Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models

Marouane Il Idrissi, Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, Jean-Michel Loubes

Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models

Marouane Il Idrissi[a,b,c,e], Nicolas Bousquet[a,b,d], Fabrice Gamboa[c], Bertrand Iooss[a,b,c], Jean-Michel Loubes[c]

[a] *EDF Lab Chatou, 6 Quai Watier, 78401 Chatou, France*
[b] *SINCLAIR AI Lab., Saclay, France*
[c] *Institut de Mathématiques de Toulouse, 31062 Toulouse, France*
[d] *Sorbonne Université, LPSM, 4 place Jussieu, Paris, France*
[e] *Corresponding Author - Email: marouane.il-idrissi@edf.fr*

Abstract

Robustness studies of black-box models is recognized as a necessary task for numerical models based on structural equations and predictive models learned from data. These studies must assess the model's robustness to possible misspecification of regarding its inputs (e.g., covariate shift). The study of black-box models, through the prism of uncertainty quantification (UQ), is often based on sensitivity analysis involving a probabilistic structure imposed on the inputs, while ML models are solely constructed from observed data. Our work aim at unifying the UQ and ML interpretability approaches, by providing relevant and easy-to-use tools for both paradigms. To provide a generic and understandable framework for robustness studies, we define perturbations of input information relying on quantile constraints and projections with respect to the Wasserstein distance between probability measures, while preserving their dependence structure. We show that this perturbation problem can be analytically solved. Ensuring regularity constraints by means of isotonic polynomial approximations leads to smoother perturbations, which can be more suitable in practice. Numerical experiments on real case studies, from the UQ and ML fields, highlight the computational feasibility of such studies and provide local and global insights on the robustness of black-box models to input perturbations.

Keywords: interpretability, machine learning, sensitivity analysis, computer model, sensitivity analysis, robustness, epistemic uncertainty, domain uncertainty, quantiles, isotonic polynomials

1. Introduction

Multiple engineering fields require models for prediction and phenomenological understanding. Machine learning (ML) and uncertainty quantification (UQ) of numerical models are essential approaches to developing and manipulating such models. Because they require, for their enlightened use, an adequate understanding of their characteristics, they share fundamental similarities. These two frameworks feed off each other through the duality of sensitivity analyses (SA), a fundamental methodological corpus in UQ, and ML interpretability methods, as highlighted by [72, 50]. In particular, recent advances in explainable ML leverage tools from SA to produce meaningful interpretations of black-box models [35, 17], and novel SA estimation schemes are heavily based on the construction of suitable ML models [15, 9]. Both SA and ML interpretability rely on studying the relationships between a black-box model behavior and its inputs [24, 59, 77]. Formally speaking, let a model f be defined as a mapping between *inputs* $X \in \mathcal{X}$ and *outputs* $Y \in \mathcal{Y}$ where $(\mathcal{X}, \mathcal{Y})$ are two metric spaces:

$$Y = f(X).$$

In an ML context, f can be understood as a *predictive model* (e.g., penalized linear regression, neural network) linking an observation x of an input X to a prediction $y = f(x)$ [46]. In the UQ framework, a

so-called *computer model* f represents the numerical implementation of a hypothetical-deductive link (e.g., by systems of ordinary differential equations, by finite element methods) between X and Y [80].

In both fields, the inputs X are generally assumed to be random, leading to random outputs Y. Let P be the distribution of X. In the ML context, P is defined implicitly by an empirical measure: given a set of observations $x^{(1)}, \dots, x^{(n)} \in \mathcal{X}$,

$$P = \frac{1}{n} \sum_{i=1}^{n} \delta_{x^{(i)}} \tag{1}$$

where δ denotes the Dirac measure. On the other hand, in the UQ setting, P is often explicitly chosen based on observations of X, expert assessment (domain knowledge), or stochastic inversion from observations of Y [85].

Quantities of interest (QoI) are key in measuring the relationships between X and Y. They are generally expressed as statistics of Y. In the SA literature, they are often referred to as *scores* [76], while ML researchers are usually interested in *predictive performance metrics* [66]. QoIs can be either global (e.g., variance [81, 35], loss metric [43, 22, 49]), or local (e.g., a prediction instance [84, 90], numerical model derivatives [24]). These quantities are usually chosen to be interpretable in that domain experts or decision-makers can understand the information they bear.

Diagnostics are evaluations of interpretable quantities related to specific QoI. They can be the evaluation of QoIs themselves or based on their decomposition [47]. For instance, evaluations of local interpretation tools such as SHAP [58] and LIME [74] are diagnostics, as well as global SA methods such as Sobol' indices [81] or Shapley effects [65].

The present paper is concerned with one particular problem: robustness to input perturbations. More precisely, the main goal is to study changes in key diagnostics of Y whenever P is perturbed. This main problem is analogous to many frameworks in both the ML field (e.g., domain adaptation [16], covariate shift [44, 88], adversarial attacks [4]) and SA (e.g., distributional sensitivity analysis [2, 62], distributional robustness [56, 41]) or distributional modifications to understand the fairness of algorithms [26, 27]. In the context of this work, the perturbations are subject to four desirability criteria:

- *Interpretable*, i.e., can be understood by domain experts and decision-makers;
- *Generic*, i.e., the overall perturbation scheme should not depend on properties of either P or f;
- *Proximity*, i.e., the perturbed distribution should be as "close" to P as possible;
- *Exploration*, i.e., the perturbed distribution should allow exploration of unobserved or low probability regions of $\mathcal{X}$.

Formally, one can define the perturbed distribution Q as the solution to the projection problem:

$$\begin{aligned} Q = \underset{G \in \mathcal{P}(\mathcal{X})}{\operatorname{argmin}} \quad & \mathcal{D}(P, G) \\ \text{s.t.} \quad & G \in \mathcal{C}. \end{aligned} \tag{2}$$

where $\mathcal{P}(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$, $\mathcal{D}$ is a discrepancy between probability measures, and $\mathcal{C} \subset \mathcal{P}(\mathcal{X})$ is a *perturbation class*, i.e., a particular subset of probability measures. Leveraging the pioneering work of [23] on entropic projections, a particular instance of this problem has been studied in the SA field by [56] and in the ML field by [6], where the chosen discrepancy is the Kullback-Leibler (KL) divergence, and $\mathcal{C}$ is defined through constraints on generalized moments. While this method produces interpretable perturbed distributions that are close (in the KL sense) to P, they do not allow for exploration and genericity: the resulting perturbed distribution is a linear reweighting of P, and the existence of particular generalized moments must be assumed, further restricting the perturbation class $\mathcal{C}$.

Motivated by these four desirability criteria and by improving the connections between the interpretability analyses conducted in UQ and ML settings, the present work is focused on the following choices:

- $\mathcal{X} \subseteq \mathbb{R}^d$ for a positive integer d and $\mathcal{Y} \subseteq \mathbb{R}$.
- The 2-Wasserstein distance as a suitable discrepancy between probability measures to ensure genericity and exploration;
- Perturbation classes $\mathcal{C}$ based on three types of constraints:
  1. Interpolation constraints on generalized quantile functions to ensure interpretability and genericity;
  2. Smoothness of the generalized quantile functions to ensure exploration;
  3. Copula-preservation to ensure interpretability.

Several results are uncovered and presented. This particular perturbation problem reduces to solving univariate constrained projections of quantiles functions in L^2 (see Lemma 5), and even admits an analytical result if no smoothness restrictions are enforced (see Proposition 1). However, this closed form does not satisfy the exploration criterion. To that extent, the use of isotonic piece-wise polynomials to ensure continuity is studied and is shown to lead to a well-posed quadratic program with convex constraints (see Theorem 1), ensuring practical feasibility. Aside from theoretical results, the computational tractability of this methodology is studied and implemented in two use cases from the ML and UQ fields, where the response of several diagnostics is studied, leading to robustness to input perturbation insights on the black-box model.

This article is organized as follows. In Section 2, preliminaries are presented, and the motivations behind the four perturbation criteria are discussed. Section 3 is dedicated to perturbation classes. The desirability criteria are discussed, as well as the three constraints introduced above. Section 4 presents the framework of probability measure projection using the 2-Wasserstein distance and its declination when constrained to the chosen perturbation class. Section 5 showcases insights on ML and UQ fields, highlighting local and global robustness insights. A discussion section ends this article, opening avenues for improvement. All proofs of technical results are postponed to a dedicated appendix.

2. Preliminaries and motivation for the perturbation criteria

2.1. Motivations for the perturbation criteria

The general question that the proposed method aims to answer is:

What are the variations of a black-box model's diagnostics induced by a given perturbation of its inputs?

Answering this question entails uncovering a causal link (in the physical sense) between a perturbation and the behavior of the black-box model. In the literature, many methods have been proposed in order to define relevant perturbations (e.g., via geodesics on Fréchet manifolds [41], adversarially [60], using empirical quantiles [6]). However, while generic and automatic, these methods often disregard the physical meaning of these perturbations. The overall aim of the proposed criterion is to ensure that the perturbations are *meaningful* to the eyes of domain experts and decision-makers. For instance, perturbations can be used as proxies for epistemic uncertainty, leading to exploratory studies on the behavior of a model induced by a lack of knowledge. Another example would be to prospectively design perturbations to anticipate future changes in the inputs (e.g., , climate change). Finally, if a gap between some observed data and domain experts' opinions is proven, perturbations can be modeled to enforce this knowledge while keeping some of the empirical information gathered on the field.

2.1.1. Interpretability

The perturbations should be meaningful to domain experts and decision-makers. It ensures that well-understood phenomena induce the uncovered variations in the model's behavior. Hence, designing perturbations should be done with practitioners and precisely reflect a domain-specific question. In-fine, perturbation interpretability ensures that the (physical) causal link one aims to draw of a perturbation on the behavior of a model is insightful on the question at stake.

2.1.2. Genericity

The perturbations should be generic because they should not depend on restrictive properties assumed to hold for either f (e.g., continuity, derivability) or P (e.g., absolute continuity). Genericity ensures the proposed methodology is *post-hoc* [8]. To emphasize the duality between SA and ML interpretability [72, 50], generic perturbation ensures that the proposed methodology is usable in both settings.

2.1.3. Proximity

The perturbed distribution should be "close" to the initial distribution P. Proximity ensures that the perturbed distribution remains somewhat similar to the initial, where similarity needs to be measured through a discrepancy. For instance, closeness in the KL divergence sense entails similar information, whereas closeness in the Wasserstein distance sense has a more geometric meaning. Either way, the initial distribution, be it empirical or chosen, bear some information on the behavior of the input, which needs to be partially preserved.

2.1.4. Exploration

The perturbation scheme should allow for exploring unobserved or low probability regions of $\mathcal{X}$. This criterion ensures that "out of distribution" scenarios can be reached. Hence, the model's behavior can be assessed on "unusual" (for P) evaluations, which is crucial when testing for robustness.

2.1.5. Assumptions and notations

In the following, X is an $\mathcal{X}$-valued random vector, where $\mathcal{X} \subseteq \mathbb{R}^d$, and $Y = f(X)$ is an $\mathbb{R}$-valued random variable. Denote by $\mathcal{P}(\mathcal{X})$ the set of probability measures defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Let p be a positive integer and denote:

$$\mathcal{P}_p(\mathbb{R}) = \left\{ P \in \mathcal{P}(\mathbb{R}) : \mathbb{E}_{|X|^p} [<] \infty, X \sim P \right\},$$

the set of univariate distributions of random variables with finite p-th moment. Suppose that $X \sim P$ for some $P \in \mathcal{P}(\mathcal{X})$, and denote $X_i \sim P_i$ the i-th univariate marginals of X, $i = 1, \ldots, d$. It is assumed that $P_i \in \mathcal{P}_2(\mathbb{R})$, i.e., every univariate marginal of X has a finite variance. In the remainder of the present work, P is the *initial* probability measure.

