- function(data){
  ##order the dataset according to x's
  data <- data[order(data[,1]),]
  ##calculate the denominator of index I
  denominator <- sum(abs(diff(data[,2])))
  ##calculate the numerator of index I
  temp <- c()
  for(i in 1:(length(data[,2])-1)){
    temp[i] <- ifelse(diff(data[,2])[i] > 0, diff(data[,2])[i],0)
```

```
  }
  numerator <- sum(temp)
  index <- numerator/denominator
  return(index)
}
```

As an illustration, we calculate the index of increase for the three pairs $(3,1)$, $(1,3)$, and $(2,0)$ by running the following code:

```
x <- c(3,1,2)
y <- c(1,3,0)
data <- data.frame(x,y)
index(data)
```

The result is 0.25, which suggests (because $0.25 < 0.5$) that the trend is more decreasing than increasing.

A.1.3 Index of increase for arbitrary discrete data

The following code is an augmentation of the previous one in order to allow for ties among the x's as well as for ties among the y's. The code follows the median-adjusted methodology described in Section 3.3.2. We start by running the following code in the R Console:

```
indexmed <- function(data){
  #order the dataset according to x's
  data <- data[order(data[,1]),]
  #find all the distinct value of x
  tempx <- unique(data[,1])
  #for each distinct x, find the median of y's
  medy <- c()
  for(i in 1:length(tempx)){
    medy <- c(medy,median(data[data[,1]==tempx[i],2]))
  }
  newdata <- data.frame(tempx,medy)
```

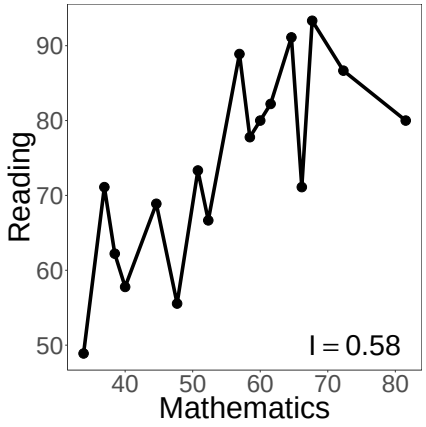
```
  #calculate the index value
  denominator <- sum(abs(diff(newdata[,2])))
  temp <- c()
  for(i in 1:(length(newdata[,2])-1)){
    temp[i] <- ifelse(diff(newdata[,2])[i] > 0,diff(newdata[,2])[i],0)
  }
  numerator <- sum(temp)
  index <- numerator/denominator
  return(index)
}
```

As an illustration, we calculate the index of increase for the data set that consists of the five pairs $(1, 1)$, $(2, 3)$, $(2, 2)$, $(3, 1)$, and $(3, 2)$. For this, we run the following code:

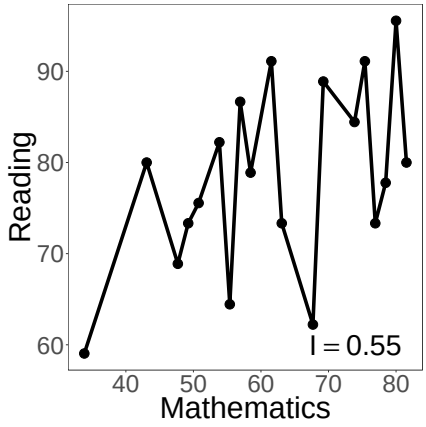
```
x <- c(1,2,2,3,3)
y <- c(1,3,2,1,2)
data <- data.frame(x,y)
indexmed(data)
```

The result is 0.6, which suggests (because $0.6 > 0.5$) that the trend is more increasing than decreasing.

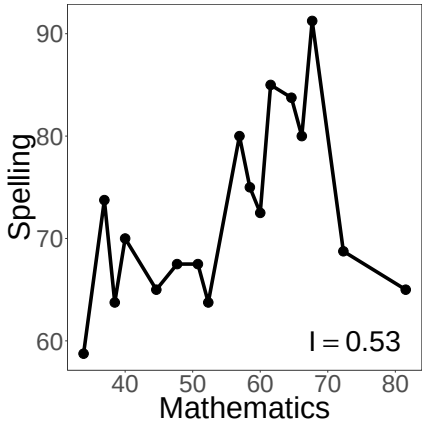
A.2 Supplementary graphs



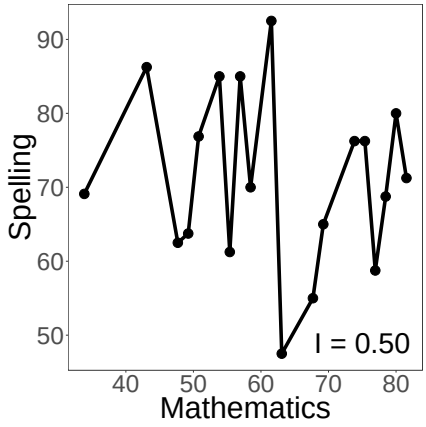
(a) M vs. R for boys.



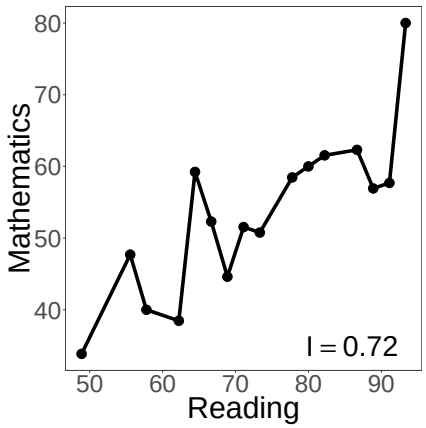
(b) M vs. R for girls.



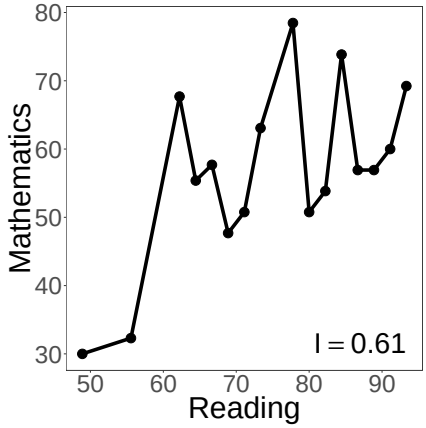
(c) M vs. S for boys.



(d) M vs. S for girls.

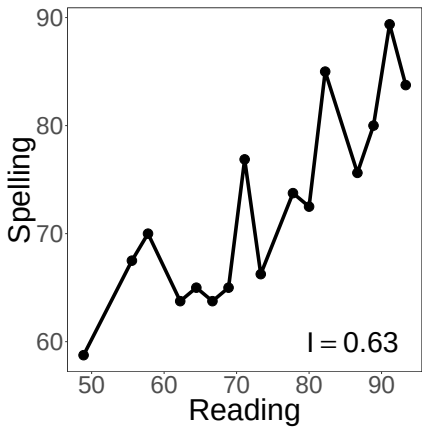


(e) R vs. M for boys.

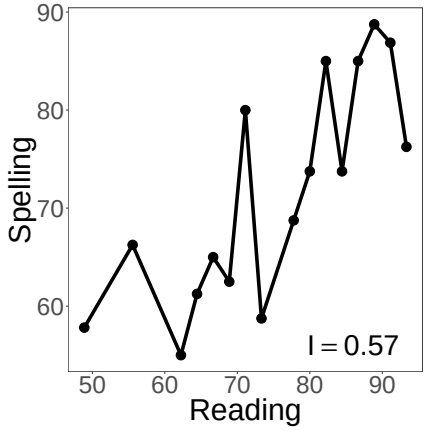


(f) R vs. M for girls.

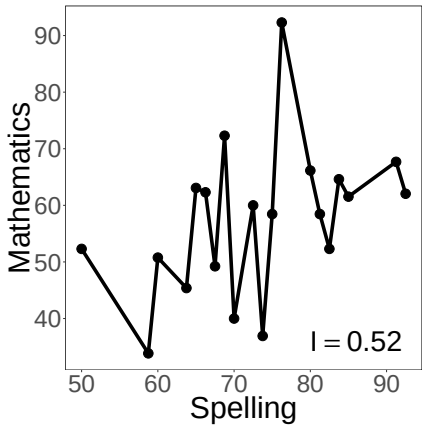
Figure A.2.1: Piece-wise linear fits and their indices of increase for both classes combined.



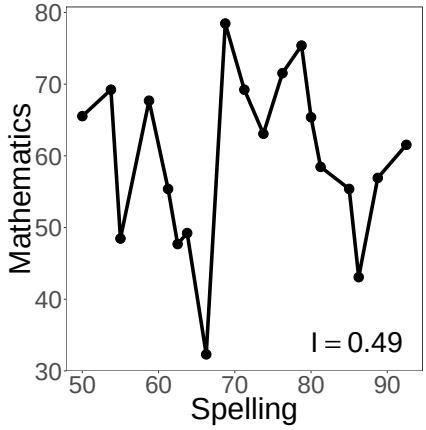
(g) R vs. S for boys.



