SAPIENZA
UNIVERSITÀ DI ROMA

Nonparametric Density Estimation with Wasserstein Distance for Actuarial Applications

Scuola di Scienze Statistiche

Dottorato di Ricerca in Scienze Attuariali – XXXII Ciclo

Candidate

Edoardo Glauco Luini
ID number 1741673

Thesis Advisor

Prof. Gian Paolo Clemente

There are no routine statistical questions,
only questionable statistical routines.

David Cox

Abstract

Density estimation is a central topic in statistics and a fundamental task of actuarial sciences. In this work, we present an algorithm for approximating multivariate empirical densities with a piecewise constant distribution defined on a hyperrectangular-shaped partition of the domain. The piecewise constant distribution is constructed through a hierarchical bisection scheme, such that locally, the sample cannot be statistically distinguished from a uniform distribution. The Wasserstein distance represents the basic element of the bisection technique, and has been used to measure the uniformity of the sample data points lying in each partition element.

Since the resulting density estimator can be efficiently and concisely represented, it can be used whenever the information contained in a multivariate sample needs to be preserved, transferred or analysed. Also, the proposed methodology is peculiar because Wasserstein distance makes it possible to establish an upper bound on the absolute deviation, in terms of tail value at risk, between original sample and estimator marginals. For these features, our algorithm can play an important role in nowadays insurance and financial environments characterised by greater complexity, increasingly interconnections and advanced quantitative approaches. Its applications range from pricing, to capital modelling and, in general, to all those contexts where multivariate problems arise.

Acknowledgements

The author is grateful to Philipp Arbenz, Luca Vincenzo Ballestra, Gian Paolo Clemente, Bernhard Elsner, William Guevara-Alarcòn, Travis A. O'Brien, Jürg Schelldorfer, Vivien Seguy, Larry Wasserman, and Diego Zappa for their helpful comments and precious support.

Contents

Introduction

1.1 Aim and motivations

The present work is aimed at introducing a nonparametric algorithm designed to be adopted within the context of actuarial science for estimating the density of multivariate samples.

The algorithm generates a piecewise constant distributed estimator defined on a hyperrectangular-shaped partition of the domain. The piecewise constant distribution is constructed by using a hierarchical bisection scheme, such that locally, the sample cannot be statistically distinguished from a uniform distribution. The Wasserstein distance has been used to measure the uniformity of the sample data points lying in each partition element. Furthermore, through the algorithm the user can set in each margin, with respect to the original sample, an upper limit upon the absolute error over a coherent risk measure: the tail value at risk.

The resulting estimator can be represented by a tree diagram, or tree structure, because of its hierarchical nature and recursive design. As asserted by [1], this aspect gives, compared to other nonparametric estimates, a computational advantage, since information retrieval can be systematized via a decision tree, and a benefit in terms of archiving thrift, since the final estimator can be qualified as a sequence of bisections from a starting hyperrectangle. Alternatively, a parsimonious representation can be implemented because instead of storing all data, one should only know the estimate for each nonempty region, the number of which is typically much smaller than the original sample size [2]. Besides, hyperrectangular-shaped boxes, which form the estimator domain partition, can be effortlessly expressed and cached, using their lower left and upper right vertices. It is for these reasons that our estimator is able to concisely represent the original sample, while keeping as much information as possible from the data. This approach, allowing the user not to maintain the initial sample, facilitates the storage, and expedites data transfer of the results by reducing memory requirements.

Figure 1.1 shows the recursive bisection procedure that the algorithm uses to expand the hyperrectangular-shaped partition. The corresponding hierarchical tree structure that can be used to schematize the estimator is displayed below. This is constituted as a collection of nodes, each of which embodies a hyperrectangle, that starts at a root node and expands at every bisection. Each node lists the lower left (m) and the upper right (M) vertices, the relative number of samples contained therein (p) and the split location (s), reported as the dimension and point. The leaf nodes at the bottom of the diagram compose the partition on which the piecewise constant estimator is defined.

The main reasons underlying this work and its application within actuarial sciences are to be found in the following considerations and evidences. In all the situations described below, our algorithm offers itself both as a modern methodology to stay abreast of the current shifts and developments affecting the insurance industry, and

as a state-of-the-art statistical apparatus to approximate the underlying probability distribution of the phenomenon under study.

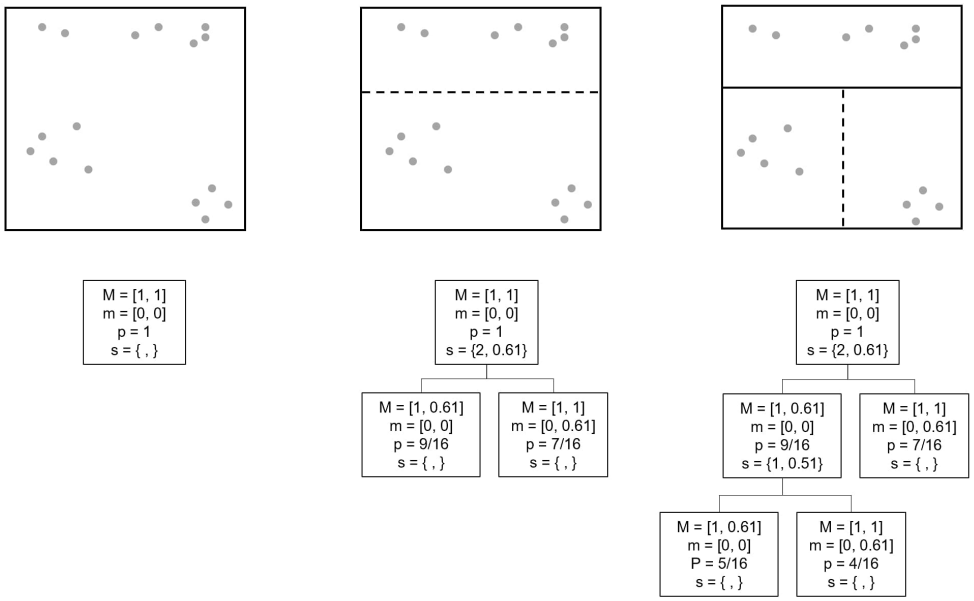

Figure 1.1. Illustration of the algorithm hierarchical bisection scheme and the tree diagram representation of the piecewise constant distribution.

Advanced Stochastic Actuarial Models

In the insurance industry, stochastic actuarial models can be found in several areas ranging from pricing, reserving, natural catastrophe modelling, to capital modelling, and many of these models use distributions characterized by non-trivial features such as heavy tails, jumps, atoms or alterations arising from the wording, terms and conditions of (re)insurance contracts. Thereby, the great majority of these random phenomena are very complex and cannot be represented in a simple manner through a parametric distribution. In such cases, empirical distributions, obtained by means of a simulated sample of realizations from a stochastic model, are widely used as an approximation of the random variable of interest [3]. In addition, in order to reasonably capture extreme events and the associated values of the risks that the insurance company is analyzing, the size of the samples involved in this process is generally large: millions or even more data points are to be handled. For such reasons, it is of crucial importance to develop statistical methodologies that are capable of operating with sample generated by advanced stochastic actuarial models.

A context in which our methodology might be adopted is the pricing of multiple lines of business and risks. Indeed, actuarial pricing activity frequently involves situations where policies can be described only through complex stochastic actuarial models, and by implication, extensive simulation exercises are required to generate an empirical distribution that is able to properly represent all the features of the risk. Even in the simplest cases where a risk to be priced is described by a single loss model, also called collective risk model, i.e. the combination of a single frequency model and a

single severity model, their probability behaviour cannot be represented through closed-form expressions, typically because of the effect of the (re)insurance cover structures. Indeed, it is common that neither the entire probability distribution nor its moments can be expressed in a simple analytical form because of policy modifications such as reinstatements and indexation clauses [4].

One of those situations regards reinsurance programs, which are usually combinations of non proportional reinsurance contracts covering multiple lines of business and risks with aggregate conditions. In order to price such contracts, the general recommendation is to keep the number of components to model at a minimum, in order to avoid a loss in credibility (there is not enough data to risk-cost each component separately) and the complications of dependency (we need to introduce assumptions about the dependency structure among the different components). However, it often happens that the separate modelling of the different components is inevitable and risks are analysed separately and then priced together [4].

Consider, by way of example, the following reinsurance combined liability policy covering employers' liability (EL), public liability (PL) and third-party motor liability (MTPL), similarly to example 28.1 in [4]. The program structure includes three excess-of-loss (XS) layers ceded to insurer ABC on top of a primary XS layer retained by a captive insurer, with an aggregate condition as outlined in Figure 1.2.

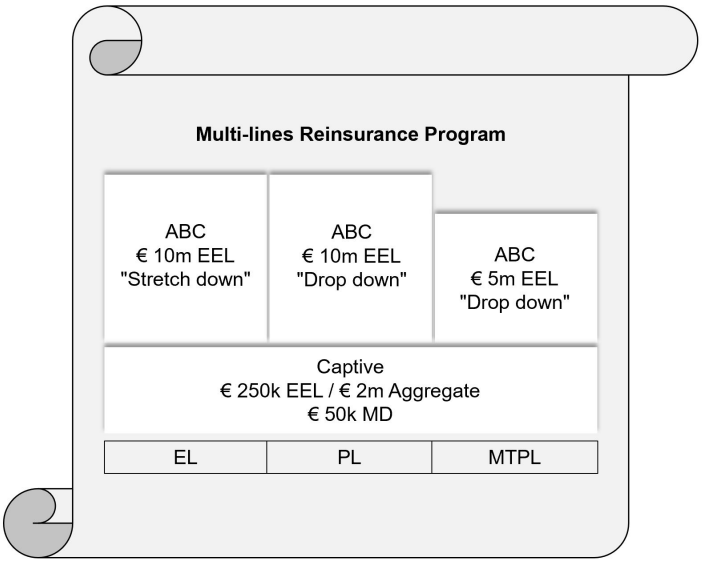

Figure 1.2. Representation of a reinsurance policy covering multiple lines of business with aggregate condition. A simulation is necessary in order to approximate the distribution of the losses and price the relative risk.

The first €250'000 of each employers' liability, public liability, or motor third-party liability loss are retained by the insured, whereas the exceeding losses are ceded to a reinsurance. Once the overall amount retained reaches €2 million, all the losses are ceded to the market. There is a maintenance deductible (MD) of €50'000. Both the EL and the PL layers have a €10 million each and every loss limit in excess of €250'000, whereas the MTPL layer has a €5 million each and every loss cover, in excess of the same amount. The EL XS layer envisages a stretch-down clause; the PL and

the MTPL XS layers have a drop-down clause. Although the loss models of the three liability risks could be regarded as mutually independent, a combined model is needed because summing the distribution of retained losses for the separated models would not give information on how likely it is that the €2 million aggregate is breached and the stretch-down clause takes affect. In addition, this models applies a common shock to the claim frequency across the three lines of business and it is assumed that upon the onset of an economic recession, the number of claims would increase by 10%.

In the case outlined above, the system describing the losses cannot be precisely reproduced using a parametric model, and requires an empirical distribution of a simulated sample generated by a stochastic procedure. Amongst other examples, Chapter 5 shows the application of our algorithm to approximate the density of the simulated loss distribution related to the above-mentioned reinsurance policy.

Complexity and Interconnection

The growing complexity and interconnection of the insurance environment have generated an increased tendency among insurance companies to store, transfer and reuse samples, models and simulated distributions [3]. The reasons behind such increasing complexity are multiple and include the following.

- Firstly, the conception and proliferation of review requirements: inputs and outputs from actuarial models are reviewed with greater frequency, from both technical and non-technical stakeholders, such as independent validation or audit. Actuarial peer reviews can form important steps in the sign-off governance. Moreover, in some jurisdictions, (re)insurance contracts need to be documented in a way that auditors can check risk transfer test requirements (see [5]).

- Secondly, the tightening of prudential legislation and regulatory standards: in the European Economic Area, the adoption of the so-called Solvency 2 regime, on the basis of the Directive 2009/138/EC [6], requires various actuarial models to be evaluated and documented for calculating the Solvency Capital Requirement (SCR), conducting the Own Risk and Solvency Assessment (ORSA), and writing actuarial function holder reports. In addition, risk management and internal control systems (Pillar 2 of Solvency 2) mandate a consistent, controlled, and transparent use of these models and their results. As stated by [7], a great amount of information needs to be managed and recorded to ensure that minimum record retention periods are adhered to, i.e. guarantee that data can be traceable and accessible for an appropriate number of years. Within the Solvency II framework each insurance company (via its actuarial and risk-management functions), together with the Supervisory Authority, is required to focus and understand the specific risk profile, as the financial requirements are effectively linked to it, and have a correct and careful management of risks. Pricing, reserving and solvency departments have therefore to intensively communicate in a dynamic and responsive way, as outlined in Figure 1.3; also, they need to deal with valuable, consistent and high-quality information.

- Thirdly, the increasingly interconnected models: information systems and applications inside insurance companies have become more convoluted, besides being mutually connected and dependent on each other. Interlinked systems entail that part of the data is accessible to various stakeholders in a manner that it must be consistent and understandable to all of them. The use test, which is mandatory

for internal models under Solvency 2 and Swiss Solvency Test [8], promotes the precise and coherent evaluation of risks, and encourage a consistent use of methods and parameters within insurance companies.

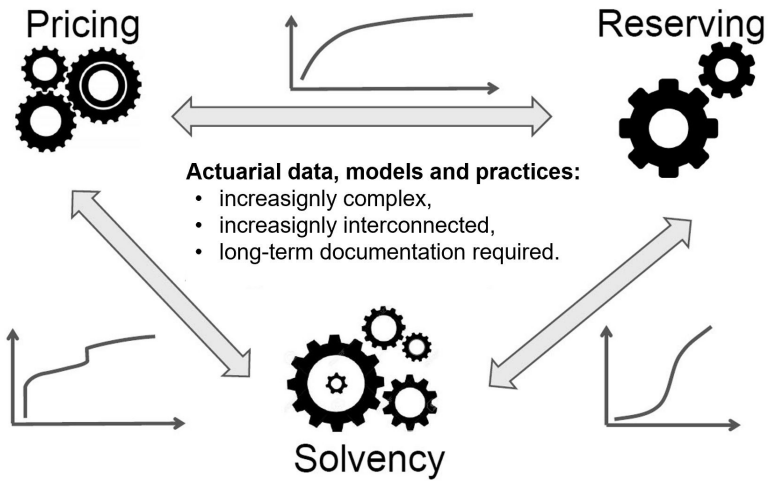

Figure 1.3. Actuarial pricing, reserving and capital modelling (solvency) functions are nowadays increasingly connected to each other.

Considering all these aspects, actuaries invariably need to have statistical and mathematical procedures designed to efficiently capture the information contained in large data set, and represent this information in such a way that can be easily handled, transferred or analysed *inter pares*. Indeed, both the market and regulators foster not to discard the resulting sample distributions, and put an increasing pressure to store them for future review, audit, or validation, in addition to redeploy them between actuarial systems.

Technological Advancements

In addition to the above observations, thanks to the significant development of technology of the last decades, the insurance sector has been witness to the outbreak of Big Data. Rising amounts of data will be collected and the data basis underlying actuarial models is set to increase.

One of the consequences of the spread of Big Data applications and simulation-driven approaches is that today, more than ever before, actuaries and actuarial science researchers are increasingly required to analyse data sets of the size of millions of entries. The constant progress in computer processing power and the expansion of available digital information, in conjunction with modern and more accurate instruments to register and store it, makes a huge plethora of insurance business related data available to be examined, and this tendency will grow continuously. Simultaneously, Monte Carlo simulation has become an ubiquitous tool to approximate probability distributions of variables that have a random behaviour. In the actuarial environment, modern computer simulation techniques opened up a wide field of practical application for risk theory concepts [9] and, as stated before, the data sets that can be obtained from

such computational intensive applications and simulations often extend to millions of observations [10].

A problem that has emerged in this context of richness of data is that even though the mechanisms to capture and store this growing influx of data are broadening and ameliorating, the amount of information created increases at a faster rate than the available storage [11]. As an example, IDC and Cisco have recently released a study report on the ever-growing datasphere, what it calls the collective world's data [12]. They predict that the collective sum of the world's data will grow from 33 zettabytes this year to a 175ZB by 2025, for a compounded annual growth rate of 61%. The 175ZB figure represents a 9% increase over last year's prediction of data growth by 2025. The study estimates that nearly 850 ZB will be generated by all people, machines, and things by 2021, up from 220 ZB generated in 2016. Most of the more than 850 ZB that will be generated by 2021 will be ephemeral in nature and will be neither saved nor stored. However, much of this ephemeral data is not useful to save, but Cisco estimates that approximately 10% is useful, which means that there will be 10 times more useful data being created (85 ZB, 10% of the 850 total) than will be stored or used (7.2 ZB) in 2021. And interestingly, useful data also exceeds dedicated space (data center) traffic (21 ZB per year) by a factor of four. Figure 1.4 shows such circumstances.

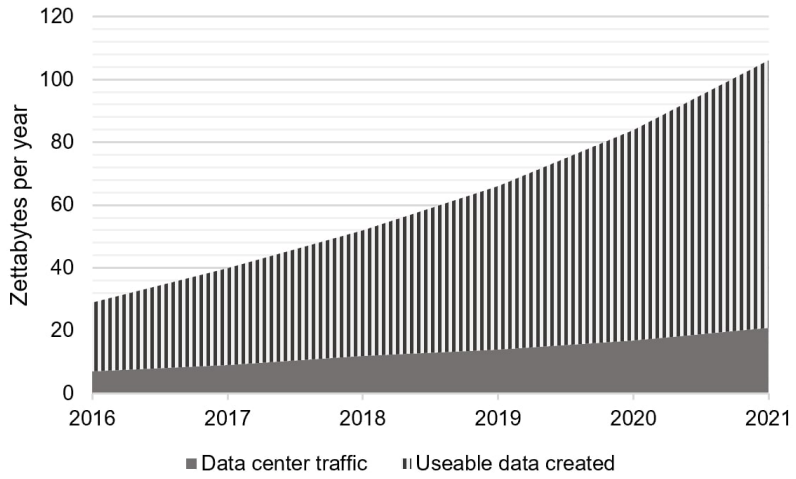

Figure 1.4. Discrepancy between the amount of stored information (data center traffic) and the amount of created data (useable data created). The rising trend is expected to escalate in the next years.

These figures clearly highlight what is reaffirmed by [13]: there has been a progressive imbalance between processor and memory technology, i.e. the speed at which instructions are executed has increased much faster than the time at which the memory is accessed. In addition, this situation emphasizes how having a large amount of data from which to mine useful and valuable information can be really problematic, not only because of methodological issue, but also because of tool and infrastructure deficiencies.

Thus, it is emphasised yet again the primary importance of having procedures capable of interacting with large samples and efficiently capturing the contained information. Such methodological schemes should be able to represent as much intelligence as possible

from the original data, and at the same time guarantee a more concise representation than the original sample.

1.2 Additional facts and considerations

Density estimation: fresh eyes on a traditional topic

Along with the transformations occurring in actuarial proceedings, new methodological challenges are being consolidated and modern technical topics are imposing themselves. These are the result of new emerging risks that arise from prevailing requests by the insurance market players like insurance for autonomous/self-driving vehicles, pricing techniques for telematics, insurance on data, cyber risk, and machine learning in individual claims. On the other hand, demands are also deriving from more traditional frameworks that need to be reconsidered, reevaluated and accordingly redesigned to embody the latest trends affecting and mutating the background they belong to. These includes for example mortality forecast and modeling, loss reserving, natural catastrophe and risk management.

Density estimation is also a long-established task in actuarial sciences and one of the fundamental topics in applied statistics. Nevertheless, in the present work, a new approach for performing density estimation in actuarial science is proposed: its purpose is to estimate the empirical density in such a way that allows an approximated and efficient representation of the original sample. Hence, in other words, our methodology can be considered a data-driven estimation technique that greedily gets as close as possible to the original sample.

In addition, our approach is intended to be consistent with the relevant frame of reference and to take into considerations the combination of recent methodological, technological and contextual variations, which manifests peculiar aspects never seen before within the financial and the insurance sectors.

In general, in probability and statistics, density estimation is any methodology aimed at constructing an estimate, based on observed data, of the unknown underlying probability density function according to which the population is distributed.

Insurance companies have to deal with several unknown sources of randomness to make their business successful: they need to estimate the characteristics of risks in order to be able set the right price (pricing), to make adequate provisions (reserving), to accurately assess its financial stability (solvency) and furthermore to write reinsurance to reduce unwanted risk. All these aspects of the insurance activity involve random variables and, as a result, the actuarial literature is interested and abundant in studies and methodologies on how to estimate the probability density function underlying an observed data sample (for example the reader may refer to [14, 15, 16]).

Density estimation techniques adopted in actuarial science, traditionally, were characterized by two aspects: they make use of parametric models and they focus on univariate approaches, as they regards a single characteristic or attribute.

Parametric methods involve a two-step model-based approach. First, an assumption about the functional form, or shape of the underlying model of the phenomenon is made. After a model has been selected, a procedure is used to estimate the parameters of the model. Therefore, an approach is referred to as parametric when it reduces the problem of estimating the model of the data down to one of estimating a set of parameters, assuming a parametric form that should simplify the task. The potential disadvantage of a parametric approaches is that the chosen model will usually not match

the true unknown one. If the former is too far from the latter, then our estimate will be poor. A way to address this problem is by choosing flexible models that can fit many different possible functional forms. But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to the problem of overfitting the data, following the noise too closely [17].

By contrast, non-parametric methods do not essentially make explicit assumptions about the functional form of the underlying model. They seek an estimate that gets as close to the data points as possible without being too erratic. Such approaches can have a major advantage over parametric ones: by avoiding the assumption of a precise functional form, they have the potential to accurately fit a wider range of possible shapes for the underlying model. As opposed to any parametric approach, non-parametric ones completely avoid the danger that the functional form of the estimate is very different from the true one. But non-parametric methodologies do suffer from a major disadvantage: since they do not reduce the problem of estimating the latent scheme to a small set of parameters, a very large number of observations (far more than is typically needed for a parametric technique) is required in order to obtain accurate estimates [17].

The one-dimensional approach is prevalent both because multivariate techniques are historically "younger" and suffer from the theoretical complexities of the multidimensional mathematics and statistics, and because actuarial phenomena were mostly regarded, analysed and maintained under a lone, singular view. However, in recent times, both these aspects have been receding.

New frontiers have been explored and are emerging in multivariate analysis [18], which is gaining in importance as concomitantly the technological innovations has permitted to handle the computational burden of modern applications of statistical theory and methods [19], and has permitted to collect massive amount of data with relatively low cost.

Notwithstanding this situation and attitude, in a context of rapid evolution and innovation as the one described above, actuarial methodologies must stay abreast of recent changes and actuarial science are challenged to ensure a rigorous e genuine assessment of risk to which the insurance sector is exposed to.

Actuaries and the new insurance landscape

The insurance industry landscape is changing at a rapid pace and a remarkable number of market tendencies and technology shifts are generating new opportunities and challenges for actuaries [20].

The actuarial profession is well known for their technical ability and knowledge of the insurance sector and its products. Actuaries were once accepted as the guardians of financial strength and equity in the insurance industry. They used analytical techniques applied to data to determine premiums and set capital and reserves [21]. However, the environment that actuaries operate in has evolved over time. There is greater emphasis on sharing and utilising insights gained from data and analytics across the entire organisation. The business tempo of motion is considerably quicker. Actuarial teams need to change to stay relevant [21].

It is no longer sufficient for an actuary to exclusively model, analyse, and estimate. Company leaders expect actuarial employees at all levels to provide business intuitions and value drivers to aid strategic decision making. Still, many actuarial organizations lack the advanced capabilities, processes, and technologies they need to meet stakeholders'

changing expectations. In response, astute chief actuaries are exploring how to modernize their organization's actuarial operating model [20].

A transition can only occur through a transformation of actuarial people, systems and procedures so that actuarial teams are in the position to undergo sustainable constructive changes that will be value adding to their management teams and their organisations. In the future, the 20/20 actuary will be a value adding business partner that communicates insight to their stakeholders, by using efficient models, data processes and technology. S/he will be supported by effective organisational structures and provide strategic leadership that will be a catalyst for change. The actuary will be involved in the traditional reserving, pricing and capital modelling areas, and other non-traditional areas related to risk management and business roles [21]. There is much to be gained from this paradigm shift. This is a unique opportunity for management teams in the organisation to maximise existing capabilities that can generate significant value and provide a competitive advantage; For chief actuaries and the actuarial professionals, to make a significant impact and reconnect more effectively with the business as a trusted advisor on issues beyond the technical [21].

Furthermore, the scope of the involvement of the actuary has significantly increased with new areas for actuarial involvement emerging. The future actuaries will be plugged into integrated data and IT systems and enabled by analytical and business intelligence tools [21]. The technological landscape of the future will have the following game changing features:

- integrated data warehousing and data access;
- robust, effective and efficient reconciliations process;
- faster close process;
- ability to review and analyse data to create insight;
- technology aligned to outcome.

It is crucial for the data and technology aspects to be tackled appropriately in order for the actuarial teams to achieve operational efficiency and success [21].

Ultimately, according to what was said by [22], actuaries can expect that all of these technologies will continue to become more interconnected. The challenge, and the promise, for the actuarial profession is managing a shrinking world in which connections among the data (and the size of the data) are expanding exponentially. This phenomenon, which is part of the onset of Big Data and is a concept related to the spread of data science, is heavily impacting actuarial science [23].

Actuarial and (Big) Data science

Data have always been a key factor for insurance companies and data analysis is an essential element of all phases of their value chain. From the product design stage, to the risk management phase, going through the underwriting process, insurers make use of techniques based on data in order to support sophisticated risk assessments and calculations for providing reliable coverages. Together with the emergence of probability theory and actuarial science as mathematical disciplines in the late 17th century, they allowed the insurance activity to evolve from "intuitive bets" on future states of the world to an industry based on rational calculus and decision making [24]. With no physical products to manufacture, data information is arguably one of insurance providers most

important assets. Financial, actuarial, claims, risk, consumer, producer/wholesaler and many other types of data form the basis for virtually every decision an insurer makes [25].

Nevertheless, until recently, the role of data remained basically the same: helping insurances to better manage their business, understand their risks, and know their customers [26]. Today however, the emergence of the so-called phenomenon of Big Data has triggered a deep transformation of the insurance industry and we are assisting to a fundamental change of the role of data in its business model [24]. Big data is often defined as "high-volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" and is a term used to describe data sets whose size is beyond the ability of traditional databases to capture, store, manage and analyse [27]. Technology experts coined the expression in early 2000 in relation to the information and data explosion driven by the ever-growing internet access and the invention of digital, or cloud, storage [28]. Moreover, the scope of the term has significantly expanded over the years and now Big Data not only refers to the data itself but also a set of technologies that capture, store, manage and analyse large and variable collections of data, to solve complex problems [29].

In the insurance field, like in many other businesses, Big Data is viewed as an inescapable imperative. In addition to providing solutions to insurance companies' long-standing business challenges, this paradigm offers the power to transform the entire sector. The technological development driving the Big Data process means additional data accessible, as well as new analytical methods applicable: the increased availability of information and new data-driven techniques allow insurers to carry out more accurate risk assessments, meaning insurance products can be better tailored to each consumer's risks and needs. Developing more refined risk models may enable companies to understand risks, offer more competitive rates, and provide insurance for risks that were previously uninsurable. As a result, established insurance providers have been investing to develop insurance products that use large amounts of data to assess, select, price, predict and prevent new risks. Going forward, access to data and the ability to derive new risk-related insights from it will be a key factor for competitiveness in the industry [24]. According to a study of IBM, 74% of insurance companies surveyed reported that the use of information (including Big Data) and analytics is creating a competitive advantage for their organizations, compared with 63% of cross-industry respondents. Compared to 35% of insurers that reported an advantage in IBM's 2010 New Intelligent Enterprise Global Executive Study and Research Collaboration, this represents a staggering 111% increase in just two years. Furthermore, the report titled "Big Data in the Insurance Industry: 2018 - 2030 - Opportunities, Challenges, Strategies & Forecasts" published by SNS Telecom & IT [30] has mentioned the following key findings:

- In 2018, Big Data vendors would have earned more than $ 2.4 billion from hardware, software and professional services revenues in the insurance industry. These investments are further expected to grow at a compound growth annual rate of approximately 14% over the next three years, eventually accounting for nearly $ 3.6 billion by the end of 2021.

- Through the use of Big Data technologies, insurers and other stakeholders are beginning to exploit their data assets in a number of innovative ways ranging from targeted marketing and personalized products to usage-based insurance, efficient

claims processing, proactive fraud detection and beyond.

- The growing adoption of Big Data technologies has brought about an array of benefits for insurers and other stakeholders. Based on feedback from insurers worldwide, these include but are not limited to an increase in access to insurance services by more than 30%, a reduction in policy administration workload by up to 50%, prediction of large loss claims with an accuracy of nearly 80%, cost savings in claims processing and management by 40-70%, accelerated processing of non-emergency insurance claims by a staggering 90%; and improvements in fraud detection rates by as much as 60%.
- In addition, Big Data technologies are playing a pivotal role in facilitating the adoption of on-demand insurance models – particularly in auto, life and health insurance, as well as the insurance of new and underinsured risks such as cyber crime.

According to [31] this new business model have the potential to generate great economic and societal benefits [24].

- Risk reduction and loss prevention. In many instances, better aligning premiums and risk has clear economic and societal benefits. It allows premiums to signal risk and encourages risk reduction. By establishing a feedback loop to policyholders, digital monitoring allows them to reduce risk by adapting their behaviour (at least as long as consumers know how to adapt their behaviour to reduce risk and their premiums). Moreover, enhanced data facilitates the establishment of advanced risk management and early-warning systems that allow for timely interventions to reduce losses and lead to additional benefits for policyholders.
- Cost reductions. A key feature of insurance markets is the prevalence of two types of informational asymmetries: moral hazard and adverse selection. They represent a market inefficiency and imply that insurers must invest considerable resources in assessing the risks of their contractual partners and verifying information provided by policyholders. In fact, a considerable fraction of premiums is spent on claims handling, acquisition and administration [31]. Accordingly, a considerable amount of employee time is spent on processing data. There is therefore a great potential for automation of data processing. [32] estimates the automation potential to be 43% of the time spent by finance and insurance employees. In non-life insurance, insurance fraud alone consumes almost 10% of premiums [33]. Automation therefore has the potential to considerably enhance market efficiency and lower costs by reducing informational asymmetries. In a competitive market environment, this will ultimately be reflected in lower premiums, boosting affordability and coverage and contributing to narrowing the protection gap [31]. Moreover, better estimation of distribution functions at the portfolio level through Big Data and artificial intelligence allows insurers to charge lower premiums by reducing the risk load.
- New and enhanced products. Data that is more granular allows insurers to offer products that are tailored to the needs of the insured, including insurance on demand or pay-as-you-use propositions. Such usage-based coverage ensures that consumers pay based on the actual risk, e.g. when they drive as opposed to when the car stays in the garage. Better understanding of risks also facilitates the development of new types of coverage and enhances the insurability of existing

and emerging risks (such as cyber risk, for example). The augmented use of data may also enable insurers to develop insurance products for high risks that so far could not be insured. For example, patients suffering from previously uninsurable diseases could share data related to their physical condition and benefit from individualised care offers [34].

To sum up, the societal and economic benefits of the enhanced use of data are highest in lines of business where the cost of moral hazard and adverse selection is high, there is great potential for risk reduction through mitigation and prevention, and there is a high degree of under insurance.

The above-mentioned benefits do not come without costs. There is a general concern that consumers do not share these benefits, or that they come at a disproportional cost to consumers, specific consumer groups or society at large. There are ethical and societal apprehensions that have been raised in regulatory and public debate. These can be grouped into three broad categories:

- Concerns about privacy and data protection. Fairness and discrimination, Intrusiveness and interference with the right of (informational) self-determination and contextual integrity.
- Concerns around an increasing individualisation of insurance. Affordability and exclusion, implications for solidarity and risk pooling and premium volatility.
- Concerns about the implications of Big Data and artificial intelligence for competition. Abuse of market power, uneven playing field and Market transparency.

It is agreed that the use of Big Data in insurance raises complex issues and trade-offs with respect to customer privacy, individualisation of products and competition. Assessing these trade-offs requires complex value judgements, and the way they are addressed leads to different scenarios for the future development of the sector. In this context, policy choices can have far-reaching consequences for the future face of the industry, its socio-economic relevance and the value it creates for its customers [24].

All these aspects and considerations exhibit how it is increasingly clear that insurance companies aim to leverage and optimize the value of their data and information assets to gain a comprehensive understanding of markets, customers, products, distribution channels, regulations, competitors, employees and more. Insurance companies will realize value by effectively managing and analyzing the rapidly growing volume, velocity and variety of new and existing data. By putting the right skills and tools in place to better understand their operations, customers and new markets, insurance organizations will be on the right track to compete and thrive in this global, dynamic marketplace [25].