Remark 1. *In practice, the initial probability measure can be defined as an explicit distribution (e.g., from a parametric family) or empirically using a dataset as in* (1)*.*

For a univariate marginal input X_i, $1 \leq i \leq d$, let $\Omega_{X_i} \subset \mathbb{R}$ be its *application domain*. It represents the range in which X_i is intended to vary in practice [75]. Figure 1 illustrates a typical situation for a univariate marginal of X.

Remark 2. *In practice, the application domains of marginal distribution can be defined in many ways. For instance, if P is empirical, it can represent the range between the smallest and largest observed value of X_i in a specific dataset. If P is part of a parametric family, it can be defined using experts' opinions, usually enforced using truncation. These domains are usually subject to uncertainties in their bounds.*

For any univariate marginal $P_i \in \mathcal{P}(\mathbb{R})$, its *cumulative distribution function* (cdf) is denoted by:

$$F_P(t) = \int_{(-\infty,t]} dP = P\big((-\infty, t]\big).$$

Furthermore, denote $\mathcal{F}$ the space of univariate distribution functions:

$$\mathcal{F} = \Big\{ F : \mathbb{R} \to [0,1] \mid F \text{ is right-continuous, non-decreasing} \\ \text{such that } \lim_{x\to\infty} F(x) = 1 \text{ and } \lim_{x\to-\infty} F(x) = 0 \Big\}. \tag{3}$$

2.2. Preliminaries

2.2.1. Generalized quantile functions

The use of generalized quantile functions (gqf) is motivated by the fact that the marginal distributions P_i can be atomic. They rely on the two generalized inverses of functions in $\mathcal{F}$. For each marginal probability measure P_i, one can define a left and right continuous generalized inverse, the former being usually called the gqf of X_i. However, in the following, both generalized inverses are of interest. They can be formally defined as follows [73, 31, 52].

Definition 1 (Generalized quantile function)**.** *Let $P \in \mathcal{P}(\mathbb{R})$ with cdf F_P.*

(i) *The gqf of P is the unique left-continuous, non-decreasing generalized inverse of F_P, defined, for every $a \in (0,1)$, as:*

$$\begin{aligned} F_P^{\leftarrow}(a) &= \sup \ \{t \in \mathbb{R} \mid F_P(t) < a\}, \\ &= \inf \ \{t \in \mathbb{R} \mid F_P(t) \geq a\}. \end{aligned} \tag{4}$$

(ii) *The unique right-continuous non-decreasing generalized inverse $F_P^{\rightarrow}$ of F_P, almost-everywhere equal to $F_P^{\leftarrow}$, is defined, for every $a \in (0,1)$, as:*

$$\begin{aligned} F_P^{\rightarrow}(a) &= \sup \ \{t \in \mathbb{R} \mid F_P(t) \leq a\}, \\ &= \inf \ \{t \in \mathbb{R} \mid F_P(t) > a\}, \\ &= F_P^{\leftarrow}\left(a^+\right) \end{aligned} \tag{5}$$

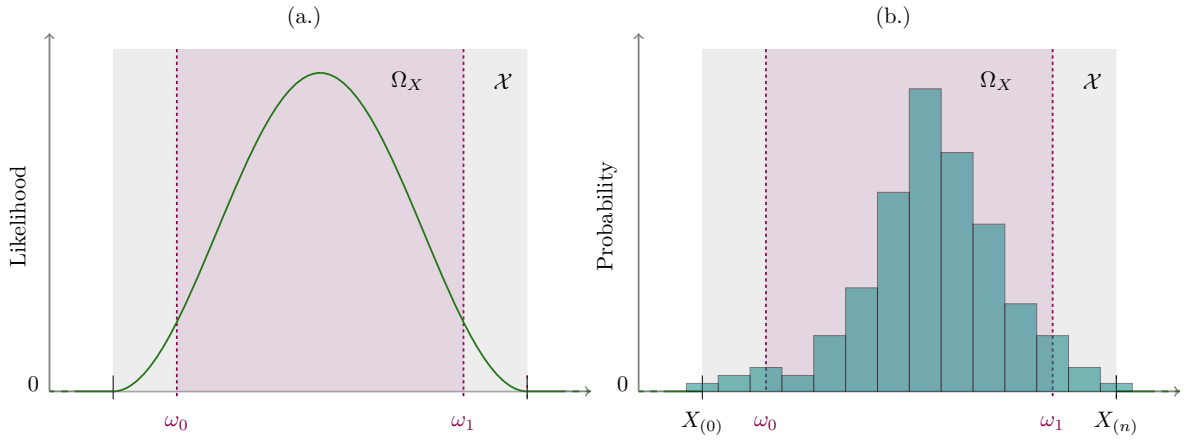

Figure 1: Application domain Ω_X of P when P admits a density (a.) and when it is empirical (b.). In (a.), $\mathcal{X}$ is the support of the density (in grey), and the application domain Ω_X (in purple) is contained in $\mathcal{X}$. In (b.), $\mathcal{X}$ is the interval between the minimum and maximum observed values (in grey), and the application domain Ω_X (in purple) is also contained in $\mathcal{X}$. In both cases, Ω_X is chosen to be strictly included in $\mathcal{X}$, although it can be bigger.

where $F_P^{\leftarrow}(a^+) = \lim_{x \to a^+} F_P^{\leftarrow}(x)$.

If the cdf F_{P_i} of X_i admits an inverse F_P^{-1} in the traditional sense (e.g., it is continuous, strictly increasing), then the following equality holds:

$$F_P^{-1} = F_P^{\leftarrow} = F_P^{\rightarrow}.$$

Furthermore, univariate probability measures are intrinsically linked to their gqf. Denote:

$$\mathcal{F}^{\leftarrow} = \left\{ F^{\leftarrow} : (0,1) \to \mathbb{R} \mid F^{\leftarrow} \text{ is left-continuous and non-decreasing} \right\}. \tag{6}$$

the space of gqfs. Recall that each probability measure in $\mathcal{P}(\mathbb{R})$ has a unique gqf in $\mathcal{F}^{\leftarrow}$ [73].

For a fixed $\alpha \in [0,1]$, an α-quantile of P is a number $p_\alpha \in \mathbb{R}$ such that, for $X \sim P$:

$$P\left(\{X < p_\alpha\}\right) \leq \alpha \quad \text{and} \quad P\left(\{X \leq p_\alpha\}\right) \geq \alpha.$$

In certain cases, α-quantiles are not unique. For instance, if P is purely atomic (e.g., an empirical measure), and its cdf F_P takes the constant value α on an open interval (t_0, t_1) (i.e., it is the case if t_0 and t_1 are both atoms of an empirical probability measure), then any $t \in (t_0, t_1)$ is an α-quantile. One can notice that $F^{\leftarrow}(\alpha)$ is the *infimum* of the α-quantiles of P, (i.e., $F_P^{\leftarrow}(\alpha) = t_0$), and $F^{\rightarrow}(\alpha)$ is the *supremum* of the α-quantiles of P (i.e., $F_P^{\rightarrow}(\alpha) = t_1$).

2.2.2. Copulas

Dependencies between random variables can be modeled using copula-based representations [63]. Let $X = (X_1, \ldots, X_d) \sim P$ be a d-dimensional $\mathbb{R}^d$-valued random vector with marginal cdfs F_{P_i}, $i = 1, \ldots, d$, assumed to be continuous. Let $U_1, \ldots, U_d$ the random variables defined as:

$$U_i = F_{P_i}(X_i)$$

and denote $U = (U_1, \ldots, U_d)^\top \sim U_P$. For any $\mathbf{u} = (u_1, \ldots, u_d) \in [0,1]^d$, denote $H_{\mathbf{u}} = \bigtimes_{i=1}^d [0, u_i]$. The copula of X is the mapping from $[0,1]^d$ to $[0,1]$, denoted C_p defined as:

$$\begin{aligned} C_P(u) &= Pr\left(U_1 \leq u_1, \ldots, U_d \leq u_d\right) \\ &= \int_{H_{\mathbf{u}}} dU_P \end{aligned}$$

If P is observed (and hence each F_{P_i} can jump), the notion of *empirical copula* characterizes the dependence structure between the inputs [63]. For $j \in \{1, \ldots, d\}$, denote $\{x_{j,i}\}_{1 \leq i \leq n}$ the jth marginal sample of observations. The empirical copula of X is defined as:

$$\hat{C}_P(u_1, \ldots, u_d) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \mathbb{1}_{\left\{ \frac{R_{j,i}}{n} \leq u_j \right\}}(u_j), \tag{7}$$

where $R_{j,k}$ denotes the rank of $x_{j,k}$ in $\{x_{j,i}\}_{1 \leq i \leq n}$.

2.2.3. Wasserstein distance

Let p be a positive integer. The p-Wasserstein distance between two univariate marginals can be defined as follows [89]:

Definition 2 (Wasserstein distance on the real line)**.** *Let $p \in \mathbb{N}^*$ and $P, Q \in \mathcal{P}_p(\mathbb{R})$ be two probability measures on $\mathbb{R}$ admitting F_P and F_Q as probability distribution functions, respectively. Then, the p-Wasserstein distance between P and Q is:*

$$W_p(P, Q) = \left(\int_0^1 \left| F_P^{\rightarrow}(x) - F_Q^{\rightarrow}(x) \right|^p dx \right)^{1/p}$$

In particular, for $p = 2$,

$$W_2(P, Q) = \sqrt{\int_0^1 \left(F_P^{\rightarrow}(x) - F_Q^{\rightarrow}(x)\right)^2 dx}.$$

3. Quantile constrained Wasserstein projections

3.1. Quantile perturbation classes

3.1.1. Motivations

First, as long as $P_i \in \mathcal{P}(\mathbb{R})$, its *generalized quantile function* $F_{P_i}^{\leftarrow}$ *always exists*. Hence, perturbing marginal quantiles do not require additional assumptions on the initial probability measure P or the shape of the target perturbed probability measure Q. It ensures that the proposed methodology is generic, in contrast to the one proposed in [55] based on generalized moments.

Second, *quantiles are interpretable*. In many applied problems, quantile specifications are often key to studying the influence of input variables on a decision-making output. Beyond the fact that quantiles have a decision-theoretical sense through pinball cost functions [20], numerous applications dealing with economic stress tests or risk mitigation against natural hazards use quantiles as influential inputs of decision-helping models. For instance, in the drought risk studies in [33], the association between soil wetness, climatic, seismic, and socioeconomic variables is often carried out using marginal quantiles that are features for predictive cost models. Input variations of daily value-at-risk percentiles, computed from legacy data, were recently required by the European Banking Authority for generating macroeconomic scenarios used for EU-wide stress tests [5]. Reverse SA studies for financial risk management, such as those conducted in [70], are primarily based on moving values-at-risk, which are quantiles.

The following examples offer additional concrete illustrations of using quantiles for influence analysis. They also illustrate two quantile perturbation schemes: quantile shifting and application domain dilatation. These schemes are formally introduced in Section 3.1.3.