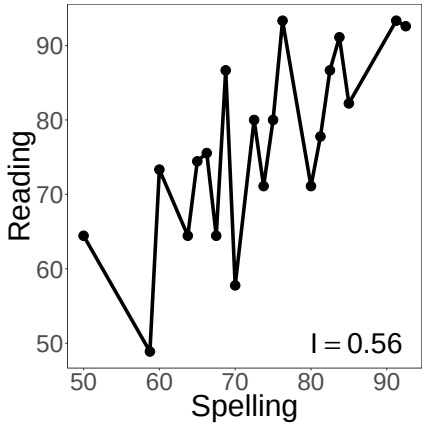
(h) R vs. S for girls.



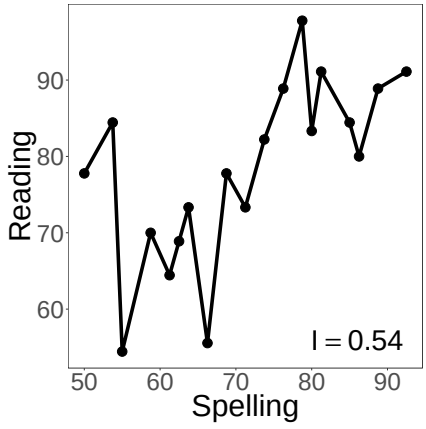
(i) S vs. M for boys.



(j) S vs. M for girls.

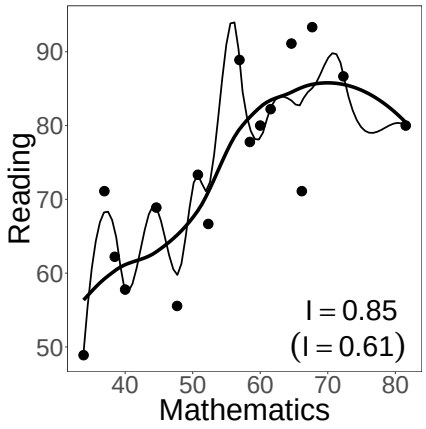


(k) S vs. R for boys.

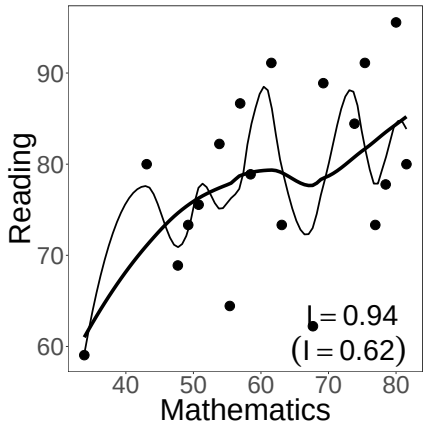


(l) S vs. R for girls.

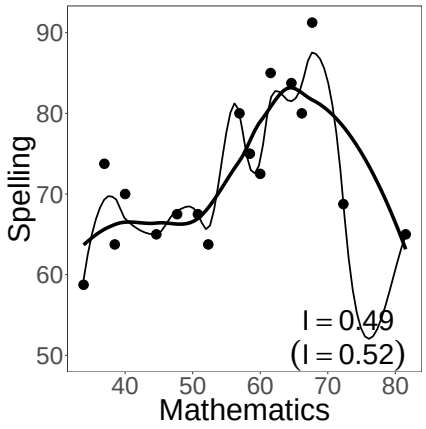
Figure A.2.2: Continuation of piece-wise linear fits and their indices of increase for both classes combined.



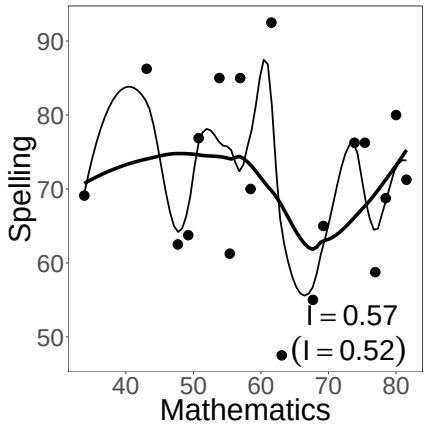
(a) M vs. R for boys.



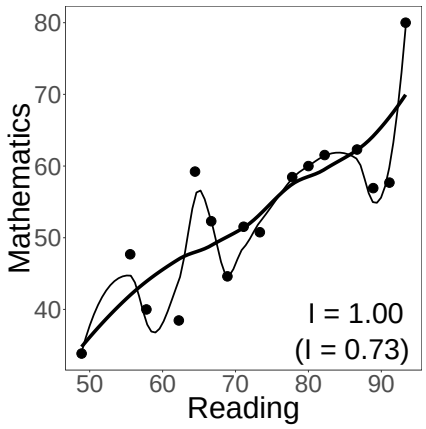
(b) M vs. R for girls.



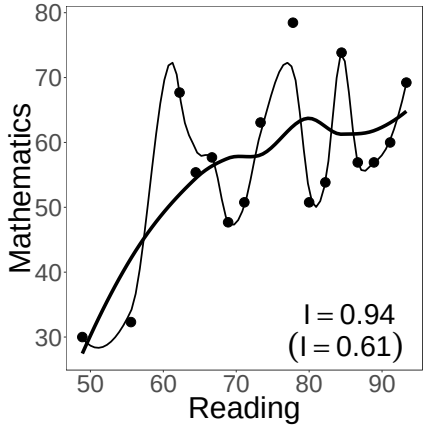
(c) M vs. S for boys.



(d) M vs. S for girls.

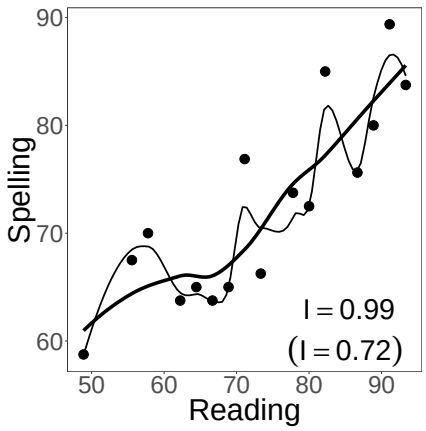


(e) R vs. M for boys.

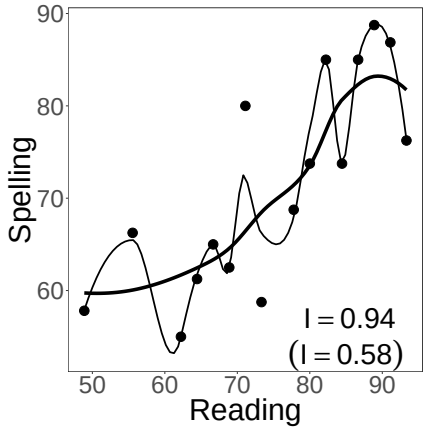


(f) R vs. M for girls.

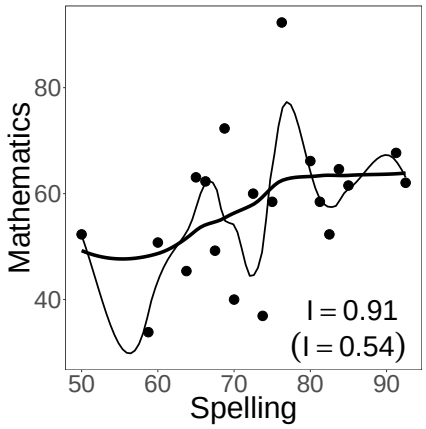
Figure A.2.3: LOESS fits when `span` = 0.75 (thicker line) and 0.35 (thinner line; the index I in parentheses) for both classes combined.



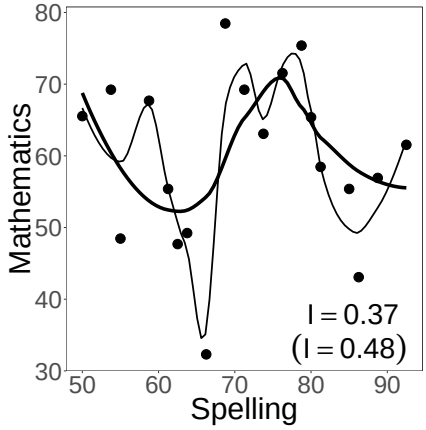
(g) R vs. S for boys.