In view of the foregoing, it appears important to question what would be the effect of this new paradigm on actuaries and how will they will face it. Actuaries have always been a key figure within insurance companies; they use their skills of analysis to design and price insurance policies, to measure the probability of occurrences of events helping insurance companies to be profitable and able to pay out any potential claims. Actuarial science has already been going through revolutionary changes since the late twentieth century due to the proliferation of high speed computers and the union of stochastic actuarial models with modern financial theory [35]. In addition to that, according to [23], it can be expected that over the next decade, all areas of actuarial practice will be significantly impacted by the use of Big Data. This have already led to the

development of a multi-billion-dollar industry referred to as InsurTech, the innovative use of technology in insurance, which is expected to have a significant impact on insurance and the work that actuaries perform. While the use of Big Data in the property and casualty insurance area is more developed than in some of the other areas of actuarial practice, significant advances have been made in recent years in the use of Big Data in health and life insurance. Similar advances in the pension area have not been as noticeable [23].

As consequence of the new vision and the focus on Big Data, there is a overall high demand on the job market of data science related professionals such as data scientists, data analysts and data engineers, who are becoming more and more necessary and important in each industry, including the insurance sector [36]. This tendency however is far from being a signal of the demise of actuaries, rather, it most certainly means that Data Science has now become one of the main constituent of today's competencies in business contexts, as well as research environments. There is no threat that data scientist will replace actuaries. It is rather a development of actuarial roles to fit the new requirements of changing world [37]. Most insurance companies already employ data science, and some actuaries have developed skills in machine learning. In the future, employers may want all actuaries to be proficient with data science techniques, and computers may do some of the tasks that actuaries do currently through pre-composed algorithms. This progressive penetration of computer science and data science into actuaries area of expertise should not alarm neither surprise if we consider that actuarial science has always been position itself at the intersection between different domains like economics, finance, mathematics, statistics and law. Therefore, these sciences will be another string to actuaries bow.

According to the Institute of Actuaries, "The world of data science continues to be both a threat and an opportunity for actuaries in traditional and new areas of work. With rapid advancements in technology, we can collect, store and draw insights from data like never before, with the potential to transform both traditional actuarial fields and open up new sectors and industries. We think that 'Big Data' is going to change the world and we want to be ready to embrace the opportunities that come along with it" [38]. Surely, Data Science holds huge potential for changing the way in which actuaries work and in fact Actuarial Associations have activated Data Science working parties and included Data Science modules the relative syllabus of the Actuarial credentialing and exams required in order to be fully qualified. As an example, the Society of Actuaries (SOA) is placing more emphasis on software proficiency. They did this by adding a new exam, the Predictive Analytics exam, which focus on software proficiency. Also, employers have been asking for ever more software and computers competency in general.

Finally, it is important to highlight that if on one hand actuaries can be regarded as data scientists of insurance, on the other hand their role go beyond and is not limited to data analytics (likewise, data scientist function is not confined to a mere elaboration of data). Actuaries and data scientists possess different and yet complementary competencies, as summarized in Figure 1.5. It would be fundamental therefore that insurances join together and match these two areas of expertise to be able to leverage the synergy arising when actuarial proficiency and know-how is supported by data analysis knowledge and intelligence. In consequence of that the hybrid notion of Actuarial Data Science is emerging as an extension of the more traditional Actuarial Science and it defines an application more focused on advanced statistics and programming to automatically extract patterns from data as well as produce robust predictions. In many applications

of Big Data in businesses in which actuaries are employed, multidisciplinary teams are utilized to efficiently and effectively complete the project. The teams are commonly composed of statisticians, computer scientists, data scientists, and actuaries. Actuaries on these teams may be thought of as the subject matter experts. But actuaries may be positioned to be the quarterbacks of the Big Data teams. With the proper background, an actuary can understand and direct the work of the Big Data multidisciplinary team based on their professionalism requirements and subject matter expertise. As the evolution of Big Data continues in the areas of practice in which actuaries provide services, the professionalism and technical expertise provided by actuaries are essential elements upon which the public and regulators can place reliance. The professionalism requirements of actuaries provide guidance for the proper application and disclosure of Big Data assumptions and methodologies. They require actuaries to adhere to the high standards of conduct, practice, and qualifications of the actuarial profession, thereby supporting the actuarial profession in fulfilling its responsibility to the public [23].

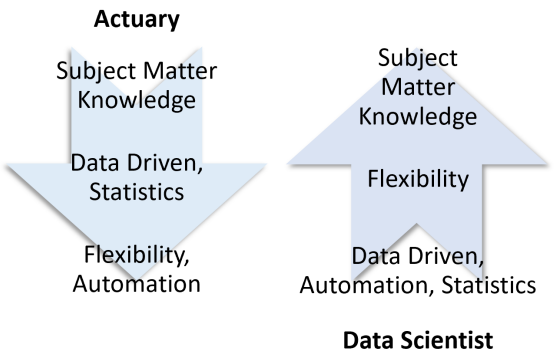

Figure 1.5. Comparison of actuary's and data scientist's knowledge and expertise according to [36].

The above-mentioned preliminary comments and introductory remarks are meant to frame the methodology herein introduced within some of the more relevant tendencies and orientations occurring in the insurance environment of today and the near-future. In the following and final part of the chapter, the plan and structure of the work are indicated.

1.3 Work synopsis

This work introduces a nonparametric algorithm for estimating an empirical density with a piecewise constant distribution defined on a hyperrectangular-shaped partition of the domain. The algorithm starts with a trivial partition, a single box containing all the observations, and recursively grows it with bisections until the sample space is divided into regions where a stopping criterion is met. Our procedure design can be classified into adaptive partitioning density estimation methods [39] and framed within Density Estimation Trees, an approach formalized by [40] that is the analogue of Classification and Regression Trees [41] for density estimation.

The aim of the work is to present a density estimator that is capable of representing in an efficient and compressed way the probabilistic information contained in a (multivariate) sample, whether this is a synthetic data set generated using a simulation-based

scheme or a data set collected and produced by information systems, computer systems, and business processes.

Initially, our procedure generates a piecewise constant distribution such that data points inside hyperrectangles are sufficiently uniformly scattered and any further partitioning of the domain does not provide additional information about the underlying density perspective. The uniformity is judged via a Wasserstein distance based hypothesis testing and a given hyperrectangle is not bisected when the hypothesis of uniformity is not rejected with a given significance level. In this respect, the underlying reasoning of our procedure is analogous to the one delineated in [42]. Unlike this, we adopt as a measure of uniformity the Wasserstein distance between the observed empirical measure and the uniform density over the hyperrectangle containing the sample. This stage stops when in all the partition elements the hypothesis of uniformity is not rejected. In a similar manner, [43] proposed a density estimation with distribution element trees method that considers, as a refinement strategy, both a statistical goodness-of-fit and pairwise independence tests to evaluate the local approximation of the distribution. This phase of our approach is compatible with the divide and conquer design paradigm: it works by repeatedly breaking down a problem into sub-problems of the same type, until all of these come to a halt. Figure 1.6 outlines the process of the algorithm, which ceases to bisect the domain when the stopping rule is met in all hyperrectangles. To improve the algorithm implementation, the uniformity of the sample points within each hyperrectangle is initially tested only on marginals, and once this condition is satisfied, also at a joint level.

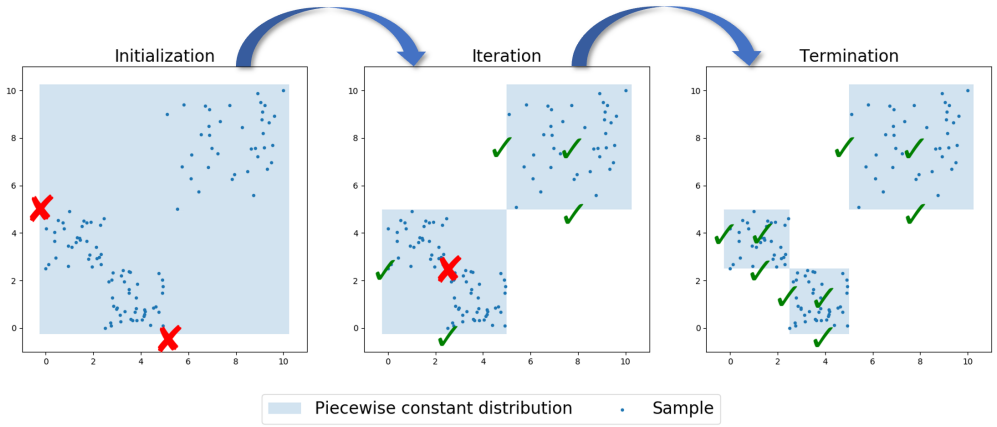

Figure 1.6. Sketch of the algorithm phases: the domain is recursively partitioned until it is divided into regions where the stopping condition is met.

In a second stage, the algorithm proceeds according to a greedy frame (see e.g. [44]) by splitting one hyperrectangle at a time, until, in all margins, a maximum allowable error level between the original sample and the final piecewise constant distribution tail values at risk has been achieved. The idea is that, under this procedure, the user knows the maximum possible discrepancy, introduced by approximating the data with our methodology, between the empirical tail value at risk and the piecewise constant estimator one. Also in this phase, Wasserstein distance is adopted, ensuring an overall consistent approach, as it implicitly sets an upper bound on the absolute difference of the tail values at risk.

This work, which has been mainly inspired by [45] and represents an extension of results of [3, 10] to the multivariate setting, is organized as follows. Chapter 2 is about piecewise constant distributions. Their representation and characteristics, together with other basic concepts which are essential to the work, are introduced. Chapter 3 focuses on the Wasserstein distance. After presenting its definition and computation, the chapter deals with the admissibility criteria, which determine the stopping rule of our partitioning algorithm. Moreover, the characteristics of the Wasserstein distance based hypothesis are detailed. In Chapter 4 the algorithm components are explained and the complete scheme for building a piecewise constant estimator is described. Finally, in Chapter 5 different examples of the algorithm run concerning benchmark and actuarial cases are presented. Conclusion and discussion follow in Chapter 6.

Piecewise constant distributions

In this chapter, piecewise constant distributions are defined and their characteristics are highlighted. Piecewise constant distributions are a flexible and concise class of distributions useful to construct summary structures for large data sets. They can approximate distributions with any shape, since the number of components scales with the complexity of the approximated distribution, and at the same time they can be represented or compressed very efficiently. The central aspect of a piecewise constant distribution is the uniformity assumption, i.e. the distribution conditioned on each hyperrectangle is a uniform density. In fact, intuitively, a distribution is said to be piecewise constant if it can be defined on a set of hyperrectangles, on which the probability is constant. In this work, this family of probability distributions is one of the two major components of our algorithm for density estimation tasks, together with the Wasserstein distance (see Chapter 3). Indeed, the estimator generated by our recursive hierarchical procedure belongs to this type of distributions.

Piecewise constant distributions can be also imagined as histograms with adaptive bandwidths for each dimension and our methodology can also be seen as an estimation technique of data-dependent multivariate histograms [44]. Other publications involving histograms and Wasserstein distance are [46] and [47]. The former deals with a clustering algorithm based on adaptive Wasserstein distance; the latter describes a strategy for constructing an optimal piecewise linear approximation of a univariate empirical distribution with a predetermined number of segments, using the Wasserstein distance of order 2 as goodness-of-fit measure. Furthermore, [48] presents a method to compute the Wasserstein distance of order 1 between a pair of two-dimensional histograms and [49] presents a novel method to compute the Wasserstein distance between a pair of d-dimensional histograms. Also, [50] offered an approach, restricted to two-dimensional settings, for image segmentation, where the authors adopt Wasserstein distance to determine the dissimilarity between two histograms. Compared with these analyses, our algorithm can handle data sets with any dimension, the number of buckets is not fixed a priori, and the Wasserstein distance is a central element in the adaptive procedure for building the histogram.

Section 2.1 illustrates the piecewise constant class of distributions adopted in this work to approximate the distribution of a given sample. After introducing the formal definition, characteristics such as the moments and covariance formulations, and the uniform mixture representation are provided. Moreover, in Section 2.2, the one-dimensional case is specifically covered. Finally, to conclude the chapter, the tail value at risk with regard to piecewise constant distributions is detailed.

2.1 Definition, characteristics and representation

In the following, we define the class of random variables with a piecewise constant (PWC) distribution that is used as an approximation to the empirical cumulative distribution function of a given sample.

2.1.1 Empirical cumulative distribution function

Definition 2.1. Consider a sample $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ of real random vectors $\mathbf{X} : \Omega \to \mathbb{R}^d$, with common distribution function $F(t) = \mathbb{P}(\mathbf{X}_i \leq t)$, where $i = 1, ..., n$. The empirical cumulative distribution function $\hat{F} : \mathbb{R}^d \to [0, 1]$ is defined as

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mathbf{X}_i \leq t\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_{i,1} \leq t_1, ..., X_{i,d} \leq t_d\}, \tag{2.1}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and $n \in \mathbb{N}$ is the sample size.

The empirical cumulative distribution function (ecdf) is the distribution function associated with the empirical measure of a sample. It is a non decreasing step function that rises by $1/n$ at each of the n sample points and its value counts the fraction of observations that are less than or equal to a specified value t.

Example 2.1. Figure 2.1 details the ecdf in $\mathbb{R}^2$ of the two-dimensional sample $(\mathbf{X}_1 = (1, 0.5), \mathbf{X}_2 = (3.5, 1), \mathbf{X}_3 = (2, 3), \mathbf{X}_4 = (4.5, 4.5))$.

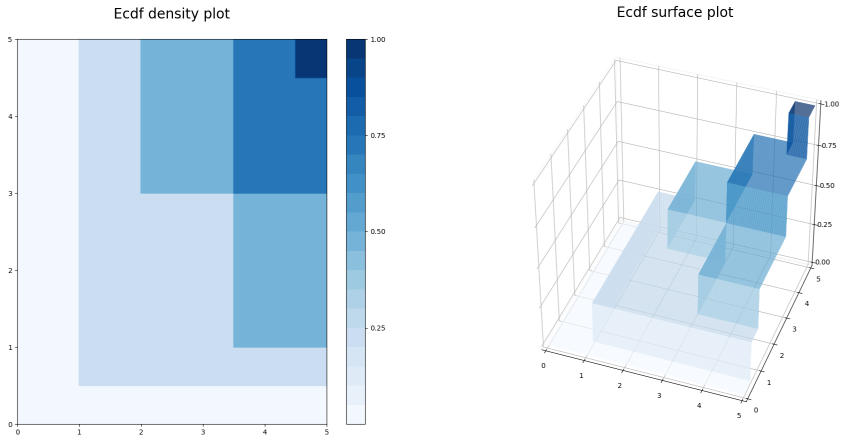

Figure 2.1. Density plot (left) and surface plot (right) of the empirical cumulative distribution function of a sample of $n = 4$ data points in $\mathbb{R}^2$.

The ecdf $\hat{F}$ enjoys small-sample and asymptotic properties. $\hat{F}$ is an unbiased and consistent estimator of F; as $n \to \infty$ it convergences almost surely, both pointwise for every value of t (by the strong law of large numbers [51]), and uniformly over t (according to the prominent Glivenko–Cantelli theorem [52]), to the true cumulative distribution function.

Example 2.2. Figure 2.2 displays the empirical cumulative distribution functions assiocated to samples of size n equal to 10 (left), 50 (center) and 250 (right) generated from a one-dimensional standard Gaussian distribution ($d = 1$). As n increases, $\hat{F}$ approaches the underlying cumulative distribution function F.

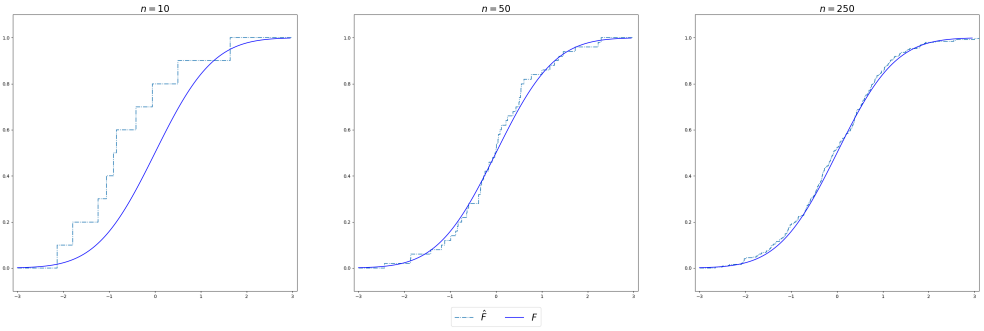

Figure 2.2. Comparison between the generating probability distribution function and the empirical cumulative distribution function, as the sample size n increases.

Throughout the remainder of the work, the notation $\mathbf{X} \sim \hat{F}$ denotes an empirical distribution function arising from a sample $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ with unknown distribution function F.

2.1.2 PWC distributions

Definition 2.2. A hyperrectangle $Q_s \subset \mathbb{R}^d$ is the Cartesian product of d intervals:

$$Q_s = I_{s,1} \times I_{s,2}, ..., \times I_{s,d}, \tag{2.2}$$

where $I_{s,j} = (a_{s,j}, b_{s,j}]$ and $-\infty < a_{s,j} \leq b_{s,j} < +\infty$, for $j = 1, ..., d$, with the convention that $I_{s,j} = \{a_{s,j}\}$ if $a_{s,j} = b_{s,j}$.

Definition 2.3. Given a set of $S \in \mathbb{N}$ disjoint hyperrectangles

$$\mathcal{Q} = \{Q_s : s = 1, ..., S\} \in \mathbb{R}^d$$

and probability weights $p = (p_s : s = 1, ..., S) \in \mathbb{R}_{\geq 0}$, such that $\sum_{s=1}^{S} p_s = 1$, a random vector $\mathbf{Y} : \Omega \to \mathbb{R}^d$ has a PWC distribution, $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$, if its cumulative distribution function $G : \mathbb{R}^d \to [0, 1]$ can be written as

$$G(t) = \mathbb{P}(\mathbf{Y} \leq t) = \sum_{s=1}^{S} p_s \prod_{j=1}^{d} H_{I_{s,j}}(t_j), \tag{2.3}$$

where the function $H_I : \mathbb{R} \to [0, 1]$ is defined as follows:

$$H_I(t) = \begin{cases} 0 & \text{if } t < a, \\ 1 & \text{if } b \leq t, \\ \dfrac{t-a}{b-a} & \text{else.} \end{cases} \tag{2.4}$$

Remark 2.1. Two special cases can be distinguished.

- Continuous case: In the case that $a_{s,j} \neq b_{s,j}$ for all j, in each s, $\mathbf{Y}$ is an absolutely continuous random vector and the probability density function $g : \mathbb{R}^d \to [0, \infty)$ can be expressed as

$$g(t) = \sum_{s=1}^{S} p_s \mathbb{1}\{t \in Q_s\} \frac{1}{\lambda(Q_s)}, \tag{2.5}$$

  where $\lambda(Q_s) = \prod_{j=1}^{d}(b_{s,j} - a_{s,j})$ denotes the d-volume of Q_s.

- Discrete case: If $a_{s,j} = b_{s,j}$ for all j, in each s, the hyperrectangles are points, and $\mathbf{Y}$ is a discrete random vector, whose probability mass function g can be represented by

$$g(t) = \mathbb{P}(\mathbf{Y} = t) = \begin{cases} p_s & \text{if } t = Q_s, \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

In the situation where only some, but not all hyperrectangles Q_s consist of points, $\mathbf{Y}$ is a mixed random vector. When $a_{s,j} = b_{s,j}$ for some s, the probability distribution does not have a density function.

Example 2.3. Figure 2.3 shows the cumulative distribution function G of a PWC distributed random vector for the continuous and the discrete cases.

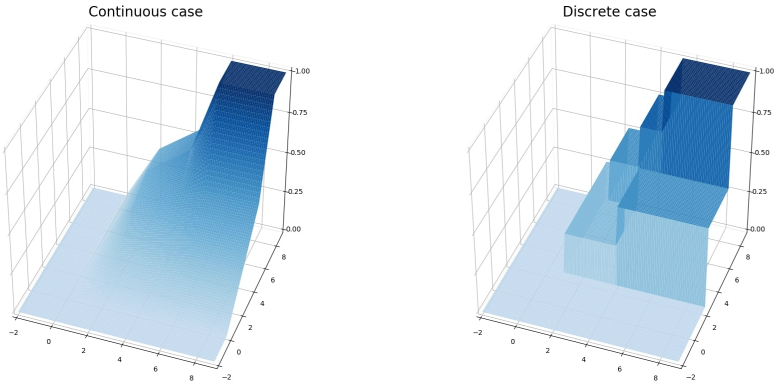

Figure 2.3. Cumulative distribution function of a two-dimensional PWC distributed random vector in the continuous case (plot on the left) and in the discrete case (plot on the right).

Definition 2.4. A PWC distribution is said to be *compatible* with the empirical distribution of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, if the following conditions hold:

a) Exists s such that $\mathbf{X}_i \in Q_s$, for all i,

b) $p_s = \dfrac{n_s}{n} = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mathbf{X}_i \in Q_s\}$, for all s.

The above requirements form the *compatibility condition* between $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ and $\mathbf{X} \sim \hat{F}$, i.e. for all s, $\mathbb{P}(\mathbf{X} \in Q_s) = \mathbb{P}(\mathbf{Y} \in Q_s)$, since p_s counts the elements of the sample that lie in Q_s.

Setting all p_s according to Definition 2.4 makes the PWC distribution a d-dimensional histogram [44].

Example 2.4. Figure 2.4 outlines the compatibility condition in $\mathbb{R}^2$ between $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ and $\mathbf{X} \sim \hat{F}$, with $n = 12$. The former is a continuous random vector and bluish shaded areas indicate rectangles Q_s, with $s = 1, ..., S$, where the density is positive; the latter has discrete support and its realizations are denoted by blue dots. Plot a): compatible. Plot b): not compatible, since there is i such that $\mathbf{X}_i \notin Q_s$ for all s, in fact two of the sample realizations lie within no hyperrectangle. Plot c): not compatible, since for example $p_4 \neq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mathbf{x}_i \in Q_4\} = 0$; according to the compatibility condition, given that none of the sample realizations is in rectangle Q_4, $\mathbf{Y}$ density should be equal to 0 in there.

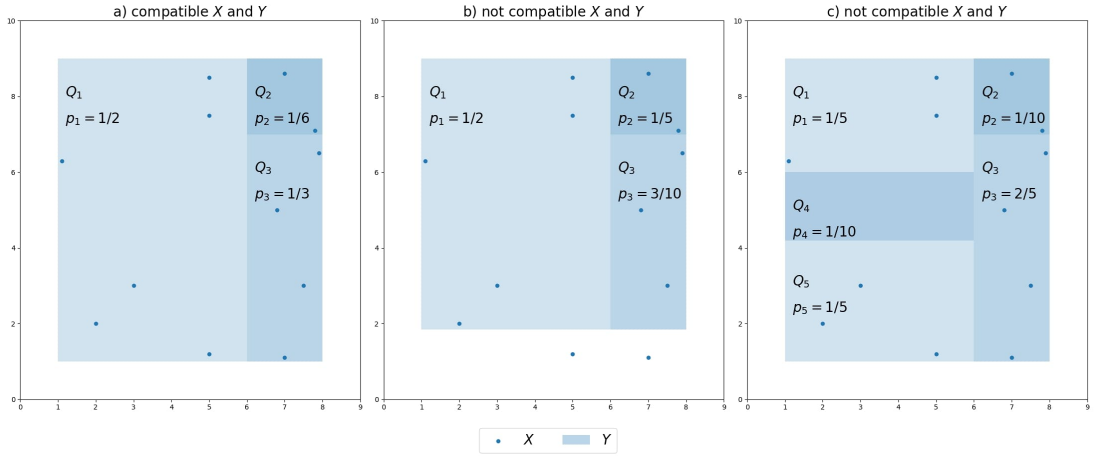

Figure 2.4. Illustration of the compatibility condition in $\mathbb{R}^2$ for a sample of $n = 12$ data points. In plot a) the PWC distribution is compatible with the observed sample. Conversely, in plot b) and plot c) the PWC distributions are not compatible.

Remark 2.2. It can be noted that if $T : \Omega \to \{1, 2, ..., S\}$ and $\mathbf{U}_s : \Omega \to Q_s$ are independent random variables, with $\mathbb{P}(T = s) = p_s$ and $\mathbf{U}_s \sim U_{Q_s}$, i.e. the (continuous) uniform random vector on Q_s, for $s = 1, 2, ..., S$, then $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ can be considered as a mixture of uniform distributions:

$$\mathbf{Y} \overset{\text{d}}{=} \sum_{s=1}^{S} \mathbb{1}\{T = s\} \mathbf{U}_s, \tag{2.7}$$

where the mixing weights are $(p_s : s = 1, ..., S)$, above-mentioned. The representation of Formula (2.7) emphasizes, how conditioned on Q_s, Y is a uniform random vector on Q_s, i.e. $\mathbb{P}(\mathbf{Y} = y | \mathbf{Y} \in Q_s) = \mathbb{P}(\mathbf{U}_s = y)$.

This representation plays a major role in the design of our algorithm and points out a useful property of the class of PWC distributions: its moments can be computed analytically. The following results are taken from [45].

Lemma 2.1. *Let $\boldsymbol{Y} \sim PWC(p, \mathcal{Q})$ and $k > 0$, then the kth raw moment $\mathbb{E}\left[\boldsymbol{Y}^k\right]$ is expressed by*

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{Y}^k\right] &= \left(\mathbb{E}\left[Y_j^k\right] : j = 1, ..., d\right), \\ \mathbb{E}\left[Y_j^k\right] &= \sum_{s=1}^{S} \frac{p_s}{k+1} \frac{b_{s,j}^{k+1} - a_{s,j}^{k+1}}{b_{s,j} - a_{s,j}}, \end{aligned} \tag{2.8}$$

where $j = 1, ..., d$.

Proof. *The result derives from the fact that Y_j can be considered as a mixture of uniform distributions with mixing weights $(p_s : s = 1, ..., S)$:*

$$\mathbb{E}\left[Y_j^k\right] = \mathbb{E}\left[\sum_{s=1}^{S} \mathbb{1}\left\{T = s\right\} U_j^k\right] = \mathbb{E}\left[\mathbb{E}\left[U_j^k | T\right]\right] = \sum_{s=1}^{S} \frac{p_s}{k+1} \frac{b_{s,j}^{k+1} - a_{s,j}^{k+1}}{b_{s,j} - a_{s,j}},$$

where $U_j \sim U_{I_{s,j}}$. In fact, the kth raw moment for a convex combination of distributions is the convex combination of the kth raw moments, provided that they exist, of the component distributions. □

Lemma 2.2. *From Lemma 2.1 it can be noted that the expected value μ_j and the variance of Y_j boil down to:*

$$\mu_j = \sum_{s=1}^{S} p_s \frac{a_{s,j} + b_{s,j}}{2}, \tag{2.9}$$

and

$$Var(Y_j) = \mathbb{E}\left[Y_j^2\right] - \mu_j^2 = \sum_{s=1}^{S} p_s \frac{a_{s,j}^2 + b_{s,j}^2 + a_{s,j} b_{s,j}}{3} - \left(\sum_{s=1}^{S} p_s \frac{b_{s,j} + a_{s,j}}{2}\right)^2. \tag{2.10}$$

Lemma 2.3. *Furthermore, the covariance between the ith and jth marginals, is equal to:*

$$Cov(Y_i, Y_j) = \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})(a_{s,j} + b_{s,j})}{4} - \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})}{2} \sum_{s=1}^{S} p_s \frac{(a_{s,j} + b_{s,j})}{2},$$

which shows that Y has a dependence structure, even though it is constructed from independent components.

Proof. *The result derives from the so-called law of total covariance:*

$$\begin{aligned} Cov(Y_i, Y_j) &= \mathbb{E}\left[Cov(Y_i, Y_j | T)\right] + Cov\left(\mathbb{E}\left[Y_i | T\right], \mathbb{E}\left[Y_j | T\right]\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[Y_i | T\right] \mathbb{E}\left[Y_j | T\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i | T\right]\right] \mathbb{E}\left[\mathbb{E}\left[Y_j | T\right]\right] \\ &= \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})(a_{s,j} + b_{s,j})}{4} - \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})}{2} \sum_{s=1}^{S} p_s \frac{(a_{s,j} + b_{s,j})}{2}, \end{aligned} \tag{2.11}$$

where the term $\mathbb{E}\left[Cov(Y_i, Y_j | T)\right]$ is equal to zero because the component distributions have independent marginals. □

Throughout the work PWC distributed random variables are indicated in formulas by their cumulative distribution function G.

2.2 One-dimensional PWC distributions

When referring to a PWC distributed random variable in $\mathbb{R}$ the parametrization simplifies. It can be noted that in this case a PWC distribution has both its cumulative distribution function and its quantile function composed by linear segments. In the one-dimensional case, a PWC distribution is effectively a piecewise linear interpolator, also referred to as linear spline and used in applications requiring data smoothing. Returning to Definition 2.3, this comes to be the following.

Definition 2.5. Given

- a set of $S \in \mathbb{N}$ disjoint intervals $\mathcal{Q} = \{I_s : s = 1, ..., S\} \in \mathbb{R}$, where $I_s = (a_s, a_{s+1}]$, for $s = 1, ..., S$, and $-\infty < a_s < a_{s+1} < +\infty$
- and probability weights $p = (p_s : s = 1, ..., S) \in \mathbb{R}_{\geq 0}$, such that $\sum_{s=1}^{S} p_s = 1$,

a random variable $Y : \Omega \to \mathbb{R}$ has a PWC distribution, $Y \sim \text{PWC}(p, \mathcal{Q})$, if its cumulative distribution function $G : \mathbb{R} \to [0, 1]$ can be written as

$$G(t) = \mathbb{P}(Y \leq t) = \sum_{s=1}^{S} p_s H_{I_s}. \tag{2.12}$$

Correspondingly to Definition 2.3, the function $H_I : \mathbb{R} \to [0, 1]$ is defined as

$$H_{I_s}(t) = \begin{cases} 0 & \text{if } t < a_s, \\ 1 & \text{if } a_{s+1} \leq t, \\ \dfrac{t - a_s}{a_{s+1} - a_s} & \text{else.} \end{cases} \tag{2.13}$$

In addition, the quantile function $G^{-1} : [0, 1] \to \mathbb{R}$ on the intervals $(\overline{p}_s, \overline{p}_{s+1}]$ is given by

$$G^{-1}(t) = \frac{(t - \overline{p}_s)}{p_s}(a_{s+1} - a_s) + a_s \text{ for } t \in (\overline{p}_s, \overline{p}_{s+1}], \tag{2.14}$$

where $\overline{p}_s = \sum_{u<s} p_u$.

Example 2.5. Figure 2.5 outlines a one-dimensional PWC distributed random variable Y, in which $\mathcal{Q} = \{I_1 = (1, 1.5], I_2 = (1.5, 2], I_3 = (2, 2.5], I_4 = (2.5, 3]\}$ and $p = (0.6, 0, 0.3, 0.1)$.

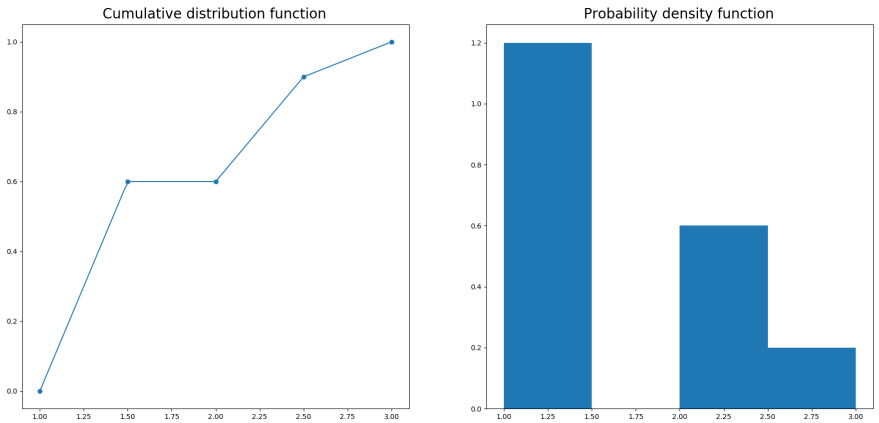

Figure 2.5. cumulative distribution function (left) and probability density function (right) of a one-dimensional PWC distributed random variable.

2.2.1 PWC distribution tail value at risk

Tail value at risk (TVaR) is a risk measure that quantifies the expected value of the tail of a distribution in the interests of taking into account the behaviour of the "more extreme" loss events. TVaR is often adopted as an alternative to other risk measure like value at risk (VaR), with which it is associated, as it is more sensitive to the shape of the tail of the loss distribution. Moreover, TVaR, although not elicitable [53, 54], is a coherent and spectral measure of risk [55]; it is calculated for a given quantile-level q, and in the financial and actuarial context can be defined as the expected loss given that a loss is occurring at or below the q-level quantile.

Definition 2.6. For a real random variable $X \sim F$, the value at risk (VaR_q), at level $q \in [0,1]$, is equal to

$$\mathrm{VaR}_q(X) = F^{-1}(q) = \inf\{x : F(x) > 1-q\}, \tag{2.15}$$

where $F^{-1}(\cdot)$ denotes the generalized inverse function of F. The VaR_q is the smallest number x such that the probability that X does not exceed x is at least q, namely it coincides with the q-level quantile of X.