Example 1 (Economic stress test (Inspired by [13]))**.** *Assume that an economic shock happens in an abstract country. Prospective analyses announce a* \$200 *drop in the population median wage. Before the shock, the population wage distribution P is known (or observed), thanks to some annual census data. This distribution has a median wage of* \$2000*. The new population wage distribution is unknown due to the lack of recent data. The economists want to know if they can be confident in their predictive macro-economic model f w.r.t. this sudden change. A way to answer this problem would be assessing the behavior of the model f on a distribution Q, such that:*

$$F_Q^{\leftarrow}(0.5) = 1800.$$

Example 2 (River water level)**.** *This example is inspired from [51] and more deeply studied in Section 5.2. The safety of an industrial site located near a river is studied through the prediction of the water level $Y = f(X)$ where f is a numerical hydrodynamic model, and X gathers the physical features of the river. A key dimension of X is the Strickler roughness coefficient for the upstream water level [40], which is modeled as a truncated Gaussian distribution on $\Omega_X = [20, 50]$. However, this application domain is tainted with epistemic uncertainties on the actual nature of the riverbed (e.g., more or less subject to shrubby vegetation). The practical use of f would require assessing its predictive power under a wider interval $\Omega_X = [5, 65]$. A way to express this prospective study is to assess the model's behavior on a distribution Q, such that:*

$$F_Q^{\rightarrow}(0) = 5, \quad F_Q^{\leftarrow}(1) = 65.$$

3.1.2. Formal Definition

Since, for a fixed $\alpha \in [0,1]$, α-quantiles are not necessarily unique, equality constraints on quantile functions seem somewhat arbitrary. It amounts to constraining the infimum of the set of α-quantitles. Arguably, given a desired α-quantile value of $b \in \mathbb{R}$, a reasonable constraint would be for b to be in the *set of α-quantils* of the perturbed distribution. Formally, the any perturbed distribution $Q \in \mathcal{P}(\mathbb{R})$ should respect the inequality:

$$F_Q^{\leftarrow}(\alpha) \geq b \geq F_Q^{\leftarrow}(\alpha^+) = F_Q^{\rightarrow}(\alpha). \tag{8}$$

In the case where F_Q is invertible, it becomes a traditional equality constraint: any α-quantile is uniquely defined (i.e., $F_Q^{\leftarrow}(\alpha) = F_Q^{\rightarrow}(\alpha)$). In the following, the inequality constraints of the form (8) are referred to as *quantile constraints*.

Definition 3 (Quantile perturbation class)**.** *Let K be an integer, and let $\alpha = (\alpha_1, \dots, \alpha_K)^\top \in [0,1]^K$ and $b = (b_1, \dots, b_K)^\top \in \mathbb{R}^K$. The* quantile perturbation class $\mathcal{Q}(\alpha, b) \subseteq \mathcal{P}(\mathbb{R})$ *is the set of probability measures defined as:*

$$\mathcal{Q}(\alpha, b) = \left\{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow}(\alpha_i) \leq b_i \leq F_Q^{\rightarrow}(\alpha_i), \quad i = 1, \dots, K \right\}.$$

An equivalent characterization, thanks to the uniqueness of gqfs, is:

$$\mathcal{Q}(\alpha, b) = \left\{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow} = L \in \mathcal{F}^{\leftarrow}, L(\alpha_i) \leq b_i \leq L(\alpha_i^+), \quad i = 1, \dots, K \right\}.$$

It is possible to derive sufficient conditions on α and b in order for $\mathcal{Q}(\alpha, b)$ to be non-empty:

Lemma 1. *Let $\alpha \in [0,1]^K$ and $b \in \mathbb{R}^K$, which are assumed to be ordered without loss of generality. If*

$$0 \leq \alpha_1 < \dots < \alpha_K \leq 1, \quad \text{and} \quad b_1 < \dots < b_K, \tag{9}$$

then $\mathcal{Q}(\alpha, b)$ is non-empty.

Quantile perturbation classes contain probability measures with discontinuous gqfs. Ensuring smooth perturbed gqfs is of practical interest (see Section 4.1.1). It entails further restricting the gqfs of the probability measures in a quantile perturbation class to respect some smoothness conditions. They can be formally defined as follows.

Definition 4 (Smooth quantile perturbation class)**.** *Let $K \in \mathbb{N}^*$, $\alpha = (\alpha_1, \dots, \alpha_K)^\top \in [0,1]^K$, $b = (b_1, \dots, b_K)^\top \in \mathbb{R}^K$ and let $\mathcal{V} \subseteq \mathcal{F}^{\leftarrow}$ be a given set of smooth non-decreasing functions. The* smooth quantile perturbation class $\mathcal{Q}_{\mathcal{V}}(\alpha, b) \subseteq \mathcal{P}(\mathbb{R})$ *is the set of probability measures defined as:*

$$\mathcal{Q}_{\mathcal{V}}(\alpha, b) = \left\{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow} \in \mathcal{V}, F_Q^{\leftarrow}(\alpha_i) \leq b_i \leq F_Q^{\rightarrow}(\alpha_i), \quad i = 1, \dots, K \right\}.$$

Note that smooth perturbation classes generalize perturbation classes since $\mathcal{Q} = \mathcal{Q}_{\mathcal{F}^{\leftarrow}}$.

3.1.3. Defining interpretable sets of quantile perturbation classes

Two sets of quantile perturbation classes are introduced: quantile shifts and application domain dilatation.

Quantile shifts.. Quantile shift perturbations defines constraints on an initial α-quantile in a predetermined range. Formally, given a quantile level $\alpha \in [0,1]$, and an initial α-quantile $p_\alpha = F_P^{\leftarrow}(\alpha)$, quantile shifts defines a set of quantile perturbations classes of probability measures having their α-quantiles ranging over a compact interval $[\eta_0, \eta_1] \subseteq \Omega_X$ such that $\eta_0 < p_\alpha < \eta_1$. In other words, for each $b_\alpha \in [\eta_0, \eta_1]$, a quantile perturbation class $\mathcal{Q}_{\mathcal{V}}(\alpha, b_\alpha)$ can be constructed. This particular type of set of quantile perturbation classes can be described by means of a *perturbation intensity* $\theta \in [-1, 1]$:

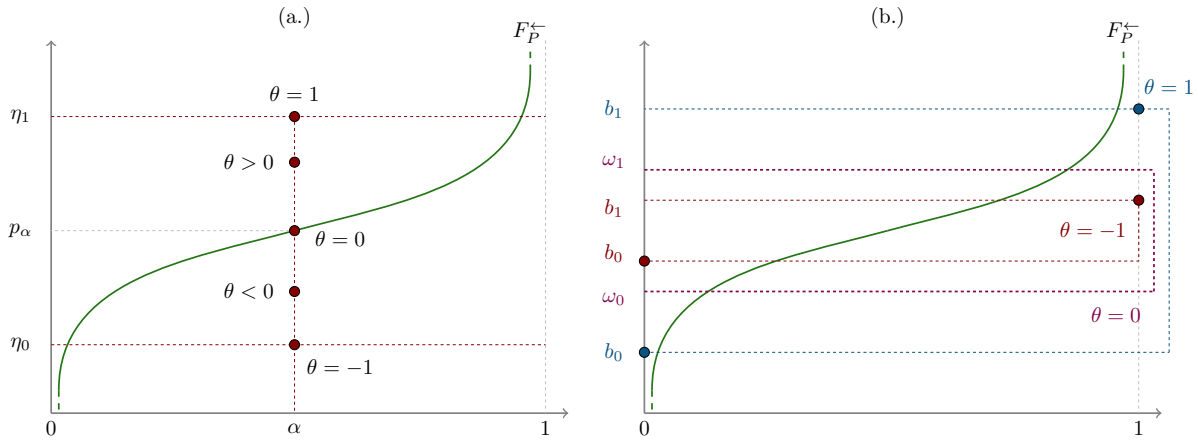

Figure 2: Quantile shift (a.) and application domain dilatation (b.) perturbation schemes. The initial quantile function is displayed in green. On the left, red points indicate different quantile shifting constraints between η_0 and η_1, leading to different intensity values θ. On the right, the application domain's width (in magenta) is up to doubled (blue points) or down to halved (red points), according to an intensity parameter $\theta \in [-1, 1]$.

Lemma 2. *Let $\Theta = [-1, 1]$ and denote $\boldsymbol{\eta} = (\eta_0, \eta_1)$ with $\eta_0 < p_\alpha < \eta_1$. For $\theta \in \Theta$, let,*

$$b_\alpha(\boldsymbol{\eta}, \theta) = \begin{cases} p_\alpha(1+\theta) - \theta\eta_0 & \text{if } -1 \leq \theta < 0, \\ p_\alpha & \text{if } \theta = 0, \\ p_\alpha(1-\theta) + \theta\eta_1 & \text{if } 0 < \theta \leq 1. \end{cases}$$

Then, for any $Q \in \mathcal{P}(\mathbb{R})$ such that

$$F_Q^{\leftarrow}(\alpha) \geq b_\alpha(\boldsymbol{\eta}, \theta) \geq F_Q^{\rightarrow}(\alpha),$$

one has that, $\forall \theta \in [-1, 1]$:

$$\begin{aligned} \theta = -1 &\Leftrightarrow b_\alpha(\boldsymbol{\eta}, \theta) = \eta_0, \\ \theta = 0 &\Leftrightarrow b_\alpha(\boldsymbol{\eta}, \theta) = p_\alpha, \\ \theta = 1 &\Leftrightarrow b_\alpha(\boldsymbol{\eta}, \theta) = \eta_1, \end{aligned} \tag{10}$$

and for any $-1 \leq \theta_1 < \theta_2 \leq 1$,

$$b_\alpha(\boldsymbol{\eta}, \theta_1) < b_\alpha(\boldsymbol{\eta}, \theta_2).$$

In other words, $b_\alpha(\boldsymbol{\eta}, \theta) \in [\eta_0, \eta_1]$ is a strictly increasing function of θ and $\theta = 0$ indicates that p_α must remain untouched (i.e., no constraint). Figure 2 (a.) illustrates this perturbation scheme. Quantile shifts are formally defined as the collection of perturbation classes $\{\mathcal{T}(\boldsymbol{\eta}, \theta)\}_{\theta \in [-1,1]}$ where,

$$\begin{aligned} \mathcal{T}(\boldsymbol{\eta}, \theta) &= \left\{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow}(\alpha) \leq b_\alpha(\boldsymbol{\eta}, \theta) \leq F_Q^{\rightarrow}(\alpha) \right\} \\ &= \mathcal{Q}\left(\alpha, b_\alpha(\boldsymbol{\eta}, \theta)\right) \end{aligned} \tag{11}$$

Application domain dilatations.. *Application domain dilatations* consists in perturbing the bounds of the application domain of a marginal input. For a univariate $X \sim P$ with $\Omega_X = [\omega_0, \omega_1]$, the dilatation process amounts to widening or narrowing the width (or diameter $\text{diam}(\Omega_X)$) of Ω_X. It amounts constraining the extreme quantiles ($\alpha \in \{0, 1\}$) while preserving the midpoint of Ω_X. The

dilatation is characterized by a parameter $\eta > 1$ controlling the rescaling magnitude of Ω_X. In other words, one aims at finding a distribution Q with support $\text{Supp}(Q) = [b_0, b_1]$ for $b_0, b_1 \in \mathbb{R}$, $b_0 < b_1$, where the midpoint of $[b_0, b_1]$ is equal to the midpoint of Ω_X, but such that $\text{diam}(Q) := \text{diam}(\text{Supp}(Q))$ is rescaled compared to $\text{diam}(\Omega_X)$. Similarly to quantile shift, the next lemma formalizes expressions for these two bounds as a function of a perturbation intensity $\theta \in [-1, 1]$.