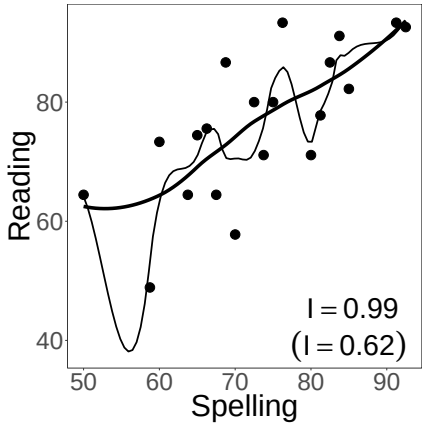
(h) R vs. S for girls.



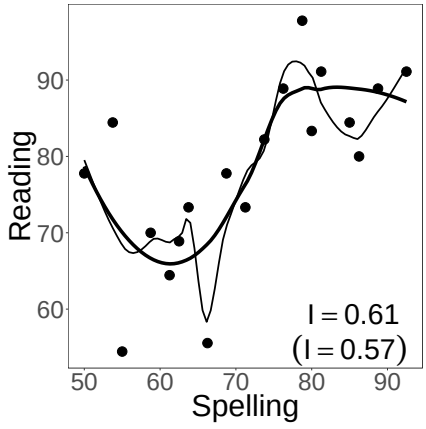
(i) S vs. M for boys.



(j) S vs. M for girls.



(k) S vs. R for boys.



(l) S vs. R girls.

Figure A.2.4: Continuation of LOESS fits when `span` = 0.75 (thicker line) and 0.35 (thinner line; the index I in parentheses) for both classes combined.

Appendix B

Supplementary materials for Chapter 4

B.1 Supplementary figures

In Figures B.1.1–B.1.9 below, panels (a) and (b) contain piecewise linear fits, and panels (c) and (d) contain LOESS fits when the `span` is 0.75 (thicker) and 0.35 (thinner). Panels (c) and (d) contain two index values: the top one is $\mathrm{I} = \mathrm{I}(h_{0.75})$ and, in parentheses, the bottom value is $\mathrm{I} = \mathrm{I}(h_{0.35})$.

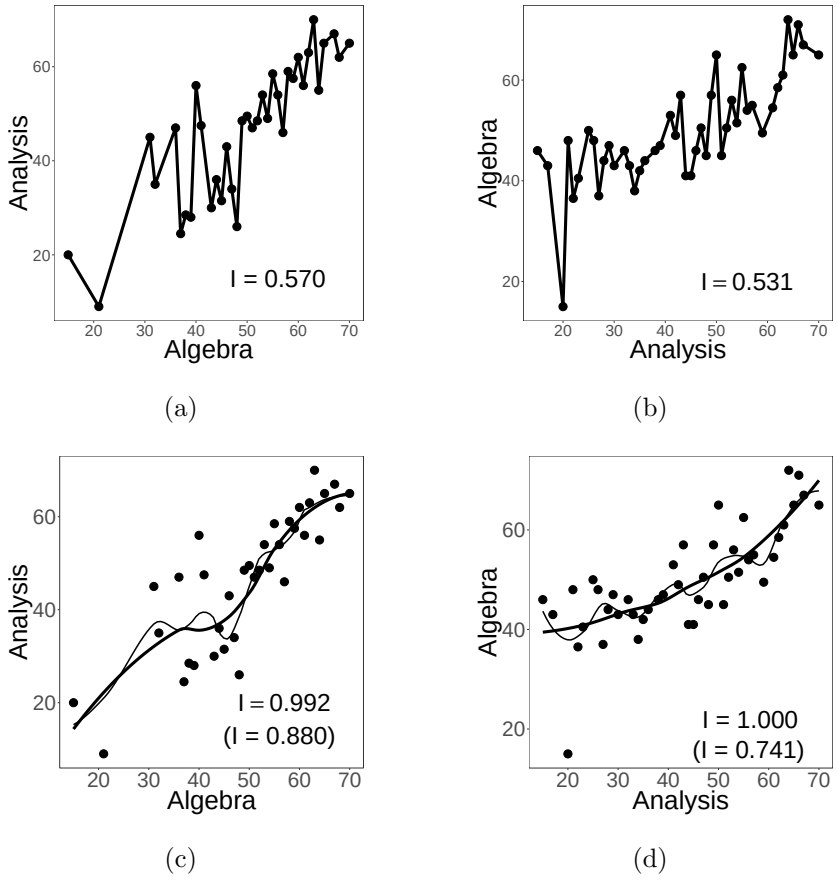


Figure B.1.1: Piecewise linear and LOESS fits for Analysis and Algebra.

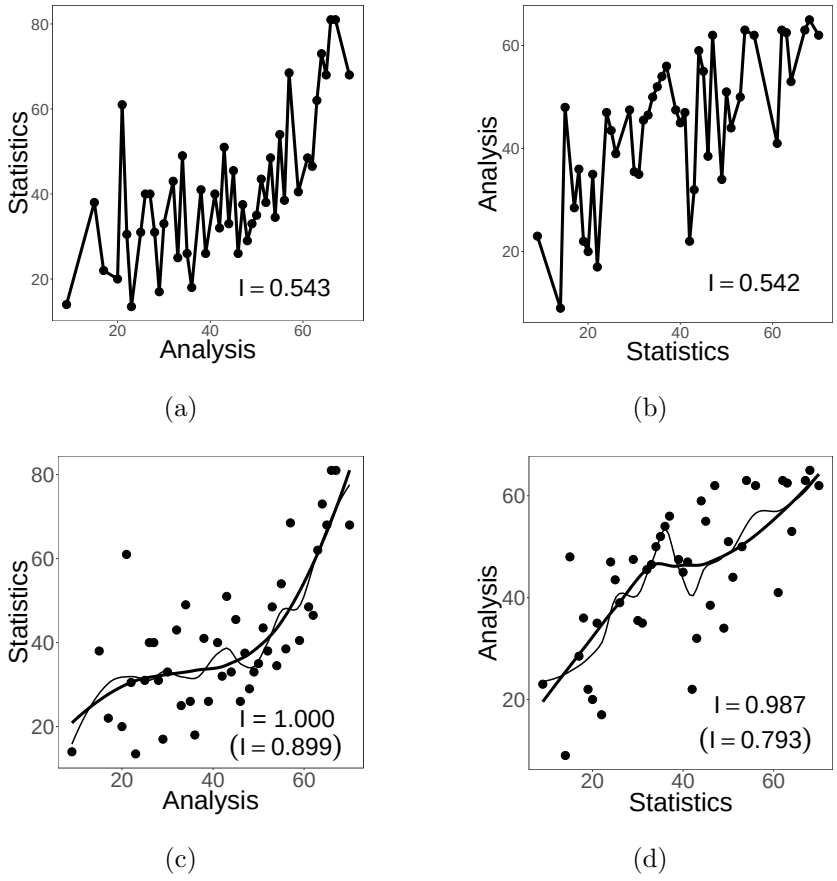


Figure B.1.2: Piecewise linear and LOESS fits for Analysis and Statistics.

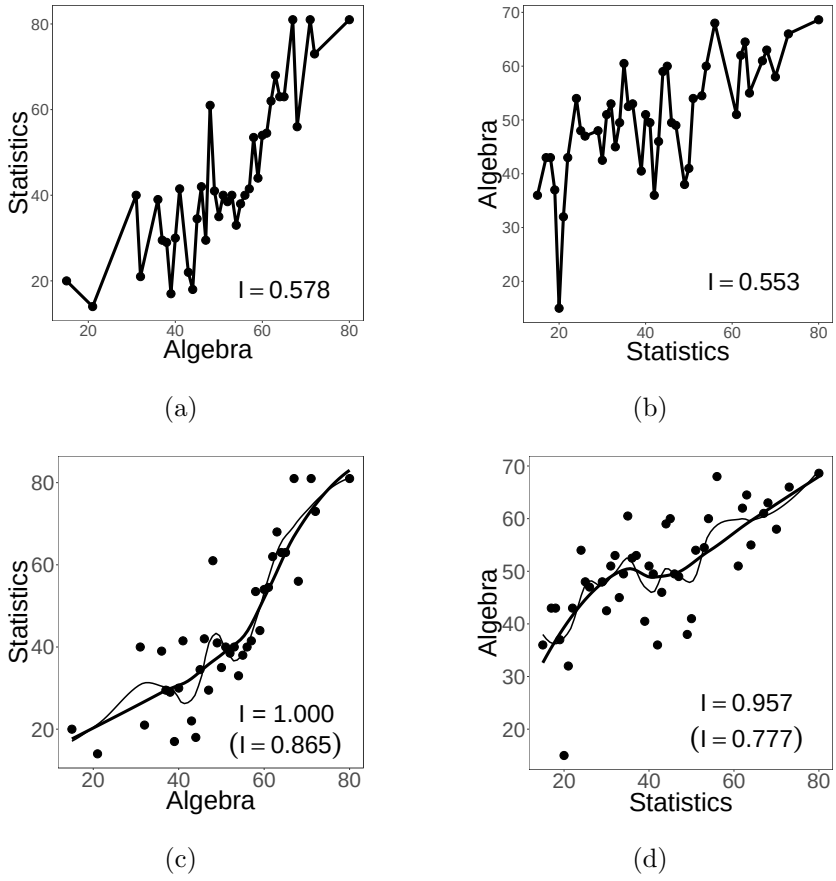


Figure B.1.3: Piecewise linear and LOESS fits for Algebra and Statistics.