Definition 2.7. For a real random variable $X \sim F$, with $\mathbb{E}[X] < \infty$, the tail value at risk (TVaR_q), at level $q \in [0,1)$, is equal to:

$$\mathrm{TVaR}_q(X) = \frac{1}{1-q}\int_q^1 \mathrm{VaR}_t(X)dt = \frac{1}{1-q}\int_{\mathrm{VaR}_q(X)}^{+\infty} tdF(t). \tag{2.16}$$

In the literature, it is often the case that different names such as average value at risk, conditional tail expectation, or expected shortfall are used to refer to the TVaR outlined in definition 2.7 above. All those quantities concur when F is an absolutely continuous distribution [56].

For PWC distributed random variable, the TVaR is always finite and it is possible to express it with the following formulation.

Lemma 2.4. *For a real random variable $Y \sim PWC(p, \mathcal{Q})$ the $TVaR_q$ is given by*

$$TVaR_q(Y) = \frac{1}{1-q}\sum_{s=1}^{S} p_s \frac{a_{s+1}+a_s}{2}\,\bar{H}_{I_s}(G^{-1}(q)). \tag{2.17}$$

Where the function $\bar{H}_{I_s} : \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$\bar{H}_{I_s}(t) = \begin{cases} 0 & \text{if } t \geq a_{s+1}, \\ 1 & \text{if } t \leq a_s, \\ \dfrac{a_{s+1}^2 - t^2}{a_{s+1}^2 - a_s^2} & \text{else.} \end{cases}$$

Proof. *$TVaR_q(Y)$ can be formalized as the conditional tail expectation, namely the expected value of Y in excess of the VaR at level q [57]. Considering the same rationale*

adopted in Remark 2.2 and Lemma 2.1, it follows that:

$$\begin{aligned}\mathbb{E}\left[Y|Y \geq G^{-1}(q)\right] &= \frac{1}{1-q}\mathbb{E}\left[Y\mathbb{1}\left\{Y \geq G^{-1}(q)\right\}\right] \\ &= \frac{1}{1-q}\mathbb{E}\left[\mathbb{E}\left[Y_s\mathbb{1}\left\{Y_s \geq G^{-1}(q)\right\}|T\right]\right] \\ &= \frac{1}{1-q}\sum_{s=1}^{S}\mathbb{E}\left[Y_s\mathbb{1}\left\{Y_s \geq G^{-1}(q)\right\}\right]p_s = \frac{1}{1-q}\sum_{s=1}^{S}\mathbb{E}\left[\bar{Y}_s\right]p_s,\end{aligned} \tag{2.18}$$

since Y can be considered as a mixture of uniform distributions $Y_s \sim U_{I_s}$, for $s = 1, 2, ..., S$, with mixing weights $(p_s : s = 1, ..., S)$ and the expected value is the combination of the expected values of the component distributions.
Hence, three cases can be distinguished for $\mathbb{E}\left[\bar{Y}_s\right]$:

- *If $G^{-1}(q) \geq a_{s+1}$, $\bar{Y}_s$ is a constant random variable equal to 0.*
- *If $G^{-1}(q) \leq a_s$, $\bar{Y}_s$ coincides with Y_s and its expected value equals*

$$\mathbb{E}\left[\bar{Y}_s\right] = \frac{a_{s+1} + a_s}{2}. \tag{2.19}$$

- *If $a_s < G^{-1}(q) < a_{s+1}$, the expected value of $\bar{Y}_s$ can be expressed as*

$$\mathbb{E}\left[\bar{Y}_s\right] = \frac{(a_{s+1} + G^{-1}(q))}{2}\frac{(a_{s+1} - G^{-1}(q))}{(a_{s+1} - a_s)}. \tag{2.20}$$

Formula (2.17) *can be derived by piecing the above results together.* □

It can be noted that Formula (2.17) is related to Formula (2.9). In fact, the TVaR of a PWC distributed random variable can be expressed by multiplying each one of the S component of the expected value formulation by the function $\bar{H}_{I_s}$, which takes into account if the value of the support considered lies to the left or to the right of the q-level quantile, and then normalizing the result for the tail level $1 - q$.

Example 2.6. Figure 2.6 gives examples of the VaR and the associated TVaR for a gamma and a PWC distributed random variables. The former has shape and scale parameters of 3.5 and 1.5 respectively, the latter is constructed on a sample of size 100 drawn from the gamma model, and has $p = (0.33, 0.55, 0.09, 0.03)$ and $\mathcal{Q} = \{I_1 = (0, 4], I_2 = (2, 8], I_3 = (8, 12], I_4 = (12, 16]\}$.

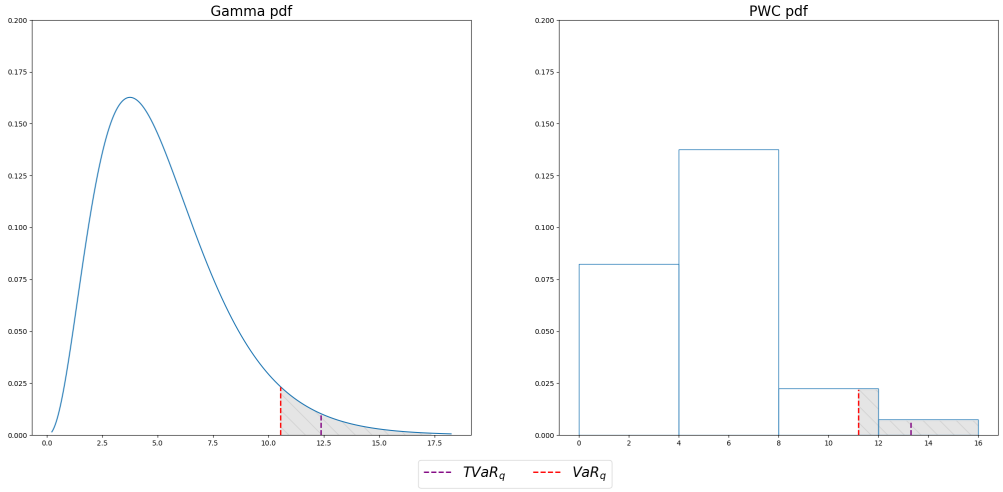

Figure 2.6. VaR and TVaR at level $q = 0.90$ for a gamma random variable (left) and for a PWC distributed random variable (right). The grey shaded area highlights the tail of the distributions.

Wasserstein distance

The Wasserstein distance constitutes, together with piecewise constant distributions, one of the cornerstones of our algorithm since it underpins the conditions required for a PWC distribution to be an admissible approximation of a sample distribution.

The Wasserstein distance, which arises from the idea of optimal transportation, has long been established as an important tool in probability theory and more recently has spread to both statistical theory and applications. Indeed, the Wasserstein distance is a powerful framework to compare two probability distributions and give a quantitative measure of their dissimilarity. It exhibits the distinctive ability to capture the geometry of the underlying space of the data, i.e. it incorporates a ground distance in comparison to other statistical distances, such as Kullback-Leibler, Hellinger, χ^2 or Total Variation, that, on the contrary, neglect how close two outcomes might be on the sample space. Moreover, the Wasserstein metric yields a map that specifies how to transform one probability distribution into the other. Last but not least, it can be applied to distributions with non-overlapping supports and compare two distributions even when one is discrete and the other is continuous.

However, the uptake of this probability distance as a statistical tool exhibits two major challenges. First, its distributional limits on spaces other than the real line are not fully known and fragmentary. Second, almost any application of the Wasserstein distance involves extensive computational effort. In the present work we address both of these matters. Indeed, the Wasserstein distance has been used as test statistic for verifying the uniformity hypothesis in a given hyperrectangle, which may contain a large multivariate sample. In literature, the so called L_2-Wassertein distance (the square root of the Wasserstein distance of order 2) has been adopted by [58, 59, 60] to introduce a goodness-of-fit hypothesis test between a fixed distribution and a location-scale family of probability distributions. It is noteworthy to mention that, using the same distance function, [61] proposed a methodology to approximate probabilities through uniform laws on convex sets. To the knowledge of the author, other publications involving the topics of Wasserstein distance and hypothesis tests are [62] and [63]. The former introduced the Wasserstein distance in nonparametric two-sample or homogeneity testing, the latter in uniformity and distributional property testing. As opposed to these works we deal with the Wasserstein distance of order 1.

Section 3.1, defines the wassestein distance. This distance function is the basis of the methodology that determines when a PWC distribution is an admissible approximation of an empirical distribution. Section 3.2 focuses on the issue of the calculation of the Wasserstein distance. After this, in Section 3.3, the admissibility criteria, which determine the stopping rule of our partitioning algorithm, are specified. In Section 3.4.2 the characteristics of the Wasserstein distance based hypothesis test are detailed. Finally, Section 3.5 concludes the chapter with supplementary aspects about the Wasserstein distance.

3.1 Wasserstein distance definition

The Wasserstein distance is a function defined on a given metric space that allows us to quantify the proximity between two probability distributions.

Definition 3.1. Given a metric space $(\mathbb{R}^d, \mathrm{c})$, for any two probability distribution function F, G on $\mathbb{R}^d$, the Wasserstein distance between F and G is defined by the formula:

$$W(F,G) = \inf_{X\sim F,\, Y\sim G} \mathbb{E}\left[\mathrm{c}(X,Y)\right]. \tag{3.1}$$

The Wasserstein distance is thus defined as the minimum expected distance among all pairs of random variables X and Y whose fixed marginal distributions are F and G respectively. Alternatively, it can also be interpreted as the minimum cost, measured by the function c, it takes to transform random variable X into Y, and vice versa. Also, an additional intuitive interpretation is that, given two distributions, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. Then, the Wasserstein distance measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of ground distance [64].

For the sake of completeness we mention that the Definition 3.1 holds true for Polish metric spaces and can be extended to the Wasserstein distance of order $q \in [1,\infty)$. In (Polish) metric spaces the minimum in Formula (3.1) is always attained (see e.g. Theorem 1.3. in [65]) and minimizers are called optimal transport plans or optimal couplings.

For convenience of notation, where it causes no confusion, we use random variables within the Wasserstein distance expression, in contrast with the above notation and most of the literature; hence, we write $W(X,Y)$ for $W(F,G)$, whenever $X \sim F$ and $Y \sim G$.

In this work we focus on the Wasserstein distance of order 1, given by Formula (3.1), and, unless otherwise stated, we restrict ourselves to the cases where c is a ℓ_p-norm, namely

$$\mathrm{c}(x,y) = \|x-y\|_p = \left(\sum_{j=1}^{d}(x_j-y_j)^p\right)^{1/p}, \tag{3.2}$$

due to its natural interpretation as the earth mover's distance [64], and implementation simplifications. Sometimes Wasserstein distance is also referred to as Mallow's distance [66], according to [67], specially if $q = 2$ and when considering distances between the distribution of random variables [68]. An exhaustive dissertation of the topic, containing also some historical context, can be found in [65], whereas [69, 70] are helpful introductions to the subject.

Definition 3.2. It can be shown that the Formula (3.1) has the following dual expression:

$$W(F,G) = \sup_{\psi\in\Psi}\left\{\int \psi(x)dF(x) - \int \psi(x)dG(x)\right\}, \tag{3.3}$$

where Ψ denotes the set of all functions $\psi : \mathbb{R}^d \to \mathbb{R}$, such that $|\psi(y)-\psi(x)| \le \mathrm{c}(x,y)$.

Lemma 3.1. *Definition 3.1 and 3.3 are equivalent.*

Proof. *See Remark* 6.5. *in [65].* □

Furthermore, it is to be noted that Wasserstein distance is strictly related to the optimal transportation problem, a topic which dates back to Monge's [71] and Kantorovic's works [72], to the extent that researchers often use the term optimal transportation (OT) distance.

Example 3.1. The concept of the optimal coupling minimizing Formula (3.1) can be readily visualized, for the two- and three-dimensional cases, when $\mathbf{X}$ and $\mathbf{Y}$ are two discrete uniform random vectors with a different support. Figure 3.1 shows the optimal coupling related to the Wasserstein distance (when c is the ℓ_2-norm) and an example of a suboptimal coupling.

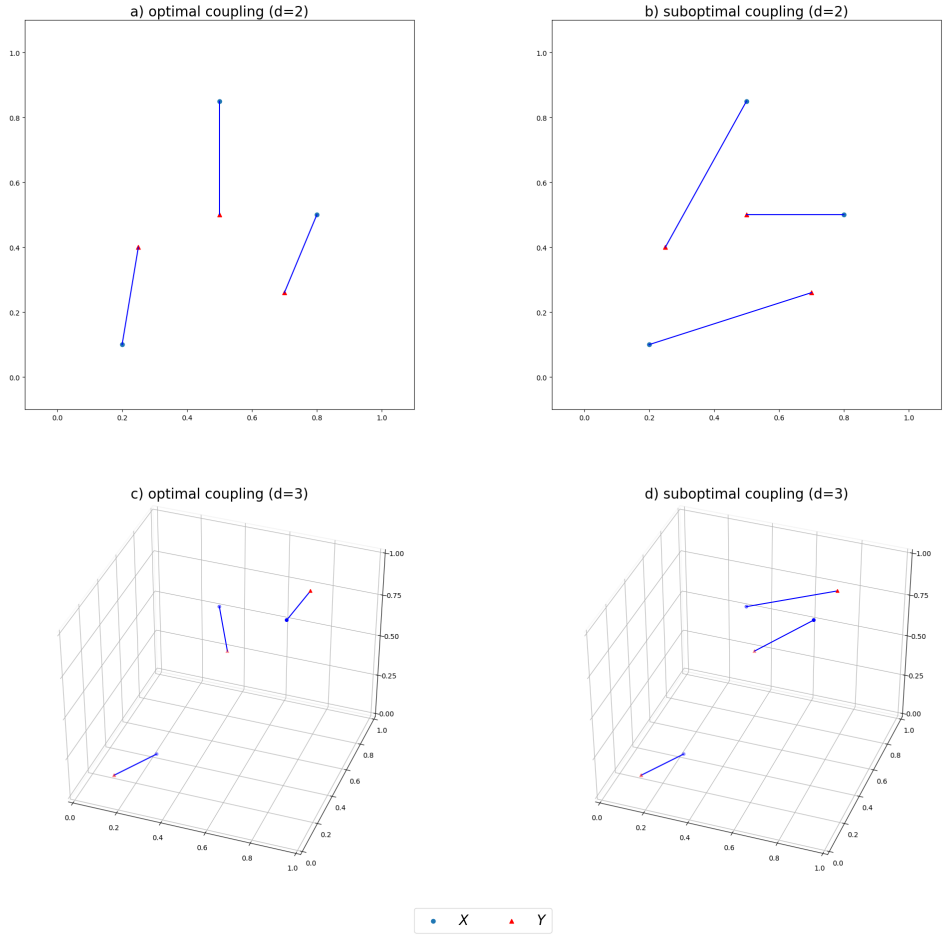

Figure 3.1. Illustration of couplings, represented by blue lines, of two discrete uniform random vectors $\mathbf{X}$ and $\mathbf{Y}$, denoted by blue dots and red triangles respectively. Plot a) and c) display the optimal coupling related to the Wasserstein distance, whereas plot b) and d) show an example of a suboptimal coupling.

3.2 Wasserstein distance computation

One-dimensional setting

In the one-dimensional setting the calculation of the Wasserstein distance can be obtained without major problems. The optimal coupling attaining the minimum in (3.1) is known explicitly [66] and the Wasserstein distance can be expressed in the following form:

$$W(F,G) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt = \int_{-\infty}^{\infty} |F(x) - G(x)|dx, \tag{3.4}$$

where the last equality is obtained by the Fubini-Tonelli Theorem. The formula shows that the Wasserstein distance between two measures on $\mathbb{R}$ then becomes the ℓ_1-norm of the difference of their quantile functions. Also, it corresponds to the area between the two quantile functions or, equivalently, to the area between the two cumulative distribution functions of the two random variables. Intuitively, this is because the minimum cost of transportation is obtained by preventing any superfluous movements of mass along the domain by preserving the order of the variables. Figure 3.2 depicts the Wasserstein distance expressed by Formula (3.4) in the one-dimensional setting.

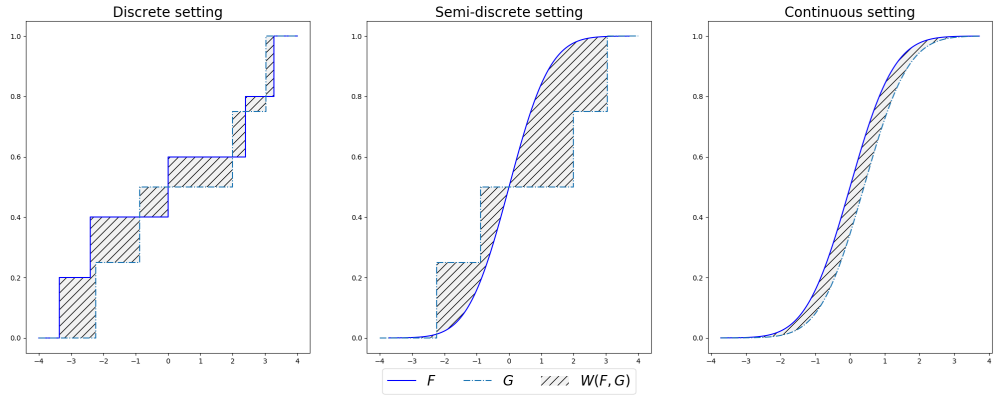

Figure 3.2. Illustration of the Wasserstein distance between two one-dimensional random variables in the discrete (left), semi-discrete (middle) and continuous (right) settings. It coincides with the area between the two quantile functions (or identically with the area between the two cumulative distribution functions).

When involving a sample and a compatible PWC distribution, the Wasserstein distance can be represented through a more specific formulation than Formula (3.4) that is given by Lemma 3.2. This is an important result in our work because it provides a closed-form expression for the specific one-dimensional cases we focus on (an additional result is given by Lemma 3.4 below).

Lemma 3.2. *The Wasserstein distance between* $X \sim \hat{F}$ *and a compatible* $Y \sim PWC(p, \mathcal{Q})$ *real random variables is given by*

$$W(\hat{F},G) = \sum_{i=1}^{n} \left[W_i + \frac{1}{2}\left(\frac{1}{n} - \frac{W_i}{\delta_i}\right) \cdot \max\left\{\delta_i - nW_i, 0\right\} \right], \tag{3.5}$$

where

$$\begin{aligned} W_i &= \frac{1}{n}\Big|X_{(i)} - G^{-1}\left(\frac{i-1/2}{n}\right)\Big|, \\ \delta_i &= \Big|G^{-1}\left(\frac{i}{n}\right) - G^{-1}\left(\frac{i-1/2}{n}\right)\Big| \end{aligned} \tag{3.6}$$

and $X_{(i)}, i = 1, ..., n$, *are the order statistics of the sample. The quantile function* $G^{-1}(t)$ *is provided by Formula* (2.14) *of Definition 2.5.*

Proof. *We have that* $\hat{F}^{-1}(t)$ *is constant for* $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right] : \hat{F}^{-1}(t) = X_{(i)}$. *Hence,*

$$nW_i = \Big|X_{(i)} - G^{-1}\left(\frac{i-1/2}{n}\right)\Big|$$

equates to the distance from $G^{-1}\left(\frac{i-1/2}{n}\right)$, *the middle point of* G^{-1} *on* $\left(\frac{i-1}{n}, \frac{i}{n}\right]$, *to the sample value* $X_{(i)}$. *Analogously, the distance from that middle point to the end of* G^{-1} *on the segment* $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ *is given by* δ_i. *We can now distinguish two cases: whether* $G^{-1}(t)$ *attains the value* $X_{(i)}$ *in the interval* $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ *(case 1) or not (case 2).*

- *Case 1:* $G^{-1}(t) \neq X_{(i)}$ *for all* $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$. *In this case* $nW_i \geq \delta_i$ *and* $\max\{\delta_i - nW_i, 0\} = 0$. *We know that* $\hat{F}^{-1}(t) - G^{-1}(t)$ *does not change sign for* $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$. *Therefore,*

$$\int_{(i-1)/n}^{i/n} |\hat{F}^{-1}(t) - G^{-1}| = W_i.$$

- *Case 2: if* $G^{-1}(t)$ *attains the value* X_i *in the interval* $\left(\frac{i-1}{n}, \frac{i}{n}\right]$, *then* $\hat{F}^{-1}(t) - G^{-1}$ *changes its sign in that interval and* $nW_i < \delta_i$. *In this case, as shown in Figure 3.3, the* i*th addend of the summation of Formula* (3.5) *is given by the area of two triangles. The areas of the smaller and larger triangles are respectively given by*

$$\frac{1}{2}(\delta_i - nW_i)\frac{1}{2n}\left(1 - \frac{nW_i}{\delta_i}\right) \quad \text{and} \quad \frac{1}{2}(\delta_i + nW_i)\frac{1}{2n}\left(1 + \frac{nW_i}{\delta_i}\right).$$

  Adding up these terms yields what was to be demonstrated:

$$W_i + \frac{1}{2}\left(\frac{1}{n} - \frac{nW_i}{\delta_i}\right)(\delta_i - nW_i).$$

□

Figure 3.3 shows the two forms, mentioned in the proof of Lemma 3.2, that the ith component of Formula (3.5) can take.

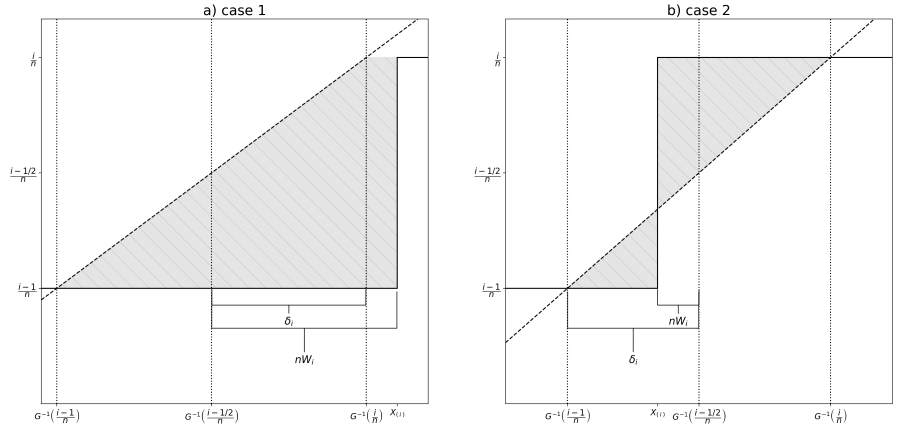

Figure 3.3. Illustration of W_i (grey shaded area) for the empirical distribution $\hat{F}$ (solid line) and the PWL approximation G (dashed line). Plot a) represents the first case in which $nW_i \geq \delta_i$; conversely plot b) displays the second case in which $nW_i < \delta_i$.

Multidimensional setting

In a multidimensional setting the calculation of the Wasserstein distance becomes more challenging and demanding.

The prevalent way to compute OT distances is by solving a large-scale linear program. In fact, when both measures involved are discretized (finite weighted sums of Dirac masses), Wassertein distance formulation fits into a *discrete* setting and becomes a Linear Assignment Problem, often shorten to LSAP [73]. This is one of the most famous problems in linear programming and in combinatorial optimization and can be exemplified as follows. Given two samples $(\mathbf{X}_i)_{i=1,\dots,n}$ and $(\mathbf{Y}_i)_{i=1,\dots,n}$ from which we compute the $n \times n$ distance (or cost) matrix $\mathcal{C} = (\mathrm{c}_{i,j})$, where $\mathrm{c}_{i,j} = \mathrm{c}(x_i, y_j)$ for all i, j, we want to select n elements of $\mathcal{C}$, so that there is exactly one element in each row and one in each column, and the sum of the corresponding distances is a minimum. Discrete solvers consists of combinatorial optimization algorithms, based on the linear programming formulation, such as the Hungarian method [74], the auction algorithm [75], and the network simplex [76]. According to [77], there are over 20 established methods for solving such problems, and at least seven software packages capable of handling one or more of these methods [78]. If on one hand, they work for almost any ground metric c and are numerically robust with regard to input data regularity, on the other, since the geometric structure of the c function is not used, this class of solvers do not scale well for large, dense problems [79]. For this reason, a wide range of approximate and dedicated methods has been applied in order to accelerate discrete solvers [80, 81, 82]. Among these, entropic-regularized approaches, introduced by the renowned work of [83], solve the OT problem by adding an entropic penalization to the original optimal transport formulation and have been shown to be extremely efficient to approximate Wasserstein distances at a low computational cost (it is worth mentioning that, because of the added regularization term, the solution will in general be different from the one of the original design). One of the most-used method for solving the resulting regularized optimization problem is represented by the Sinkhorn algorithm and its recent refinements [84, 85, 86, 87]. In addition to the above-mentioned procedures, it has to be noted that multi-scale schemes have also been proposed to boost discrete linear solvers for the computation of Wasserstein distances. The underlying idea is to approximate the original problem by a sequence of gradually finer estimations. However,

most of these techniques can be applied only in Euclidean spaces (see e.g. [88, 79, 89]). [90] carefully summarizes the topic of discrete optimal transport problem. It is worth recalling that, despite being exclusively discrete, these algorithms can also be exploited to approximate Wasserstein distance when distributions involved are not both discrete, as long as one can sample from them.

The OT entropic-regularized approach paved the way to the application of stochastic optimization methodologies [91], like stochastic gradient descend, averaged stochastic gradient descend and stochastic averaged gradient, to solve optimal transportation problems. Even though the Sinkhorn algorithm speeds up the computation and ensures convergence properties, given the more regular form of the regularized optimization problem, it does not scale well to measures supported on a large number of samples, since it needs to handle the "heavy" $n \times n$ distance matrix, where n is the largest size of the supports of the two input distributions. A solution to this issue is represented by stochastic optimization methods, which enabled the computation of Wasserstein distance also when large samples are involved. [92], in fact, showed that the dual formulation of the optimal transport problem corresponds to the maximization of an expectation, and thus it can be tackled using standard stochastic programming methods in all three possible settings. In addition, other recent works proposed different approaches to solve the dual OT problem through stochastic gradient methods: [93] introduced the so called L^2 regularization, [94] suggested a stochastic gradient based procedure to compute large-scale optimal transport, and [95] outlined a formulation that is specific to the OT using the Euclidean distance as a ground metric.

Semi-discrete optimal transport regards the setting when one of the two input measures is discrete and the other one has a density. The solution to the semi-discrete transport problem has a nice geometrical interpretation and is linked with (generalized) Voronoi diagrams and power diagrams, as shown in [96, 97, 98] and in [99]. This connection allows the enforcement of tools from computational geometry to obtain fast computational schemes. Nonetheless, this approach is (again) restricted to the Euclidean and the squared Euclidean distance, and can be implemented only in low dimensions, because of its intractable form when $d \geq 3$. For theory and application see [100, 101], [102] and [103]. Other relevant works, which try to provide more general methods in a semi-discrete setting, are the ones of [104, 78], and [105].

The *continuous setting*, i.e. when two continuous probability distributions are compared, is not of primary interest in the present work. Two publications regarding this setting are [106] and [107].

3.3 Admissibility criteria

Admissibility criteria specify the conditions, based on Wasserstein distance, required for a PWC distribution to be an admissible approximation of $\hat{F}$. The marginal and the joint admissibility rules are consistent with the essential assumption of piecewise constant distributions that the distribution conditioned on each hyperrectangle is a uniform, and arise from the fact that a uniform distribution on a hyperrectangle has two peculiar attributes: it has uniform marginals and these are mutually independent. The TVaR admissibility condition is principally connected the idea that by means of the Wasserstein distance is possible to control the error of approximating the original sample, measured using the TVaR.

3.3.1 Marginal admissible approximation

Definition 3.3. Given a hyperrectangle Q_s, let $F_{s,j}$ and $\hat{F}_{s,j}$ denote, respectively, the cumulative distribution function of X_j in $I_{s,j}$, and the empirical cumulative distribution function of the sample projection on $I_{s,j}$. We define the null hypothesis

$$\mathbb{H}_0^{*\,s,j} : F_{s,j} = U_{I_{s,j}}, \tag{3.7}$$

and the test statistic

$$W(\hat{F}_{s,j}, U_{I_{s,j}}),$$

which is the Wasserstein distance in Q_s between the sample projection on the kth dimension, i.e the kth marginal, and the uniform density on $I_{s,j}$.

The above hypothesis test is aimed at verifying, using a Wasserstein distance based test statistic, that the kth margin of the sample contained in Q_s is uniformly spread over $I_{s,j}$.

Definition 3.4. We define $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ to be a *marginal admissible approximation* of $\mathbf{X} \sim \hat{F}$, if it is compatible and none of the null hypotheses

$$\left\{\mathbb{H}_0^{*\,s,j} : s = 1, ..., S, j = 1, ..., d\right\} \tag{3.8}$$

is rejected. This means that with a significance level $\alpha \in [0, 1]$, for all s and j:

$$\mathbb{P}\left(W(\hat{F}_{s,j}, U_{I_{s,j}}) > w_{s,j} \mid \mathbb{H}_0^{*\,s,j}\right) > \alpha \tag{3.9}$$

where $w_{s,j}$ denotes the observed value of the test statistic in $I_{s,j}$.

In other terms, Definition 3.4 states that $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ is a marginal admissible approximation of $\hat{F}$ when $\hat{F}_{s,j}$ cannot be distinguished in a statistically significant manner from a uniform distribution on $I_{s,j}$, with a predefined significance level α, in each and every hyperrectangle Q_s and dimension j.

When $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ is a marginal admissible approximation of $\hat{F}$ the first distinctive feature of uniform distributions is met: in each hyperrectangle, the margins of the sample contained in it do not significantly differ from the uniform distribution.

We introduce the following hypothesis test in addition to the one above.

3.3.2 Admissible approximation

Definition 3.5. Given a hyperrectangle Q_s, consider the random vector $\widetilde{\mathbf{U}}_s : Q_s \to C$, where $C = [0, 1]^d$, as the following transformation of $\mathbf{X}$ in Q_s:

$$\widetilde{\mathbf{U}}_s = \left(H_{I_{s,1}}(X_1), H_{I_{s,2}}(X_2), ..., H_{I_{s,d}}(X_d)\right). \tag{3.10}$$

Furthermore, let F_s and $\hat{F}_s$ denote, respectively, the cumulative distribution function of $\widetilde{\mathbf{U}}_s$ and the empirical cumulative distribution function of the transformed sample. We define the null hypothesis

$$\mathbb{H}_0^s : F_s = U_C, \tag{3.11}$$

and the test statistic

$$W(\hat{F}_s, U_C),$$

which is the Wasserstein distance between the transformed sample and the uniform density on set C.

The hypothesis test introduced by Definition 3.5 is aimed at verifying, using a Wasserstein distance based test statistic, that the transformed sample is uniformly spread over C.

Definition 3.6. We define $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ to be an *admissible approximation* of $\hat{F}$, if it is a marginal admissible approximation, and none of the null hypotheses $\{\mathbb{H}_0^s : s = 1, ..., S\}$ is rejected. This means that with a significance level $\alpha \in [0, 1]$, for all s:

$$\mathbb{P}\left(W(\hat{F}_s, U_C) > w_s \mid \mathbb{H}_0^s\right) > \alpha \tag{3.12}$$

where w_s denotes the observed value of the test statistic in Q_s.

In other terms, Definition 3.6 states that $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ is an admissible approximation of $\hat{F}$ when the transformed sample cannot be distinguished in a statistically significant manner from a uniform distribution on C, with a predefined significance level α, in each and every hyperrectangle.

According to Definition 3.6, the marginal admissibility condition is required for a piecewise constant distribution to be also an admissible approximation of $\hat{F}$. Given this fact, in each Q_s, it stands to reason that sample margins are uniformly distributed on $I_{s,j}$, for all j, and consequently, by applying the transformation $H_{I_{s,j}}$, they are uniform on $[0, 1]$. Hence, the joint uniformity assumption can be tested against the transformed sample, which lies in C and retains the dependence structure of the original sample. Intuitively, we want to detect distributions with uniform marginals at first, and secondarily exclude, between these, those that have not a joint uniform distribution. Evidently, not all distributions with uniform marginals have mutually independent marginals, and not all distributions with mutually independent marginals have uniform marginals.

Example 3.2. Figure 3.4 contrasts the marginal admissible and the admissible conditions in $\mathbb{R}^2$. Realizations are denoted by blue dots and the bluish shaded square indicates the support of $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$. Plot a) displays a sample uniformly spread out over the square. Conversely, plots b) and c) depict samples whose margins are close to uniforms, but joint distributions are not. In these cases, the PWC distribution is only a marginal admissible approximation of the empirical distribution.