Lemma 3. *Let $\eta > 1$. For $\theta \in [-1, 1]$, let:*

$$b_0(\eta, \theta) = \begin{cases} \frac{1}{2}\left(\omega_0\big(2 - \theta(\eta^{-1} - 1)\big) + \theta\omega_1(\eta^{-1} - 1)\right) & \text{if } -1 \leq \theta < 0, \\ \omega_0 & \text{if } \theta = 0, \\ \frac{1}{2}\left(\omega_0\big(2 + \theta(\eta - 1)\big) - \theta\omega_1(\eta - 1)\right) & \text{if } 0 < \theta \leq 1, \end{cases}$$

$$b_1(\eta, \theta) = \begin{cases} \frac{1}{2}\left(\omega_1\big(2 - \theta(\eta^{-1} - 1)\big) + \theta\omega_0(\eta^{-1} - 1)\right) & \text{if } -1 \leq \theta < 0, \\ \omega_1 & \text{if } \theta = 0, \\ \frac{1}{2}\left(\omega_1\big(2 + \theta(\eta - 1)\big) - \theta\omega_0(\eta - 1)\right) & \text{if } 0 < \theta \leq 1. \end{cases}$$

Then, $\forall(\theta, \eta) \in [-1, 1] \times [1, \infty)$,

$$b_0(\eta, \theta) + b_1(\eta, \theta) \quad = \quad \omega_0 + \omega_1 \quad \textit{(midpoints equality)}.$$

Denote $\boldsymbol{b}(\eta, \theta) = [b_0(\eta, \theta), b_1(\eta, \theta)]$, and notice that

$$\begin{aligned} \theta = -1 \Leftrightarrow \textit{diam}\left(\boldsymbol{b}(\eta, \theta)\right) &= \frac{\textit{diam}(\Omega_X)}{\eta}, \\ \theta = 0 \Leftrightarrow \textit{diam}\left(\boldsymbol{b}(\eta, \theta)\right) &= \textit{diam}(\Omega_X), \\ \theta = 1 \Leftrightarrow \textit{diam}\left(\boldsymbol{b}(\eta, \theta)\right) &= \eta\,\textit{diam}(\Omega_X), \end{aligned} \tag{12}$$

and for any $-1 \leq \theta_1 < \theta_2 \leq 1$,

$$\textit{diam}\left(\boldsymbol{b}(\eta, \theta_1)\right) < \textit{diam}\left(\boldsymbol{b}(\eta, \theta_2)\right).$$

In other words, $\text{diam}\left(\boldsymbol{b}(\eta, \theta)\right) \in \left[\eta^{-1}\text{diam}(\Omega_X), \eta\text{diam}(\Omega_X)\right]$ is a strictly increasing function of θ, and for $\theta = 0$, one has that $\boldsymbol{b}(\eta, \theta) = \Omega_X$, i.e., the application domain is not perturbed.

Figure 2 (b.) illustrates this perturbation scheme. The initial application domain is displayed in magenta and is subject to a dilatation of parameter $\eta = 2$. The red constraints halve its width, and the blue constraints double it. One can additionally check that in both cases, the midpoint of the original validity domain is preserved. Application domain dilatations are formally defined as the collection of perturbation classes $\{\mathcal{T}(\eta, \theta)\}_{\theta \in [-1,1]}$ where,

$$\begin{aligned} \mathcal{T}(\boldsymbol{\eta}, \theta) &= \left\{Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow}(m) \leq b_m(\eta, \theta) \leq F_Q^{\rightarrow}(m), m \in \{0, 1\}\right\} \\ &= \mathcal{Q}\left((0, 1)^\top, (b_0(\eta, \theta), b_1(\eta, \theta))^\top\right) \end{aligned} \tag{13}$$

Many perturbation settings can be defined by combining quantile shifts and domain dilatations. However, for the sake of simplicity, quantile shifts and domain dilatations are studied independently in Section 5.

3.2. Copula preservation and marginal quantile perturbations

3.2.1. Motivations

Regarding multivariate perturbations in general, independence assumptions are often required [56]. While it facilitates mathematical calculations, it is questionable in practice. One of the main challenges

in ML interpretability and SA is to account for the potential dependence structure between the inputs (or features) [71].

Dependencies provide helpful information on the global behavior of the inputs. In SA, the dependence structure is often chosen after extensive studies [28], and expresses the physical relationship between the uncertainties on the inputs. In ML, it can be argued that preserving dependencies avoids creating meaningless patterns [10] and is critical in some practical studies [57, 69]. Dependencies between random variables can be modeled using copula-based representations [63], which relaxes the framework of probabilistic graphical models usually encountered in ML [34].

From the *interpretability* standpoint, in practice, the intricacies of multivariate insights due to stochastic dependence are much more complicated to grasp. Moreover, many of the properties presented above do not hold regarding multivariate quantile functions. The definition itself of multivariate quantile functions is a highly non-trivial task. Many interesting approaches have been recently proposed [19, 45]. However, they lack the broad adoption of their univariate counterpart in practice, which makes them less interpretable.

In order to ensure the *interpretability*, the proposed methodology is restricted to:

- Quantile perturbations on marginal inputs.
- Perturbed probability measures having the same copula as the initial probability measure.

3.2.2. Marginal perturbation maps and copula preservation

Let $X \sim P$ and for $i = 1, \dots, d$, let each marginal input $X_i \sim P_i$ and $(F_i^{\leftarrow})_{i=1,\dots,d}$ be a collection of quantile functions in $\mathcal{F}^{\leftarrow}$. A *marginal perturbation map* is a mapping:

$$T : \begin{array}{ccc} \mathcal{X} & \to & \mathcal{X} \\ \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} & \mapsto & \begin{pmatrix} T_1(x_1) \\ \vdots \\ T_d(x_d) \end{pmatrix} \end{array} \tag{14}$$

where

$$T_j = \left[F_j^{\leftarrow} \circ F_{P_j}\right], \quad j = 1, \dots, d.$$

Denote $\widetilde{X} := T(X)$ the *perturbed inputs*.

Lemma 4. *Suppose that each $F_i^{\leftarrow}$, $i = 1, \dots, d$ is strictly increasing:*

(i) *If P is an empirical measure then X and $\widetilde{X}$ have the same empirical copula.*

(ii) *If P is atomless then X and $\widetilde{X}$ have the same copula.*

Hence, perturbation maps composed of compositions of marginal cdfs and strictly increasing quantile functions preserve the copula. For instance, if P is an empirical measure related to an observed dataset, applying T to every observation results in a perturbed dataset with the same Spearman correlation matrix.

3.2.3. Copula-preserving multivariate perturbation classes

Combining quantile perturbation classes with marginal perturbation maps allows for defining multivariate perturbation classes, which are *generic* and *interpretable*. Let $X \sim P$, and for $i = 1, \dots, d$, let $\theta_i \in [0,1]^K \times \mathbb{R}^K$ and $\theta = (\theta_1, \dots, \theta_d)$. Finally, let $\mathcal{Q}^{(i)} := \mathcal{Q}(\theta_i)$ be the perturbation class associated with the input X_i. For $Q \in \mathcal{P}(\mathbb{R}^d)$, and denote $Q_1, \dots, Q_d$ its marginal distributions. Denote the set:

$$\mathcal{Q}_d(\theta) = \left\{ Q \in \mathcal{P}(\mathbb{R}^d) \mid Q_i \in \mathcal{Q}^{(i)} \right\},$$

and for any $Q \in \mathcal{P}(\mathbb{R}^d)$, denote T_Q the marginal perturbation map defined as:

$$\begin{array}{rccc} T_Q: & \mathcal{X} & \to & \mathcal{X} \\ & \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} & \mapsto & \begin{pmatrix} \left[F^{\leftarrow}_{Q_1} \circ F_{P_1}\right](x_1) \\ \vdots \\ \left[F^{\leftarrow}_{Q_d} \circ F_{P_d}\right](x_d) \end{pmatrix} \end{array} \tag{15}$$

Marginal quantile perturbation classes are defined as the set:

$$\mathcal{Z}(P, \theta) = \{Q \in \mathcal{Q}_d(\theta) \mid T_Q(X) \sim Q, X \sim P\},$$

and, from Lemma 4, *copula-preserving marginal quantile perturbation classes* are defined as:

$$\widetilde{Z}(P, \theta) = \left\{Q \in \mathcal{Z}(P, \theta) \mid F^{\leftarrow}_{Q_i} \text{ is strictly increasing}, i = 1, \ldots, d\right\}.$$

3.3. Wasserstein projections

3.3.1. Motivations

The Wasserstein distance is deeply rooted in optimal transportation theory [89] and has been used successfully in many ML and deep learning applications [38, 3]. It has also been extensively studied as a tool for guaranteeing distributional robustness to adversarial attacks in ML [30]. It has been used in SA to produce novel sensitivity indices [36, 14].

The 2-Wasserstein distance is *interpretable*. The choice of transportation cost as the squared distance is intrinsically linked to notions of the L^2 norms, which can be interpreted as lengths, analogous to the well-known Euclidean geometry [89]. It becomes natural and intuitive to quantify transportation costs as distances between points. It becomes even more natural in one dimension since the 2-Wasserstein distance can be interpreted as the absolute difference in areas between two quantile functions. Hence, *proximity* between two univariate probability measures, in the 2-Wasserstein sense, is rather natural.

Moreover, the 2-Wasserstein distance ensures *genericity*. The only requirement for two probability measures to be comparable is the finiteness of their variance. This assumption is classical in SA and ML interpretability. Compared to the KL divergence, which requires the absolute continuity of one probability measure versus the other and the existence of logarithmic moments, it appears less restrictive. In practice, it allows for more flexible perturbations: if P is an empirical measure (i.e., purely atomic), then Q is not restricted to be purely atomic; conversely, if P admits a density, then it does not restrict Q to admit a density. These benefits are key in unifying the frameworks of SA and ML interpretability: the flexibility of the 2-Wasserstein distance allows for greater explicit control (e.g., through smoothing restriction) on the resulting perturbed measure Q, independently of the properties of P.

Additionally, the 2-Wasserstein distance allows for *exploration*. Optimal transport maps between two probability measures w.r.t. the 2-Wasserstein distance are (usually) not linear [78]. In other words, perturbed solutions are not limited to the support of the initial probability measures: atoms can be added, and ranges with 0 probability can be made relevant.

3.3.2. Marginal quantile constrained Wasserstein projections

The problem of finding a probability measure Q closest to P, but $Q \in \widetilde{Z}(P, \theta)$ can be formalized as follows:

$$Q = \underset{G \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \; W_2^2(P, G) \quad \text{s.t.} \quad G \in \widetilde{Z}(P, \theta) \tag{16}$$

However, since the set of probability measures in $\widetilde{Z}(P, \theta)$ share the same copula as P, this problem can be simplified:

Lemma 5. *The perturbation map $T : \mathbb{R}^d \to \mathbb{R}^d$ that minimizes (16) is defined, for any $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, as:*

$$T(x) = \begin{pmatrix} \left[F^{\leftarrow}_{Q_1} \circ F_{P_1}\right](x_1) \\ \vdots \\ \left[F^{\leftarrow}_{Q_d} \circ F_{P_d}\right](x_d) \end{pmatrix}$$

where, for $i = 1, \dots, d$:

$$\begin{aligned} F^{\leftarrow}_{Q_i} = \underset{L \in L^2([0,1])}{argmin} & \left\{ \int_0^1 \left(L(x) - F^{\rightarrow}_{P_i}(x)\right)^2 dx \right\} \\ s.t. \quad & L(\alpha_j) \leq b_j \leq L\left(\alpha_j^+\right), \quad i = 1, \dots, K, \\ & L \text{ is strictly increasing.} \end{aligned} \tag{17}$$

where for $\alpha = (\alpha_1, \dots, \alpha_k)^\top$, $b = (b_1, \dots, b_k)^\top$, $\theta_i = (\alpha, b)$.

Hence, solving the projection problem in (16) is equivalent to solving the d problems of the form of (17).