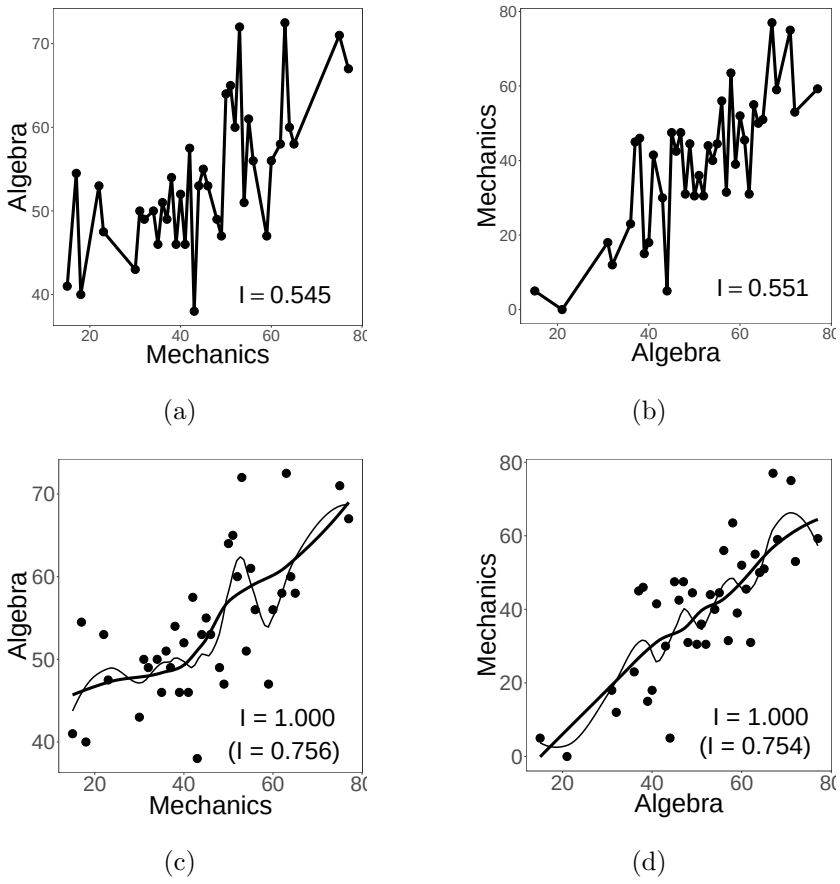


Figure B.1.4: Piecewise linear and LOESS fits for Algebra and Mechanics.

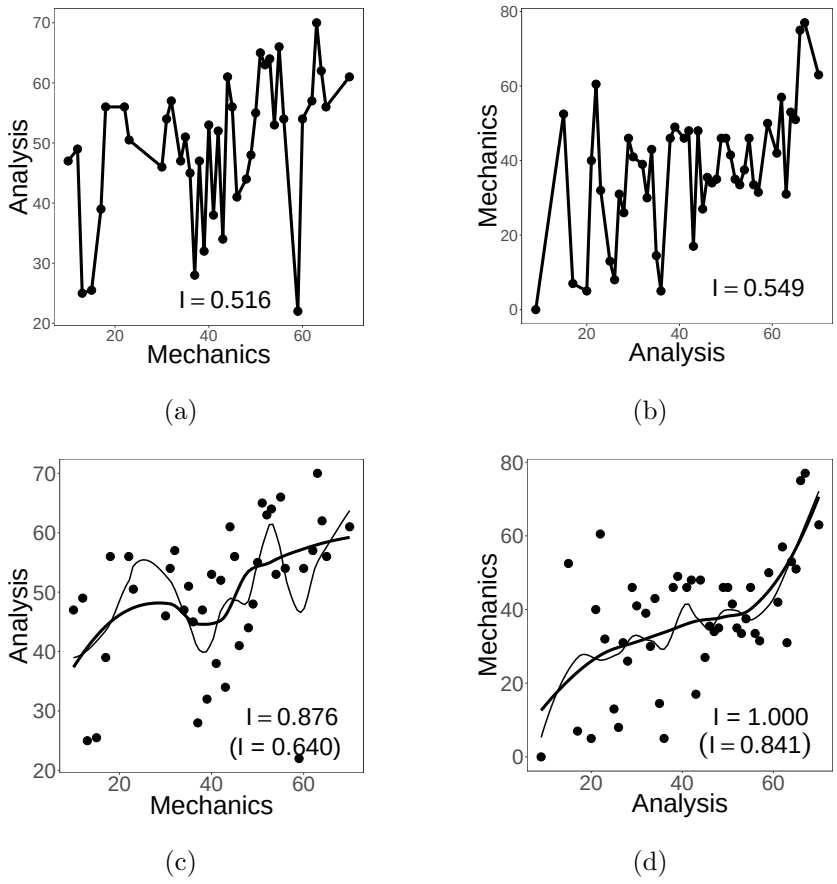


Figure B.1.5: Piecewise linear and LOESS fits for Analysis and Mechanics.

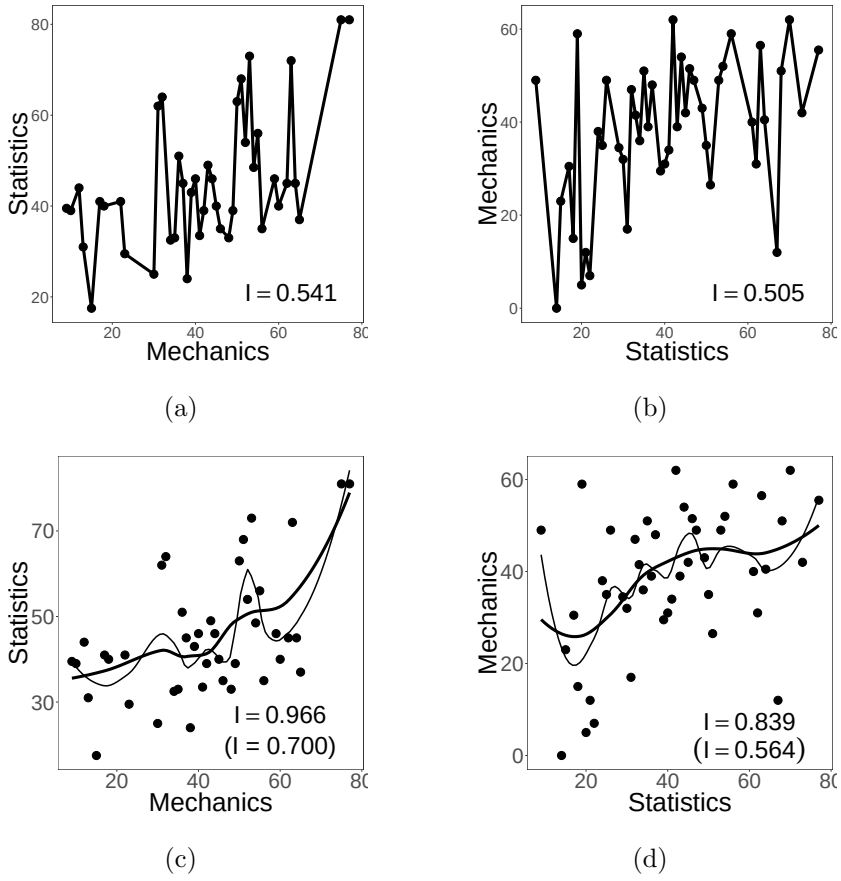


Figure B.1.6: Piecewise linear and LOESS fits for Mechanics and Statistics.