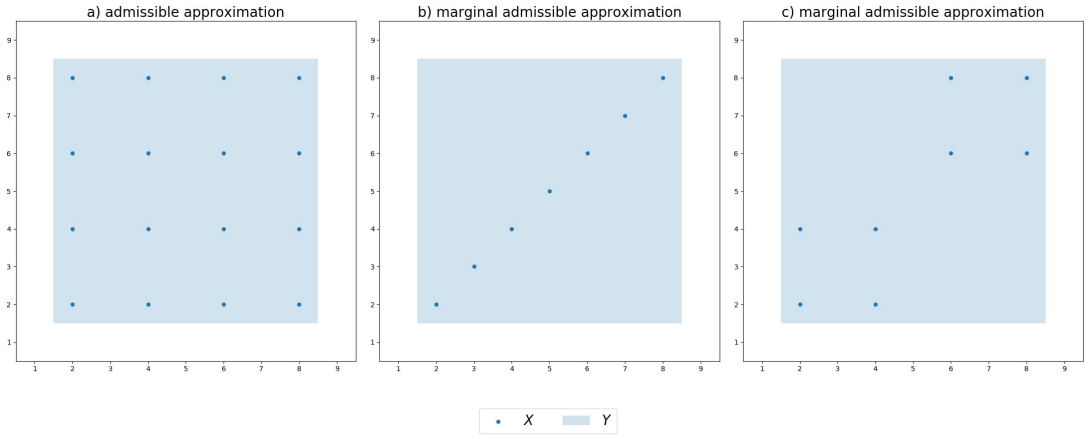

Figure 3.4. Illustration of the admissibility conditions in $\mathbb{R}^2$. Plot a) represents an admissible PWC distribution, whereas plots b) and c) show PWC distributions which are only marginal admissible.

Finally, it has to be noted that testing the marginal admissibility condition in a particular Q_s involves a set of statistical inferences simultaneously, and hence the multiple comparisons or multiple testing problem occurs, i.e. the more inferences are made, the more likely erroneous inferences are to occur. Different procedures are available in statistics for adjusting p-values and controlling the so-called Family Wise Error Rate (FWER), namely the probability of at least one false positive (type I error). Whilst on one hand this approach seems appropriate, since a single significant p-value across $\mathbb{H}_0^{*\,s,j}$ establishes that Q_s is not marginal admissible, on the other hand it would increase the probability of false negative (type II error). In our context, we consider more important to ensure that non-uniformity is identified, rather than to limit uniformity to not being detected.

3.3.3 TVaR admissible approximation

In the financial context using a fixed accuracy parameter approach can be meaningful since a Wasserstein constraint implies a bound on certain risk measures. In fact, [10] demonstrate in a one dimensional setting ($d = 1$) the existence of a relation between the Wasserstein distance and some risk measures, including the tail value at risk. In particular, the authors show that the Wasserstein distance between two random variables implicitly determines an error bound of the tail value at risks of the two random variables in question.

Definition 3.7. We define $\mathbf{Y} \sim \mathrm{PWC}(p, \mathcal{Q})$ to be a *TVaR admissible approximation*, at q-level, of $\mathbf{X} \sim F$, if it is admissible, and for all $j = 1, ..., d$ it holds that:

$$W(X_j, Y_j)/(1-q) \leq \epsilon_j, \tag{3.13}$$

where $\epsilon_j \in [0, \infty)$, $q \in [0, 1)$.

The reason for a TVaR admissible approximation is made clear by the following theorem which relates the Wasserstein distance and the above-mentioned risk measure.

Theorem 3.1. *Given any two real random variables (is in some ℓ_p-space, where $p \geq 1$, such that the TVaR exists) $X \sim F$ and $Y \sim G$, for any $q \in [0, 1)$, it follows that:*

$$W(X, Y) \geq (1-q)|\,TVaR_q(X) - TVaR_q(Y)|. \tag{3.14}$$

Proof. *The theorem results from Definition 2.7 of TVaR and from considering the following.*

$$\begin{aligned} W(X,Y) &= \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt \geq \int_{1-q}^1 |F^{-1}(t) - G^{-1}(t)|dt \\ &\geq \Big| \int_{1-q}^1 F^{-1}(t) - G^{-1}(t)dt \Big| = (1-q)|\,TVaR_q(X) - TVaR_q(Y)|. \end{aligned}$$

The reader can also refer to Theorem 3.8 in [10]. □

Necessarily then, by combining Definition 3.7 and Theorem 3.1, once it holds that $W(X, Y)/(1-q) \leq \epsilon_j$, we also know that $|\mathrm{TVaR}_q(X) - \mathrm{TVaR}_q(Y)| \leq \epsilon_j$. In this regard,

the Wasserstein distance also determines an upper bound on the the (one dimensional) marginal distributions approximation error, in terms of tail value at risk. Put another way, if an admissible piecewise constant distribution of a sample is also a TVaR admissible approximation at q-level, then in each dimension j, the difference (in module) between the q-level TVaR of the PWC distribution and the q-level TVaR of the original sample is no greater than ϵ_j. This threshold measures the maximum error allowed by the PWC distribution approximation in terms of TVaR. Hereafter, the term $W(X,Y)/(1-q)$ is referred to as the Wasserstein distance bound, $|\text{TVaR}_q(X_j) - \text{TVaR}_q(Y_j)|$ as the absolute TVaR deviation, and ϵ_j as the tolerance threshold (error) parameter.

Example 3.3. Figure 3.5 depicts the relationship that exists between the Wasserstein distance bound and the absolute TVaR deviation. The former is indicated with a dashed line; the latter is denoted by a solid line. In the left side of the graph two lognormal distributions with parameters $(\mu = 0, \sigma = 1)$ and $(\mu = 0, \sigma = 1.5)$ are considered, whereas, on the right, two Pareto distributions with shape parameters respectively equal to 1.5 and 2 are confronted.

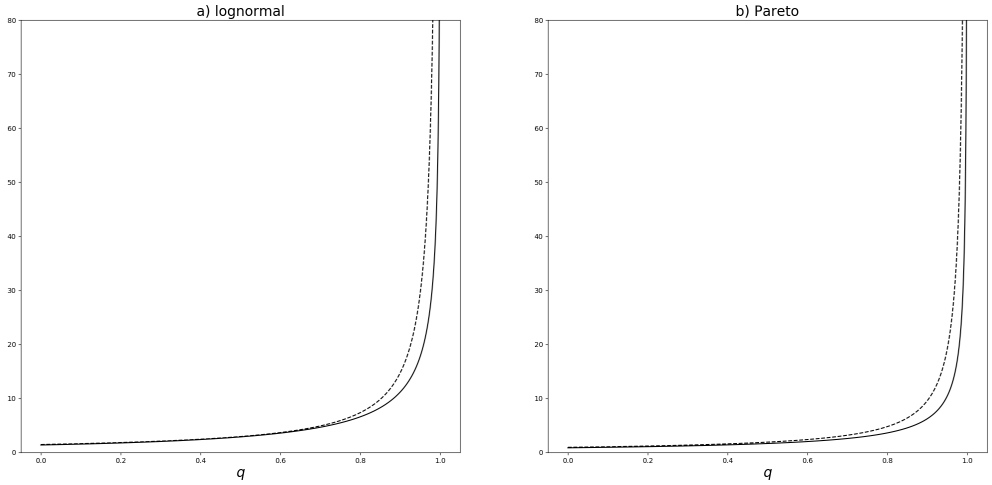

Figure 3.5. Bound on the absolute TVaR deviation (solid line) determined by the Wasserstein distance bound (dashed line). Plot a) regards two lognormal distributions, both with a parameter μ of 0, one with σ equal to 1, the other with a value of 1.5. Plot b) involves two Pareto distributions with shape 1.5 and shape 2.

In general, the behaviour of the absolute TVaR deviation curve might not correspond to the one depicted in Figure 3.5 and should be determined on a case-by-case basis, since it depends upon the characteristics of the two underlying distributions. For this reason, it is good to adopt the bound provided by the Wasserstein distance, whose shape is smooth and hinges on q.

More to the point, in this work, we are dealing with the TVaR of the sample and of the PWC distribution. They are both finite and tend, as q goes to 1, in one case to the maximum observation, and to the upper boundary of the support, in the other case. The sample TVaR is usually estimated via the so-called empirical TVaR (the reader may refer to [108] for a more comprehensive discussion on the topic) and the latter using Formula 2.17.

Lastly, it can be noted that when $q = 0$ Theorem 3.1 tells us that the absolute value of the difference of the expectations of two random variables is always less than or equal to their Wasserstein distance. A supplementary proof of this result is provided by the following lemma.

Lemma 3.3. *Given any two real random variables $X \sim F$, $Y \sim G$, such that $\mathbb{E}[X] < \infty$ and $\mathbb{E}[Y] < \infty$, it follows that*

$$|\mathbb{E}[X] - \mathbb{E}[Y]| \leq W(X, Y). \tag{3.15}$$

Proof. *By definition, in this case:*

$$W(X, Y) = \inf_{X \sim F, Y \sim G} \mathbb{E}[|X - Y|].$$

and since for any coupling of X and Y, we have that

$$|\mathbb{E}[X] - \mathbb{E}[Y]| \leq \left|\mathbb{E}[X - Y]\right| \leq \mathbb{E}[|X - Y|],$$

the inequality holds also for the infimum over the set of all couplings of X and Y. Hence, we can conclude that

$$|\mathbb{E}[X] - \mathbb{E}[Y]| \leq W(X, Y),$$

which proves the statement. An alternative yet similar proof is provided by [10] (see Theorem 4.1 pag. 10). □

3.4 Wasserstein distance hypothesis testing

The Wasserstein distance has been receiving increasing attention from the research community and has found different utilization in statistics and machine learning, starting from the work of [109] and including generative adversarial networks (GAN) [95, 110, 111], Wasserstein barycenter [112], clustering [113, 114], and principal component analysis [115]. Nevertheless, practical applications remain tentative because its numerical calculation is very arduous, especially when $d > 1$: explicit coupling results are only known for multivariate Gaussian and elliptic distributions [116]. Recent publications have proposed a wide range of approaches to find efficient solvers that address the Wasserstein distance computation problem. For a comprehensive work on computational OT topic the reader can refer to [117]. When $d = 1$, by contrast, the Wasserstein distance has the closed-form expression and is much more easy to handle.

In the algorithm we propose, for practical purposes, we need an efficient and agile scheme: the overall computation can be extremely demanding, considering both the complexity of the Wasserstein distance calculation in itself, and the fact that it may need to be determined possibly numerous times. A decisive factor, in this respect, is that we verify the marginal admissibility condition at first, and the (joint) admissibility condition only after the former is already met. Hence, during the initial phase of the algorithm, we deal with Wassertein distance between one-dimensional distributions: this aspect considerably lightens the load of the calculation, especially when the number S of partitioning hyperrectangles Q_s is still limited and large samples could be situated within these. The algorithm successively move to the hypothesis tests of Definition 3.5 regarding the joint uniformity when $\mathrm{PWC}(p, \mathcal{Q})$ is already a marginal admissible

approximation and the initial sample should be sufficiently partitioned into smaller data sets within each bucket.

When testing $\mathbb{H}_0^{*\,s,j}$ and $\mathbb{H}_0^s$ hypotheses, the distribution of the test statistic under the null hypothesis entails the so-called empirical Wasserstein distance $W(\hat{F}, F)$, i.e. the distance between the empirical measure $\hat{F}$ of a sample drawn from F and F itself, and especially the case where F is a uniform density on a hyperrectangle. It is a consequence of the strong law of large numbers that if the first moment of F is finite, then $W(\hat{F}, F)$ converges to 0, almost surely, as the sample size approaches infinity (see [65], Cor. 6.11).

Example 3.4. The Wasserstein distance between a sample drawn from the standard Gaussian and the standard Gaussian distribution has been calculated for 1000 samples of increasing size. The box plots for each group is depicted in Figure 3.6. The median is highlighted with the red dash-dotted line. The above-mentioned convergence behaviour of the empirical Wasserstein distance can be noted as the interquartile range narrows and moves closer to 0.

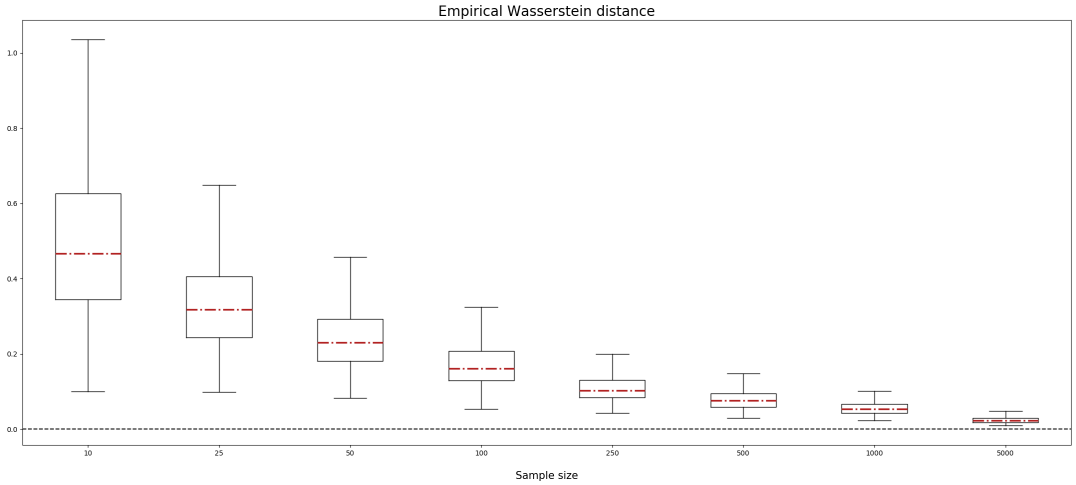

Figure 3.6. Box plots of the Wasserstein distance between the empirical distribution of a random samples drawn from the standard Gaussian and the generating distribution itself. The distance converges to 0 with the increasing of the sample size. The median of the box diagram is highlighted with the red dash-dotted line.

However, getting hypothesis tests based on the (empirical) Wasserstein distance is severely hampered by a lack of inferential tools. Determining the exact rate of convergence and distributional limits, which give a genuine perspective for practicable inference, is the subject of a large body of literature (see e.g. [118, 119, 120, 121, 122, 123]), but despite the considerable interests in the topic, results have remained elusive and the problem of constructing confidence intervals for Wasserstein distance is in general unsolved.

3.4.1 One-dimensional setting

In the one-dimensional setting, as outlined in Section 3.2, the Wasserstein distance can be expressed in a closed formulation (see Formula (3.4)). In addition, for measures on $\mathbb{R}$, a rather complete theory regarding rates of convergence and distributional limit is

available [124] and the following result applies for testing the marginal admissibility condition.

Theorem 3.2. *Consider a sequence of n independent random variables uniformly distributed on $I = (a, b]$, $a < b \in \mathbb{R}$, then as $n \to \infty$:*

$$\sqrt{n}\ \frac{W(\hat{F}, U_I)}{(b-a)} \xrightarrow{d} \int_0^1 |B(t)|dt \tag{3.16}$$

where $B(t), 0 \leq t \leq 1$ denotes a Brownian bridge process, that is, a centred Gaussian process with continuous sample paths and covariance: $\mathbb{E}(B(s)B(t)) = \min\{s,t\} - st$ [125].

Proof. *The statement follows from Theorem 1.1 in [124] by substituting the quantile function associated to U_I and applying the scaling property of the Wasserstein distance [126], i.e. $W(vX, vY) = |v|W(X, Y)$ for any scale $v \in \mathbb{R}$ and random variables X and Y.* □

Example 3.5. Figure 3.7 depicts the result stated by Theorem 3.2. For a given sample size, 50000 random samples uniformly distributed on $I = [0, 1]$ have been drawn. The simulated distribution of the test statistic has been obtained by evaluating for each sample the left side of Formula (3.16). As the sample size increases, the limiting distribution of the Wasserstein distance test statistic approaches the distribution of the integral of a Brownian bridge process absolute value. Plot a) shows, for different sample sizes, the box-and-whisker chart of the simulated test statistic distribution. Plot b) is focused on the upper quantiles, which are of interest for the hypothesis test. The limiting distribution quantiles are indicated with dashed lines.

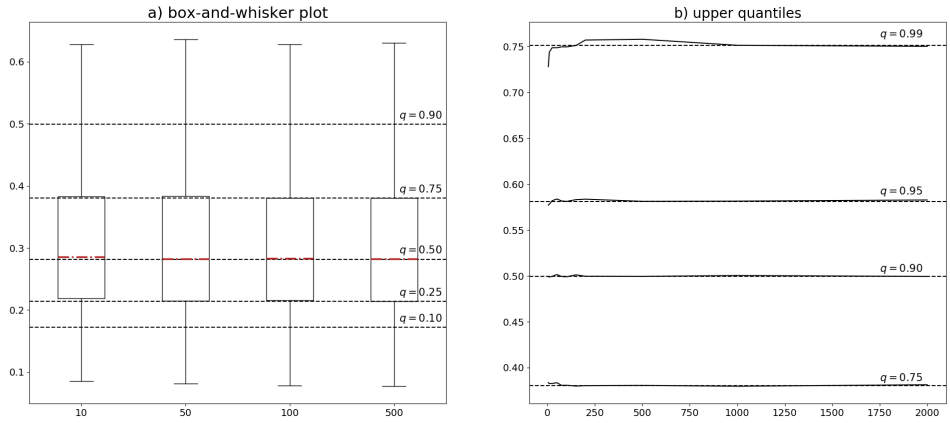

Figure 3.7. Illustration of the convergence, as the sample size increases, of the Wasserstein distance test statistic distribution. Plot a) shows the test statistic box-and-whisker plot. Plot b) focuses on the upper quantiles. In both graphs quantiles of the limiting distribution highlighted with dashed lines.

Theorem 3.2 provides us with a theoretical instrument for testing any $\mathbb{H}_0^{*\,s,j}$ hypothesis. The observed value of the test statistic, to compare with the critical one, is derived from the Wasserstein Distance between a uniform distribution on $I_{s,j}$ and

the respective sample margin. In this circumstance Formula (3.4) has the following closed-form expression.

Lemma 3.4. *Given a sample $X \sim \hat{F}$ of size n, such that $a \leq X_{(1)} \leq ... \leq X_{(n)} \leq b$, and a uniform random variable on $I = (a, b]$, then $W(\hat{F}, U_I)$ is equal to:*

$$W(\hat{F}, U_I) = \sum_{i=1}^{n} \left[\widehat{W}_i + \frac{1}{2} \left(\frac{1}{n} - \frac{2n\widehat{W}_i}{b-a} \right) \cdot \max \left\{ \frac{b-a}{2n} - n\widehat{W}_i, 0 \right\} \right], \tag{3.17}$$

where

$$\widehat{W}_i = \frac{1}{n} \Big| X_{(i)} - a - \left(\frac{i - 1/2}{n} \right) (b-a) \Big|, \tag{3.18}$$

and $X_{(i)}, i = 1, ..., n$, are the order statistics of the sample.

Equivalently, it also holds that:

$$W(\hat{F}, U_I) = \sum_{i=1}^{n+1} W_i^* = \sum_{i=1}^{n+1} \left[\overline{W}_i + \frac{1}{2} \left(\beta_i - \frac{\overline{W}_i}{\delta_i} \right) \cdot \max \left\{ \delta_i - \frac{\overline{W}_i}{\beta_i}, 0 \right\} \right]. \tag{3.19}$$

Where

$$\begin{aligned} \overline{W}_i &= \beta_i \Big| \frac{i-1}{n} - \frac{X_{(i-1/2)} - a}{b-a} \Big|, \\ \beta_i &= X_{(i)} - X_{(i-1)}, \\ \delta_i &= \frac{X_{(i-1/2)} - X_{(i-1)}}{b-a}, \end{aligned} \tag{3.20}$$

and it has been set $X_{(0)} = a$, $X_{(n+1)} = b$ and $X_{(i-1/2)} = (X_{(i)} + X_{(i-1)})/2$.

Proof. *These results can be readily derived from previous Lemma 3.2.*

□

We have therefore the analytical tools required to check the marginal admissible condition.

Algorithm 3.1. ***Marginal admissible hypothesis testing.***
For each hyperrectangle Q_s where the marginal admissibility condition is checked:

1. *For each dimension $j = 1, ..., d$, compute the observed test statistic w_j and compare it with the respective α critical value $w_{1-\alpha}$ using Theorem 3.2 and Formula* (3.19).
2. *If any $w_j \geq w_{1-\alpha}$, ${\mathbb{H}_0^*}^{s,j}$ is rejected, split Q_s and go back to (1), else the sample is considered to have uniform marginals.*

It can be noted that Formula 3.17 and Formula 3.19 represent, respectively, the specific closed-from expressions, when $G = U_I$ and $X \in I$, of the middle and the right side of the equations in Formula 3.4. In an intuitive way, the former captures the "horizontal" discrepancy and the latter measures the "vertical" distance. Figure 3.8 displays these interpretations.

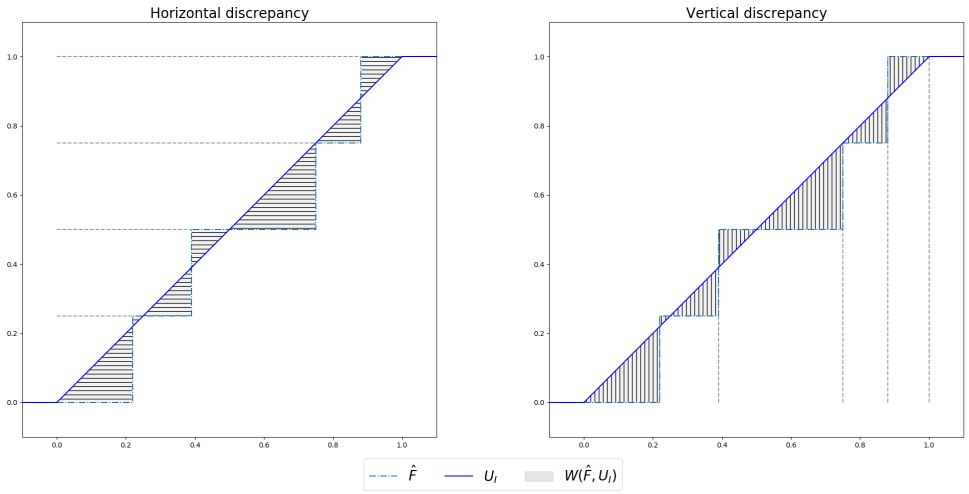

Figure 3.8. Illustration of the two notions of the Wasserstein Distance as "horizontal" (left) and "vertical" (right) discrepancy between a uniform distribution on I and the empirical distribution of the sample lying in I.

3.4.2 Multidimensional setting

When testing $\mathbb{H}_0^s$ hypotheses we are in a multidimensional regime. As mentioned above, inferential tools for Wasserstein distances are elusive when $d \geq 2$, and hence complications arise for checking the joint admissibility condition. For measures on $\mathbb{R}^d$, there are, indeed, only few distributional results, none of which can be successfully employed in our framework (see e.g. [127, 128, 129]). With regard to the question of quantifying the rate of convergence of $W(\hat{F}, F)$, the major findings regarding our context are given by the works of [130], who considers the uniform distribution on the unit square, and [131, 132, 133, 134], for the uniform distribution in higher dimensions on a d-dimensional unit cube.

The solution we propose to obtain critical values of the test statistic distribution under $\mathbb{H}_0^s$ is meant to guarantee a feasible implementation of the algorithm. It would be, in fact, possible to test the joint admissibility condition by simulating an approximated distribution of the test statistic under the null hypothesis for each hyperrectangle Q_s. However, according to the authors, this *modus operandi* comes at too high a (computational) cost, which can undermine the whole scheme workability, especially when n_s is not a small value. The idea we propose combines two components: a reference simulated distribution of the test statistic under the null hypothesis and the results of rate of convergence for empirical Wasserstein distances concerning uniform densities. The procedure steps are detailed below.

Algorithm 3.2. ***Reference test statistic critical value.***
Before the algorithm is initiated, for $\alpha \in [0, 1]$ and $d \in \mathbb{Z}^+$:

1. *Draw N samples of size m from the uniform distribution on $C = [0, 1]^d$.*
2. *Compute the Wasserstein distances $\mathcal{W} = (w_1, w_2, \ldots, w_N)$ between each sample and U_C.*
3. *Determine $w_{1-\alpha}$, namely the $(1 - \alpha)$-level empirical quantile of $\mathcal{W}$.*

Algorithm 3.3. ***Admissible hypothesis testing.***
For each hyperrectangle Q_s where the admissibility condition is checked:

1. *Compute the observed test statistic w_s and scale the value with the appropriate order of convergence of the empirical Wasserstein distance (see Lemma 3.5) for considering the actual number of data points n_s lying in Q_s.*

2. *Compare the value w thus obtained with $w_{1-\alpha}$; if $w \geq w_{1-\alpha}$, $\mathbb{H}_0^s$ is rejected and Q_s has to be split, else the sample is considered as uniform on Q_s.*

With the above steps, an approximation of the α-level critical value of the test statistic under $\mathbb{H}_0^s$ is obtained, without simulating for each hyperrectangle sample size n_s the test statistic distribution under the null hypothesis.

A heuristic rule for setting the sample size m of Algorithm 3.2 is $m = \lfloor \log n \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function and log the natural logarithm. The reason for this is that m should be an arbitrary positive integer that allows for a relatively fast computation of the simulated distribution of the test statistic. The number of simulation N is determined in such a way that there is at least 90% confidence that the estimated quantile does not differ by more than 1% from the true value (see [135] Section 5.2).

The order of convergence adjustments, previously disclosed, come from the following lemma.

Lemma 3.5. *For a uniform random vector defined on $C = [0,1]^d$, where $d \geq 2$, the limiting behaviour of the empirical Wasserstein distance is given by*

$$W(\hat{F}, U_C) = \begin{cases} \mathcal{O}(n^{-1/2} \log n^{1/2}), & \text{if } d = 2, \\ \mathcal{O}(n^{-1/d}), & \text{if } d > 2, \end{cases} \tag{3.21}$$

where $\mathcal{O}$ is the Big O(micron) notation, log *is the natural logarithm, and n is the sample size.*

Proof. *For the case $d = 2$, the proof is provided by [130]; for $d > 2$, we refer to [134].* □

The authors have confirmed through a simulation study the reliability of the above-mentioned approach. The results of this procedure for obtaining critical values of the test statistic are reported in the Appendix A.

It can be noted that, as the algorithm proceeds, the power of the hypothesis tests tends to diminish, i.e. the number of true positive correct inferences reduces. The reason for this is that the partitioning set $\mathcal{Q}$ grows in size and consequently the sample size n_s in its every element affected by the bisection technique decreases. This situation leads to a design less susceptible to overfitting, since makes it more likely for the algorithm to stop partitioning hyperrectangles.

With regard to the Wasserstein distances in Algorithm 3.2 and Algorithm 3.3, we make the computation a discrete problem: the uniform density is approximated with a quasi-random low-discrepancy sequence to evenly cover the d-dimensional hypercubic space. The number of elements in the sequence scales with d and is set equal to 10^d. As both measures involved are discrete the Sinkhorn algorithm (see [86]) has been adopted to solve the Wasserstein distance optimization problem and obtain the distance value. We adopted one of the recent refinements of this approach [86], which has a complexity of $\mathcal{O}(n^2 \log n)$, i.e. in nearly linear time in the input size n^2, for approximating the optimal transportation distance.

Example 3.6. Figure 3.9 illustrates approximation of the Wasserstein distance in $\mathbb{R}^2$, with ℓ_2-norm as ground metric, between a sample of size $n = 5$ and the uniform

distribution on $[0,1]^2$. The latter is represented with a quasi-random low-discrepancy sequence that covers the square area and is indicated with black dots. The approximated value of the Wasserstein distance is represented by the average length of the azure lines mapping the observations to the target points.

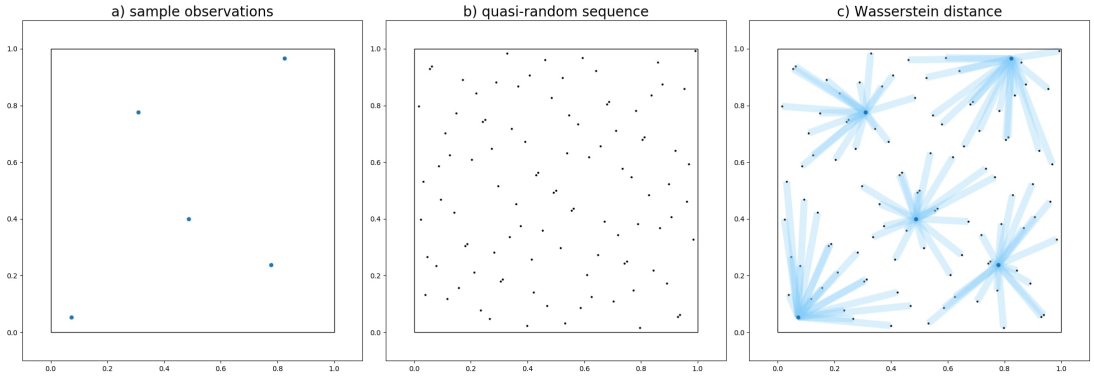

Figure 3.9. Illustration of the approximated Wasserstein distance between a sample of size $n = 5$ and the uniform distribution on the unit square, when the ground metric is the ℓ_2-norm. Plots a) and b) highlight the sample and the quasi-random low-discrepancy sequence serving as the uniform distribution, respectively. The optimal coupling associated to the (approximated) Wasserstein distance is displayed in plot c).

3.5 Wasserstein distance additional aspects

In the remaining part of the chapter additional aspects of the Wasserstein Distance are presented. In general, these are not strictly necessary for the algorithm itself, and hence they may not be adopted in the PWC estimator fitting scheme. Nevertheless, these results might be helpful in specific situations such as in the early stage of the algorithm, when the calculation of the Wasserstein distance can be computationally demanding and approximated solutions are required.

The following results provide us with some upper and lower bounds to the Wasserstein dostance. For the sake of completeness, we mention to the reader that in [136] are presented additional relationships among probability and statistical distances, including the Wasserstein metric.

Since a PWC distribution is by definition uniform within each hyperrectangle Q_s in which the domain is partitioned, we consider the possibility of estimating the Wasserstein distance between $\mathbf{X} \sim \hat{F}$ and $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ by separating its calculation for each hyperrectangle Q_s, and then aggregating all the resulting components to revert to the original overall distance. In this respect, it applies the following statement.

Theorem 3.3. *Given $\boldsymbol{X} \sim \hat{F}$ and a compatible $\boldsymbol{Y} \sim PWC(p, \mathcal{Q})$, it holds that:*

$$W(\boldsymbol{X}, \boldsymbol{Y}) \leq \sum_{s=1}^{S} W(\boldsymbol{X}_s, \boldsymbol{Y}_s) p_s, \tag{3.22}$$

where $\boldsymbol{X}_s \sim \hat{F}_s$ *and* $\boldsymbol{Y}_s \sim U_{Q_s}$ *denote the random vectors* $\boldsymbol{X}|\boldsymbol{X} \in Q_s$ *and* $\boldsymbol{Y}|\boldsymbol{Y} \in Q_s$ *respectively.*

Proof.

$$\begin{aligned} W(\boldsymbol{X}, \boldsymbol{Y}) &= \inf_{\boldsymbol{X}\sim\hat{F}, \boldsymbol{Y}\sim G} \mathbb{E}\left[\mathrm{c}(\boldsymbol{X}, \boldsymbol{Y})\right] = \inf_{\boldsymbol{X}\sim\hat{F}, \boldsymbol{Y}\sim G} \mathbb{E}\left[\mathbb{E}\left[\mathrm{c}(\boldsymbol{X}, \boldsymbol{Y})|T\right]\right] \\ &\leq \inf_{\substack{\boldsymbol{X}\sim\hat{F}, \boldsymbol{Y}\sim G \\ \boldsymbol{X}\in Q_s \implies \boldsymbol{Y}\in Q_s}} \mathbb{E}\left[\mathbb{E}\left[\mathrm{c}(\boldsymbol{X}, \boldsymbol{Y})|T\right]\right] = \sum_{s=1}^{S} \inf_{\boldsymbol{X}_s\sim\hat{F}_s, \boldsymbol{Y}_s\sim U_{Q_s}} \mathbb{E}\left[\mathrm{c}(\boldsymbol{X}_s, \boldsymbol{Y}_s)\right] p_s \\ &= \sum_{s=1}^{S} W(\boldsymbol{X}_s, \boldsymbol{Y}_s) p_s. \end{aligned}$$

Where $T = \{s : X \in Q_s\}$, $\mathbb{P}(\boldsymbol{X} \in Q_s) = \mathbb{P}(\boldsymbol{Y} \in Q_s) = p_s$ *for all* s *because of the compatibility condition, and the inequality holds because* $\{\boldsymbol{X}, \boldsymbol{Y} : X \sim \hat{F}, \boldsymbol{Y} \sim G\} \supseteq \{\boldsymbol{X}, \boldsymbol{Y} : \boldsymbol{X} \sim \hat{F}, \boldsymbol{Y} \sim G, \boldsymbol{X} \in Q_s \implies \boldsymbol{Y} \in Q_s\}$. □

The overall Wasserstein distance $W(\mathbf{X}, \mathbf{Y})$ is then obtained by simply aggregating each Wasserstein distance obtained in the partitioning rectangles. Separating the comprehensive Wasserstein distance into mutually independent components can lead to a significant relief from the computational burden. In fact, this approach allows us to:

- Deal with the "simpler" local component of the PWC distribution, i.e. uniform densities, because of conditioning to each Q_s.
- Parallelize the calculation of the overall Wasserstein distance.

Theorem 3.4 below regards Wasserstein distance when an ℓ_p-norm is used as ground cost.