3.3.3. Relaxed projection problem

Imposing that the resulting optimally perturbed marginal gqf be strictly increasing guarantees preserving the initial copula of the inputs. However, such constraints can lead to the non-existence of an optimum of (17) due to the non-closure of the set of strictly increasing functions [12]. To that extent, this work focuses on a relaxation of the problem in (17) to increasing functions, namely:

$$\begin{aligned} F^{\leftarrow} = \underset{L \in L^2([0,1])}{\text{argmin}} & \left\{ \int_0^1 \left(L(x) - F^{\rightarrow}_{P_i}(x)\right)^2 dx \right\} \\ \text{s.t.} \quad & L(\alpha_j) \leq b_j \leq L\left(\alpha_j^+\right), \quad i = 1, \dots, K, \\ & L \in \mathcal{V} \subseteq \mathcal{F}^{\leftarrow}. \end{aligned} \tag{18}$$

where $\mathcal{V}$ can be understood as a set of "smooth quantile functions" (see , Definition 4). Notice that this problem is indeed a relaxation of the initial problem. Indeed, if $\mathcal{V}$ is chosen as the set of strictly increasing functions, this problem becomes equivalent to (17).

Remark 3. *In practice, the relaxed problem (18) is frequently computationally easier to solve and can still lead to strictly increasing solutions.*

4. Solving the relaxed quantile perturbation problem

4.1. Relaxed problem with no smoothing

4.1.1. Analytical solution

The following proposition provides a convenient way to solve the perturbation problem (18) in the particular case of $\mathcal{V} = \mathcal{F}^{\leftarrow}$.

Proposition 1. *Let P be a probability measure in $\mathcal{P}_2(\mathbb{R})$. Let $\alpha \in [0,1]^K$ and $b \in \mathbb{R}^k$, such that $\alpha_1 < \dots < \alpha_K$ and $b_1 < \dots < b_K$, and $\mathcal{Q}(\alpha, b)$ the associated quantile perturbation class. Define the intervals $A_i = (c_i, d_i]$ for $i = 1, \dots, K$, such that:*

$$\begin{aligned} c_1 = \min(\beta_1, \alpha_1), \quad c_i = \min\Big[\max(\alpha_{i-1}, \beta_i), \alpha_i\Big], i = 2, \dots, K, \\ d_K = \max(\beta_K, \alpha_K), \quad d_j = \max\Big[\min(\beta_j, \alpha_{j+1}), \alpha_j\Big], j = 1, \dots, K-1. \end{aligned}$$

Let $A = \bigcup_{i=1}^{K} A_i$ *and* $\overline{A} = [0,1] \setminus A$. *Then the problem* (18) *where* $\mathcal{V} = \mathcal{F}^{\leftarrow}$ *has a unique solution which can be written as, for any* $y \in [0,1]$*:*

$$F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A}, \\ b_i & \text{if } y \in A_i, \quad i = 1, \dots, K. \end{cases} \tag{19}$$

4.1.2. Interpretation and solution

In order to interpret this result, illustrated in Figure 3, let us recall that when a quantile function is constant on an interval, it implies that its related probability measure admits an atom at the constant value taken by the gqf. Moreover, the mass allocated to this atom is equal to the length of the interval. Additionally, each jump of the quantile function induces an interval with no mass. The solution displayed in (19) shows that both initial and perturbed quantile functions are equal on $\overline{A}$. However, they differ on every interval A_i in the following fashion:

- Q have atoms at each constraint point b_i, $i = 1, \dots, K$;
- Each of these atoms have mass $Q(\{b_i\}) = d_i - c_i$, for $i = 1, \dots, K$;
- Each open interval $I_i \subset \mathbb{R}$ defined as

$$I_i = \begin{cases} \Big(\max(F_P^{\leftarrow}(\alpha_i), b_{i-1}), b_i\Big), & \text{when } b_i > F_P^{\leftarrow}(\alpha_i), \\ \Big(b_i, \min\left(b_{i+1}, F_P^{\leftarrow}(\alpha_i)\right)\Big), & \text{when } b_i < F_P^{\leftarrow}(\alpha_i) \end{cases} \tag{20}$$

  with, by convention, $b_0 = -\infty$ and $b_{K+1} = \infty$, has no mass. To put it briefly, $Q(I_i) = 0$ for every $i = 1, \dots, K$.

In other words, whenever an α-quantile p_α is shifted up to a value b, the perturbation entails sending every possible value in the range (p_α, b) to b. Hence, every value in (p_α, b) cannot be sampled according to Q. Moreover, the singleton $\{b\}$ now admits a probability of being observed equal to the initial probability of this interval, i.e., $Q(\{b\}) = P\big((p_\alpha, b)\big)$. When an α-quantile is shifted to b, the interval becomes (b, p_α), and the same reasoning can be done.

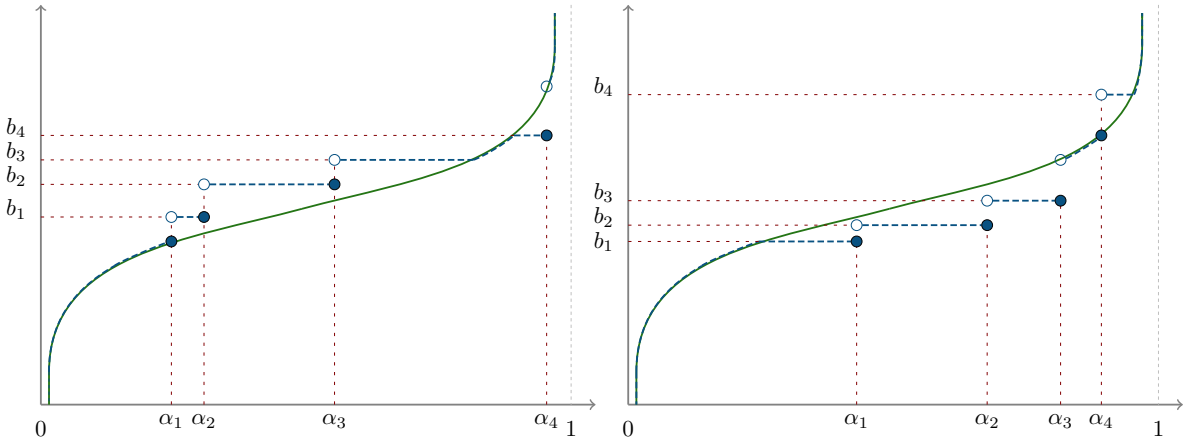

Figure 3: Characterizing quantile function of the solution of the perturbation problem (dashed blue). The initial quantile function (i.e., $F_P^{\leftarrow}$) is displayed in green, and dashed red lines identify the quantile constraints. (a.) and (b.) illustrate different possible perturbation configurations, increasing or decreasing several initial quantile values.

The statement of Proposition 1 is intuitive. Indeed, the Wasserstein distance quantifies the amount of *work* needed to transform a probability measure into another one [78]. When using W_2, the amount of work is quantified using the Euclidian distance, i.e., transporting a point x_0 to x_1 requires $(x_0 - x_1)^2$ units of work. This intrinsic *point-wise* way of quantifying similarities can be sensed in the previous result: perturbing an α-quantile entails giving the initial mass of an interval adjacent to b to the singleton $\{b\}$ in order to satisfy the constraint.

4.2. Isotonic piece-wise interpolating polynomial smoothing

The analytical solution provided in Proposition 1 presents a significant drawback: part of the application domain Ω_X of the perturbed input receives no mass, which hurts the perturbation *exploration* criteria. This result is because $\mathcal{F}^{\leftarrow}$ contains discontinuous functions. Ensuring continuity through a smooth perturbation class $\mathcal{Q}_{\mathcal{V}}$ where $\mathcal{V}$ is a set of continuous, non-decreasing functions can solve this issue.

4.2.1. Characterization of the problem

This section studies the projection of $F_P^{\leftarrow}$ onto a space of piece-wise continuous polynomials. It implies that the support of Q must be bounded. These bounds are made explicit using extremal quantile constraints (i.e., $F_Q^{\leftarrow}(0)$ and $F_Q^{\leftarrow}(1)$ are constrained to take finite values). Formally, the goal is to find a piece-wise polynomial of the form

$$G(x) = \begin{cases} G_0(x) & \text{if } \alpha_0 := 0 \leq x < \alpha_1, \\ \vdots & \\ G_i(x) & \text{if } \alpha_i \leq x < \alpha_{i+1}, \\ \vdots & \\ G_K(x) & \text{if } \alpha_K \leq x \leq 1 =: \alpha_{K+1}. \end{cases} \tag{21}$$

under the continuity constraints at each knot on the grid $\alpha_1 < \cdots < \alpha_K$, i.e.,

$$G_i(\alpha_{i+1}) = G_{i+1}(\alpha_{i+1}), \quad i = 0, \ldots, K-1.$$

Here, each $G_j \in \mathbb{R}[x]_{\leq p}$, for $j = 0, \ldots, K$, where $\mathbb{R}[x]_{\leq p}$ denotes the set of all real polynomials of degree at most equal to p. Let $\mathcal{S}_p$ denote the space of functions defined by (21). Restricting the solution of the perturbation problem (18) leads to the following optimization problem

$$\begin{aligned} F_Q^{\leftarrow} = \underset{L \in L^2([0,1])}{\operatorname{argmin}} & \left\{ \int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \right\} \\ \text{s.t.} \quad & L(\alpha_i) = b_i, \quad i = 1, \ldots, K, \\ & L \in \mathcal{F}^{\leftarrow} \cap \mathcal{S}_p. \end{aligned} \tag{22}$$

or, in other words, $\mathcal{V} = \mathcal{F}^{\leftarrow} \cap \mathcal{S}_p$ in the initial relaxed problem. Due to the piece-wise nature of polynomials in $\mathcal{S}_p$ defined on the $\alpha_0 < \alpha_1 < \cdots < \alpha_K < \alpha_{K+1} = 1$, solving (22) reduces to solve sub-problems on each sub-interval $[\alpha_i, \alpha_{i+1}], i = 0, \ldots, K$ of $[0, 1]$. (22) is indeed separable into $K+1$ independent optimization sub-problems. Each defines an optimal component G_i of the piece-wise polynomial G as defined in (21).

Any of these problems can be formulated generically as follows. Let $[t_0, t_1] \subset [0, 1]$, and $z_0, z_1 \in \mathbb{R}$ be interpolation values at t_0 and t_1 respectively. The goal is to find the solution to the optimization sub-problem

$$\begin{aligned} S = \underset{L \in \mathbb{R}[x]_{\leq p}}{\operatorname{argmin}} & \left\{ \int_{t_0}^{t_1} (F_P^{\leftarrow}(x) - L(x))^2 dx \right\} \\ \text{s.t. } & L(t_0) = z_0, L(t_1) = z_1, \\ & L'(x) \geq 0, \quad \forall x \in [t_0, t_1]. \end{aligned} \tag{23}$$

This optimization sub-problem is nothing more than the L^2 isotonic (i.e., monotonic, in this case non-decreasing) polynomial approximation on a compact interval [61], with interpolation constraints at the boundaries. The interpolating polynomials have been extensively studied in the literature [37], as well as isotonic polynomial regression and approximation [79, 91]. However, to our knowledge, this specific optimization problem does not seem to have been particularly studied.

A strategy for solving (23) is to use the *sum-of-squares* (SOS) [54] representation of nonnegative polynomials. These SOS representations can then be characterized using semi-definite positive (SDP) matrices [67, 68, 82]. A similar characterization of isotonic polynomials has been proposed in [82]. The following result shows that this optimization problem fits into the category of strictly convex programs: the solution of (26) is unique [12].