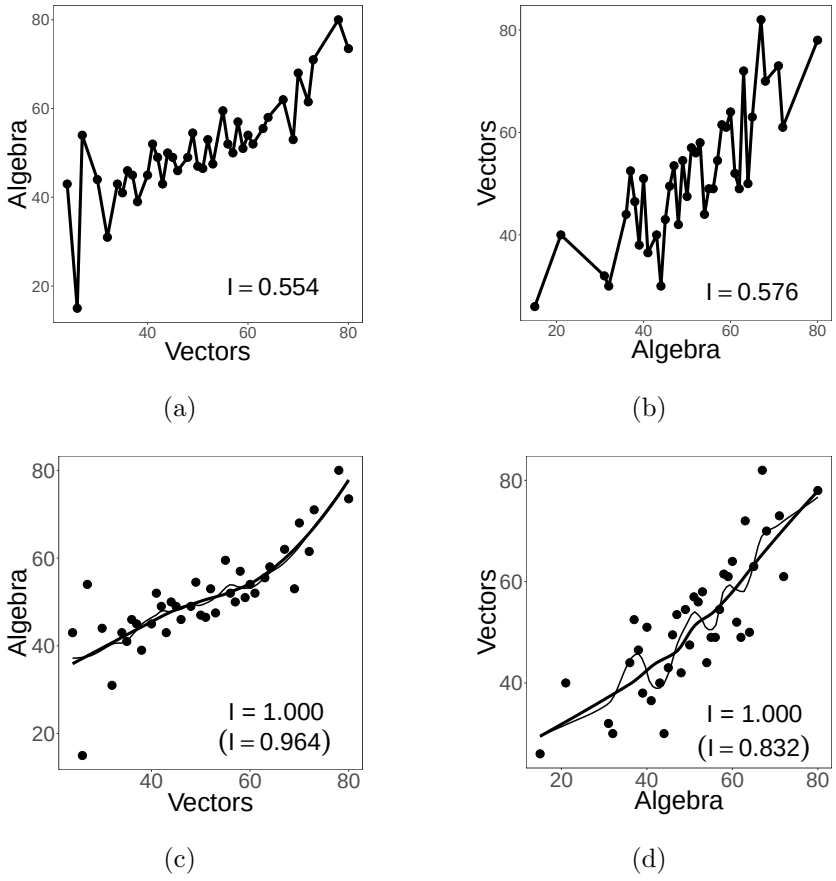


Figure B.1.7: Piecewise linear and LOESS fits for Algebra and Vectors.

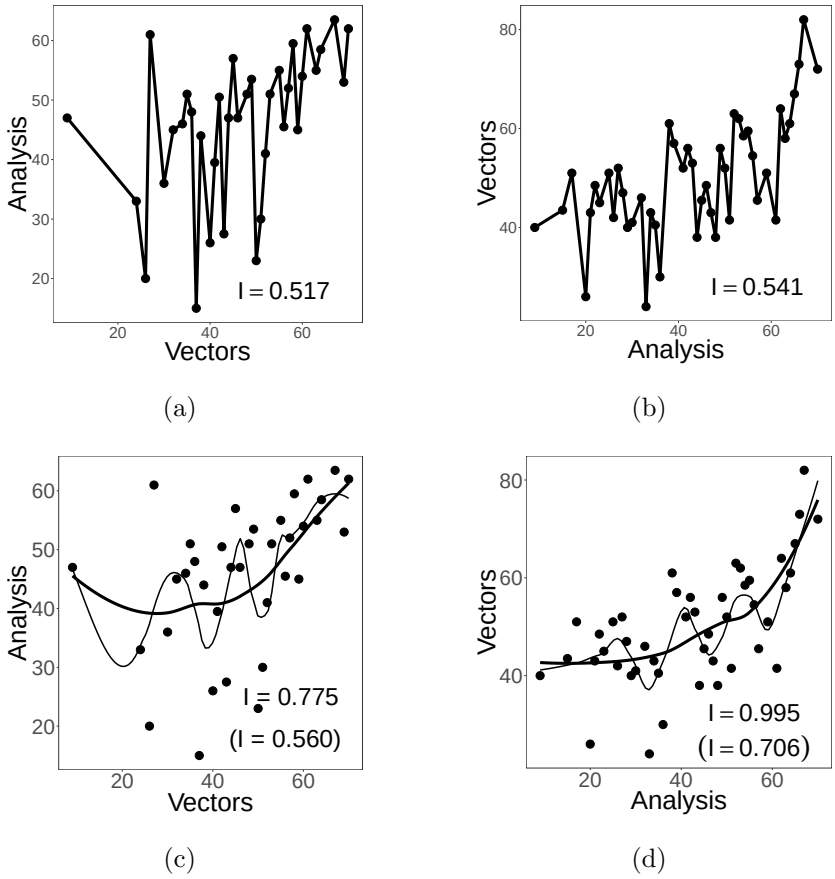


Figure B.1.8: Piecewise linear and LOESS fits for Analysis and Vectors.

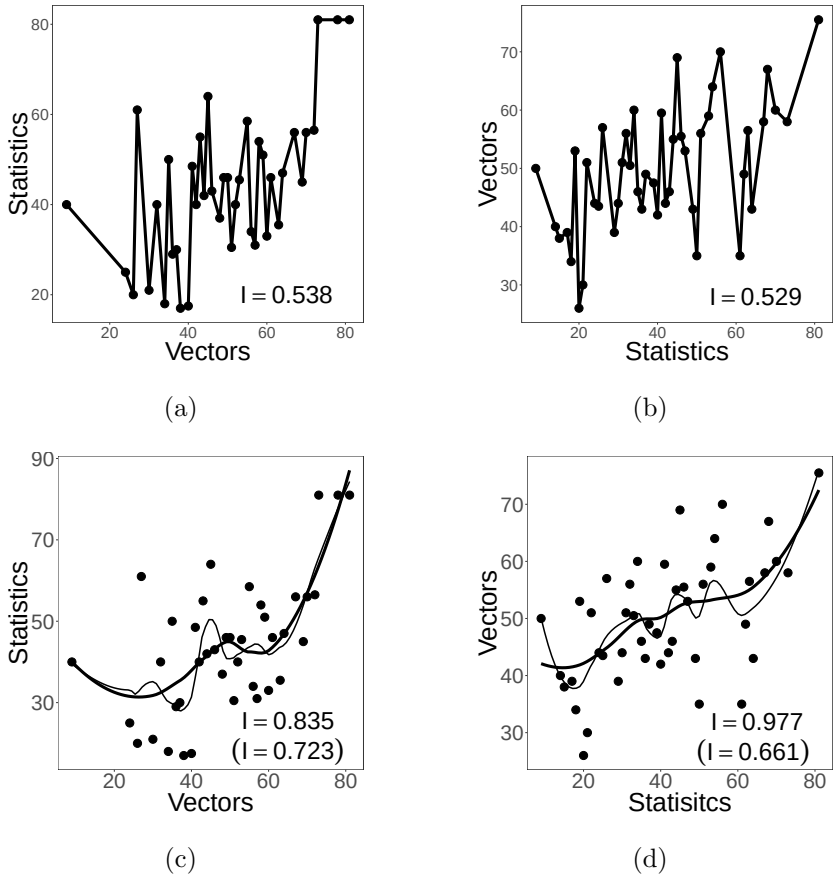


Figure B.1.9: Piecewise linear and LOESS fits for Statistics and Vectors.

Bibliography

Abramo, G., Cicero, T., and D'Angelo, C.A. (2012). Revisiting size effects in higher education research productivity. *Higher Education, 63*, 701-717.

Agarwal, P.K., Karpicke, J.D., Kang, S.H.K., Roediger,H.L., and McDermott, K.B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861-876.

Alexander, N.A., Jang, S.T., and Kankane, S. (2017). The performance cycle: The association between student achievement and state policies tying together teacher performance, student achievement, and accountability. *American Journal of Education, 123*(3), 413-446. https://doi.org/10.1086/691229

Anaya-Izquierdo, K., Critchley, F., and Vines, K. (2011). Orthogonal simple component analysis: a new, exploratory approach. *Annals of Applied Statistics, 5*, 486-522.

Anscombe, F.J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17-21. https://doi.org/10.1080/00031305.1973.10478966

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40-79.

Avendano, M., Jürges, H., and Mackenbach, J.P. (2009). Educational level and changes in health across Europe: longitudinal results from SHARE. *Journal of European Social Policy, 19*, 301-316.

Bebbington, M., Lai, C.D. and Zitikis, R. (2007). Modeling human mortality using mixtures of bathtub shaped failure distributions. *Journal of Theoretical Biology, 245*, 528-538.

Bebbington, M., Lai, C.D. and Zitikis, R. (2011). Modelling deceleration in senescent mortality. *Mathematical Population Studies, 18*, 18-37.

Bickel, P.J., Götze, F. and van Zwet, W.R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica, 7*, 1-31.

Bickel, P.J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica, 18*, 967-985.

Brazauskas, V., Jones, B.L., and Zitikis, R. (2007). Robustification and performance evaluation of empirical risk measures and other vector-valued estimators. *Metron - International Journal of Statistics, 65*(2), 175-199.