Theorem 3.4. *Given two real numbers* $p, q \in [1, \infty)$*, and random vectors* $\boldsymbol{X} \sim F$ *and* $\boldsymbol{Y} \sim G$*, then*

$$p \geq q \implies W_p(\boldsymbol{X}, \boldsymbol{Y}) \leq W_q(\boldsymbol{X}, \boldsymbol{Y}), \tag{3.23}$$

where

$$W_p(\boldsymbol{X}, \boldsymbol{Y}) := \inf_{\boldsymbol{X}\sim F, \boldsymbol{Y}\sim G} \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{Y}\|_p\right], \tag{3.24}$$

and $\mathrm{c} = \|\cdot\|_p$ *denotes the* ℓ_p*-norm.*

Proof. *Let* $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ *be the optimal coupling when* $\mathrm{c} = \|\cdot\|_q$ *and hence*

$$W_q(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\|\boldsymbol{X}^* - \boldsymbol{Y}^*\|_q\right].$$

Besides, since $p \geq q$*,*

$$\mathbb{E}\left[\|\boldsymbol{X}^* - \boldsymbol{Y}^*\|_p\right] \leq \mathbb{E}\left[\|\boldsymbol{X}^* - \boldsymbol{Y}^*\|_q\right],$$

and from the definition of the Wasserstein distance, it follows that

$$W_p(\boldsymbol{X}, \boldsymbol{Y}) \leq \mathbb{E}\left[\|\boldsymbol{X}^* - \boldsymbol{Y}^*\|_p\right],$$

where the equality holds only if $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ *is the optimal coupling also when* $\mathrm{c} = \|\cdot\|_p$. □

Moreover, it is also worth noting that the following result holds between the Wasserstein distance between two random vectors and the Wasserstein distances between the relative marginals.

Theorem 3.5. *Given any two random vectors $\boldsymbol{X} \sim F$ and $\boldsymbol{Y} \sim G$ in $\mathbb{R}^d$, and ground metric $\mathrm{c} = \|\cdot\|_p^p$, namely the pth power of the ℓ_p-norm, with $p \in [1, \infty)$, then:*

$$W(\boldsymbol{X}, \boldsymbol{Y}) \geq \sum_{j=1}^{d} W(X_j, Y_j), \tag{3.25}$$

where $X_j \sim F_j$ and $Y_j \sim G_j$ are the jth marginal random variables of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively.

Proof.

$$\begin{aligned} W(\boldsymbol{X}, \boldsymbol{Y}) &= \inf_{\boldsymbol{X}\sim F,\, \boldsymbol{Y}\sim G} \mathbb{E}\left[\mathrm{c}(\boldsymbol{X}, \boldsymbol{Y})\right] \\ &= \inf_{\boldsymbol{X}\sim F,\, \boldsymbol{Y}\sim G} \mathbb{E}\left[\sum_{j=1}^{d} \mathrm{c}(X_j, Y_j)\right] = \inf_{\boldsymbol{X}\sim F,\, \boldsymbol{Y}\sim G} \sum_{j=1}^{d} \mathbb{E}\left[\mathrm{c}(X_j, Y_j)\right] \\ &\geq \sum_{j=1}^{d} \inf_{X_j\sim F_j, Y_j\sim G_j} \mathbb{E}\left[\mathrm{c}(X_j, Y_j)\right] = \sum_{j=1}^{d} W(X_j, Y_j). \end{aligned}$$

□

Theorem 3.5 extends the one in [137] (where a different proof is available, dedicated to the squared Euclidean Wasserstein distance case) and holds, in a more general context, whenever the argument of the expected value contained within the definition of the Wassestein distance can be split into marginal independent components (this is true, for example, also when the order of the Wasserstein distance is equal to the order of the p-norm used as ground metric), the reader can refer also to [138].

Algorithm

In this chapter our algorithm design and all its component are described. For obtaining an admissible PWC approximation of the sample we opted for a top-down algorithm with recursive layout that starts with a single axis-aligned hyperrectangle enclosing the entire observations, and builds a hierarchical partition by splitting an existing hyperrectangle into two non-overlapping ones. The recursive partitioning is repeated until the admissibility condition is met in each region of the partition. The initial phase is aimed at verifying the condition that sample data are uniformly distributed within each hyperrectangle. The later stage is intended to ensure that the marginal distributions of the resulting PWC distribution do not exceed the TVaR error limits.

The dimension along which the split is executed, is chosen according to a specific bisection technique. As each the bisection concerns only a single dimension, the regions in the resulting partition always have axis-parallel boundaries. In addition, since the bisection of each hyperrectangle is independent from the bisection of the other partition elements, the algorithm enables a high degree of parallelism. It should be also noted that the algorithm is compatible with the divide-and-conquer design paradigm: it works by repeatedly breaking down a problem into sub-problems of the same type, until all of these come to a halt.

In Section 4.1 the initialization step is presented. Section 4.2 follows and is about the bisection technique. The algorithm unit to ensure TVaR admissibility is detailed in Section 4.3. Finally, the full algorithm is summarised in 4.4 and some asymptotic properties are outlined in 4.5.

4.1 Initialization

The algorithm starts with a single box whose sides are parallel to the d coordinate axes and containing all the observations, and hence the piecewise constant distribution simply coincides with a uniform density.

Algorithm 4.1. ***Initialization procedure.***

1. *Set* $S = 1$, $\mathcal{Q} = \{Q_1\}, p = (1), PWC(p, \mathcal{Q}) = U_{Q_1}$. *Where*

$$Q_1 = [\hat{\mathtt{a}}_1, \hat{\mathtt{b}}_1] \times [\hat{\mathtt{a}}_2, \hat{\mathtt{b}}_2] \times ... \times [\hat{\mathtt{a}}_d, \hat{\mathtt{b}}_d], \tag{4.1}$$

and $\hat{\mathtt{a}}_j$, $\hat{\mathtt{b}}_j$, *for* $j = 1, ..., d$ *are given by*

$$\begin{aligned} \hat{\mathtt{a}}_j &= \frac{nX_{(1)} - X_{(n)}}{n-1} = X_{(1)} - \frac{X_{(n)} - X_{(1)}}{n-1}, \\ \hat{\mathtt{b}}_j &= \frac{nX_{(n)} - X_{(1)}}{n-1} = X_{(n)} + \frac{X_{(n)} - X_{(1)}}{n-1}. \end{aligned} \tag{4.2}$$

$X_{(1)}$ and $X_{(n)}$ are, respectively, the first and last order statistics of the sample jth marginal.

Formula (4.2) represents the minimum-variance unbiased estimators for the two parameters $\mathtt{a}_j$ and $\mathtt{b}_j$ of a uniform on $[\mathtt{a}_j, \mathtt{b}_j]$ [139].

Thereafter, during algorithm iterations, the partition (and the PWC distribution consequently) is grown by splitting each partition member into two sub-hyperrectangles, until the stopping condition is met and $\mathrm{PWC}(p, \mathcal{Q})$ is an admissible approximation of $\hat{F}$. The choice of using hyperrectangular shaped buckets is also driven by the data compression intent of the algorithm: this type of shape can be represented concisely, allowing a large number of buckets to be stored efficiently.

4.2 Bisection technique

The aim of the bisection technique is to build and shape the PWC distribution. The bisection scheme selects in each hyperrectangle Q_s, where either $\mathbb{H}_0^s$ null hypothesis or at least one of ${\mathbb{H}_0^*}^{s,j}$ null hypotheses is rejected, the dimension to split and the relative split point. More specifically, it operates as follows.

Algorithm 4.2. ***Bisection technique.***

1. ***Dimension selection****. In a given hyperrectangle Q_s to bisect, split the dimension in which the Wasserstein distance between $\hat{F}_{s,j}$ and $U_{I_{s,j}}$, i.e. the jth marginal distribution of the sample in Q_s and the uniform density on jth dimension range respectively, is associated to the smallest p-value. Namely:*

$$j^* = \mathrm{argmax}_{j \in \{1,2,\ldots,d\}} \sqrt{n_s}\ \frac{W(\hat{F}_{s,j}, U_{I_{s,j}})}{(b_{s,j} - a_{s,j})}. \tag{4.3}$$

2. ***Split point selection and bisection****. For the selected dimension j^*, let the index k denote:*

$$k = \mathrm{argmax}_{i \in \{1,2,\ldots,n+1\}} W_i^*, \tag{4.4}$$

where W_i^* is defined in Formula (3.19). Bisect at $(X_{(k)} + X_{(k-1)})/2$, where $X_{(k)}$ is the marginal k th-order statistic of the sample. By implication, the initial hyperrectangle Q_s is split in two non-overlapping hyperrectangles and S is augmented of one.

Both steps of the bisection technique relies on Formula (3.19). In particular, the split point selection phase uses the fact that above-mentioned formula is composed by $n + 1$ areas, each of which measures the vertical difference, occurring between two consecutive data points.

Example 4.1. Figure 4.1 exemplifies how the bisection procedure works. Each group of three plots depicts $\mathbf{X} \sim \hat{F}$ and $\mathbf{Y} \sim \mathrm{PWC}(p, \mathcal{Q})$ random variables in a given Q_s (lower-left graph), and their marginal cumulative distribution functions (upper and lower-right graphs). In the right group, W_k^* and the selected split are highlighted.

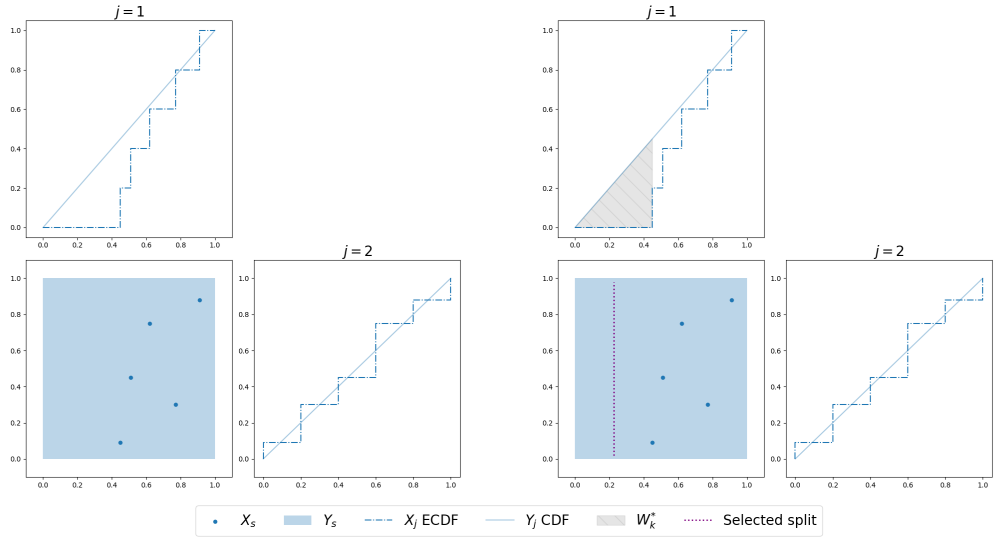

Figure 4.1. Illustration of the way in which bisection technique operates in $\mathbb{R}^2$. The algorithm is able to detect the dimension that is the most in need of partitioning ($j = 1$) and bisect the area where the largest vertical discrepancy from uniformity occurs.

Furthermore, the following bisection technique, similar to the one mentioned above, is adopted during the last stage of the algorithm, when the TVaR admissibility of the PWC distribution is tested.

Algorithm 4.3. ***TVaR admissibility bisection technique.***

1. *Select the dimension associated with the largest Wasserstein distance bound relative to its tolerance threshold parameter:*

$$j^* = \text{argmax}_{j\in\{1,2,\ldots,d\}} \frac{W(\hat{F}_j, G_j)}{(1-q)\,\epsilon_j}. \tag{4.5}$$

2. *In the dimension j^*, split the interval in which the Wasserstein distance between $\hat{F}_{s,j^*}$ and $U_{I_{s,j^*}}$ is largest:*

$$s^* = \text{argmax}_{s\in\{1,2,\ldots,S\}}\ W(\hat{F}_{s,j^*}, U_{I_{s,j^*}}). \tag{4.6}$$

3. *For the selected interval I_{s^*,j^*}, let the index k denote:*

$$k = \text{argmax}_{i\in\{1,2,\ldots,n+1\}}\, W_i^*, \tag{4.7}$$

where W_i^* is defined in Formula (3.19)*. Bisect at $(X_{(k)} + X_{(k-1)})/2$, where $X_{(k)}$ is the marginal kth-order statistic of the sample. By implication, the hyperrectangle Q_s^* is split in two non overlapping hyperrectangles and S is augmented of one.*

It can be noted that margins where Formula (3.13) holds are *de facto* excluded from the maximum search of point 1 of Algorithm 4.3. In these margins, indeed, one has that $W(\hat{F}_j, G_j)/(1-q) \leq \epsilon_j$, and hence the ratio in Formula (4.5) is less than, or equal to 1, unlike in margins where Formula (3.13) is not fulfilled.

4.3 Ensuring TVaR admissibility

After the algorithm generates an admissible PWC approximation of the original sample, the next and final stage is aimed at ensuring the TVaR admissibility of the PWC distribution. This results in verifying that for each dimension j the Wasserstein distance bound does not exceed the tolerance threshold parameter ϵ_j. To this end, we introduce two methods for selecting such threshold. In the first approach, which is referred to as the explicit approach, ϵ_j is set so that the user knows which will be, at level q, the absolute TVaR deviation value at most. In the other approach, the implicit approach, ϵ_j is set according to a statistical rule that solely considers the empirical Wasserstein distance. It consequently controls the absolute TVaR deviation value as a consequence of Theorem 3.1.

4.3.1 Explicit approach

Algorithm 4.4. *For a given sample marginal distribution $\hat{F}_j$, and $q \in [0,1)$, $\xi \in [0,1]$ set*

$$\epsilon_j = \frac{\xi \; TVaR_q(Y_j)}{(1-q)}. \tag{4.8}$$

This logic of setting the tolerance threshold parameters ensures that, in all dimensions, the absolute deviation between the q-level TVaR of the PWC estimator and of the sample is not bigger than ξ times the original sample TVaR.

By means of this pragmatic manner, the user knows that the error, measured in terms of absolute difference of TVaR, made by using the PWC estimator in place of the original sample is no more than a fixed percentage (i.e. ξ) of the original sample TVaR.

4.3.2 Implicit approach

The implicit approach for determining the tolerance threshold parameter echoes what proposed in [10]. We herein consider the empirical Wasserstein distance and the convergence of $\hat{F}_j$ to F_j in its terms. It is reiterated that what follows refers to the one-dimensional setting.

The parameter ϵ_j, for each j, is set such that the approximation error, measured as the Wasserstein distance between the sample and the PWC estimator, is significantly smaller than the expected sampling error, measured as the Wasserstein distance between the sample and its generating probability model. For further details on the following results and the statistical properties, the reader should refer to Section 4.5 of [10]. Most of the convergence theory is based on [124].

Definition 4.1. An estimator of the expected empirical Wasserstein distance between $\hat{F}_j$ and F_j, is defined as:

$$\widehat{W}(\hat{F}_j, F_j) = \sqrt{\frac{2}{n\pi}} \int_{-\infty}^{+\infty} \sqrt{\hat{F}_j(t)(1-\hat{F}_j(t))}dt. \tag{4.9}$$

Note that $\widehat{W}(\hat{F}_j, F_j)$ depends only on $\hat{F}_j$, i.e. it is independent of the (unknown) true underlying model F_j and can therefore be calculated from the sample margin. The following theorem provides us with the conditions and mathematical formulation of $\widehat{W}(\hat{F}_j, F_j)$ as a consistent estimator of $\mathbb{E}[W(\hat{F}_j, F_j)]$.

Theorem 4.1. *Suppose* $\mathbb{E}[|X|^{\xi}] < \infty$ *for* $X \sim F_j$ *and some* $\xi > 2$*. Then,*

$$\lim_{n\to\infty} |\sqrt{n}\widehat{W}(\hat{F}_j, F_j) - \sqrt{n}\mathbb{E}[W(\hat{F}_j, F_j)]| = 0. \tag{4.10}$$

Proof. *Refer to Theorem 4.12 in [124]: for the expectation we have that* $\mathbb{E}|B(F)| = 2/\pi \int F_j(1-F_j)$ *(pag. 1038); Equation (1.6) provides the limiting distribution and the variance of of* $W(\hat{F}_j, F_j)$ *around the mean in terms of a weighted Brownian motion. It can be noted that Theorem 4.12 requires* $\mathbb{E}[|X|^{\xi}] < \infty$ *for some* $\xi > 2$*; [124] also provides the convergence behaviour in the other cases.* □

We propose to use the result of Theorem 4.1 in order to set the the tolerance threshold parameter as follows.

Algorithm 4.5. *For a given sample marginal distribution* $\hat{F}_j$ *and* $q \in [0, 1)$ *set*

$$\epsilon_j = \frac{0.1}{(1-q)}\,\widehat{W}(\hat{F}_j, F_j) = \frac{0.1}{(1-q)}\,\sqrt{\frac{2}{n\pi}}\sum_{i=1}^{n-1}\sqrt{\frac{i(n-i)}{n^2}}(X_{(i+1)} - X_{(i)}), \tag{4.11}$$

where $X_{(i)}$ *is the marginal* i*th-order statistic of the sample* j*th margin.*

Using Algorithm 4.5 implies that asymptotically the error introduced through the approximation via the PWC distribution is at least one order of magnitude smaller than the sampling error. Moreover, if the marginal distributions F_j satisfies certain conditions, it is possible to establish an upper bound on the probability that the difference between these two errors is bigger than a fixed value δ. Readers who are interested in examining this question more thoroughly should refer to Section 4.5 of [10].

4.3.3 TVaR admissibility verification

Algorithm 4.6. ***TVaR admissibility testing.***

1. *Compute for each dimension* $j = 1, ..., d$*, the tolerance threshold* ϵ_j *using Algorithm 4.4 or 4.5.*
2. *Verify that* $W(\hat{F}_j, G_j)/(1-q) \leq \epsilon_j$*, for all* j*, i.e. that the Wasserstein distance bounds are less than or equal to the corresponding tolerance threshold parameters. If not, apply the bisection technique provided by Algorithm 4.3 and go back to step 1.*

It is noted that during the phase of the TVaR admissibility testing, the algorithm deals only with the Wasserstein distance, and the TVaR of the original sample and of the PWC distribution do not need to be calculated.

4.4 Full algorithm

In this section, after having presented the details of all scheme components, we resume the full algorithm.

Algorithm 4.7. ***Complete Algorithm.***
Require: $\alpha \in (0, 1)$*,* $\epsilon_j \in [0, \infty)$ *,* ℓ_p*-norm.*
Input: *An observed sample* $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *of a* d*-dimensional random vector.*
Output: *An admissible PWC distribution.*

1. ***Reference critical value:*** *Determine the test statistic α-level critical value using Algorithm 3.2.*
2. ***Initialize:*** *Start the PWC distribution as stated by Algorithm 4.1.*
3. ***First step:*** *Test the marginal admissible condition using Algorithm 3.1. Bisect $\mathcal{Q}$ elements that require it, according to Algorithm 4.2, until a marginally admissible PWC$(p, \mathcal{Q})$ is found.*
4. ***Second step:*** *Test the admissible condition using Algorithm 3.3. Further bisect $\mathcal{Q}$ elements that require it, according to Algorithm 4.2, until an admissible PWC$(p, \mathcal{Q})$ is found.*
5. ***Third step:*** *Test the TVaR admissible condition using Algorithm 4.6, until a TVaR admissible PWC$(p, \mathcal{Q})$ is found.*

It can be noted that the uniformity is not tested during the third step for the novel additional hyperrectangles. This is justified by the fact that, from a theoretical point of view, uniform distributions remain such when truncated. In other words, given a uniform density on a hyperrectangle, the two distributions generated by the bisection are still uniforms on the resulting hyperrectangles.

4.5 Asymptotic properties

In this section, we present some asymptotic properties of the estimator generated by our algorithm.

Definition 4.2. An admissible PWC distribution resulting from Algorithm 4.7 is said to be a PWC estimator of the unknown probability distribution of the observed sample.

Firstly, it can be shown that a PWC estimator tends to the empirical cumulative distribution function of a given a sample as the significance level parameter approaches 1.

Theorem 4.2. *Given an empirical distribution $\hat{F}$ of a sample of size n, for a PWC estimator G, it holds that:*

$$\lim_{\alpha \to 1} G(t) = \hat{F}(t). \tag{4.12}$$

Proof. *As $\alpha \to 1$, by the definition of significance level, the probability of rejecting the null hypothesis approaches one. Therefore, from the algorithm construction, the partition grows until only a single point x_i is left in each hyperrectangle Q_i, for $i = 1, ..., n$. At this stage, in any Q_i, the null hypothesis is still rejected and the bisection technique continues to shrink all intervals $I_{i,j}$, $j = 1, ..., d$, by alternately splitting at either $(a_{i,j} + x_j)/2$ or $(x_j + b_{i,j})/2$. As a result, the non-empty hyperrectangles are reduced to the observations.* □

It can also be noted that a PWC estimator is a consistent estimator under mild assumptions on the generating model distribution. In fact, the algorithm output converges towards the true underlying probability model if the number of training samples goes towards infinity, and the variance of the error between true and estimated values reduces to zero.

Mathematically, convergence of a nonparametric estimator like PWC estimator can be proven by showing that the following holds [140].

$$\mathbb{P}\left(\lim_{n\to\infty}\int_{\mathcal{X}}\left(g(t)-f(t)\right)^2 dt = 0\right) = 1, \tag{4.13}$$

where $f(t)$ denotes the probability density function of the unknown model generating the sample defined on $\mathcal{X} \subset \mathbb{R}^d$ and $g(t)$ is provided by Remark 2.1.

The proof is based on [40] and is reported below. Indeed, the authors introduced in their work a group of estimators that they called Density Estimation Trees, to which our PWC estimator belong.

Theorem 4.3. *A PWC estimator G satisfies* (4.13).

Proof. *Let $\mathcal{B}$ be the collection of all sets $r \in \mathcal{X}$ that can be interpreted as the solution set to a system of k inequalities of the form $b^T t \leq c$ where $b \in \mathcal{X}$ and $c \in \mathbb{R}$. Each hyperrectangle $Q \in \mathcal{Q}$ in the PWC estimator, is the solution set of a system with k inequalities of the form $b^T t \leq c$ with $b \in \mathcal{X}$ and b has exactly one entry equal to 1 and the others equal to 0. Hence, the partitions that are generated in a PWC estimator are a subset of all possible solutions to a system of inequalities in the above form, i.e. $\mathcal{Q} \subset \mathcal{B}$. According to a general version of the Glivenko-Cantelli theorem [141].*

$$\mathbb{P}\left(\lim_{n\to\infty}\sup_{r\in\mathcal{B}}|G(t)-F(t)| = 0\right) = 1 \tag{4.14}$$

This means that the probability that the largest possible error between the PWC estimated and the real distribution of any possible partition t is equal to 0, converges to 1, if the sample size n increases towards infinity.

By the definition of F and G, and Remark 2.1 we get

$$\begin{aligned} &\mathbb{P}\left(\lim_{n\to\infty}\sup_{r\in\mathcal{B}}\Big|\int_r g(t)dt - \int_r f(t)dt\Big| = 0\right) = 1 \\ \implies &\mathbb{P}\left(\lim_{n\to\infty}\sup_{r\in\mathcal{B}}\int_r\Big|g(t)dt - f(t)dt\Big| = 0\right) = 1. \end{aligned} \tag{4.15}$$

Now we make the assumption that if the sample size diverges towards infinity, the volumes of the partitioning hyperrectangles get infinitesimally small, because each leaf can only have a bounded number of points. This means in our algorithm that as $n \to \infty$, the significance level α goes to 0, in such a way as to ensure that $n\alpha \to \infty$. Hence, the following conclusion can be drawn with probability 1:

$$\begin{aligned} &\lim_{n\to\infty}\sup_{r\in\mathcal{B}}\int_r\Big|g(t)dt - f(t)dt\Big| \leq \\ &\lim_{n\to\infty}\Big|g(t') - f(t')\Big|\int_r dt \ \textit{ for some } t' \in r = 0 \end{aligned} \tag{4.16}$$

This assumption is commonly used for the consistency of data-partitioning estimators [142] and is justified since as $n \to \infty$, the diameter of any leaf node would become smaller and smaller since the leaf node can only have a bounded number of points. Therefore, it comes to what was to be demonstrated

$$\begin{aligned} &\mathbb{P}\left(\lim_{n\to\infty}\sup_{r\in\mathcal{B}}\int_r |g(t)-f(t)|dt=0\right)=1 \\ \implies &\mathbb{P}\left(\lim_{n\to\infty}\int_{\mathcal{X}}\left(g(t)-f(t)\right)^2 dt=0\right)=1. \end{aligned} \tag{4.17}$$

□

4.6 Cross-validation

The main parameter of the algorithm can be considered to be α, since it drives the Wasserstein distance based hypothesis tests, on which depends the learning of the PWC estimator.

Because of α meaning as significance level in the *reductio ad absurdum* argument within the hypothesis testing rationale, natural candidates would be 0.01, 0.05 and 0.10. However, nothing should prevent the user from following another rule for determining this parameter. For example, a different approach can be favored in consideration of the fact that α controls the degree of adaptation to the data, and overly large values could lead to overtrained model (see Theorem 4.2). Therefore, the significance level parameter can be decided according to the principle of cross-validation, which is an important statistical framework for the purposes of selecting a good estimator (i.e. model selection) and assessing its performance (i.e. estimating generalization error) [143]. The value is chosen through minimizing the unbiased estimator of the integrated squared error, often referred to as the cross-validation criterion (see [144], pag. 18):

$$\begin{aligned} \alpha^* &= \text{argmin}_{\alpha\in(0,1)}\, J(\alpha), \\ J(\alpha) &= \int_{\mathcal{X}} g(t)^2 dt - \frac{2}{n}\sum_{i=1}^{n} g_{-i}(X_i), \end{aligned} \tag{4.18}$$

where X_i is the ith observation of the sample, and g_{-i} denotes the probability density function of the PWC estimator calibrated by removing the ith observation from the sample.

4.7 Ensemble learning: bootstrap aggregating

Ensemble approaches combine multiple learning algorithms into an aggregated model with the aim to obtain better performance than could be obtained from any of the single constituent algorithms. Typically, ensemble methods construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions [145], embracing the claim, known as the "Wisdom of Crowds" [146], that "the collective knowledge of a diverse and independent body of people typically exceeds the knowledge of any single individual, and can be harnessed by voting" [147].

These techniques have been broadly studied from the theoretical viewpoint and increasingly adopted in the supervised framework. The success of ensemble learning methodologies comes from their strong performances when tested over several data sets selected from machine learning benchmarks. In contrast, fewer developments exists within the unsupervised context [148].

Similarly to what has been done for other density estimators such as histograms and kernel density estimators (see e.g. [149, 150, 151]), an aggregation technique that can be applied to our scheme is bootstrap aggregating, often called bagging [152], which has its origin in bootstrapping (see [153, 154] for a comprehensive treatise). Algorithm 4.8 provides an overview of the PWC estimator enhanced by the bootstrap aggregating ensamble method.

Algorithm 4.8. ***Bagging PWC estimator.***

Require: $B \in \mathbb{N}^+$.

Input: *An observed sample* $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ *of a* d*-dimensional random vector.*

Output: *The bagging PWC estimator is the simple pointwise average of the* B *individual estimators*

$$G_{Bag}(x) = \frac{1}{B} \sum_{b=1}^{B} G_b(x).$$

1. ***Bootstrap samples:*** *From the observed sample, obtain boostrap samples* $X_b = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_n^*)$*, for* $b = 1, \ldots, B$.
2. ***PWC estimators learning:*** *Over each boostrap sample* X_b*, calibrate a PWC estimator* G_b *by using Algorithm 4.7.*

Bagging helps both reducing the variance of the estimator (while leaving, in general, bias unchanged) and eliminating the issue of overfitting. However, despite ease of implementation and a parallel design, Algorithm 4.8 entails a considerable increase in computation time, which is not always beneficial in practise, especially given that the accuracy gain can be modest [147].

Implementation and illustrations

In this chapter, the performance of our methodology has been investigated on several examples involving data in different multi-dimensional spaces.

In Section 5.1 our algorithm has been tested in two- and three-dimensions on reference samples that have already been adopted by the literature as standards for density estimation tasks. To complete the analysis, Section 5.2 documents the results of our procedure applied on insurance-related data sets; it focuses on samples recalling realistic actuarial cases and specific situations. This part also concentrates on the application of PWC estimator as a tool for estimating the dependence contained in the data.

All experiments and analyses were run on a computer with an Intel® Core™ i7-6700HQ processor with 16GB RAM, running at 809.549 MHz, on Ubuntu Linux distribution version 18.04.2. An implementation of our algorithm in Python is available under the permissive free software MIT license. It can be obtained through the authors.

5.1 Benchmark data sets

In the first instance, as reported in [45], we evaluate our methodology against a non-trivial two-dimensional distribution. In the literature, the same bivariate distribution has been originally adopted by [155] for the testing of their work. The reader should refer to the original paper for a more detailed explanation. Moreover, we considered a group of simulated data sets, in three dimensions, with known density to evaluate the ability of the PWC estimator to recover the underlying distributions [156].

5.1.1 Two-dimensional space

The distribution is defined as a random sample from any of 350 Gaussian distributions. The first 349 normal distributions are sampled with equal probability of 1/698 and have all variances of 0.3 and no covariance. The last distribution is also a normal distribution, but it is sampled with probability 349/698 and its dispersion matrix is defined by the covariance matrix of the means of the above-mentioned 349 distributions. The first 349 normal distributions are located in a manner that reproduces the (rotated) logo of the original paper lead author's home institution. The last component is centred on the origin, with a width and height that traverse the other component distributions and with principal axes parallel to the x-y axes. This produces a mixture distribution with a long-wave feature combined with a sophisticated structure of comparatively shortwave elements aligned with different axes.

Figure 5.1 outlines the application of our algorithm to approximate a sample of size 1 million drawn from a two-dimensional mixture model. Plot a) depicts the probability

density function of the mixture distribution, plot b) illustrates the underlying sample and plot c) shows the resulting probability density function of the PWC estimator.

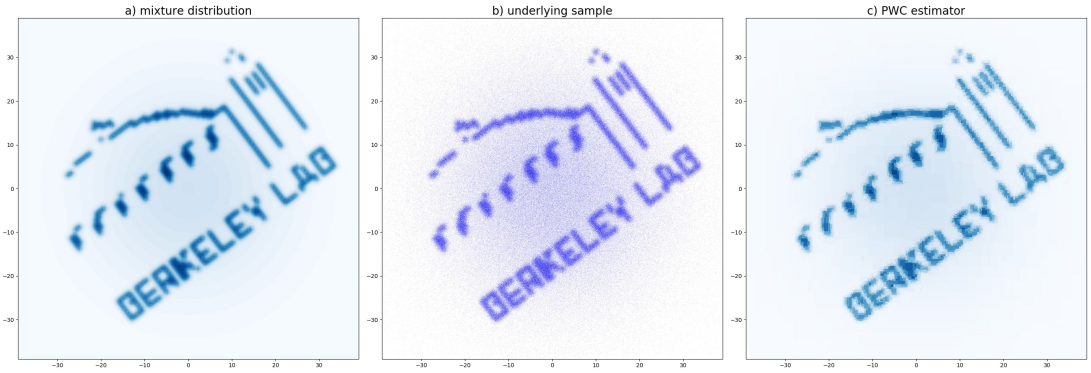

Figure 5.1. Illustration of the PWC estimator of a non-trivial mixture of Gaussian distributions in $\mathbb{R}^2$.

Figure 5.2 sets out in more detail part of the sample realizations (blue dots) and the partitioning of the domain forming the PWC estimator.

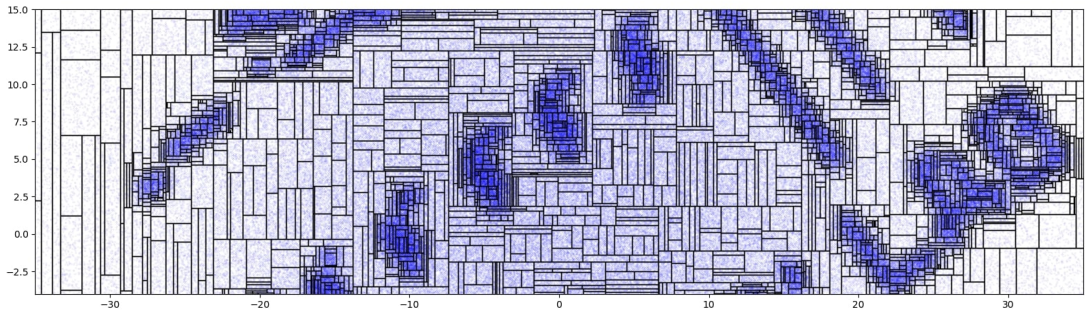

Figure 5.2. Detail of the PWC estimator partitioning rectangles dividing the domain and encapsulating the sample observations.

Table 5.1 summarizes some features regarding the PWC estimator such as the number of rectangles S partitioning the domain, the value of the parameter α and the selected ℓ_p-norm as ground distance.

A total of 7103 bisections have been performed by the algorithm on the starting trivial partition to obtain a PWC estimator that is equipped with 6880 rectangles, containing on average approximately 145 data points (ranging from a minimum of a single data point to a maximum of 675 observations). This aspect exhibits one of the PWC estimator attributes: it is efficient in terms of information needed for storing purposes.