Theorem 1. *Let $[t_0, t_1] \subset [0, 1]$. Let M be the symmetric positive definite $((d+1) \times (d+1))$ moment matrix of the Lebesgue measure on $[t_0, t_1]$, i.e. for $i, j = 1, \ldots, d+1$,*

$$M_{ij} = \int_{t_0}^{t_1} x^{i+j-2} dx = \frac{(t_1)^{i+j-1} - (t_0)^{i+j-1}}{i+j-1}, \tag{24}$$

and denote $r \in \mathbb{R}^{d+1}$ the moment vector of $F_P^{\rightarrow}(x)$, i.e., for $i = 0, \ldots, d$

$$r_i = \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx. \tag{25}$$

Then, the vector $s^* = (s_0, \ldots, s_d)^\top \in \mathbb{R}^{d+1}$ of coefficients characterizing the polynomial S in (23) is the solution of the following convex constrained quadratic program

$$\begin{aligned} s^* = \underset{s \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \; & s^\top M s - 2 s^\top r \\ s.t. \; & s \in \mathcal{K}, \end{aligned} \tag{26}$$

where $\mathcal{K}$ is an identifiable closed convex subset of $\mathbb{R}^{p+1}$ (for the sake of conciseness, $\mathcal{K}$ is characterized within the proof).

Remark 4. *Constraining the polynomials in (23) to be strictly increasing (i.e., $L'(x) > 0$) would ensure copula preservation. However, the set $\mathcal{K}$ in Theorem 1 would be open, and the existence of an optimal solution would not be guaranteed.*

4.2.2. Solving strategy and empirical computational cost

As solving for s^* in (26) is a convex-constrained quadratic program, it can be addressed efficiently using devoted solvers. The problem (22) amounts to solving $K+1$ optimization problems of the form (26). Furthermore, computations can be done in parallel. The problem (26) can be formulated and solved using `CVXR`, an `R` package for disciplined convex programming [39]. The optimization scheme is illustrated in Algorithm 1.

While computing the Lebesgue moment matrix M on each sub-interval of $[0, 1]$ is straightforward, computing strategies for r, the moment vector of $F_P^{\leftarrow}$, can vary depending on whether P is empirical or not. Additional computational details are given in Appendix Appendix B. The set-up of the `CVXR` constraints is detailed in the accompanying GitLab repository[1].

To provide a frame of reference for the practical usage of this method, the empirical computational time of solving one element of G, w.r.t. the polynomial degree is studied. Values $t_0, t_1 \in [0, 1]$, and $z_0, z_1 \in \Omega_X$ are randomly selected, and an isotonic interpolating piece-wise continuous polynomial is fitted (i.e., solving (26)). Polynomials of degrees ranging from 2 to 50 are fitted for each experiment, repeated 150 times. The execution time has been recorded and is displayed in Figure 4. The

[1] https://gitlab.com/milidris/qcWasserteinProj

Algorithm 1 Isotonic interpolating piece-wise continuous polynomial optimization strategy

Require: α, b, $F_P^{\rightarrow}$, p

1: **for** $i = 0, \ldots, K$ **do** (in parallel)
2: Compute M on $[\alpha_i, \alpha_{i+1}]$ (24).
3: Compute r on $[\alpha_i, \alpha_{i+1}]$ (25).
4: Setup `CVXR` constraints.
5: $s^{(i)} \leftarrow$ Solve (26).
6: $G_i(x) \leftarrow \sum_{j=0}^{p} s_j^{(i)} x^j$
7: **end for**
8: **return** $G(x) \leftarrow \sum_{i=0}^{K} G_i(x) \mathbb{1}_{[\alpha_i, \alpha_{i+1}]}(x)$

mean computational time seems to be linear w.r.t. the polynomial degree. However, the higher the degree, the wider the 90% time coverage seems to be, which may be caused by the complexity of the underlying optimization problem. In our limited testing, further numerical experiments showed that small polynomial degrees (≤ 7) often appear sufficient to obtain good approximations. Moreover, the approximation error tends to stabilize, w.r.t. the polynomial degree, rather rapidly.

Remark 5. *The numerical solver used is* `SCS V3.2.1` *[64]. The quantile functions have been mapped to take values between* $[-1, 1]$ *to improve numerical stability. All the figures and all obtained optimal perturbations have been computed by performing this pre-processing step first.*

5. Robustness diagnostics to distributional perturbations

The perturbation method is applied to two use cases to illustrate the robustness insights one can gather regarding black-box models. First, the robustness to feature perturbations of a classification model (i.e., a one-layer neural network) trained on an acoustic fire extinguisher dataset is studied. Local and global diagnostics are showcased, leading to tangible insights. The second use case deals

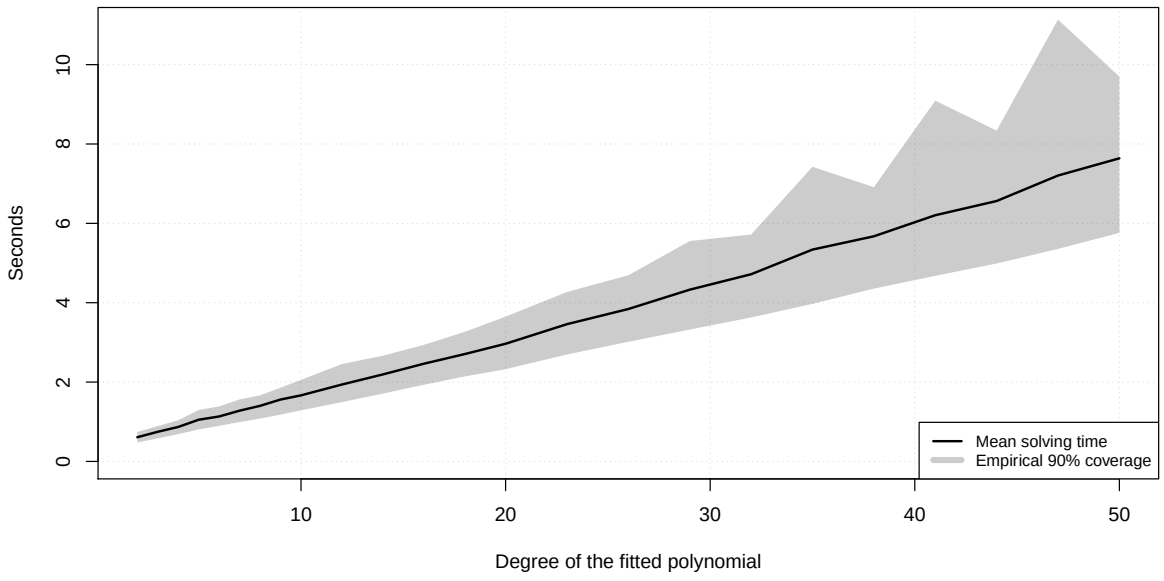

Figure 4: Computational solving time in seconds of the optimization problem (26) using `CVXR`, w.r.t. the chosen degree of the polynomial.

with a numerical hydrological model from the UQ literature. The perturbation methodology allows going beyond classical metrics for surrogate model validation.

Remark 6. *The following applications apply optimal perturbations using an isotonic polynomial smoothing with an arbitrarily high degree. The degree is chosen based on an empirical inspection of the solutions and ensuring that the approximation error remains relatively the same w.r.t. higher degrees.*

Particular attention has been put on copula preservation. Even though the relaxed problem (18) *is solved in the following applications, the solutions are composed of strictly increasing marginally perturbed quantile functions.*

5.1. ML application: Acoustic fire extinguisher dataset

The acoustic fire extinguisher dataset comprises 15390 experiments of fire extinguishing tests of three different liquid fire fuels. Amplified subwoofers are placed in a collimator with an opening. When activated at different frequencies, the acoustic waves produce an air escape through the opening, which is used to extinguish fires. Three features are set using a design of experiment (DoE), and two are measured using appropriate equipment. One can refer to the in-depth descriptions in [53, 87] for more details on the experiment's settings. Table 1 gives additional details on the nature of the features.

Feature	Unit	Mode of measure	Description
TankSize	cm	DoE	Discrete feature (5 levels) describing the tank size containing the fuel.
Fuel		DoE	Fuel type used (3 levels: Gasoline, Kerosene, Thinner).
Distance	cm	DoE	Distance of the flame to the collimator opening.
Frequency	Hz	DoE	Sound frequency range.
Decibel	dB	Measured	Sound pressure level.
Airflow	m/s	Measured	Airflow created by the sound waves.

Table 1: Description of the features of the acoustic fire extinguisher dataset.

For each experiment, a binary output variable Y is measured, representing the result of the experiment, i.e., , whether the fire has been put out ($Y = 1$) or not ($Y = 0$). The two output classes are relatively balanced (i.e., 48.97% of the observations describe effectively extinguished fires). The distribution, correlation structure, and relationship between the features and the output are represented in Figure 5. Some variables seem fairly correlated (in Spearman's sense, i.e., the linear correlation of the rank-transformed data), such as Frequency and Decibel, as well as Distance and Airflow.

The classification black-box model is a one-layer neural network (composed of 100 neurons), trained on 500 epochs, with a learning rate of 10^{-4}, similar to the study conducted in [86]. 5% of the data has been randomly selected for validation. The model resulted in a good prediction accuracy: 95.15% of the training data and 94.26% of the validation data are correctly classified. Figure 6 depicts the trained black-box model's ROC curve and confusion matrix. The model's predictive performance can be validated globally with an AUC of 0.992 and less than 3% of type 1 and 2 prediction errors.

However, global predictive performance only focuses on effectively observed data points. Studying the model's behavior on predictions outside these points is mandatory to improve confidence in its usage. Hence, one can be interested in the robustness of the model w.r.t. perturbations on its inputs.

Note that ground truths cannot be observed for perturbed data. However, the impact, either globally or locally, of these perturbations on the predictive behavior of the model can still be assessed using predictions on the perturbed data. The feature perturbation scheme is detailed and motivated in the following, and then the model's behavior is studied under these perturbations.

5.1.1. Perturbation strategy

A straightforward perturbation strategy is proposed for the Airflow feature. The perturbation is composed of the $K = 14$ constraints:

- The application domain of the feature is preserved by setting both the 0 and 1-quantiles to the dataset's minimum and maximum observed value.
- The left tail of the distribution is preserved by constraining every quantile of level 0.1 to 0.6 with a step of 0.05 to interpolate the empirical quantile function of the feature.
- A quantile shift perturbation is put on the 0.8-quantile of the feature, with an initial value of $F_P^{\leftarrow}(0.8) = 12$, being shifted between 9.5 ($\theta = -1$) and 14.5 ($\theta = 1$).

In addition to these perturbations, piece-wise continuous isotonic polynomials smoothing is enforced. The degree of each increasing polynomial has been arbitrarily chosen to be up to 9. The constraints and the resulting quantile-constrained Wasserstein projections are illustrated in Figure 7 for intensity values -1, 0, and 1.