Brazauskas, V., Jones, B.L., and Zitikis, R. (2009). Robust fitting of claim severity distributions and the method of trimmed moments. *Journal of Statistical Planning and Inference, 139*(6), 2028-2043. https://doi.org/10.1016/j.jspi.2008.09.012

Bresgi, L., Alexander, D., and Seabi, J. (2017). The predictive relationships between working memory skills within the spatial and verbal domains and mathematical performance of grade 2 South African learners. *International Journal of Educational Research, 81*, 1-10. https://doi.org/10.1016/j.ijer.2016.10.004

Brezger, A. and Steiner, W. (2008). Monotonic Regression Based on Bayesian P-Splines: An Application to Estimating Price Response Functions From Store-Level Scanner Data. *Journal of Business & Economic Statistics, 26*(1), 90-104. https://doi.org/10.1198/073500107000000223

Bühlmann, H. (1980). An economic premium principle. *ASTIN Bulletin, 11*, 52-60.

Bühlmann, H. (1984). The general economic premium principle. *ASTIN Bulletin, 14*, 13-21.

Castellano, K.E. and Ho, A.D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics, 38*(2), 190-215. https://doi.org/10.3102/1076998611435413

Celisse, A. (2008). *Model Selection via Cross-Validation in Density Estimation, Regression, and Change-Points Detection.* Université Paris Sud – Paris XI, Paris. HAL Id: tel-00346320. `https://tel.archives-ouvertes.fr/tel-00346320`

Chen, L. (2020). *Price Elasticities and Promotion Cannibalization Effect on Promotion Activities.* Technical Report (in progress). Unilever Canada, Toronto, Canada.

Chen, L. and Zitikis, R. (2017). Measuring and comparing student performance: a new technique for assessing directional associations. *Education Sciences, 7*, 1-27.

Chen, L., Davydov, Y., Gribkova, N., and Zitikis, R. (2018). Estimating the index of increase via balancing deterministic and random data. *Mathematical Methods of Statistics, 27*, 83-102.

Corrigan, J., Hackenberry, B., Lambert, V., Rousu, M., Thrasher, J., and Hammond, D. (2021). Estimating the price elasticity of demand for JUUL E-cigarettes among teens. *Drug and Alcohol Dependence, 218*, 108406-108406. https://doi.org/10.1016/j.drugalcdep.2020.108406

Cover, T. and Thomas, J. (2006). *Elements of information theory.* (2nd ed.). Hoboken, N.J: Wiley-Interscience.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (2001). Overconfidence, Arbitrage, and Equilibrium Asset Pricing. *The Journal of Finance, 56*(3), 921-965. https://doi.org/10.1111/0022-1082.00350

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, New York.

Davison, A.C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge, UK.

David, H.A. and Nagaraja, H.N. (2003). *Order Statistics* (3rd ed.). Wiley: Hoboken, NJ, USA; ISBN 978-0-47-138926-2.

Davydov, Y. and Zitikis, R. (2004). The influence of deterministic noise on empirical measures generated by stationary processes. *Proceedings of the American Mathematical Society, 132*, 1203-1210.

Davydov, Y. and Zitikis, R. (2005). An index of monotonicity and its estimation: A step beyond econometric applications of the Gini index. *Metron - International Journal of Statistics, 63*(3); (special issue in memory of Corrado Gini), 351-372.

Davydov, Y. and Zitikis, R. (2007). Deterministic noises that can be statistically distinguished from the random ones. *Statistical Inference for Stochastic Processes, 10*, 165-179.

Davydov, Y. and Zitikis, R. (2017). Quantifying non-monotonicity of functions and the lack of positivity in signed measures. *Modern Stochastics: Theory and Applications, 4*(3), 219-231. https://doi.org/10.15559/17-VMSTA84

Davydov, Y., Moldavskaya, E., and Zitikis, R. (2018). Searching for, and quantifying, non-convexity of functions. *Lithuanian Mathematical Journal, 59*, 507-518.

De Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*; Springer: New York, NY, USA; ISBN 978-0-38-723946-0.

Dehbi, H., Cortina-Borja, M., and Geraci, M. (2015). Aranda-Ordaz quantile regression for student performance assessment. *Journal of Applied Statistics, 43*(1), 58-71. https://doi.org/10.1080/02664763.2015.1025724

Duzhin, F., and Gustafsson, A. (2018). Machine learning-based app for self-evaluation of teacher-specific instructional style and tools. *Education Sciences, 8*, 1-15. (Article #7.)

Dunford, N. and Schwartz, J.T. (1988). *Linear Operators, Part 1: General Theory*. Wiley, New York.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics, 7*(1), 126. https://doi.org/10.1214/aos/1176344552

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall/CRC, Boca Raton, FL.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Annals of Applied Statistics, 6*, 1971-1997.

Egozcue, M., Fuentes García, L., Wong, W.K. and Zitikis, R. (2011). The covariance sign of transformed random variables with applications to economics and finance. *IMA Journal of Management Mathematics, 22*, 291-300.

Eichner, T. and Wagener, A. (2009). Multiple risks and mean-variance preferences. *Operations Research, 57*, 1142-1154.

Eisenbeiss, M., Kauermann, G., and Semmler, W. (2007). Estimating Beta-Coefficients of German Stock Data: A Non-Parametric Approach. *The European Journal of Finance, 13*(6), 503-522. https://doi.org/10.1080/13518470701201405

Fama, E. and French, K. (2004). The Capital Asset Pricing Model: Theory and Evidence. *The Journal of Economic Perspectives, 18*(3), 25-46. https://doi.org/10.1257/0895330042162430

Fama, E. and French, K. (2006). The Value Premium and the CAPM. *The Journal of Finance, 61*(5), 2163-2185. https://doi.org/10.1111/j.1540-6261.2006.01054.x

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 70*(5), 849911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Feder, G., Just, R.E. and Schmitz, A. (1980). Futures markets and the theory of the firm under price uncertainty. *Quarterly Journal of Economics, 94*, 317-328.

Friedman, M. and Savage, L.J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy, 56*, 279-304.

Furman, E. and Zitikis, R. (2008). Weighted premium calculation principles. *Insurance: Mathematics and Economics, 42*, 459-465.

Furman, E. and Zitikis, R. (2017). Beyond the Pearson correlation: Heavy-tailed risks, weighted Gini correlations, and a Gini-type weighted insurance pricing model. *ASTIN Bulletin, 47*(3), 919-942. https://doi.org/10.1017/asb.2017.20

Gamoran, A., and Hannigan, E.C. (2000). Algebra for everyone? Benefits of college-preparatory mathematics for students with diverse abilities in early secondary school. *Educational Evaluation and Policy Analysis, 22*, 241-254.

Gauthier, T.D. (2001). Detecting trends using Spearman's rank correlation coefficient. *Environmental Forensics, 2*(4), 359-362. https://doi.org/10.1006/enfo.2001.0061

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., Wang, G., and Zhou, F. (2016). McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinformatics, 17*(1), 142-142. https://doi.org/10.1186/s12859-016-0990-0

Giorgi, G.M. (1990). Bibliographic portrait of the Gini concentration ratio. *Metron - International Journal of Statistics, 48*, 183-221.

Giorgi, G.M. (1993). A fresh look at the topical interest of the Gini concentration ratio. *Metron - International Journal of Statistics, 51*, 83-98.

Gini, C. (1914) On the measurement of concentration and variability of characters (English translation from Italian by Fulvio de Santis) *Metron, 63*, 3-38.

Gini, C. (1921). Measurement of Inequality of Incomes. *The Economic Journal, 31*(121), 124-126. doi:10.2307/2223319

Gonçalves, A., Xue, C., and Zhang, L. (2020). Aggregation, Capital Heterogeneity, and the Investment CAPM. *The Review of Financial Studies, 33*(6), 2728-2771. https://doi.org/10.1093/rfs/hhz091

Greselin, F. and Zitikis, R. (2018). From the Classical Gini Index of Income Inequality to a New Zenga-Type Relative Measure of Risk: A Modellers Perspective. *Econometrics, 6*(1), 4-. https://doi.org/10.3390/econometrics6010004

Gribkova, N., and Zitikis, R. (2019a). Assessing transfer functions in control systems. *Journal of Statistical Theory and Practice, 13*, Article #35.

Gribkova, N., and Zitikis, R. (2019b). Statistical detection and classification of background risks affecting inputs and outputs. *Metron – International Journal of Statistics, 77*, 1-18.

Gribkova, N., and Zitikis, R. (2018). A user-friendly algorithm for detecting the influence of background risks on a model. *Risks* (Special Issue on "Risk, Ruin and Survival: Decision Making in Insurance and Finance"), 6, Article #100.

Gribkova, N.V. and Helmers, R. (2007). On the Edgeworth expansion and the M out of N bootstrap accuracy for a Studentized trimmed mean. *Mathematical Methods of Statistics, 16*, 142-176.