Table 5.2 matches PWC estimator moments against their empirical ones. It can be noted that the shape of the sample distribution is preserved and the PWC estimator,

Table 5.1. PWC estimator characteristics.

PWC estimator	
Number of rectangles	6880
Number of split along $j = 1$	3594
Number of split along $j = 2$	3509
Ground metric	ℓ_1-norm
α	0.05
Execution time (in sec.)	196.31
Sample size	1000000

Table 5.2. Comparison of PWC estimator and sample statistics.

	Mean	Variance	Covariance	Skewness
Sample	(1.481, 0.403)	(238.119, 244.734)	44.856	(0.009, -0.102)
PWC estimator	(1.481, 0.402)	(237.487, 244.303)	45.404	(0.010, -0.102)

despite losing part of the information contained in the data, has characteristics very similar to the sample ones.

The impact on the PWC estimator of different values of α has been assessed on the same sample. When this parameter value increases, as stated by Lemma 4.2, the partition naturally becomes finer, and the final PWC estimator grows nearer to each single observation. Figure 5.3 shows how the PWC estimator changes by varying the significance level, in terms of execution time, number of rectangles, average number of points per rectangle, and total number of splits executed. With an increasing α, the total number of splits and the execution time escalate, the number of rectangles rises and tends to the number of observations, the average number of points decreases to the value 1.

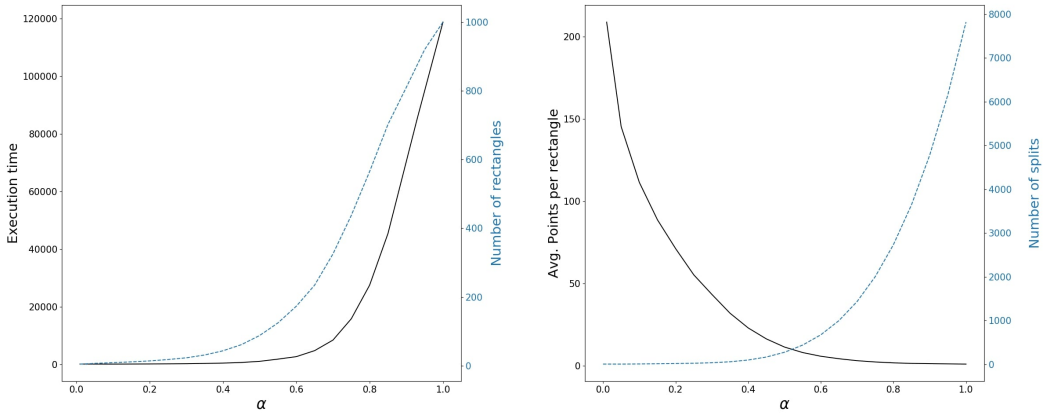

Figure 5.3. Algorithm sensitivity analysis of the variation of the significance level α. The left plot reports the execution time (in seconds) and the number of rectangles (in thousands) of the resulting PWC estimator. The right plot indicates the average numbers of data points per rectangle and the total number of splits executed (in thousands).

5.1.2 Three-dimensional space

Similarly to [156], the following four samples, each of size 600 000, have been examined.

- Data set 1 resembles a wall-like and a filament-like structure. The first and the second dimensions of the wall-like structure are both uniform on $[0, 100]$, the third dimension is drawn from a Gaussian distribution with mean 50 and variance 5. The filament-like structure, conversely, is created with a bivariate Gaussian distribution, with location (50, 50), variances 5 and 0 covariance, in the first and second coordinates, and a uniform distribution on $[0, 100]$ in the third dimension.
- Data set 2 mirrors three wall-like structures. Each wall consists of uniform distributions on $[0, 100]^2$ and a Gaussian distribution. In one wall-like structure the Gaussian has mean 10 and variance 5, in the others it has mean 50 and variance 5.
- Data set 3 is generated from a three-dimensional distribution with independent and identically distributed lognormal components, whose mean and variance are equal to 3 and 4, respectively.
- Data set 4 contains points drawn from two trivariate Gaussian distributions. Each has independent and identically distributed marginals, one with locations 25 and variances 5, the other with location 65 and variance 20. To these is added a uniform noise on $[0, 100]^3$.

The three-dimensional scatter plots of the above-mentioned samples and the relative PWC estimators are displayed in Figure 5.4.

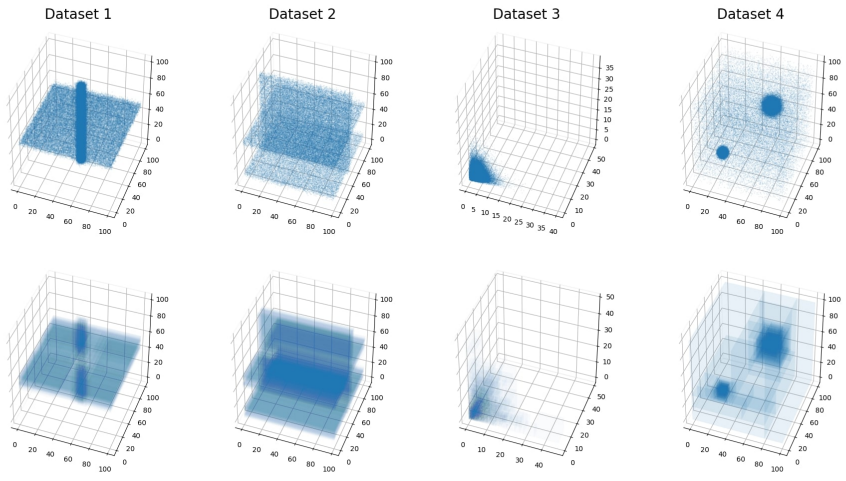

Figure 5.4. Scatter graphs of the simulated data sets 1-4 from left to right (top) and their corresponding PWC estimators (bottom).

Table 5.3 reports the characteristics of the PWC estimators for the four data sets. It can be noted that, as expected, in data sets 1 and 2 the dimensions with uniform components are affected by a lower number of bisections. In data sets 3 and 4 the number of splits are similar across the dimensions.

Table 5.3. PWC estimators characteristics for data sets 1-4.

	Data set 1	Data set 2	Data set 3	Data set 4
Number of rectangles	1793	1062	3968	2944
Number of split along $j = 1$	631	57	1373	984
Number of split along $j = 2$	583	653	1456	1040
Number of split along $j = 3$	683	379	1404	988
Ground metric	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm
α	0.05	0.05	0.05	0.05
Execution time (in sec.)	783.41	341.36	283.14	1658.95
Sample size	600000	600000	600000	600000

5.2 Insurance data sets

In the following section we investigate the performance of PWC estimator on three insurance-related examples. The first has been inspired by [157, 158] and deals with a bivariate distribution used as a model for the losses from windstorms in France and Germany. The second resumes the reinsurance policy exemplification referred to in the introduction of Chapter 1. Lastly, the third examples concerns indemnity claims data, formerly investigated by [159], comprising general liability claims payments and the relative allocated loss adjustment expenses.

5.2.1 Pareto-Clayton windstorm model

The bivariate model of losses from windstorms in France and Germany is composed of two identically distributed Pareto random variables with scale parameter 3 and shape parameter 4, coupled according to an Archimedean Clayton copula with the tail dependency on the upper side. Its parameter is set such that there is a Kendall rank correlation of $\tau = 0.50$ between the two marginal distributions [158].

The PWC estimator calibration has been run on a random sample generated by the above-mentioned distribution of size 1 million. Different values of α have been set, in order to perform a sensitivity analysis of the significance level parameter of the algorithm.

Since we know the analytical form of the underlying model, i.e. the two-dimensional meta-Clayton distribution, we are in a position to carry-out a comparison of the PWC estimator with both the sample on which it has been constructed, and the true model behind the data. Thus, this fact allows us to ascertain in detail the following points:

- To what extent is our algorithm able to represent and compress the information contained in the original sample.
- How does this result compare with the true underlying model.
- How well is our methodology capable of mirroring the joint comportment introduced by the Clayton copula and grasping the behaviour of the Pareto-distributed marginals.

Table 5.4 contains the characteristics of the PWC estimators for different values of the hypothesis test significance level. As expected, consistently with Theorem 4.2, the number of rectangles in the partition rises with the increasing α which has the effect

Table 5.4. PWC estimators characteristics at different α.

α	0.01	0.05	0.1	0.25	0.5
Number of rectangles	2478	3759	5001	10362	71374
Number of split along j = 0	1273	2032	2796	7684	107877
Number of split along j = 1	1277	1974	2765	7685	108413
Ground metric	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm
Execution time (in sec.)	185.05	220.51	234.39	310.85	1484.44
Sample size	1000000	1000000	1000000	1000000	1000000

of making the null hypotheses more likely to be rejected. This aspect also emerges from Figure 5.5, which shows the scatter plot of the sample (plot a) and the partitions of the PWC estimator with increasing α (plots b-c). Moreover, it can be noted that the number of splits is similar among the two dimensions. This is desirable since the marginal distributions are equal and the adopted copula, since it is Archimedean, is exchangeable.

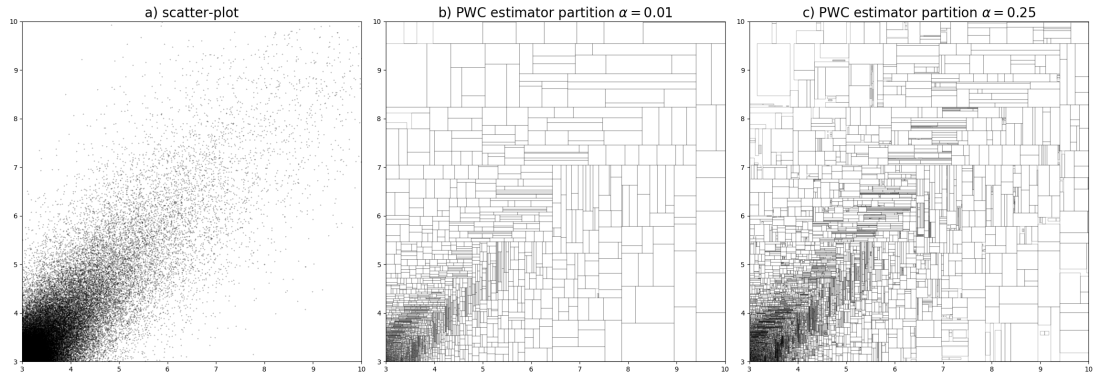

Figure 5.5. Plot a) depicts the scatter plot of the sample generated from the meta-Clayton distribution. Plots b-c) show the partition the PWC estimator is built upon: with the increase of the significance level, the partition becomes finer.

An important element of our procedure is that, through the final stage of our algorithm, it is possible to achieve an upper bound on the absolute TVaR deviation of the original sample and the PWC estimator. The tolerance threshold has been set according to Algorithm 4.5, so as to ensure that the approximation error introduced by the PWC estimator, in terms of TVaR at level $q = 0.99$, is at least one order of magnitude smaller than the expected Wasserstein distance sampling error. Figure 5.6 plots the behaviour stated by 3.1 for the PWC estimators calibrated with a different significance level parameter.

Consider now the following exercise intended to perform a sanity check of our methodology. An actuary has the task of analysing the two-dimensional sample in question, which is referred to as the original sample, and decides to estimate its density using a PWC estimator. Assume that this actuary does not know the actual model underlying the data, but s/he also believes that the marginals should be modeled using the (two-parameter) Pareto distribution (e.g. because of an established practice of the business he is dealing with). Hence, in this hypothetical situation, after the calibration,

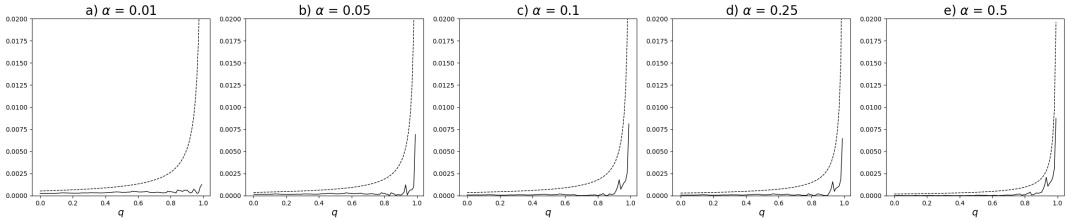

Figure 5.6. The absolute TVaR deviation (solid line) and the Wasserstein distance bound (dashed line) for the PWC estimators calibrated with different α.

the PWC estimator can be used to generate random samples, herein referred to as synthetic samples, from which, in line with the parametric assumption made, the Pareto parameters are estimated. In particular, given a sample $(X_1, X_2, \ldots, X_n)$ of n independent identically Pareto distributed random variables of scale x_m and shape θ, the maximum likelihood (ML) estimators are (see e.g. [160])

$$\hat{x}_m = \min_{i=\{1,2,\ldots,n\}} X_i, \quad \hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log\left(X_i/\hat{x}_m\right)}. \tag{5.1}$$

The ML method of estimating parameters, as well as being one of the most common approach, is proven to be adequate for the Pareto distribution when the sample sizes are not too small [161]. This helps us to examine the estimation variability of the Pareto distribution fitted on the synthetic samples, and to estimate the quality of approximation made by using the PWC estimator instead of the original sample. To that end, we simulated 5000 synthetic samples of the same size as the original sample from PWC estimators with different significance level α, and then, by means of Formula (5.1), we computed the estimates of the Pareto parameters. Figure 5.7 depicts the box-and-whisker diagrams of θ parameter estimates constructed around the above-mentioned synthetic samples $\hat{\theta}_{ss}$, the shape parameter estimates deriving from the original sample $\hat{\theta}_{os}$, and the true value are also highlighted for comparison purposes.

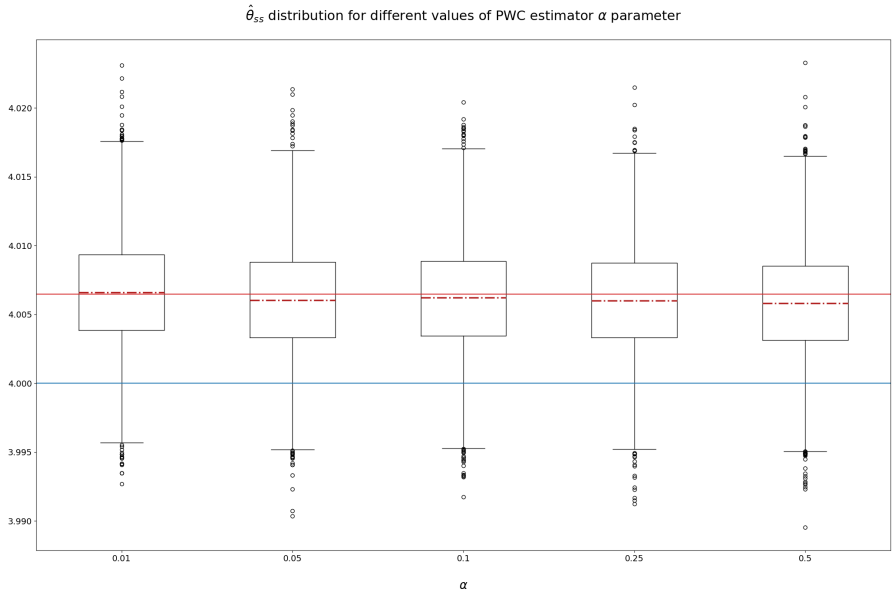

Figure 5.7. Box-and-whisker plots of $\hat{\theta}_{ss}$ for different PWC estimator significance levels. The dash-dotted lines within the box denotes the median. The red and blue solid lines indicate, respectively, the parameter value $\hat{\theta}_{os}$ resulting from the original sample and its true value θ.

The reach of the whiskers is determined such that the upper whisker extends to last datum less than the third quartile + 1.5 IQR, where IQR is the interquartile range. Similarly, the lower whisker spreads to the first datum greater than the first quartile - 1.5 IQR. This apply also to the forthcoming graphs.

It can be seen that $\hat{\theta}_{os}$ lies close to the median of $\hat{\theta}_{ss}$ distribution. This result shows that the model estimated using solely PWC estimator information is very similar to the one that we would have achieved by using the original sample. Furthermore, the lower whisker of the diagram attains the true value of θ in all cases.

As it typically happens in actuarial practices, the actuary is also concerned to retrieve risk measures and "extreme" scenarios related to the shape of the tail of distributions. It is therefore interesting to investigate how the TVaR, estimated on the basis of synthetic data generated from the PWC estimator, behaves when balanced against its analogue estimated on the original sample and the true one. In our case, for a Pareto distributed random variable, the TVaR_q can be expressed by the following formula [162]:

$$\mathrm{TVaR}_q = \frac{x_m \theta}{(1-q)^{1/\theta}(\theta - 1)}. \tag{5.2}$$

Figure 5.8 shows the distributions of the estimated TVaR at level $q = 0.99$ in the form of box-and-whisker plot. Each TVaR realization is calculated from a synthetic sample observation on the parametric Pareto assumption.

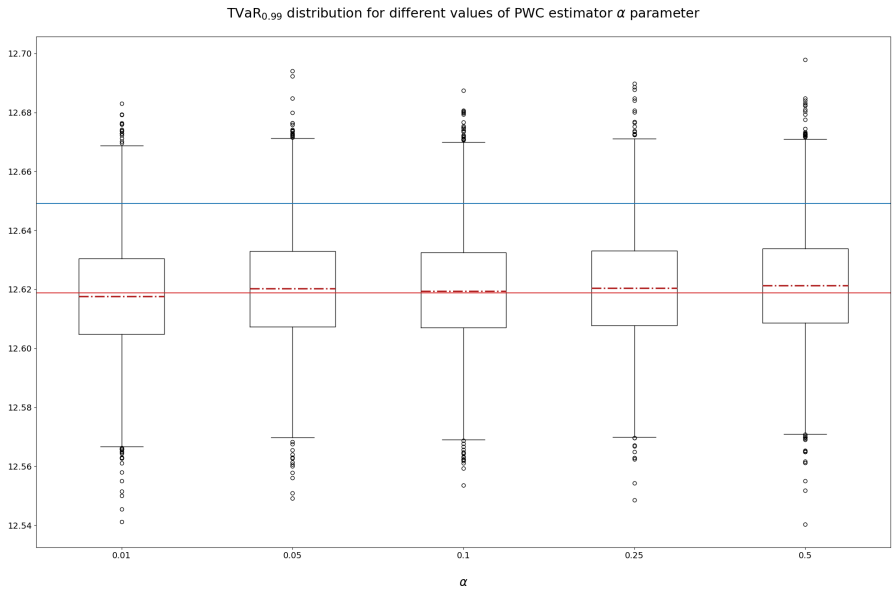

Figure 5.8. Box-and-whisker plot of $\mathrm{TVaR}_{0.99}$, under the Pareto assumption, estimated on samples drawn from PWC estimators. The dash-dotted lines within the box denotes the median. The red and blue solid lines indicate the TVaR resulting from the original sample and its true value respectively.

Moreover, Figure 5.9 provides us with further details about the TVaR behaviour from a different perspective. The chart shows the box-and-whisker plots of the estimated

TVaR at levels 0.90, 0.95, 0.99, 0.995, keeping fixed the PWC estimator significance level parameter at $\alpha = 0.05$.

The findings of the analysis confirm that our algorithm is able to replicate the information contained in the original sample, with the benefit of a more succinct representation. It has been shown that, when the user decides to adopt a specific parametric model for describing the marginal distribution, the parameter estimates obtained by a synthetic sample drawn from the PWC estimator are analogous with the parameter estimates determined from the original sample.

Finally, to complement the sanity test, we look at the dependence structure embedded within the PWC estimator as against that of the Clayton copula generating the data. Indeed, since the one-dimensional marginal distributions of the PWC estimator are one-dimensional PWC distributions (see Section 2.2 of Chapter 2) and their distribution function can be represented by Formula (2.12), it is possible to separate the implicit copula out of the PWC estimator distribution function, according to the renowned Sklar's theorem [163]. Figure 5.10 illustrates the (rotated) Clayton copula embedded in the two-dimensional model, and compares it with the implicit copulas derived from the PWC estimators. It can be noticed that these latter accurately replicate the original association implemented in the data.

The outcome of the above analysis evidences that our methodology succeeds in capturing the dependence introduced by the Clayton copula, and it shows that the PWC estimator provides a good approximation of the empirical association between margins observed in the data.

On top of that, given that it is possible to decompose a PWC estimator into its marginals and its coupling function, our approach can also be utilized to estimate multivariate copulas. Therefore, PWC estimators can bring improvements to situations in which the dependence is modeled using approaches that do not allow straightforward formulations (e.g. when the dependence structure is introduced via explicit models like common factors/shocks, or complex explicit models) and only simulated samples can be generated from it.

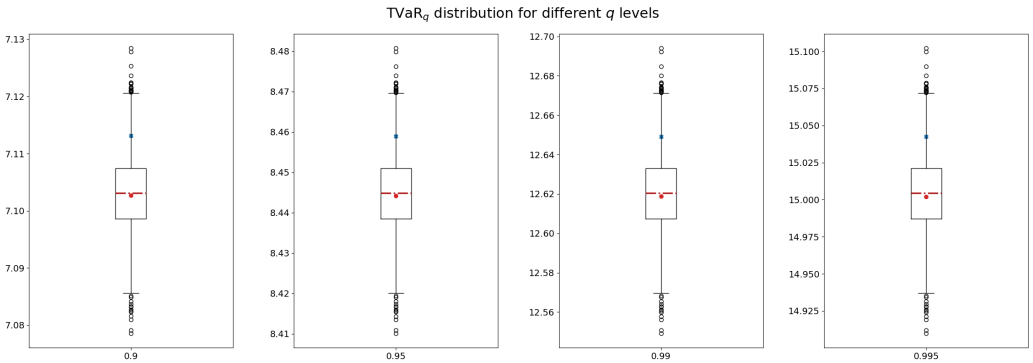

Figure 5.9. Box-and-whisker plots of TVaR$_q$, under the Pareto assumption, estimated on samples drawn from the PWC estimator with $\alpha = 0.05$. The dash-dotted lines within the box denotes the median. The red dot indicates the TVaR resulting from the original sample, the blue cross specifies the true TVaR value.

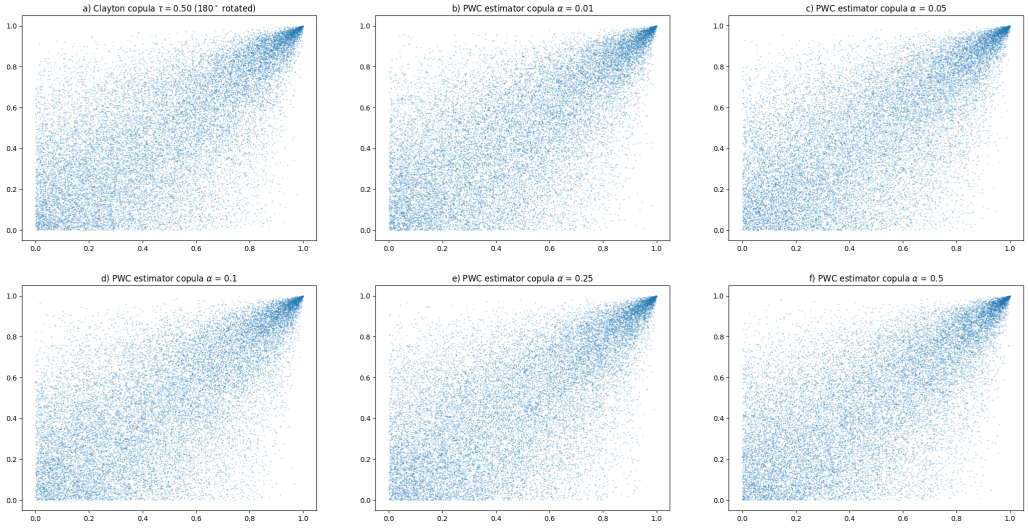

Figure 5.10. Scatter plots generated from the the true underling Clayton copula (plot a) and from the PWC estimator implicit copulas with different values of α (plots b-f).

5.2.2 Multi-lines reinsurance program

Let's consider the reinsurance program presented in Section 1.1 of Chapter 1. Our algorithm can be applied to the simulated empirical distribution of the layers losses with the objective of approximating and "compressing" the relevant sample information.

It should be recalled that, as opposed to the previous example, it is not possible to express neither the joint distribution nor the marginal behaviour of the layer losses because of the policy modifications. In addition, it is not possible either to disclose the closed-form expressions of the moments of the distributions, such as the expected value, which lies at the heart of any pricing evaluation, due to the aggregate condition, and the drop-down and stretch down clauses.

Recalling what was said at the beginning of Chapter 1, the policy covers multiple lines of liability business: employers' liability (EL), public liability (PL) and third-party motor liability (MTPL). The structure includes three XS layers, one for each risk, on top of a shared XS layer, whose losses are retained by the cedant insurance, with an aggregate limit condition. In particular, the reinsurance program we are considering has the following aspects:

- Primary layer: retained by the cedant's captive insurance. It is a €250'000 Each and every loss layer with an aggregate limit of €2 million. The policy provides a maintenance deductible of €50'000.
- First layer EL: €10 million each and every loss limit in excess of €250'000 with a "Stretch down" clause.
- First layer PL: €10 million each and every loss limit in excess of €250'000 with a "Drop down" clause.
- First layer MTPL: €5 million each and every loss limit in excess of €250'000 with a "Drop down" clause.

The three loss models describing the underlying risks have been assumed to have all lognormal distributed severity, whereas the frequency is modeled as a Poisson random

variable for both EL and PL, and as a Polya random variable for MTPL. Model parameters are summarized in Table 5.5.

Table 5.5. Loss models parameters.

	Frequency	Severity
EL	Poisson($\lambda = 20$)	Lognormal($\mu = 8, \sigma = 1.5$)
PL	Poisson($\lambda = 10$)	Lognormal($\mu = 9.1, \sigma = 1.2$)
MTPL	Polya($r = 50$, $\beta = 1.5$)	Lognormal($\mu = 8, \sigma = 1.4$)

Although the loss models of the three liability risks might be initially regarded as mutually independent, a dependency follows implicitly from the policy layer structures, more specifically from the €2 million aggregate limit and from the stretch-down and drop-down clauses. Additionally, dependency is induced by the fact that we have introduced a common shock scenario to the claim frequency. This foresees, with a 10% probability, the rise by 10% of the number of claims in the three lines of business, on account of an economic recession.

In order to obtain the loss distribution of the layers and its characteristics, since this cannot be obtained using a parametric model or a closed-form expression, a Monte Carlo simulation experiment has been developed and deployed. In this regard, we adopted the empirical distribution of such simulated sample generated from the stochastic model, and then applied our algorithm. Figure 5.11 supplies a visualization of the four-dimensional resulting PWC estimator by showing all its six two-dimensional marginal distributions.

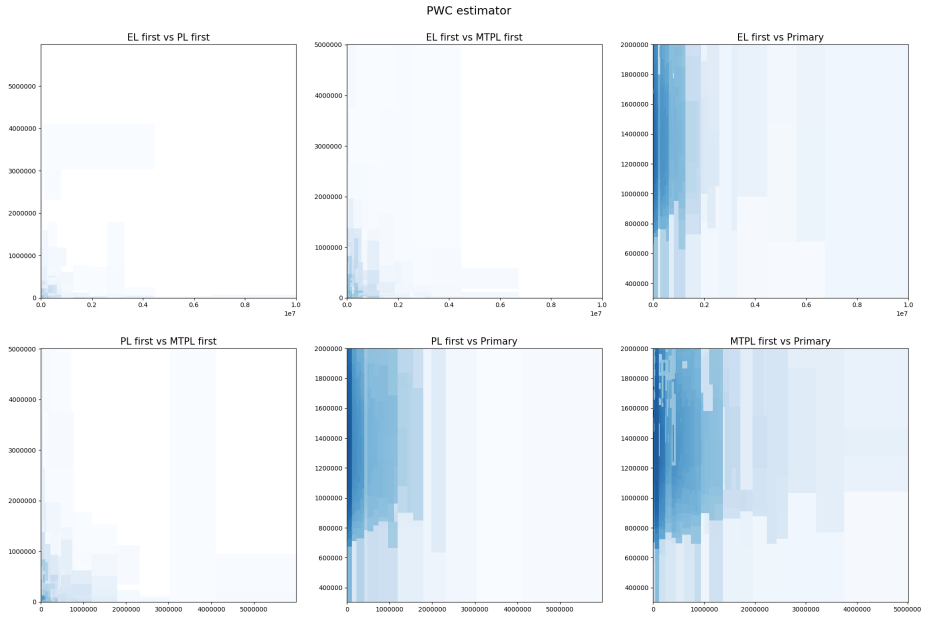

Figure 5.11. Two-dimensional marginal distributions of the (four-dimensional) PWC estimator

In more detail, Table 5.6 summarizes some aspects of the resulting PWC estimator. According to our technique, the original sample of 2.5 million observations can be effectively represented through 2563 four-dimensional boxes, wherein data are judged to

be uniformly scattered. In addition, Table 5.7 highlights the number of splits occurred in each dimension. The largest number of splits have been executed in Primary, the dimension corresponding to the primary xs layer. This results from the fact that basically all loss process realizations are relevant to the structure of the primary layer, as opposed to the other layers on top of it, which are affected only by a small portion (the tail) of the losses.

Table 5.6. PWC estimator characteristics.

PWC estimator	
Number of rectangles	2563
Ground metric	ℓ_1-norm
α	0.05
Execution time (in sec.)	242.98
Sample size	2500000

Table 5.7. Number of splits per dimension.

Dimension	EL first	PL first	MTPL first	Primary
Split	398	337	880	2141

We can see that a total of 3756 bisection have been executed during the algorithm process run. Therefore, the final diagram tree associated to the PWC estimator is capable of summarizing in a sequence of 3756 bisections the original sample of a much larger size.

Furthermore, as shown in Table 5.8, which compares PWC estimator moments and statistics with their empirical counterparts, the PWC estimator is able, for all four layers, to capture the information contained in the one-dimensional margins of the data. Both for the means, standard deviations, coefficients of variation and the skewness, results are similar and close to their associated values.

Focusing on the final stage of our algorithm, aimed at ensuring the TVaR admissibility of the PWC distribution, the tolerance threshold ϵ has been set according to the explicit approach (Algorithm 4.4), such that, the error introduced by the PWC estimator, in terms of TVaR at level $q = 0.99$, is at most equal to 1% of the TVaR measured on the original sample, the so-called empirical TVaR. Figure 5.12 details the resulting absolute TVaR deviation and the Wasserstein distance bound.

At last, it is worth investigating the extent to which our methodology is capable of catching the dependence ingrained in the sample. Figure 5.13, illustrates the dependence structure of the PWC estimator dimensions, as expressed by the pairwise (Pearson) correlations, and confronts it with the corresponding ones observed in the sample. Evident is that the PWC estimator, besides approximating the sample marginal behaviour, is also capable of recognizing the relationships intervening between the dimensions of the data set. In addition, the ability to recognize correlation patterns within the sample dimensions is corroborated by the result of the comparison of the ordinal association. Figure 5.14 shows the Kendall and the Spearman rank correlation matrices, checking the empirical values against the PWC estimator ones. It can be noted that, for both measures, the algorithm preserves and does not impair the rankings in the dimensions.

Table 5.8. Comparison of PWC estimator and sample statistics.

	EL first	PL first	MTPL first	Primary
			Mean	
Sample	5663.19	3799.69	11764.19	1163196.50
PWC estimator	5704.19	3829.51	11823.48	1162991.43
			Standard Deviation	
Sample	59624.04	38557.06	74973.35	262436.99
PWC estimator	59918.27	38664.68	75330.28	261036.98
			Coefficient of Variation	
Sample	10.53	10.15	6.37	0.23
PWC estimator	10.50	10.10	6.37	0.22
			Skewness	
Sample	33.47	24.44	16.29	0.44
PWC estimator	33.30	24.30	16.51	0.59

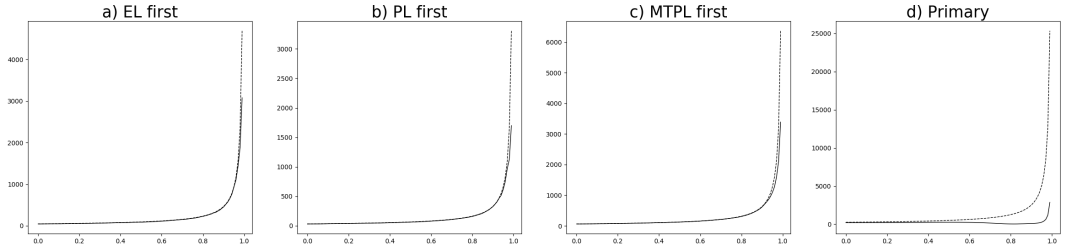

Figure 5.12. The absolute TVaR deviation (solid line) and the Wasserstein distance bound (dashed line) for each layer.