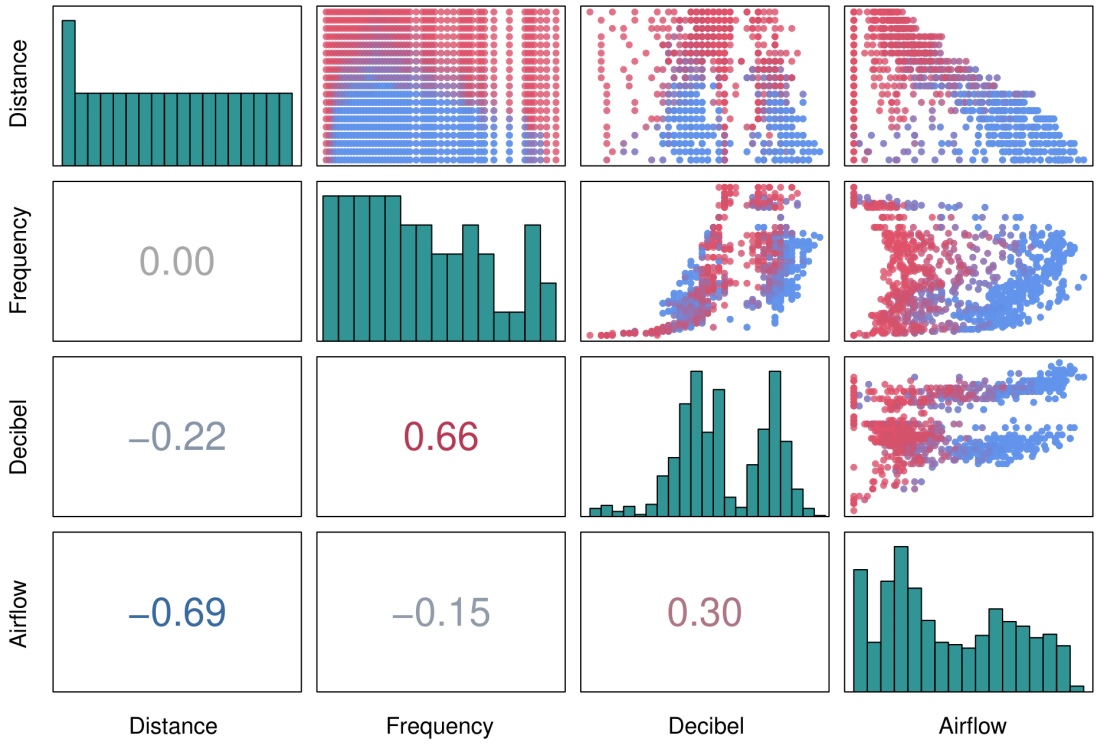

Figure 5: Histogram, cross-scatterplot, and Spearman's correlation coefficient of the input features. Red dots represent observations resulting in $Y = 0$, and blue dots for $Y = 1$.

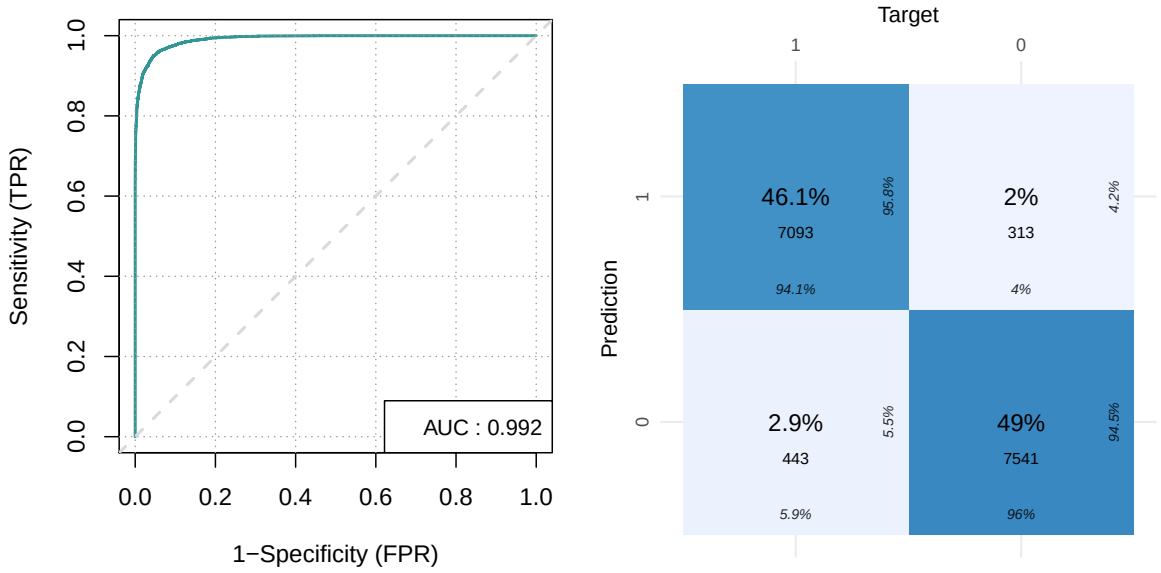

Figure 6: ROC curve (left) and confusion matrix (right) of the neural network model trained on the acoustic fire extinguisher dataset.

The perturbed quantile level has been chosen with the model's decision boundary in mind: no observation in the initial dataset with an Airflow value exceeding 12.3m/s is classified by the model as not extinguishing the fire, regardless of the values taken by the other features. Perturbing the 0.8-quantile of the Airflow variable allows for exploring the model's behavior in regions close to this decision boundary. More importantly, it allows for assessing the predictive robustness of the neural network in this region under perturbations of varying magnitude. Generally, this quantile shift regime can be understood as a perturbation on the right tail of the initial distribution, i.e., on values higher than the 0.6-quantile.

5.1.2. Model robustness assessment

First, global robustness insights are highlighted. The left plot of Figure 8 presents the proportion of perturbed observations with predictions of 1 w.r.t. to the intensity of the perturbation. Notice that the proportion is increasing, along with θ. Hence, decreasing the value of the initial 0.8-quantile tend to result in a lower number of predicted put-out fires, and increasing its value results in an increasing number of predicted put-out fires. This interpretation is rather intuitive: all other things being equal, a higher Airflow value entails a higher chance of predicting $Y = 1$. The right plot of Figure 8 presents the proportion of prediction shift w.r.t. θ. Notice that the higher the magnitude of the perturbation (positively or negatively), the more predictions tend to change, and the closest θ is to 0, the fewer predictions shift. This observation informs on the predictive stability in the vicinity of the decision boundary of the model: small perturbations tend to result in fewer prediction shifts than bigger perturbations.

Figure 9 presents the target Shapley effects [48], a global SA input importance measure for binary black-box model outputs with dependent inputs, w.r.t. the quantile shift intensity parameter θ. These indices have been computed using the nearest-neighbor (KNN) approach proposed in [15] (with an arbitrarily chosen number of neighbors equal to 6). Studying the behavior of importance measures informs on the stability of this diagnostic (i.e., feature importance order) w.r.t. input perturbation, i.e., if the importance hierarchy between the inputs changes due to perturbations around the model's decision boundary. The left barplot presents the initial target Shapley effects, computed on the model's prediction on the observed data, and the right plot presents their behavior under the airflow perturbation. One can notice that the importance indices remain stable w.r.t. θ. This result indicates that

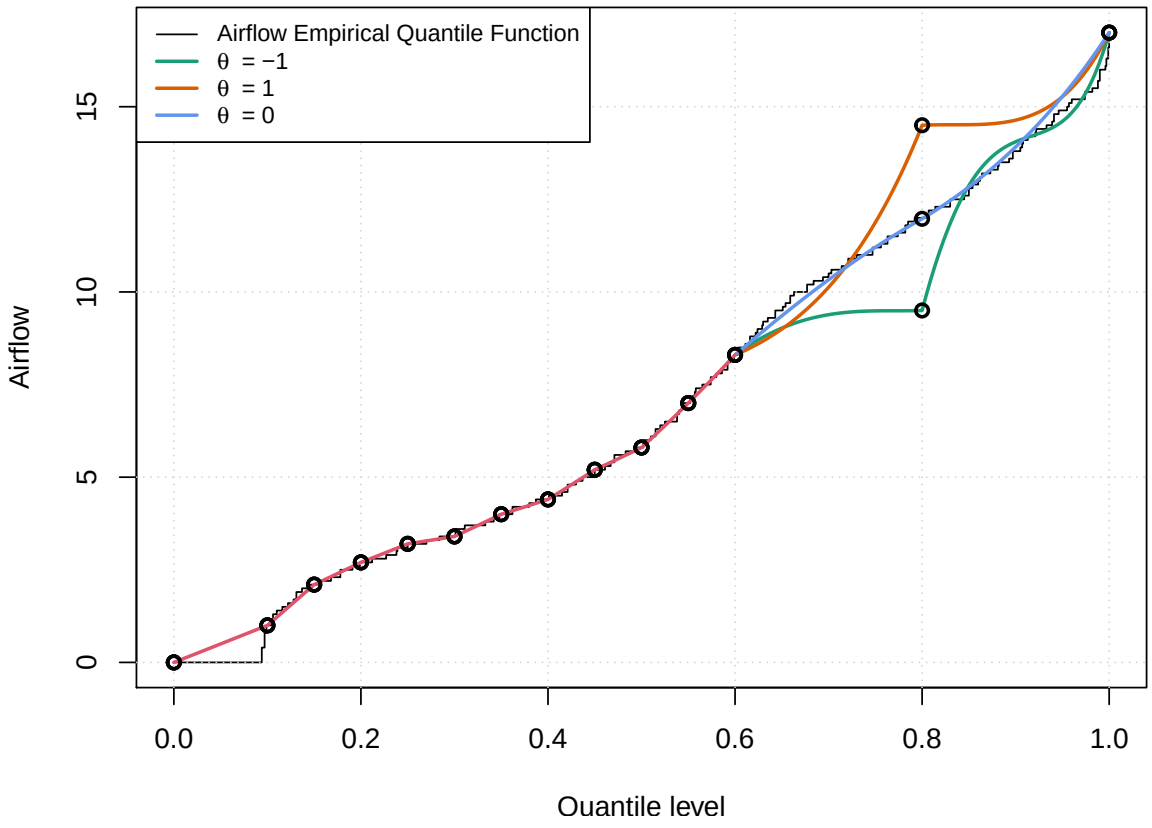

Figure 7: Quantile functions of the optimally perturbed Airflow feature, with a chosen polynomial degree equal to 9. The red line represents the preserved tail; meanwhile, the green, blue, and yellow lines represent various quantile shift intensity levels ($\theta = -1$, $\theta = 0$, and $\theta = 1$, respectively).

the global SA of the neural network is robust to the distributional perturbations driven by θ. Hence, one can be confident in those diagnostics under uncertainties in the region near the model's decision boundary.

Finally, the robustness of the neural network can also be assessed locally. Figure 10 allow visualizing whether a prediction has shifted w.r.t. to the effective magnitude of the perturbation. The black line indicates no perturbation change: the airflow value of an observation has been mapped to itself. For a fixed initial airflow datapoint, its vertical distance to the black line indicates the (signed) magnitude of the applied perturbation. Red points indicate that the prediction has shifted w.r.t. the initial dataset, and blue points indicate no predictive change. One can note the presence of red dots close to the black line around the prediction boundary of the model. Small perturbations for observations with airflow values around 12, all other features being equal, can lead to a prediction change. Hence, the confidence in predictions on observations in this region can be questioned. However, notice the lack of red dots near the black line for airflow values on the interval $[13, 17]$ and on the interval $[7, 10]$. Hence, one can be confident in the model's predictions for Airflow values on these intervals, which seem robust w.r.t. the quantile shift.

One may notice the presence of small perturbations resulting in prediction changes for Airflow values around $[0, 5]$. However, since the perturbation scheme focuses on exploring the model's behavior around the decision boundary, their interpretation is voluntarily omitted: a different perturbation scheme involving perturbing the left tail of the airflow distribution would be advised.

In summary, besides its good prediction accuracy, the model is globally robust to distributional perturbation focused around the decision boundary of its Airflow feature. Moreover, one can be confident in the feature importance indices since they remain relatively similar under perturbation. Locally, the model prediction seems stable w.r.t. small perturbations, except on a small interval around its decision boundary (a behavior generally expected in ML applications). In conclusion, this robust interpretability analysis further assesses the model's behavior beyond classical accuracy metrics and provides additional arguments for its validation.

5.2. SA application: Simplified hydrological model and surrogate model validation

This use case focuses on a simplified model of the water level of a river. This model has been extensively used in the safety and reliability of industrial sites, where the occurrence of a flood can lead to dramatic human and ecological consequences. It consists of a substantial simplification of the one-dimensional Saint-Venant equation, with a uniform and constant flow rate, inspired from [51, 40]. The maximal annual water level from sea level is modeled as follows:

$$Y = Z_v + \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{3/5},$$

where the description of each input variable and their explicit marginal probabilistic structure is detailed in Table 2.