Gribkova, N.V. and Helmers, R. (2011). On the consistency of the $M \ll N$ bootstrap approximation for a trimmed mean. *Theory of Probability and Its Applications, 55*, 42-53.

Groenen, P.J.F., and Meulman, J.J. (2004). A comparison of the ratio of variances in distance-based and classical multivariate analysis. *Statistica Neerlandica, 58*, 428-439.

Haile, G.A. and Nguyen, A.N. (2008). Determinants of academic attainment in the United States: A quantile regression analysis of test scores. *Education Economics, 16*(1), 29-57. https://doi.org/10.1080/09645290701523218

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Härdle, W. (1991). *Smoothing Techniques, with Implementation in S.* Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (Second Edition.) Springer, New York.

Havranek, T., Irsova, Z., and Janda, K. (2012). Demand for gasoline is more price-inelastic than commonly thought. *Energy Economics, 34*(1), 201-207. https://doi.org/10.1016/j.eneco.2011.09.003

Hey, J.D. (1981). Hedging and the competitive labor-managed firm under price uncertainty. *American Economic Review, 71*, 753-757.

Heijne-Penninga, M., Kuks, J.B.M., Schönrock-Adema, J., Snijders, T.A.B., and Cohen-Schotanus, J. (2008). Open-book tests to complement assessment-programmes: analysis of open and closed-book tests. *Advances in Health Sciences Education, 13*, 263-273.

Heijne-Penninga, M., Kuks, J.B.M., Hofman, W.H.A., and Cohen-Schotanus, J. (2010). Influences of deep learning, need for cognition and preparation time on open- and closed-book test performance. *Medical Education, 44*, 884-891.

Hedges, L.V. and Hedberg, E.C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. https://doi.org/10.3102/0162373707299706

Hill, H.C., Rowan, B., and Loewenberg Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371-406.

Huxham, M., Campbell, F., and Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Assessment and Evaluation in Higher Education, 37*(1), 125-136. https://doi.org/10.1080/02602938.2010.515012

Inoue, J., Ghosh, A., Chatterjee, A., and Chakrabarti, B. (2015). Measuring social inequality with quantitative methodology: Analytical estimates and empirical data analysis by Gini and k indices. *Physica A*, 429, 184-204. https://doi.org/10.1016/j.physa.2015.01.082

Jacoby, W.G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies, 19*, 577-613.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* (1st ed.). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jovanovic, J. and King, S.S. (1998). Boys and girls in the performance—Based science classroom: Who's doing the performing? *American Educational Research Journal, 35*(3), 477-496. https://doi.org/10.3102/00028312035003477

Jurečková, J., Picek, J., and Schindler, M. (2019). *Robust Statistical Methods with R.* (Second Edition.) Chapman and Hall/CRC, Boca Raton, FL.

Kahneman, D. and Tversky, A. (1979). Prospect theory of decisions under risk. *Econometria, 47*, 263-291.

Kamps, U. (1998). On a class of premium principles including the Esscher principle. *Scandinavian Actuarial Journal, 1998*(1), 75-80.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461. https://doi.org/10.3102/00346543064003425

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *ournal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kirk, T., Malak, R., and Arroyave, R. (2021). Computational Design of Compositionally Graded Alloys for Property Monotonicity. *Journal of Mechanical Design, 143*(3). https://doi.org/10.1115/1.4048627

Kitchen, A., Savage, M., and Williams, J. (1997). The continuing relevance of mechanics in A-level mathematics. *Teaching Mathematics and its Applications, 16*, 165-170.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). *Handbook of Quantile Regression*; Chapman and Hall/CRC: Boca Raton, FL, USA; ISBN 978-1-49-872528-6.

Kolmogorov, A.N. and Fomin, S.V. (1970). *Introductory Real Analysis*. Dover, New York.

Kong, J., Wang, S., and Wahba, G. (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine, 34*(10), 17081720. https://doi.org/10.1002/sim.6441

Krasne, S., Wimmers, P.F., Relan, A., and Drake, T.A. (2006). Differential effects of two types of formative assessment in predicting performance of first-year medical students. *Advances in Health Sciences Education, 11*, 155-171.

Lehmann, E.L. (1966). Some concepts of dependence. *Annals of Mathematical Statistics, 37*, 1137-1153.

Leedy, M.G., LaLonde, D., and Runk, K. (2003). Gender equity in mathematics: beliefs of students, parents, and teachers. *School Science and Mathematics, 103*, 285-292.

Lee, S., Harrison, M.C., and Robinson, C.L. (2006). Engineering students' knowledge of mechanics upon arrival: expectation and reality. *Engineering Education, 1*, 32-38.

Li, D. X. (2000). On Default Correlation: A Copula Function Approach. *The Journal of Fixed Income, 9*(4), 43-54. https://doi.org/10.3905/jfi.2000.319253

Li, R., Zhong, W., and Zhu, L. (2012). Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association, 107*(499), 11291139. https://doi.org/10.1080/01621459.2012.695654

Liao, T. (2006). Measuring and Analyzing Class Inequality with the Gini Index Informed by Model-Based Clustering. *Sociological Methodology, 36*(1), 201-224. https://doi.org/10.1111/j.1467-9531.2006.00179.x

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics, 47*(1), 13-37. https://doi.org/10.2307/1924119

Looney, M.A. (2000). When is the intraclass correlation coefficient misleading? *Measurement in Physical Education and Exercise Science, 4*(2), 73-78. https://doi.org/10.1207/S15327841Mpee0402_3

Lorenz, M. (1905). Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association, 9*(70), 209-219. https://doi.org/10.1080/15225437.1905.10503443

Low, R., Alcock, J., Faff, R., and Brailsford, T. (2013). Canonical vine copulas in the context of modern portfolio management: Are they worth it? *Journal of Banking & Finance, 37*(8), 30853099. https://doi.org/10.1016/j.jbankfin.2013.02.036

Low, R., Faff, R., and Aas, K. (2016). Enhancing meanvariance portfolio selection by modeling distributional asymmetries. *Journal of Economics and Business, 85*, 49-72. https://doi.org/10.1016/j.jeconbus.2016.01.003

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Education Measurement, 38*(1), 1-18. https://doi.org/10.1111/j.1745-3984.2001.tb01114.x

Masci, C., Leva, F., Agasisti, T., and Paganoni, A.M. (2017). Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *Journal of Applied Statistics, 44*(7), 1296-1317. https://doi.org/10.1080/02664763.2016.1201799

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy, 60*, 151-156.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

Meyer, J. and Robison, L.J. (1988). Hedging under output price randomness. *American Journal of Agricultural Economics, 70*, 268-272.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press, Cambridge, MA

Natanson, I.P. (2016). *Theory of Functions of a Real Variable.* Dover, New York.

Nadaraya, E. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability and Its Applications, 10*(1):186190. https://doi.org/10.1137/1110024

McCornack, R. and McLeod, M. (1988). Gender bias in the prediction of college course performance. *Journal of Educational Measurement, 25*(4), 321-331. https://doi.org/10.1111/j.1745-3984.1988.tb00311.x

Nelsen, R. (2006). *An introduction to copulas*. (2nd ed.). New York : Springer.

Newman, R. and Stevenson, H. (1990). Childrens achievement and causal attributions in mathematics and reading. *Journal of Experimental Education, 58*(3), 197-212. https://doi.org/10.1080/00220973.1990.10806535

Nguyen, N.T., Allen, L.C., and Fraccastoro, K. (2005). Personality predicts academic performance: exploring the moderating role of gender. *Journal of Higher Education Policy and Management, 27*, 105-117.

Mokros, J.R. and Koff, E. (1978). Sex-stereotyping of childrens success in mathematics and reading. *Psychological Reports, 42*(3), 1287-1293. https://doi.org/10.2466/pr0.1978.42.3c.1287

Osorio, F., and Galea, M. (2015). Statistical inference in multivariate analysis using the t-distribution. https://cran.r-project.org/web/packages/MVT/

Perera, S. and Tan, D. (2019). In search of the Right Price for air travel: First steps towards estimating granular price-demand elasticity. *Transportation Research. Part A, Policy and Practice, 130*, 557-569. https://doi.org/10.1016/j.tra.2019.09.013

Politis, D. and Romano, J. (1994). Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions.*The Annals of Statistics,22*(4), 20312050.

Putwain, D.W. (2008). Test anxiety and GCSE performance: the effect of gender and socio-economic background. *Educational Psychology in Practice, 24*, 319-334.

Puth, M.-T., Neuhäuser, M., and Ruxton, G.D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour, 102*, 77-84. https://doi.org/10.1016/j.anbehav.2015.01.010

Qoyyimi, D.T. and Zitikis, R. (2014). Measuring the lack of monotonicity in functions. *The Mathematical Scientist, 39*, 107-117.