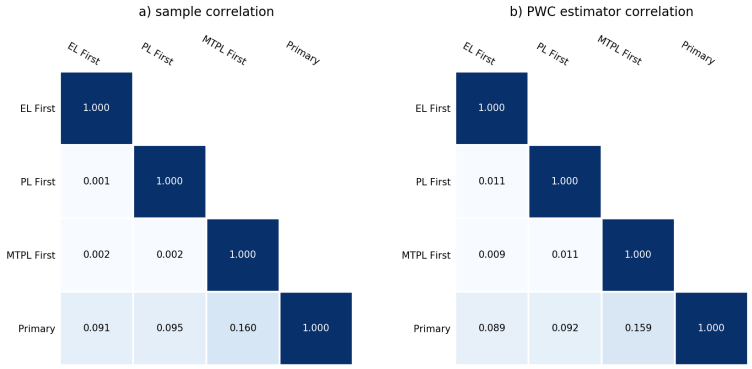

Figure 5.13. Comparison of the dependence structure: plot a) reports the entries of the (Pearson) correlation matrix measured on the sample, whereas plot b) indicates the same quantities calculated for the PWC estimator.

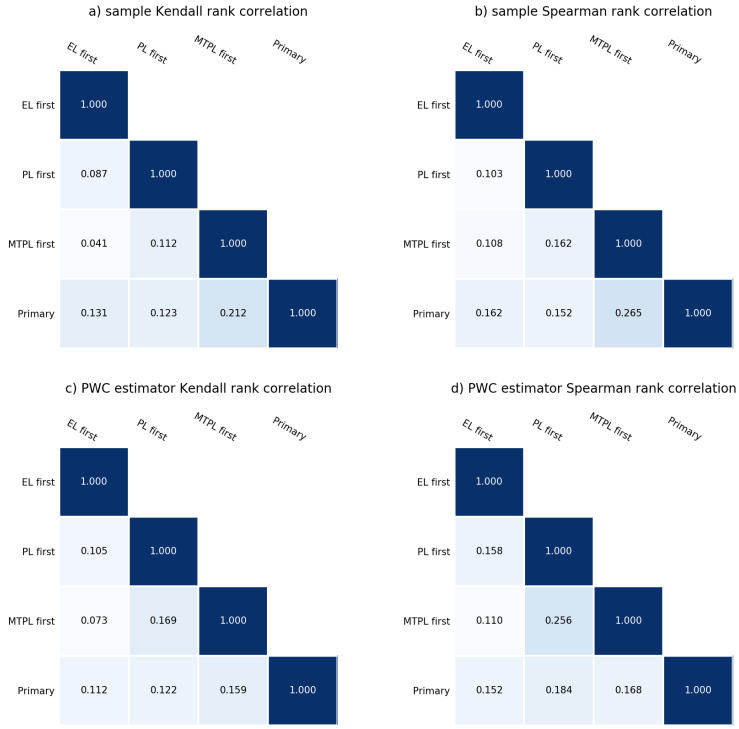

Figure 5.14. Comparison of the dependence structure: plots a) and b) report the entries of the Kendall and the Spearman rank correlation matrices measured on the sample. Plot c) and d) indicates the same quantities calculated for the PWC estimator.

5.2.3 Loss and ALAE

The following analysis illustrates the application of our algorithm to indemnity claims data. The data set is available at [164] and has been originally investigated by [159]. As described in these works, the data comprise 1500 general liability claims from late settlement lags and were provided by Insurance Services Office, Inc. (frequently referred to as ISO). Each claim consists of an indemnity payment, indicated as the loss, and an allocated loss adjustment expense (ALAE). ALAE are categories of insurance expenses that are specifically attributable to the settlement of individual claims (e.g. lawyers' fees and claims investigation expenses).

The objective is to describe the joint distribution of losses and expenses using the PWC estimator. The data set, as its size amounts to only 1500 observations, would not require *per se* the application of such a procedure with the characteristics and purposes of our algorithm. However, performing this exercise is useful to see the result of the PWC estimator calibration on a smaller sample, and compare our output with the findings of [159].

It can be noted that the estimation of the joint distribution of losses and expenses is complicated, from a parametric point of view, by the presence of censoring (see e.g. [165]). Specifically, in addition to loss and expense information, for each claim, the policy limit and the maximal claim amount are recorded. With the presence of a limit in the policy, the loss variable is censored because the amount of claim cannot exceed the stated cap. No policy limit was assumed for claims coming from policies where the policy limit was not available. Such elements do not pose a particular problem to the

PWC estimator.

Table 5.9 summarizes the main characteristics regarding the PWC estimator fitted with three different values of the parameter α.

Table 5.9. PWC estimators characteristics at different α.

α	0.01	0.05	0.1
Number of rectangles	39	46	55
Number of split along loss	19	21	33
Number of split along ALAE	28	36	56
Ground metric	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm
Execution time (in sec.)	3.78	5.16	16.78
Sample size	1500	1500	1500

The partitions of the final PWC estimators count 39, 46 and 55 (non-empty) rectangles for significance level α equal to 0.01, 0.05 and 0.1 respectively. Each rectangle contains on average 38, 32 and 27 data points, approximately.s This aspect exhibits that also with smaller inputs the PWC estimator is capable of efficiently approximating the original data set and provide a more parsimonious representation.

Table 5.10 matches PWC estimator mean, standard deviation (SD), coefficient of variation (CV) and skewness against their empirical counterparts. It can be argued that the shape of the sample distribution is preserved and the PWC estimator has characteristics very similar to the sample, although it necessarily drops part of the information contained in the data.

Table 5.10. Comparison of PWC estimator and sample statistics.

	Mean	SD	CV	Skewness
Sample	(41208.42, 12588.16)	(102713.46, 28136.26)	(2.49, 2.24)	(9.15, 9.24)
$\alpha = 0.01$	(44512.43, 13632.95)	(108024.03, 29768.55)	(2.43, 2.18)	(8.78, 9.31)
$\alpha = 0.05$	(43787.90, 13265.79)	(103002.29, 28971.82)	(2.35, 2.18)	(8.53, 9.43)
$\alpha = 0.1$	(43019.04, 13187.12)	(100557.51, 28944.82)	(2.34, 2.19)	(8.55, 9.45)

Moreover, Figure 5.15 illustrates the comparison of the cumulative distribution function of the PWC estimator marginals with the ecdf of the sample and the cumulative distribution function of the models fitted by [159]. In this work, the authors fitted the indemnity claims marginals with Pareto distributions. The parameter estimates obtained are $\theta = 1.122$ and $x_m = 14.036$ for the loss component, and $\theta = 2.118$ and $x_m = 14.219$ for ALAE. In Figure 5.15, Losses are displayed in the chart on the left side of the plot; ALAE are shown on the right side. The two bottom charts zoom in on a part of the probability distribution curves. It can be noted that, for both dimensions, the three different models, i.e. the PWC estimator margin, the Pareto, and the empirical distribution functions, lie close to each other and have a very similar behaviour, also considering the distinct and different aspects of each single approach.

In the final stage of our algorithm aimed at ensuring the TVaR admissibility of the PWC distribution, the tolerance threshold has been set using Algorithm 4.4, i.e. the explicit approach, so that the discrepancy between the TVaR, at level $q = 0.99$, of the PWC estimator and the sample is at most equal to 0.1% of the original sample TVaR. Figure 5.16 depicts the absolute TVaR deviation and the Wasserstein distance bound.

Finally, we investigate the copulas extracted from the PWC estimators and, similarly to what has previously done, we check them against both the empirical copula and the copulas fitted in [159]. The authors of this work selected a Gumbel-Hougaard's copula with parameter 1.453 and a Frank copula with parameter 3.158. Figure 5.17 shows the scatter plots of the copula models. The bottom charts refer to the implicit PWC estimator copulas. Each plot considers 1500 points to be consistent with the sample size on which is based the empirical copula. Axes of abscissa represent loss quantiles, whereas ordinates report ALAE quantiles. It can be noted that the PWC estimator captures the dependence behavior appearing in the data set between loss and ALAE in a manner similar to what the Archimedean copula does.

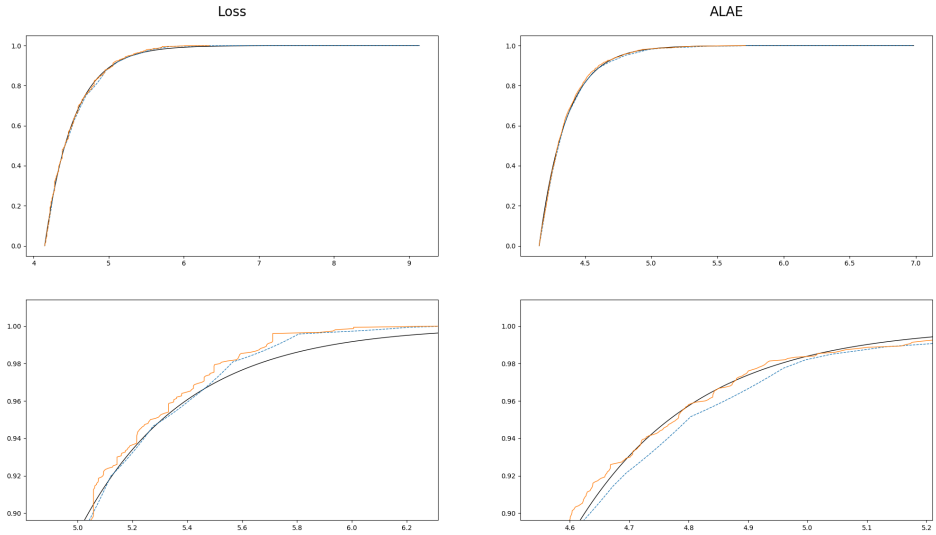

Figure 5.15. Comparison, on a marginal level, of the PWC estimator (blue dashed line) with the sample (orange solid line) and the Pareto models (black solid line). The top charts show the cdf and the bottom charts give a more detailed focus on a part of it. A base-10 log scale is used for all x-axes.

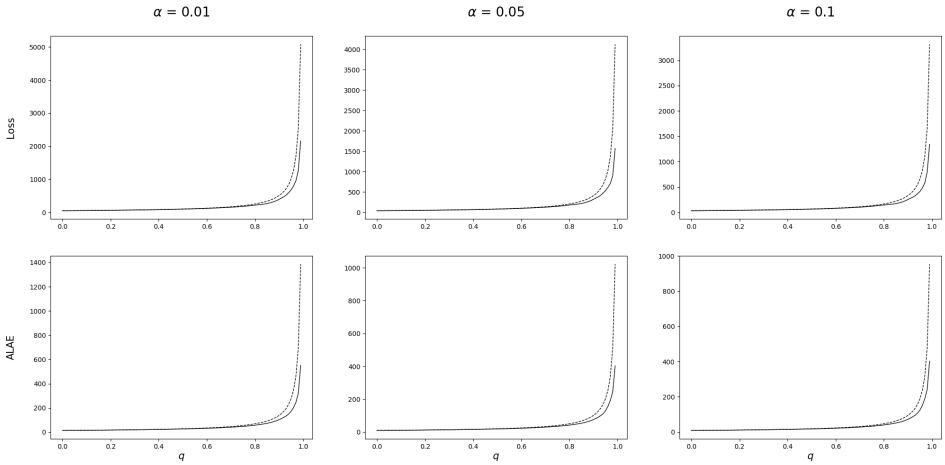

Figure 5.16. The absolute TVaR deviation (solid line) and the Wasserstein distance bound (dashed line) for loss and ALAE at different α.

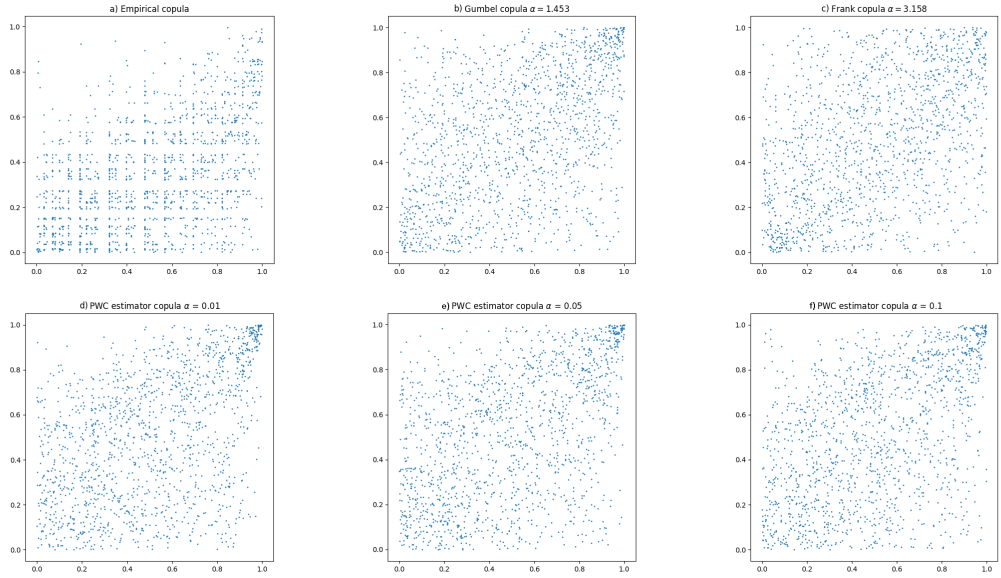

Figure 5.17. Scatter plots of copula models derived from the data set. Loss quantiles are reported on the x-axis, ALAE quantiles on the y-axis. Plot a) shows the empirical copula of the sample, plots b) and c) display the Gumbel and the Frank copulas with parameter 1.453 and 3.158 respectively. Plots d-f) illustrate PWC estimator implicit copulas with different values of α.

Conclusion and discussion

The work moves away from more traditional actuarial research topics such as Property & Casualty General Insurance, Pension, Mortality and Longevity, Risk Management and Predictive Analytics & Modeling [166] and lies between the areas of expertise of actuarial science, computational & data science, and statistics. The purpose is to introduce a nonparametric density estimation algorithm that finds application in all those situations where multivariate problems arise, and that can be adopted within the insurance context, which has recently undergone major changes, as a result of constant technological breakthroughs, business developments and financial regulatory framework modifications. As compared with other long-established methods for estimating the density generating a sample, the methodology presented hereby is focused on some of the latest scientific challenges and factual problems encountered by actuaries. The density estimation algorithm is specially intended to:

- Provide an efficient scheme that works in case of large data sets and considers more flexible distributions to model empirical data without being constrained within the set of parametric models.
- Compress the original sample, i.e. to be able to concisely represent the information contained in the data (without losing too much of this information).

These aspects are emerging in today's actuarial science environment characterized by an increasing complexity, interconnection and availability of data, all of which require having procedures designed for situations where the information contained in large-sized multivariate samples needs to be preserved, conveyed or analysed.

Additionally, our algorithm can be linked to dependence analysis and integrated with the application of copula methodology (see e.g. [167]), since it is possible to separate out of the resulting PWC estimator distribution function its implicit copula. In this respect, from an actuarial point of view, the major advantage of our procedure is that it can be embedded within the standard approach of analyzing multivariate phenomena, whereby the dependence structure and the one-dimensional marginal behaviours are modeled separately, by using (although not necessarily) parametric models, which are often assumed by practitioners for a matter of both convention, regulatory constrains, and historical evidence, to be then unified in a second phase. The PWC estimator can comply with the common actuarial practice inasmuch as the user who decides to adopt specific parametric models for describing the margins and the dependence structure, might:

a) Calibrate the PWC estimator and derive its implicit copula.

b) Find the parametric copula which best approximate the resulting implicit copula of the PWC estimator.

c) Use the parametric marginals in combination with the selected parametric copula to jointly represent the data.

Figure 6.1 schematizes the above-mentioned plan, which allows us to reconcile the PWC estimator based proposal with the standard approach of modelling dependence.

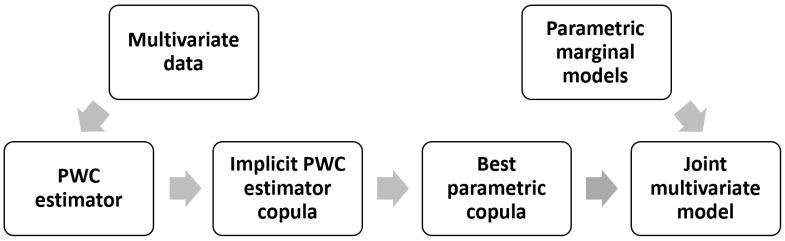

Figure 6.1. Scheme to consolidate the PWC estimator with the standard application of copula methodology and parametric marginals to model multivariate probabilistic phenomena.

Further to the above, the algorithm also ensures a similarity at the marginal level, as it enforces a bound on the error (measured as absolute deviation) between the estimator and original sample marginal distributions, in terms of tail value at risk.

6.1 Discussion

From a technical and statistical point of view, as opposed to other piecewise constant density estimators with a tree structure (refer to e.g. [40, 168, 43]), our methodology centres on the idea of assessing uniformity within each partition element using a hypothesis test based on the Wasserstein distance. As a result, the learning of the tree is implicitly defined by the hypothesis test, and its significance level establishes the stopping rule controlling the partition growth process. On top of that, our algorithm is not prone to overfit the data, because, as the cardinality of the partition grows, the power of the hypothesis test in each new partition element, resulting from the bisection, decreases with lower sample size. Hence, in the calibration phase, there is no arbitrary threshold to select or error function to minimize, and it is not necessary to consider any techniques to penalize complexity and lessen the chance of overfitting, such as regularization, cross-validation, early stopping and pruning.

In addition, our Wasserstein distance based hypothesis test for assessing uniformity provides also a comprehensive method, in sense that it works for any dimension and for any type of distribution. Other candidates that would fit in a scheme similar to ours are the Kolmogorov-Smirnov (K-S) and Pearson's chi-squared (χ^2) goodness-of-fit tests. The former, however, poses non-trivial obstacles in more than one dimension, and there is currently no single approach which is universally applicable (see e.g. [169]). The χ^2 test, although it may theoretically be applied for testing any multivariate distribution, is sensitive to the set of non-overlapping bins chosen to reduce the observations to a set of counts. The greater the number of bins, the more accurately the local data behaviour will be quantified. Nevertheless, the test may give invalid results if not all expected frequencies are sufficiently large (greater than 5 is the usual rule of thumb) and, as a result, the test cannot be trusted in high dimensions [170]. On the other hand, when using a small number of bins, the test loses power. This, in our algorithm, would

increase the chance of having hyperrectangles where data are not adequately uniform. The authors have compared the empirical rejection rates of the Wasserstein distance based hypothesis test with K-S and χ^2 goodness-of-fit tests. Results are displayed in Appendix B and show the value of our approach.

Finally, referring to the comparison of density estimation methodologies carried out in [40] when introducing Density Estimation Trees (DET), we summarize some properties of our approach in Table 6.1.

Table 6.1. Qualitative characteristics of the PWC estimator.

Methodology	Accuracy	Interpretability			Adaptability		Speed	
		COD	VI	Rules	ABD	AWD	Calibration	Query
PWC estimator	medium	✓	✓	✓	✓	✓	slow $\mathcal{O}(n^2 \log(n))$	fast $\mathcal{O}(D)$

Despite the cost of having relatively less accuracy in prediction, our PWC estimator enjoys adaptability, interpretability and the efficient querying like other DET-based approaches [147]. Flexibility applies in terms of adaptability between dimensions (ABD) and within dimension (AWD). The former means that dimensions are treated differently according to their impact on the density; the latter implies that the estimator, in a given dimension, adjusts to the local behaviour of the observations. The estimator we present also benefits from interpretability since:

- It is able to detect clusters and outliers (COD).
- It provides variable importance (VI), i.e. identify dimensions that significantly affect the density.
- It produces rules for specifying subsets of the data which might represent a cluster or outliers.

Although the calibration phase is more onerous than other methodologies, this cost is compensated by the efficient queries. The computational cost for fitting the PWC estimator is dependent on the complexity of the algorithm adopted for computing the Wasserstein distance in the admissibility condition verification phase; in our case $\mathcal{O}(n^2 \log(n))$. The query time for the PWC estimator is $\mathcal{O}(D)$, the same of DET, where D is the depth of the diagram tree representation.

6.2 Implementation

An implementation of the algorithm in Python is available under the permissive free software MIT license. It can be obtained through the author.

The Python implementation considers a class that symbolizes PWC distributions, and its main method to run the algorithm generating the PWC estimator on the observed data. An instance of the class is initiated by providing the sample as a list, Pandas DataFrame or numpy array. The calibrating method takes as a primary argument the significance level α. Other arguments allow the user to select other algorithm specifications and parameters such as the ground metric c within the Wasserstein distance, the methodology for setting the tolerance threshold ϵ and level q of the TVaR on which to get the requested tolerance limit. For all these arguments, a default value is provided. Listing 6.1 outlines a basic code example of our algorithm run.

```python
# import modules
import pwc as pwc
import numpy as np

# create sample
mean = [0, 0]
cov = [[1, 0], [0, 100]]
sample = np.random.multivariate_normal(mean=mean, cov=cov, size=10000)

# run algorithm
my_pwc = pwc.PWCDistribution(sample)
my_pwc.PWCEstimator(alpha=0.01)
```

Listing 6.1. Minimal code example in Python

6.3 Summary

This work introduces an algorithm that computes a piecewise constant estimator to approximate the underlying probability density of a multivariate sample, with possibly hundred of thousands or millions data points.

The PWC estimator is determined using a recursive procedure that generates a partition of the sample domain constituted by hyperrectangular regions where the sample is sufficiently uniformly distributed. Uniformity is assessed using a Wasserstein distance based hypothesis testing. In addition, the PWC estimator, as a result of the final phase of the calibration, is such that, in all margins, a tail value at risk error, compared to the original sample, is not exceeded. This means that our algorithm allows the user to decide the maximum possible discrepancy, introduced by approximating the data with our methodology, between the empirical tail value at risk and the one given by the piecewise constant estimator. The Wasserstein distance is adopted also in this phase, ensuring a consistent approach. In fact, it is possible through this distance function to set an upper bound on the absolute difference of the tail values at risk.

The algorithm is memory efficient since the resulting distribution can be concisely represented and requires a significantly smaller number of element to be saved than the original sample. Instead of storing all data, one can only know the estimate for each nonempty hyperrectangles, which are typically fewer in number than the sample size [2]. Moreover, because of the hierarchical and recursive bisection scheme, the PWC estimator can be conveniently represented through a tree diagram, in which the root is the starting bounding box, each node represents a bisection of the domain, and the leaves are the final partition elements associated to the resulting admissible PWC distribution.

Lastly, as highlighted in [10], using PWC distributions constitutes a favourable approach when information on empirical distributions should be preserved or transferred between systems, because of its memory and bandwidth efficiency, and because it does not distort shape or statistics of the sample. Therefore, our algorithm can be advantageous in applied environments where empirical distributions are repeatedly used and transferred among different users.

Bibliography

[1] J. Friedman. A Recursive Partitioning Decision Rule for Nonparametric Classification. *IEEE Trans. Computers*, 26:404–408, 05 1977.

[2] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[3] P. Arbenz and W. Guevara-Alarcón. Risk Measure Preserving Piecewise Linear Approximation of Empirical Distributions. *European Actuarial Journal*, 1(6):113–148, 05 2016.

[4] P. Parodi. *Pricing in General Insurance.* A Chapman & Hall book. Taylor & Francis, 2014.

[5] International Accounting Standards Board (IASB). International financial reporting standard (IFRS) 4 insurance contracts., 2004. `https://www.ifrs.org/issued-standards/list-of-standards/ifrs-4-insurance-contracts/`.

[6] European Parliament and Council of the European Union (EP-CEU). Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II), 2009. `https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32009L0138`.

[7] De Brauw Blackstone Westbroek. European document retention guide: a comparative view across 16 countries to help you better understand legal requirements and records management best practice. *Iron Mountain Publications*, 2014.

[8] Eidgenössische Finanzmarktaufsicht (FINMA). Rundschreiben 2008/44 Schweizer Solvenz-test (SST), 2008.

[9] C.D. Daykin, T. Pentikäinen, and M. Pesonen. *Practical Risk Theory for Actuaries.* Chapman & Hall, London - New York, 1996.

[10] P. Arbenz and W. Guevara-Alarcón. Piecewise linear approximation of empirical distributions under a Wasserstein distance constraint. *Journal of Statistical Computation and Simulation*, pages 1–24, 2018.

[11] The Economist. Data, data everywhere, A special report on managing information. February 27th, 2010; Available from: http://www.economist.com/node/15557443. *A special report on managing information*, February 2010.

[12] D. Reinsel, J. Gantz, and J. Rydning. The Digitization of the WorldFrom Edge to Core. *Data Age 2025, an IDC White Paper - US44413318*, pages 1–28, 2018.

[13] E. Hoffman. Evoluzione e prospettive dell'high performance computing. *mondodigitale.net*, 1:51–65, 2003.

[14] M. Guillen, C. Bolancé, and J. Nielsen. Kernel Density Estimation of Actuarial Loss Functions. *Insurance: Mathematics and Economics*, 32:19–36, 02 2003.

[15] C. Bolance, M. Guillen, and D. Pitt. Non-parametric Models for Univariate Claim Severity Distributions - an approach using R. Working Papers 2014-01, Universitat de Barcelona, UB Riskcenter, February 2014.

[16] K.M. Sakthivel and C.S. Rajitha. Kernel Density Estimation for Claim Size Distributions Using Shifted Power Transformation. *International Journal of Science and Research*, 5:2319–7064, 2016.

[17] J. Gareth, W. Daniela, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[18] I.M. Johnstone and D.M. Titterington. Statistical challenges of high-dimensional data. philosophical transactions. *Series A, Mathematical, physical, and engineering sciences*, 367(1906):4237–4253, 2009.

[19] I. Olkin and A.R. Sampson. Multivariate Analysis: Overview. In N.J. Smelser and P.B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 10240 – 10247. Pergamon, Oxford, 2001.

[20] M. Clark, J. Morton, B. Fell, D. Wagner, and T. Johnson. Modernizing the insurance actuarial operating model. Preparing insurance companies for the future of work. *Deloitte publications*, 2017.

[21] C. Bennet, S. Keane, M. Clark, K. Stephan, A. Panayi, and S. Morton. Actuarial 20/20. The power of clarity. *Deloitte publications*, 2014.

[22] A.G. Baribeau. Fast Forward, Emerging Technology and Actuarial Practice. *Contingencies*, 08:37–41, August 2014.

[23] Big Data Task Force. Big Data and the Role of the Actuary. *American Academy of Actuaries*, June 2018.

[24] B. Keller. Big data and insurance: Implications for innovation, competition and privacy. *The Geneva Association-International Association for the Study of Insurance Economics*, March 2018.

[25] P. Corbett, M. Schroeck, and R. Shockley. Analytics: The real-world use of big data in insurance. how innovative insurance organizations extract value from uncertain data. *IBM Global Business Services. Business Analytics and Optimization. Executive Report*, May 2013.

[26] P. Bharal and A. Halfon. Making sense of big data in insurance making sense of big data in insurance. *ACORD and MarkLogic publication*, pages 1–13, 2013.

[27] M.A. Beyer and D. Laney. The Importance of "Big Data": A Definition. *Gartner report*, pages 1–9, 2012.

[28] D. Laney. 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group, February 2001.

[29] H. Forest, E. Foo, D. Rose, and D. Berenzon. Big Data. How it can become a differentiator. *White Paper*, pages 1–28, October 2012.

[30] SNS Telecom & IT. Big Data in the Insurance Industry: 2018 - 2030 - Opportunities, Challenges, Strategies & Forecasts. *Market Research Report*, August 2018.

[31] K-U. Schanz and F. Sommerrock. Harnessing technology to narrow the insurance protection gap. *The Geneva Association-International Association for the Study of Insurance Economics*, December 2016.

[32] J. Manyika. What's now and next in analytics, AI, and automation. *McKinsey Executive briefing*, May 2017.

[33] J. McWaters. The Future of Financial Services. *World Economic Forum*, June 2015.

[34] Chief Risk Officer Forum. Big Data & Privacy: unlocking value for consumers. *CROs in a changing environment*, June 2017.

[35] E.W. Frees. Stochastic life contingencies with solvency considerations. *Transactions of the Society of Actuaries*, 42:91–148, 1990.

[36] D.W. Liang and G. Guan Wang. The Data Scientists: Actuary 2.0? *Swiss Re, AAC2018*, September 2018.

[37] Quantee: Actuarial Data science. `https://quantee.ai/actuarial-data-science/`. Accessed: 2019-10-09.

[38] Data science and its potential for Actuaries. `https://www.actuaries.org.uk/learn-and-develop/lifelong-learning/data-science`. Accessed: 2019-10-08.

[39] G.A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, May 1999.

[40] P. Ram and A.G. Gray. Density Estimation Trees. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 627–635, New York, NY, USA, 2011. ACM.

[41] L. Breiman, J. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. The Wadsworth statistics / probability series. CRC, 1984.

[42] D. Li, K. Yang, and W.H. Wong. Density Estimation via Discrepancy Based Adaptive Sequential Partition. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1091–1099. Curran Associates, Inc., 2016.

[43] D.W. Meyer. Density estimation with distribution element trees. *Statistics and Computing*, 28(3):609–632, May 2018.

[44] J. Klemelä. *Smoothing of Multivariate Data: Density Estimation and Visualization.* Wiley, New York, 2009.

[45] E. Luini and P. Arbenz. Density estimation of multivariate samples using wasserstein distance. *Journal of Statistical Computation and Simulation*, 90(2):181–210, 2020.

[46] A. Irpino, R. Verde, and F. De Carvalho. Dynamic Clustering of Histogram Data Based on Adaptive Squared Wasserstein Distances. *Expert Systems with Applications*, 41:3351–3366, 2014.

[47] A. Irpino and E. Romano. Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. *Extraction et Gestion des Connaissances, EGC 2007*, E-9, 01 2007.

[48] F. Bassetti, S. Gualandi, and M. Veneroni. On the Computation of Kantorovich-Wasserstein Distances between 2D-Histograms by Uncapacitated Minimum Cost Flows. *ArXiv: 1804.00445.*, 2018.

[49] G. Auricchio, F. Bassetti, S. Gualandi, and M. Veneroni. Computing Kantorovich-Wasserstein Distances on d-dimensional histograms using $(d+1)$-partite graphs. *ArXiv: 1805.07416*, 05 2018.

[50] K. Ni, X. Bresson, T. Chan, and S. Esedoglu. Local Histogram Based Segmentation Using the Wasserstein Distance. *International Journal of Computer Vision*, 84(1):97–111, Aug 2009.

[51] W. Feller. The Strong Law of Large Numbers. In W. Feller, editor, *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed.*, volume 84, pages 243–245, New York, 1968. Wiley.

[52] H.G. Tucker. A Generalization of the Glivenko-Cantelli Theorem. *Ann. Math. Statist.*, 30(3):828–830, 09 1959.

[53] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[54] J.F. Ziegel. Coherence and Elicitability. *Mathematical Finance*, 26(4):901–918, 2016.

[55] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228, 1999.

[56] R. Kaas, M. Goovaerts, J. Dhaene, and M. Denuit. *Modern Actuarial Risk Theory: Using R.* SpringerLink : Bücher. Springer Berlin Heidelberg, 2008.

[57] S.Y. Chun, A. Shapiro, and S. Uryasev. Conditional Value-at-Risk and Average Value-at-Risk: Estimation and Asymptotics. *Operations Research*, 60(4):739–756, 2012.

[58] E. Del Barrio, J A. Cuesta-Albertos, C. Matrán, and J.M. Rodriguez-Rodriguez. Tests of Goodness of Fit Based on the L_2 Wasserstein Distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[59] E. Del Barrio, J.A. Cuesta-Albertos, C. Matrán, S. Csörgö, C.M. Cuadras, T. de Wet, E. Giné, R. Lockhart, A. Munk, and W. Stute. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, 9(1):1–96, Jun 2000.

[60] E. Del Barrio, E. Giné, and F. Utzet. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189, 01 2005.

[61] J.A. Cuesta-Albertos, C. Matrán, and J. Rodríguez-Rodríguez. Approximation to Probabilities Through Uniform Laws on Convex Sets. *Journal of Theoretical Probability*, 16(2):363–376, Apr 2003.

[62] A. Ramdas, N. García Trillos, and M. Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, 2017.

[63] S. Deng, W. Li, and X. Wu. Wasserstein Identity Testing. *CoRR*, abs/1710.10457, 2017.

[64] Y. Rubner, L. Guibas, and C. Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668, 1997.

[65] C. Villani. *Optimal Transport, Old and New*. Springer-Verlag, Berlin, 2009.

[66] C.L. Mallows. A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43:508–515, 1972.

[67] E. Levina and P. Bickel. The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics. *Proceedings of the Eighth IEEE International Conference on Computer Vision*, pages 251–256, 2001.

[68] D.J.H. Garling. *Analysis on Polish Spaces and an Introduction to Optimal Transportation*. London Mathematical Society Student Texts (89). Cambridge University Press, Cambridge, 2017.

[69] F. Santambrogio. Optimal Transport for Applied Mathematicians. *Progress in Nonlinear Dierential Equations and Their Applications.*, 87, 2015.

[70] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks. Lecture Notes in Mathematics*, volume 2062. Springer Berlin Heidelberg, 2013.

[71] G. Monge. Memoire sur la theorie des deblais et des remblais. *Histoire de l'Acadmie Royale des Sciences (1781)*, pages 666–704, 1784.

[72] L. Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[73] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, 2012.

[74] H.W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics*, 2:83–97, 1995.