Qoyyimi, D.T. and Zitikis, R. (2015). Measuring association via lack of co-monotonicity: The LOC index and a problem of educational assessment. *Dependence Modeling, 3*(1), 83-97. https://doi.org/10.1515/demo-2015-0006

Qoyyimi, D.T. (2015). *A Novel Method for Assessing Co-monotonicity: an Interplay between Mathematics and Statistics with Applications.* Electronic Thesis and Dissertation Repository Nr. 3322. `https://ir.lib.uwo.ca/etd/3322`

Quenouille, M. (1949). Problems in Plane Sampling. *The Annals of Mathematical Statistics, 20*(3), 355375. https://doi.org/10.1214/aoms/1177729989

Rad, H., Low, R., and Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance, 16*(10), 1541-1558. https://doi.org/10.1080/14697688.2016.1164337

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Reimherr, M. and Nicolae, D.L. (2013). On quantifying dependence: A framework for developing interpretable measures. *Statistical Science, 28*(1), 116-130. https://doi.org/10.1214/12-STS405

Reiss, R.-D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields* (3rd ed.). Birkhäuser: Basel, Switzerland; ISBN 376-4-37-230-3.

Ren, J., Sendova, K. and Zitikis, R. (2019). Editorial of "Risk, Ruin and Survival: Decision Making in Insurance and Finance." *Risks* (Special Issue on "Risk, Ruin and Survival: Decision Making in Insurance and Finance"), 7, Article #96.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean. G, Turnbaugh, P., Lander, E., Mitzenmacher, M., and Sabeti, P. (2011). Detecting Novel Associations in Large

Data Sets. *Science (American Association for the Advancement of Science), 334*(6062), 1518-1524. https://doi.org/10.1126/science.1205438

Reshef, Y., Reshef, D., Finucane, H., Sabeti, P., and Mitzenmacher, M. (2016). Measuring Dependence Powerfully and Equitably. *Journal of Machine Learning Research, 17*(211), 1-63. http://jmlr.org/papers/v17/15-308.html

Reshef, D., Reshef, Y., Sabeti, P., and Mitzenmacher, M. (2018). An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics, 12*(1), 123-155. https://doi.org/10.1214/17-AOAS1093

Ross, J.A. (1992). Teacher efficacy and the effects of coaching on student achievement. *Canadian Journal of Education / Revue canadienne de l'éducation, 17*, 51-65.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J. GGally: Extension to 'ggplot2' (R Package Version 1.3.1). Available online: http://CRAN.R-project.org/package=GGally (accessed on 11 August 2017).

Schweizer, B. and Wolff, E. (1981). On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics, 9*(4), 879-885. https://doi.org/10.1214/aos/1176345528

Scott, D.W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization.* (Second edition.) Wiley, New York.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

Sharpe, W. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance, 19*(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M. and Despouy, P. (2017). *plotly: Create Interactive Web Graphics via 'plotly.js'*. https://CRAN.R-project.org/package=plotly

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall/CRC, London.

Sinn, H.-W. (1990). Expected utility, μ-σ preferences, and linear distribution classes: a further result. *Journal of Risk and Uncertainty, 3*, 277-281.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika, 9*, 449-460.

Sun, G., Li, J., Dai, J., Song, Z., and Lang, F. (2018). Feature selection for IoT based on maximal information coefficient. *Future Generation Computer Systems, 89*, 606-616. https://doi.org/10.1016/j.future.2018.05.060

Synge, J.L., and Griffith, B.A. (1949). *Principles of Mechanics.* (Second Edition.) McGraw-Hill, New York.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics, 35*(6), 27692794. https://doi.org/10.1214/009053607000000505

Székely, G. and Rizzo, M. (2009). BROWNIAN DISTANCE COVARIANCE. *The Annals of Applied Statistics, 3*(4), 12361265. https://doi.org/10.1214/09-AOAS312

Taylor, P. (2019). Mathematics is about wonder, creativity and fun, so let's teach it that way. *The Conversation*. https://theconversation.com/mathematics-is-about-wonder-creativity-and-fun-so-lets-teach-it-that-way-120133

Thorndike, R.M. and Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education* (8th edition). Prentice Hall: Boston, MA, USA; ISBN 978-0-13-240397-9.

Tucker, M., Laugesen, M., and Grace, R. (2018). Estimating Demand and Cross-Price Elasticity for Very Low Nicotine Content (VLNC) Cigarettes Using a Simulated Demand Task. *Nicotine & Tobacco Research, 20*(7), 843-850. https://doi.org/10.1093/ntr/ntx051

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297-323.

Wade, C.H., Sonnert, G., Wilkens, C.P., and Sadler, P.M. (2017). High school preparation for college calculus: Is the story the same for males and females? *The High School Journal, 100*, 250-263.

Wang, S. (1995). Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics, 17*, 43-54.

Wang, S. (1998). An actuarial index of the right-tail risk. *North American Actuarial Journal, 2*, 88-101.

Wang, W., Lesner, C., Ran, A., Rukonic, M., Xue, J., and Shiu, E. (2020). Using Small Business Banking Data for Explainable Credit Risk Scoring. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(08), 13396-13401. https://doi.org/10.1609/aaai.v34i08.7055

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference.* Springer, New York.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A, 26*(4), 359-372. http://www.jstor.org/stable/25049340

Wen, T., Dong, D., Chen, Q., Chen, L., and Roberts, C. (2019). Maximal Information Coefficient-Based Two-Stage Feature Selection Method for Railway Condition Monitoring. *IEEE Transactions on Intelligent Transportation Systems, 20*(7), 2681-2690. https://doi.org/10.1109/TITS.2018.2881284

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*; Springer, New York, NY, USA; ISBN 978-0-38-798140-6.

Wilcox, R.R. (2001). Detecting nonlinear associations, plus comments on testing hypotheses about the correlation coefficient. *Journal of Educational and Behavioral Statistics, 26*(1), 73-83. https://doi.org/10.3102/10769986026001073

Wong, W.K. (2006). Stochastic dominance theory for location-scale family. *Journal of Applied Mathematics and Decision Sciences, 2006*, 1-10.

Wu, C. F. J. (1990). On the Asymptotic Properties of the Jackknife Histogram. *The Annals of Statistics,18*(3), 14381452.

Yitzhaki, S. and Schechtman, E. (2013). *The Gini Methodology: A Primer on a Statistical Methodology*. New York, NY: Springer-Verlag.

Young, D.S. (2017). *Handbook of Regression Methods*; Chapman and Hall/CRC: Boca Raton, FL, USA; ISBN 978-1-49-877529-8.

Yuan, K.H., and Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika, 77*, 803-826.

Zhang, L. (2017). The Investment CAPM. *European Financial Management: the Journal of the European Financial Management Association, 23*(4), 545-603. https://doi.org/10.1111/eufm.12129

Zhou, H., Muellerleile, P., Ingram, D., and Wong, S. (2011). Confidence intervals and F tests for intraclass correlation coefficients based on three-way mixed effects models. *Journal of Educational and Behavioral Statistics, 36*(5), 638-671. https://doi.org/10.3102/1076998610381399

Curriculum vitae

Name: Lingzhi Chen

Education: Ph.D. Statistics, University of Western Ontario, 2017 - 2021

M.Sc. Statistics, University of Western Ontario, 2016 - 2017

B.Sc. Applied Mathematics (Statistics), South China University of Technology, 2012 - 2016

Awards: *Price Elasticities and Promotion Cannibalization Effect on Promotion Activities.* MITACS Accelerate Award, in partnership with Unilever Canada Inc. 2019 - 2020

Demand Estimation in Consumer-Packaged Goods Market Using BLP Methods. Joint MITACS Accelerate Award between Southern Ontario Smart Computing Innovation Platform (SOSCIP) and the MITACS, in partnership with Unilever Canada Inc. 2020

Western Graduate Research Scholarship, 2017 - 2021

Work Experience: Teaching Assistant, Department of Statistical and Actuarial Sciences, University of Western Ontario, 2016 - 2021

Research Assistant, Department of Statistical and Actuarial Sciences, University of Western Ontario, 2017 - 2021

Publications:

1. Chen, L. and Zitikis, R. (2017). Measuring and comparing student performance: a new technique for assessing directional associations. *Education Sciences, 7*, 1-27.

2. Chen, L., Davydov, Y., Gribkova, N., and Zitikis, R. (2018). Estimating the index of increase via balancing deterministic and random data. *Mathematical Methods of Statistics, 27*, 83-102.

3. Chen, L. and Zitikis, R. (2020). Quantifying and analyzing nonlinear relationships with a fresh look at a classical dataset of student scores. *Quality & Quantity, 54*(4), 1145-1169.