[75] M.M. Zavlanos, L. Spesivtsev, and G.J. Pappas. A distributed auction algorithm for the assignment problem. In *2008 47th IEEE Conference on Decision and Control*, pages 1212–1217, Dec 2008.

[76] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[77] P. Kovács. Minimum-cost flow algorithms: an experimental evaluation. *Optimization Methods and Software*, 30(1):94–127, 2015.

[78] L. Dieci and J. D. Walsh III. The boundary method for semi-discrete optimal transport partitions and wasserstein distance computation. *ArXiv: 1702.03517.*, 02 2017.

[79] B. Schmitzer. A Sparse Multiscale Algorithm for Dense Optimal Transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, Oct 2016.

[80] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007.

[81] O. Pele and M. Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, Sept 2009.

[82] L. Li, M. Ma, P. Lei, X. Wang, and X. Chen. A Linear Approximate Algorithm for Earth Mover's Distance with Thresholded Ground Distance. *Mathematical Problems in Engineering*, 2014 (Article ID 406358), 2014.

[83] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[84] B. Schmitzer. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *ArXiv: 1610.06519.*, 2016.

[85] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis. Overrelaxed Sinkhorn-Knopp Algorithm for Regularized Optimal Transport. NIPS'17 Workshop on Optimal Transport & Machine Learning, November 2017.

[86] J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Proceedings of NIPS 2017.*, 2017.

[87] B.K. Abid and R. Gower. Stochastic algorithms for entropy-regularized optimal transport problems. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1505–1512. PMLR, 09–11 Apr 2018.

[88] A. Oberman and Y. Ruan. An efficient linear programming method for Optimal Transportation. *ArXiv: 1509.03668.*, 2016.

[89] S. Gerber and M. Maggioni. Multiscale Strategies for Computing Optimal Transport. *Journal of Machine Learning Research*, 18(72):1–32, 2017.

[90] J. Solomon. Optimal Transport on Discrete Domains. Notes for AMS Short Course on Discrete Differential Geometry. *ArXiv: 1801.07745.*, 2018.

[91] L. Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Lechevallier Y., Saporta G. (eds) Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.

[92] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic Optimization for Large-scale Optimal Transport. *ArXiv e-prints*, May 2016.

[93] A. Dessein, N. Papadakis, and J.L. Rouas. Regularized Optimal Transport and the Rot Mover's Distance. *arXiv:1610.06447*, 2016.

[94] V. Seguy, B.B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. In *Proceedings of the International Conference in Learning Representations*, 2018.

[95] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.

[96] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[97] L. Rüschendorf and L. Uckelmann. Numerical and analytical results for the transportation problem of Monge-Kantorovich. *Metrika*, 51:245–258, 09 2000.

[98] Q. Mérigot. A Multiscale Approach to Optimal Transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[99] R. Geiß, D.and Klein, R. Penninger, and G. Rote. Optimally solving a transportation problem using Voronoi diagrams. *Computational Geometry*, 46(8):1009 – 1016, 2013.

[100] P. Mullen, P. Memari, F. De Goes, and M. Desbrun. HOT: Hodge-optimized Triangulations. *ACM Trans. Graph.*, 30(4):103:1–103:12, July 2011.

[101] F. De Goes, K. Breeden, V. Ostromoukhov, and M. Desbrun. Blue Noise through Optimal Transport. *ACM Trans. Graph. (SIGGRAPH Asia)*, 31, 2012.

[102] B. Lévy. A numerical algorithm for L_2 semi-discrete optimal transport in 3D. *ESAIM Math. Modeling and Analysis*, 49(6), 2015.

[103] B. Lévy and E.L. Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

[104] J.D. Walsh. Uniqueness of optimal solutions for semi-discrete transport with p-norm cost functions. *ArXiv: 1705.09383.*, 05 2017.

[105] V. Hartmann and D. Schuhmacher. Semi-discrete optimal transport - the case p=1. *ArXiv: 1706.07650.*, 06 2017.

[106] J. D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[107] N. Papadakis, G. Peyré, and E. Oudet. Optimal Transport with Proximal Splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[108] Bangwon Ko, Ralph P. Russo, and Nariankadu D. Shyamalkumar. A Note on Nonparametric Estimation of the CTE. *ASTIN Bulletin*, 39(2):717–734, 2009.

[109] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T.A. Poggio. Learning with a Wasserstein Loss. In *Advances in Neural Information Processing Systems (NIPS) 28*, 2015.

[110] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville. Improved Training of Wasserstein GANs. *arXiv:1704.00028*, 2017.

[111] H. Petzka, A. Fischer, and D. Lukovnicov. Wasserstein GAN. *arXiv:1709.08894*, 2017.

[112] M. Cuturi and D. Arnaud. Fast Computation of Wasserstein Barycenters. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–685–II–693. JMLR.org, 2014.

[113] N. Ho, X. Nguyen, M Yurochkin, H.H. Bui, V. Huynh, and D.Q. Phung. Multilevel Clustering via Wasserstein Means. In *ICML*, 2017.

[114] M. Staib and S. Jegelka. Wasserstein k-means++ for Cloud Regime Histogram Clustering. In *Proceedings of the Seventh International Workshop on Climate Informatics: CI 2017*, 2017.

[115] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques.*, 53(1):1–26, 2017.

[116] M. Gelbrich. On a Formula for the L_2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[117] G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[118] J. Horowitz and R.L. Karandikar. Mean rates of convergence of empirical measures in the Wasserstein metric. *Journal of Computational and Applied Mathematics*, 55(3):261–273, 1994.

[119] E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(2):539–563, 05 2014.

[120] N. Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015.

[121] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv:1707.00087*, 2017.

[122] J. Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *arXiv:1804.10556*, 2018.

[123] S. Singh and B. Póczos. Minimax Distribution Estimation in Wasserstein Distance. *ArXiv e-prints*, February 2018.

[124] E. Del Barrio, E. Giné, and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, 27(2):1009–1071, 1999.

[125] O.C. Ibe. *Markov Processes for Stochastic Modeling*. Elsevier, Oxford, 2013.

[126] P.J. Bickel and D.A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.

[127] T. Rippl, A. Munk, and A. Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90 – 109, 2016.

[128] E. Del Barrio and J.M. Loubes. Central Limit Theorems for Empirical Transportation Cost in General Dimension. *ArXiv: 1705.01299*, 2017.

[129] M. Sommerfeld and A. Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

[130] M. Ajtai, J. Komlós, and G. Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, Dec 1984.

[131] M. Talagrand. Matching Random Samples in Many Dimensions. *Ann. Appl. Probab.*, 2(4):846–856, 11 1992.

[132] M. Talagrand and J.E. Yukich. The integrability of the square exponential transportation cost. *The Annals of Applied Probability*, 3(4):1100–1111, 1993.

[133] M. Talagrand. The Transportation Cost from the Uniform Measure to the Empirical Measure in Dimension ≥ 3. *Ann. Probab.*, 22(2):919–959, 04 1994.

[134] V. Dobrić and J.E. Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, Jan 1995.

[135] W.Q. Meeker, G.J. Hahn, and L.A. Escobar. *Statistical Intervals: A Guide for Practitioners and Researchers*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd edition, 2017.

[136] A.L. Gibbs and F.E. Su. On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435, 2002.

[137] J.A. Cuesta-Albertos, C. Matrán, and A. Tuero-Diaz. On lower bounds for the L2-Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability*, 9:263–283, 1996.

[138] J.A. Carrillo. *Lecture notes: Main Models and Basics of Wasserstein Distance*. ICREA - Methods and Models of Kinetic Theory, Universitat Autònoma de Barcelona, 2006.

[139] J.D. Gibbons. Estimation of the unknown upper limit of a uniform distribution. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 36(1):29–40, 1974.

[140] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[141] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971.

[142] L. Devroye, L. Györfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. In *Stochastic Modelling and Applied Probability*, 1996.

[143] S. Dudoit and M.J. Van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2:131–154, 2005.

[144] J.S. Simonoff. *Smoothing Methods in Statistics*. Springer series in statistics. Springer-Verlag, New York, 1996.

[145] T.G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer, Berlin, 2000.

[146] J. Surowiecki. *The Wisdom of Crowds*. Doubleday, New York, 2004.

[147] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

[148] M. Bourel and J. Cugliari. Bagging of density estimators. *Computational Statistics*, 34(4):1849–1869, Dec 2019.

[149] D.W. Scott. Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *The Annals of Statistics*, 13(3):1024–1040, 1985.

[150] G. Ridgeway. Looking for Lumps: Boosting and Bagging for Density Estimation. *Computational Statistics & Data Analysis*, 38(4):379–392, 2002.

[151] P. Rigollet and A.B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16:260–280, 2007.

[152] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, Aug 1996.

[153] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, Springer Series in Statistics, New York, 1992.

[154] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[155] T.A. O'Brien, K. Kashinath, N.R. Cavanaugh, W.D. Collins, and J.P. O'Brien. A fast and objective multidimensional kernel density estimation method: fastKDE. *Computational Statistics & Data Analysis*, pages 148 – 160, 2016.

[156] Ferdosi, B.J., Buddelmeijer, H., Trager, S.C., Wilkinson, M.H.F., and Roerdink, J.B.T.M. Comparison of density estimation methods for astronomical datasets. *A&A*, 531:A114, 2011.

[157] R. Bürgi, M.M. Dacorogna, and R. Iles. Risk Aggregation, Dependence Structure and Diversification Benefit. *Stress Testing for Financial Institutions*, September 2009.

[158] P. Fonseca. Modeling Dependencies. *Lectures in Quantitative Finance, Universität Zürich UZH*, April 2015.

[159] E.W. Frees and E.A. Valdez. Understanding Relationships Using Copulas. *North American Actuarial Journal*, 2(1):1–25, 1998.

[160] M. Rytgaard. Estimation in the Pareto Distribution. *ASTIN Bulletin*, 20(2):201–216, 1990.

[161] J.H.T. Kim, S. Ahn, and S. Ahn. Parameter estimation of the Pareto distribution using a pivotal quantity. *Journal of the Korean Statistical Society*, 46(3):438–450, 2017.

[162] M. Norton, V. Khokhlov, and S. Uryasev. Calculating CVaR and bPOE for Common Probability Distributions With Application to Portfolio Optimization and Density Estimation, 2018.

[163] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

[164] M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan. *copula: Multivariate Dependence with Copulas*, 2018. R package version 0.999-19.1.

[165] R.V. Hogg and S.A. Klugman. *Loss distributions*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, New York, 1984.

[166] Society of Actuaries (SOA) research topics. `https://www.soa.org/research/research-topic-list/`. Accessed: 2019-09-11.

[167] H. Joe. *Dependence Modeling with Copulas*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, London - New York, 2014.

[168] D. Li, K. Yang, and W.H. Wong. Density Estimation via Discrepancy Based Adaptive Sequential Partition. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1091–1099. Curran Associates, Inc., 2016.

[169] R. Lopes, I Reid, and P. Hobson. The two-dimensional Kolmogorov-smirnov test. *Proc. XI Int. Workshop on Advanced Computing and Analysis Techniques in Physics Research*, 2007.

[170] Maydeu-Olivares A. and Garcia-Forero C. Goodness-of-fit testing. *International Encyclopedia of Education*, 7:190–196, 2010.

[171] J.J. Liang, K.T. Fang, F.J. Hickernell, and L. Runze. Testing multivariate uniformity and its applications. *Mathematics of Computation*, 70(233):337–355, 1 2001.

[172] J.R. Berrendero, A. Cuevas, and F. Vázquez-Grande. Testing Multivariate Uniformity: The Distance-to-Boundary Method. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):693–707, 2006.

Appendices

A Order of convergence check

The following analysis examines the order of convergence of the empirical Wasserstein distance outlined in Lemma 3.5, for ℓ_1-norm and ℓ_2-norm cases.

Table A1. Order of convergence check

	ℓ_1-norm					ℓ_2-norm				
	$d = 2$					$d = 2$				
sample size	5	25	50	100	500	5	25	50	100	500
mean	0.0679	0.0534	0.0513	0.0497	0.0502	0.0549	0.0443	0.0424	0.0414	0.0431
std	0.0112	0.0079	0.0071	0.0063	0.0051	0.0092	0.0067	0.0060	0.0054	0.0043
median	0.0657	0.0526	0.0500	0.0486	0.0493	0.0533	0.0431	0.0413	0.0404	0.0423
0.80 qnt.	0.0762	0.0598	0.0564	0.0543	0.0539	0.0620	0.0493	0.0468	0.0454	0.0463
0.90 qnt.	0.0830	0.0645	0.0606	0.0581	0.0570	0.0675	0.0534	0.0505	0.0487	0.0489
0.95 qnt.	0.0894	0.0691	0.0646	0.0617	0.0598	0.0724	0.0572	0.0538	0.0517	0.0513
0.99 qnt.	0.1031	0.0787	0.0737	0.0694	0.0659	0.0825	0.0649	0.0609	0.0579	0.0564
	$d = 3$					$d = 3$				
sample size	5	25	50	100	500	5	25	50	100	500
mean	0.1267	0.1254	0.1247	0.1239	0.1236	0.0802	0.0742	0.0752	0.0776	0.0882
std	0.0119	0.0082	0.0069	0.0057	0.0036	0.0079	0.0054	0.0046	0.0040	0.0029
median	0.1248	0.1243	0.1238	0.1231	0.1231	0.0790	0.0734	0.0746	0.0770	0.0878
0.80 qnt.	0.1356	0.1316	0.1298	0.1282	0.1264	0.0863	0.0783	0.0787	0.0806	0.0904
0.90 qnt.	0.1426	0.1363	0.1337	0.1314	0.1284	0.0908	0.0813	0.0814	0.0829	0.0920
0.95 qnt.	0.1489	0.1404	0.1373	0.1344	0.1302	0.0948	0.0840	0.0838	0.0849	0.0935
0.99 qnt.	0.1631	0.1499	0.1451	0.1410	0.1342	0.1032	0.0899	0.0889	0.0894	0.0967
	$d = 6$					$d = 6$				
sample size	5	25	50	100	500	5	25	50	100	500
mean	0.6505	0.6090	0.5950	0.5828	0.5624	0.1403	0.1613	0.1802	0.2050	0.2908
std	0.0297	0.0129	0.0091	0.0063	0.0029	0.0069	0.0040	0.0032	0.0027	0.0018
median	0.6480	0.6080	0.5942	0.5823	0.5623	0.1399	0.1610	0.1800	0.2048	0.2908
0.80 qnt.	0.6745	0.6194	0.6023	0.5879	0.5648	0.1460	0.1645	0.1829	0.2071	0.2923
0.90 qnt.	0.6900	0.6260	0.6068	0.5911	0.5662	0.1495	0.1665	0.1845	0.2085	0.2931
0.95 qnt.	0.7037	0.6320	0.6109	0.5939	0.5674	0.1523	0.1683	0.1859	0.2097	0.2939
0.99 qnt.	0.7306	0.6442	0.6190	0.5996	0.5698	0.1583	0.1719	0.1887	0.2120	0.2953

Table A1 compares the test statistic distribution derived from samples of size 500 to the test statistic distributions derived from samples of sizes 5, 25, 50 and 100, adjusted according to the order of convergence. The values of mean, standard deviation, median and quantiles (0.80, 0.90, 0.95, 0.99 levels) are reported. All distributions concern the hypercube $[0, 1]^d$, where d is equal to 2, 3 and 6, and each one has been achieved via a 50000 run simulation.

B Rejection rate analysis

The performance of the Wasserstein distance based test, herein referred to as Wasserstein test, has been analyzed through a simulation study, whose design is similar to those in [171, 172]. We quantified the empirical type I error rate and the power of our test

statistic, and compared with other procedures used to test the uniformity of random samples.

We evaluated rejection percentages along 5000 independent runs based on samples drawn both from the null and from the alternative hypothesis; this provide us with the empirical type I error rate (false positive) and the power (true positive) of the tests. The analysis considers increasing sample sizes of 25, 50, 100 and 500, and nominal significance levels of 0.01, 0.05 and 0.10. Simulation outputs are displayed in Tables B1, B2 and B3.

All the models have support on the hypercube $[0,1]^d$, where $d = 1, 2, 3$. The null hypotheses refer to samples drawn from the corresponding uniform measure. The alternative hypothesis models are devised for taking into account complementary aspects of non-uniformity: the one-dimensional case especially evaluates the ability to ascertain marginal deviations from uniformity; the multidimensional setting is useful to assess the capacity of detecting departures from independence, keeping the uniformity of marginals.

With regard to the one-dimensional setting, the Wasserstein test (WASS) has been compared with the K-S and the χ^2 goodness-of-fit tests, and the following alternative hypothesis models have been considered.

- The (univariate) uniform contamination models (UC): the observed sample is drawn from a mixture of type $(1 - v)\, U_{[0,1]} + v\, U_{[1/2,1/2]}$. The parameter v has been set equal to 0.1 and 0.2 respectively.
- The arcsine distribution (AS).
- The beta models (B) with parameters (1.3, 1.3), (5, 2) and (0.8, 0.8) respectively.
- The truncated Standard Normal distribution (TZ).

In the multidimensional context, three statistics have been considered: the Wasserstein distances with ℓ_1-norm and ℓ_2-norm cost functions, and the χ^2 statistics. As for alternative distributions, we have checked the following distribution families:

- The (multivariate) uniform contamination models (UC): The observed sample is drawn from a mixture of type $(1 - v)\, U_{[0,1]^d} + v\, U_{D_d}$, where D_d denotes a cube with the same centre as $[0,1]^d$ and measure $1/2$. The parameter v has been set equal to 0.1 and 0.2 respectively.
- Meta-type uniform distributions: These are the distributions obtained from transformed $\mathbb{R}^d$-supported elliptically distributed random variables. The transformation applied is the relevant probability integral transform, to guarantee that marginals are uniformly distributed. MUT is derived from a multivariate Student's random variable with 5 degrees of freedom, MUC is obtained from a Cauchy variable and finally MUN results from a multivariate Gaussian centred at the origin $N(0, \Sigma)$, where $\Sigma = (\sigma_{i,j})$, $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.5$ for $i \neq j$.
- The beta independent models (B): All the marginals are independent, identically distributed according to a beta models with parameters (1.3, 1.3) and (0.8, 0.8).

The uniform contamination and the beta models, with parameters (1.3, 1.3) and (0.8, 0.8), are considered in both experiments as they should represent circumstances of non-uniformity hard to detect. The former specifically represents a deviation from the

theoretical model associated with the presence of "inliers" [172]. Either one of them has independent marginals, as in the case of the null distribution.

The results of rejection rates simulation indicate that:

- The empirical type I errors of the Wasserstein test have converged to the significance levels.
- In both the one-dimensional and multidimensional setting, the Wasserstein test hardly detects the uniform contamination models.
- In the one-dimensional setting, the K-S test shows overall a better performance than the other two tests. Nevertheless, the Wasserstein test is the most powerful for the arcsine and the truncated normal distributions. In the multivariate case the Wasserstein test exhibited, in general, the highest power figures. The Wasserstein statistics with ℓ_1-norm and ℓ_2-norm cost functions exhibit a similar behaviour.
- In the multivariate setting the Wasserstein test is able to classify samples from multivariate distributions with uniform marginals. This, along with the favourable power revealed by Wasserstein test when $d = 1$, suggests that the process of testing marginal admissibility condition at first, and subsequently the joint admissibility condition, should succeed in picking out non-uniformly distributed sample.

Table B1. One-dimensional setting rejection rates; sample size n = 25, 50, 100, 500.

Test	α	U	$UC(0.1)$	$UC(0.2)$	AS	$B(1.3, 1.3)$	$B(5, 2)$	$B(0.8, 0.8)$	TZ
					$n = 25$				
WASS	0.01	0.0082	0.0066	0.0046	0.0600	0.0062	0.9928	0.0146	0.0244
WASS	0.05	0.0508	0.0428	0.0348	0.2356	0.0400	1.0000	0.0748	0.095
WASS	0.10	0.0992	0.0872	0.0872	0.4024	0.0982	1.0000	0.1400	0.1610
K-S	0.01	0.0078	1.0000	1.0000	0.0694	1.0000	1.0000	1.0000	0.0168
K-S	0.05	0.0424	1.0000	1.0000	0.2044	1.0000	1.0000	1.0000	0.0722
K-S	0.10	0.0848	1.0000	1.0000	0.3360	1.0000	1.0000	1.0000	0.1324
χ^2	0.01	0.0078	0.0108	0.0154	0.2244	0.0182	0.9062	0.0164	0.0128
χ^2	0.05	0.0452	0.0554	0.0824	0.3688	0.0762	0.9910	0.0702	0.0640
χ^2	0.10	0.0930	0.1068	0.1434	0.4978	0.1402	0.9986	0.1274	0.1216
					n = 50				
WASS	0.01	0.0100	0.0070	0.0048	0.1426	0.0064	1.0000	0.0176	0.0442
WASS	0.05	0.0558	0.0466	0.0466	0.4748	0.0538	1.0000	0.0814	0.1632
WASS	0.10	0.1042	0.0922	0.1144	0.6754	0.1148	1.0000	0.1552	0.2508
K-S	0.01	0.0080	1.0000	1.0000	0.1476	1.0000	1.0000	1.0000	0.0362
K-S	0.05	0.0462	1.0000	1.0000	0.3952	1.0000	1.0000	1.0000	0.1270
K-S	0.10	0.0914	1.0000	1.0000	0.5574	1.0000	1.0000	1.0000	0.2088
χ^2	0.01	0.0118	0.0140	0.0256	0.5056	0.0242	0.9998	0.0262	0.0222
χ^2	0.05	0.0468	0.0620	0.0996	0.6572	0.0940	1.0000	0.0952	0.0856
χ^2	0.10	0.0998	0.1168	0.1888	0.7394	0.1810	1.0000	0.1758	0.1528
					n = 100				
WASS	0.01	0.0108	0.0070	0.0100	0.4904	0.0088	1.0000	0.0180	0.1118
WASS	0.05	0.0512	0.0472	0.0730	0.8618	0.0698	1.0000	0.1036	0.2802
WASS	0.10	0.0984	0.1054	0.1772	0.9484	0.1676	1.0000	0.1986	0.3914
K-S	0.01	0.0082	1.0000	1.0000	0.4014	1.0000	1.0000	1.0000	0.0852
K-S	0.05	0.0412	1.0000	1.0000	0.7522	1.0000	1.0000	1.0000	0.2300
K-S	0.10	0.0896	1.0000	1.0000	0.8872	1.0000	1.0000	1.0000	0.3404
χ^2	0.01	0.0094	0.0196	0.0526	0.8276	0.0406	1.0000	0.0468	0.0248
χ^2	0.05	0.0504	0.0738	0.1586	0.9192	0.1410	1.0000	0.1374	0.0918
χ^2	0.10	0.0982	0.1312	0.2506	0.9514	0.2284	1.0000	0.2278	0.1588
					n = 500				
WASS	0.01	0.0092	0.0118	0.1672	1.0000	0.1622	1.0000	0.1220	0.7192
WASS	0.05	0.0470	0.0930	0.6096	1.0000	0.6148	1.0000	0.4666	0.8838
WASS	0.10	0.0976	0.2172	0.8112	1.0000	0.8144	1.0000	0.6810	0.9334
K-S	0.01	0.0080	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6220
K-S	0.05	0.0452	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8324
K-S	0.10	0.0970	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9012
χ^2	0.01	0.0100	0.0260	0.1806	1.0000	0.1470	1.0000	0.2088	0.0536
χ^2	0.05	0.0502	0.0962	0.3878	1.0000	0.3432	1.0000	0.4106	0.1720
χ^2	0.10	0.0932	0.1758	0.5206	1.0000	0.4864	1.0000	0.5418	0.2824

Table B2. Two-dimensional setting rejection rates; sample size $n = 25$, 50, 100, 500.

Test	α	U	$UC(0.1)$	$UC(0.2)$	MUT	MUC	MUN	$B(1.3, 1.3)$	$B(0.8, 0.8)$
					$n = 25$				
WASS ℓ_1	0.01	0.0110	0.0084	0.0086	0.0400	0.045	0.0418	0.0062	0.0244
WASS ℓ_1	0.05	0.0488	0.0402	0.0384	0.1738	0.2212	0.1790	0.0420	0.0986
WASS ℓ_1	0.10	0.0960	0.0838	0.0904	0.3132	0.3888	0.3224	0.0876	0.1674
WASS ℓ_2	0.01	0.0112	0.0090	0.0076	0.0674	0.0916	0.0666	0.0054	0.0252
WASS ℓ_2	0.05	0.0446	0.0430	0.0354	0.2346	0.2758	0.2316	0.0364	0.0944
WASS ℓ_2	0.10	0.0956	0.0920	0.0840	0.3684	0.4334	0.3744	0.0768	0.1644
χ^2	0.01	0.0084	0.0076	0.0080	0.092	0.0798	0.0848	0.0088	0.0090
χ^2	0.05	0.0398	0.0386	0.0386	0.2262	0.2136	0.2288	0.0416	0.0432
χ^2	0.10	0.0864	0.0851	0.0846	0.3604	0.3406	0.3562	0.0934	0.0918
					n = 50				
WASS ℓ_1	0.01	0.0098	0.0068	0.0072	0.1396	0.2292	0.1452	0.0058	0.0272
WASS ℓ_1	0.05	0.0514	0.0412	0.0432	0.4765	0.602	0.4658	0.0522	0.1112
WASS ℓ_1	0.10	0.1012	0.0916	0.0943	0.6632	0.7692	0.6502	0.1178	0.1981
WASS ℓ_2	0.01	0.0118	0.0066	0.0054	0.2402	0.3254	0.2384	0.0056	0.0262
WASS ℓ_2	0.05	0.0504	0.0402	0.0402	0.5466	0.6386	0.5408	0.0456	0.1036
WASS ℓ_2	0.10	0.1036	0.0902	0.0936	0.7094	0.7864	0.7064	0.1004	0.1936
χ^2	0.01	0.0114	0.0126	0.0134	0.1212	0.3762	0.0908	0.0202	0.0244
χ^2	0.05	0.0464	0.0485	0.0568	0.2612	0.5796	0.2218	0.0762	0.0716
χ^2	0.10	0.1162	0.1234	0.1346	0.4078	0.7273	0.3658	0.1664	0.1556
					n = 100				
WASS ℓ_1	0.01	0.0125	0.0114	0.0114	0.6475	0.8178	0.6384	0.0188	0.0436
WASS ℓ_1	0.05	0.0534	0.0482	0.0636	0.9163	0.9716	0.9120	0.1026	0.1581
WASS ℓ_1	0.10	0.1062	0.099	0.1256	0.9661	0.9913	0.9684	0.1978	0.2732
WASS ℓ_2	0.01	0.0122	0.0082	0.0104	0.7592	0.8572	0.7578	0.0158	0.0392
WASS ℓ_2	0.05	0.0576	0.0458	0.0542	0.9346	0.9744	0.9340	0.0862	0.1438
WASS ℓ_2	0.10	0.1106	0.096	0.1118	0.9713	0.9912	0.9766	0.1736	0.2576
χ^2	0.01	0.0182	0.0194	0.0310	0.2140	0.6534	0.1450	0.0381	0.0376
χ^2	0.05	0.0484	0.0492	0.0776	0.3410	0.7782	0.2608	0.0846	0.0898
χ^2	0.10	0.1256	0.1242	0.1722	0.4998	0.8748	0.4170	0.1774	0.1880
					n = 500				
WASS ℓ_1	0.01	0.0094	0.0118	0.0671	1.0000	1.0000	1.0000	0.4866	0.3618
WASS ℓ_1	0.05	0.0424	0.0656	0.2566	1.0000	1.0000	1.0000	0.8224	0.6874
WASS ℓ_1	0.10	0.0844	0.1302	0.4182	1.0000	1.0000	1.0000	0.9162	0.8268
WASS ℓ_2	0.01	0.0114	0.0138	0.0548	1.0000	1.0000	1.0000	0.4124	0.3151
WASS ℓ_2	0.05	0.0470	0.0648	0.2412	1.0000	1.0000	1.0000	0.7932	0.6684
WASS ℓ_2	0.10	0.0992	0.1386	0.4104	1.0000	1.0000	1.0000	0.9163	0.8226
χ^2	0.01	0.0176	0.0226	0.0280	0.3128	0.9218	0.1564	0.0344	0.0362
χ^2	0.05	0.0486	0.0554	0.0724	0.4736	0.9648	0.2928	0.0921	0.0982
χ^2	0.10	0.0798	0.0860	0.1084	0.5532	0.9774	0.3754	0.1286	0.1452

Table B3. Three-dimensional setting rejection rates; sample size $n = 25$, 50, 100, 500.

Test	α	U	$UC(0.1)$	$UC(0.2)$	MUT	MUC	MUN	$B(1.3, 1.3)$	$B(0.8, 0.8)$
					$n = 25$				
WASS ℓ_1	0.01	0.0108	0.0078	0.0065	0.2062	0.2916	0.2093	0.0034	0.0342
WASS ℓ_1	0.05	0.0488	0.0412	0.0372	0.4584	0.5732	0.4551	0.0243	0.1314
WASS ℓ_1	0.10	0.0948	0.0848	0.0766	0.6028	0.7158	0.5998	0.0584	0.2202
WASS ℓ_2	0.01	0.0108	0.0078	0.0066	0.2655	0.3522	0.2512	0.0028	0.0334
WASS ℓ_2	0.05	0.0492	0.0368	0.0322	0.5074	0.6128	0.5011	0.0214	0.1308
WASS ℓ_2	0.10	0.0986	0.0858	0.0722	0.6456	0.7414	0.6374	0.0492	0.2204
χ^2	0.01	0.0096	0.0092	0.0081	0.2748	0.2776	0.2728	0.0086	0.0068
χ^2	0.05	0.0410	0.0431	0.0458	0.4862	0.4902	0.4818	0.0486	0.0452
χ^2	0.10	0.0856	0.0871	0.0882	0.5942	0.5968	0.5828	0.0932	0.0872
					n = 50				
WASS ℓ_1	0.01	0.0098	0.0066	0.0054	0.6778	0.8174	0.6534	0.0058	0.0402
WASS ℓ_1	0.05	0.0472	0.0426	0.0346	0.8824	0.9572	0.8798	0.0355	0.1610
WASS ℓ_1	0.10	0.1008	0.0872	0.0761	0.9432	0.9812	0.9374	0.0782	0.2688
WASS ℓ_2	0.01	0.0102	0.0066	0.0078	0.7308	0.8358	0.7172	0.0036	0.0390
WASS ℓ_2	0.05	0.0496	0.0421	0.0346	0.9032	0.9614	0.9022	0.0288	0.1536
WASS ℓ_2	0.10	0.1016	0.0883	0.0771	0.9524	0.9814	0.9494	0.0648	0.2568
χ^2	0.01	0.0116	0.0112	0.0106	0.6532	0.7966	0.6252	0.0236	0.0202
χ^2	0.05	0.0480	0.0523	0.0506	0.8098	0.9074	0.7938	0.0826	0.0705
χ^2	0.10	0.0992	0.0962	0.0891	0.8766	0.9456	0.8666	0.1438	0.1222
					n = 100				
WASS ℓ_1	0.01	0.008	0.0094	0.0072	0.9908	0.9992	0.9894	0.0114	0.0624
WASS ℓ_1	0.05	0.0484	0.0444	0.0497	0.9992	1.0000	0.999	0.0776	0.2244
WASS ℓ_1	0.10	0.0934	0.0902	0.0864	1.0000	1.0000	1.0000	0.1606	0.3626
WASS ℓ_2	0.01	0.0088	0.0062	0.0084	0.9932	0.9992	0.9938	0.0086	0.0584
WASS ℓ_2	0.05	0.0498	0.0434	0.0394	0.9992	1.0000	0.9996	0.0578	0.2146
WASS ℓ_2	0.10	0.0954	0.086	0.0864	1.0000	1.0000	1.0000	0.1264	0.3454
χ^2	0.01	0.0124	0.0112	0.0116	0.9414	0.9975	0.9282	0.0432	0.0266
χ^2	0.05	0.0465	0.0524	0.0598	0.9842	0.9994	0.9796	0.1362	0.1046
χ^2	0.10	0.0976	0.1122	0.1192	0.9934	1.0000	0.9898	0.2242	0.1858
					n = 500				
WASS ℓ_1	0.01	0.0082	0.0108	0.0214	1.0000	1.0000	1.0000	0.5408	0.5759
WASS ℓ_1	0.05	0.0424	0.0538	0.1074	1.0000	1.0000	1.0000	0.8338	0.8436
WASS ℓ_1	0.10	0.0839	0.1028	0.1932	1.0000	1.0000	1.0000	0.9218	0.9274
WASS ℓ_2	0.01	0.0114	0.01205	0.0192	1.0000	1.0000	1.0000	0.4564	0.5343
WASS ℓ_2	0.05	0.0474	0.0574	0.0944	1.0000	1.0000	1.0000	0.8019	0.8322
WASS ℓ_2	0.10	0.0954	0.1124	0.1794	1.0000	1.0000	1.0000	0.9106	0.9184
χ^2	0.01	0.0136	0.0172	0.0330	1.0000	1.0000	1.0000	0.1266	0.0876
χ^2	0.05	0.0571	0.0636	0.1122	1.0000	1.0000	1.0000	0.2972	0.2378
χ^2	0.10	0.1006	0.1144	0.1818	1.0000	1.0000	1.0000	0.4122	0.3483