Additionally, similarly to [18], a dependence structure is modeled using a Gaussian copula, with the covariance matrix

$$R_P = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0 & 0 & 0.3 & 1 \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} Q \\ K_s \\ Z_v \\ Z_m \\ L \\ B \end{pmatrix} \sim P.$$

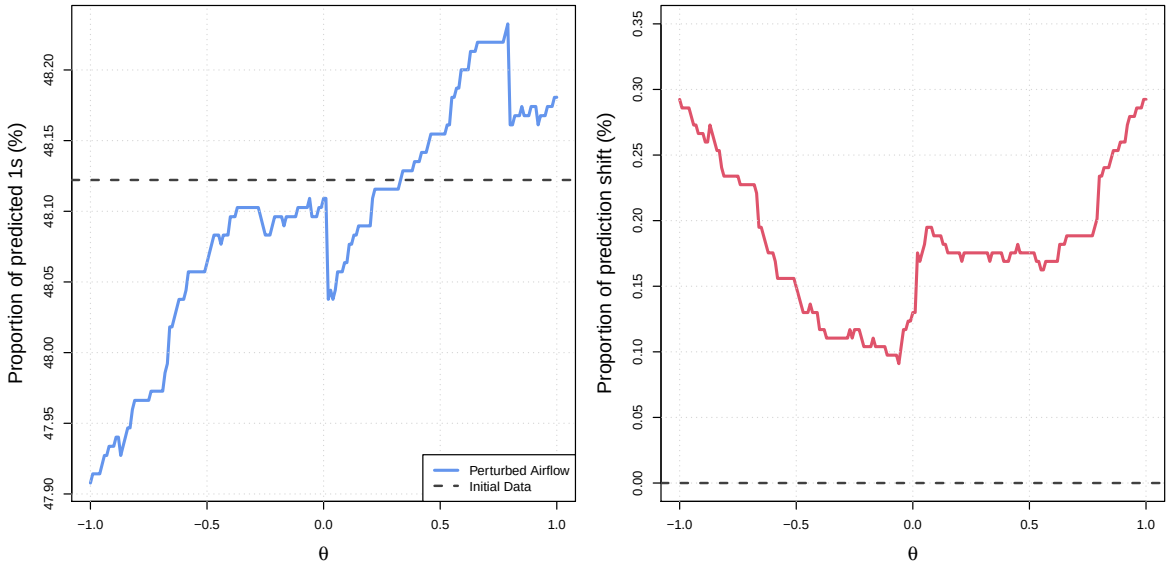

Figure 8: Proportion of predictions $Y = 1$ (left) and proportion of classification prediction shift (right) compared to the initial data, w.r.t. the perturbation intensity parameter θ.

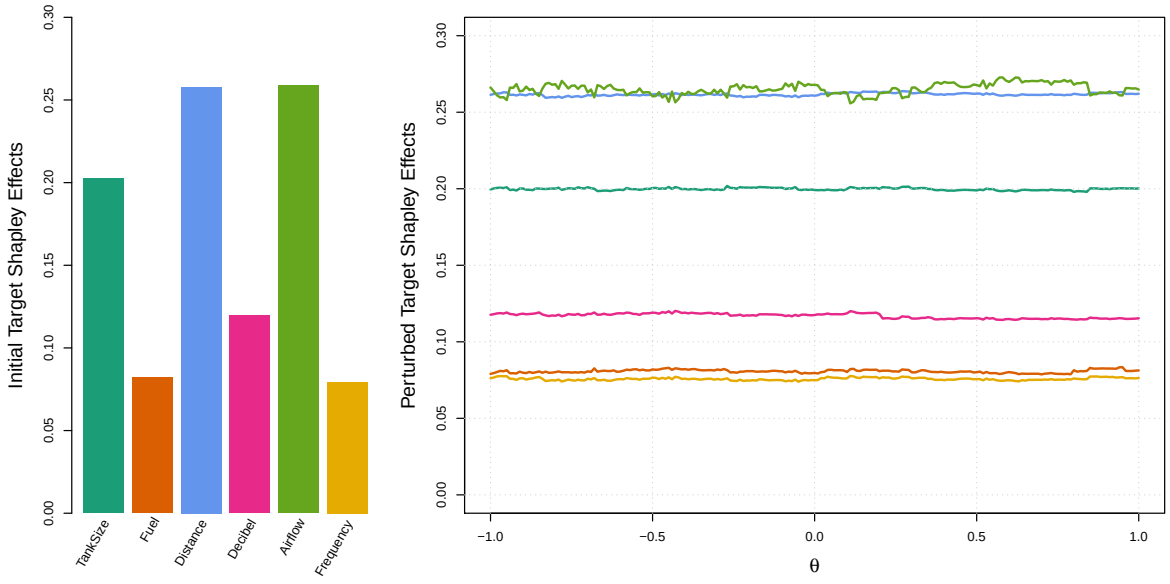

Figure 9: Initial (left) and perturbed (right) target Shapley effects, w.r.t. the intensity parameter θ, using the same color panel.

Input	Unit	Distribution	Application Domain	Description
Q	m^3/sec	$\mathcal{G}(1013, 558)$ trunc.	$[500, 3000]$	River maximum annual water flow rate.
K_s		$\mathcal{N}(35, 5)$ trunc.	$[20, 50]$	Strickler riverbed roughness coefficient.
Z_v	m	$\mathcal{T}(49, 50, 51)$	$[49, 51]$	Downstream river level.
Z_m	m	$\mathcal{T}(54, 55, 56)$	$[54, 56]$	Upstream river level.
L	m	$\mathcal{T}(4990, 5000, 5010)$	$[4990, 5010]$	River length.
B	m	$\mathcal{T}(295, 300, 305)$	$[295, 305]$	River width.

Table 2: Inputs of the simplified river water level model and their explicit marginal distributions. $\mathcal{G}, \mathcal{N}, \mathcal{T}$ denote Gumbel, Normal and Triangular distributions, respectively (trunc means truncated).

Echoing Example 2, one is interested in uncertainties on the application domain of the K_s input, i.e., the Strickler riverbed roughness coefficient (which is the inverse of the Manning coefficient). Its value can range from around 3 (proliferating algae) to 90 (smooth concrete). We refer the interested reader to the in-depth study in [40] for more details on the determination and inference of the Strickler coefficient for realistic rivers. In this use-case, the application domain Ω_X of the Strickler coefficient is initially set between the values of 20 and 50, corresponding to situations from very cluttered riverbeds to earthen channels. However, epistemic uncertainties are assumed to affect this application domain to illustrate our robustness method.

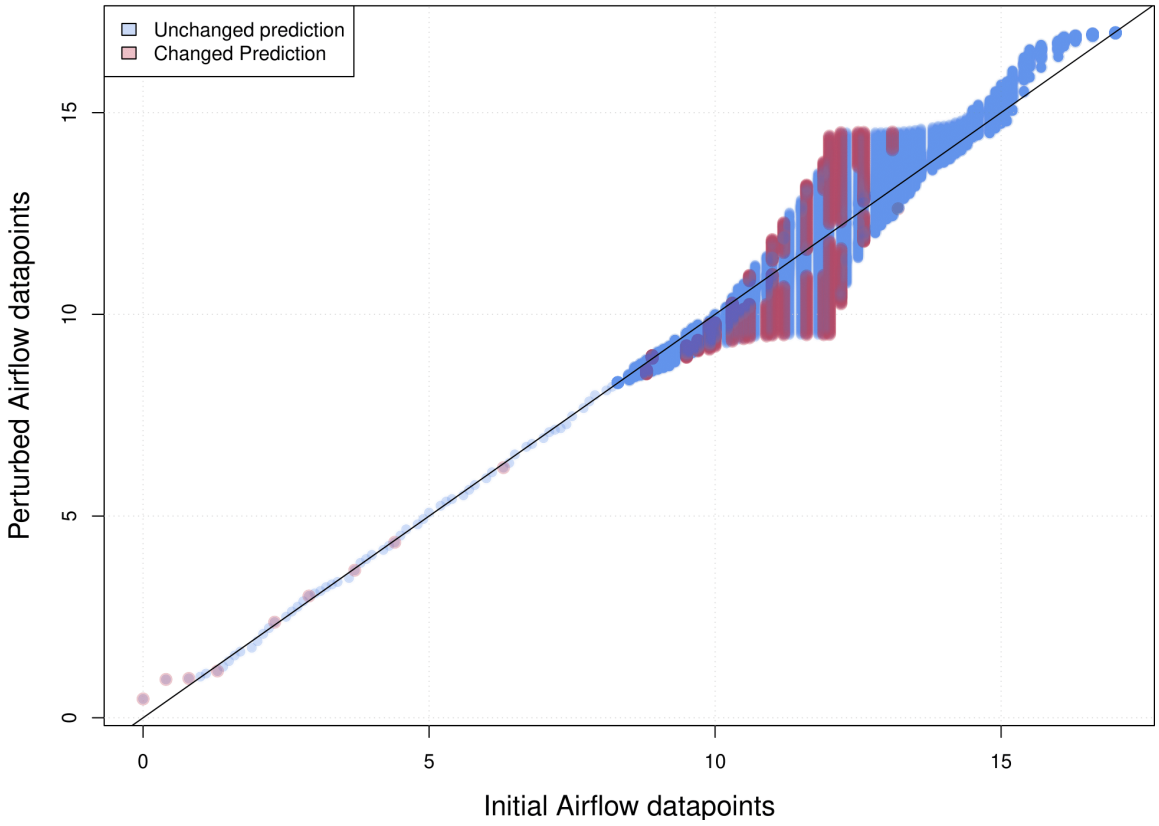

Figure 10: Perturbed datapoints w.r.t. their initial values. The black line represents no perturbation. The red and blue dots represent either a classification shift due to the perturbation or no classification shift.

5.2.1. Perturbation strategy

In this use case, the three following inputs are perturbed. The river's maximum annual water flow rate Q, the river length L, and the upstream river level Z_m are subject to the following quantile constraints:

- Quantile perturbations on Q:
  - Shift of the application domain from $[500, 3000]$ to $[500, 3200]$;
  - Preserve the median of the distribution;
  - Increase the initial 0.15-quantile by 75;
  - Decrease the initial 0.75-quantile by 125;
- Quantile perturbations on L:
  - Shift the application domain from $[4990, 5010]$ to $[4988, 5012]$;
  - Preserve the median of the distribution;
- Quantile perturbations on Z_m:
  - Preserve the application domain and the median of the initial distribution;
  - Increase the 0.8 and 0.9-quantiles by 0.1;

– Decrease the 0.25-quantile by 0.05.

The initial input distributions, their application domain, and the optimally perturbed results are illustrated in Figure 11. These constraints are mainly enforced to illustrate that multiple inputs can be perturbed simultaneously while preserving their dependence structure. They can be interpreted, for instance, as domain experts' knowledge injection into the initial probabilistic structure of the inputs (e.g., to study a specific river arm).

In addition to these constraints, the Strickler coefficient K_s is subject to an application domain dilatation perturbation, with a scaling parameter $\eta = 2$. Each perturbation intensity represents a degree of uncertainty on the type of riverbed roughness. When $\theta = -1$, the width of the initial application domain is halved, i.e., from $[20, 50]$ to $[27.5, 42.5]$, which can be interpreted in a situation where the epistemic uncertainty on the riverbed roughness is narrower, between a slow winding natural river, up to a plain river without shrub vegetation. When $\theta = 1$, the epistemic uncertainty on the riverbed is much wider. The application domain equals $[5, 65]$, depicting a range of riverbed roughness from proliferating algae to smooth concrete. Figure 12 illustrates the initial K_s distribution and the optimally perturbed quantile functions for θ equal to -1 and 1. Hence θ can be interpreted as a proxy for the "amount" of epistemic uncertainty on the riverbed roughness.

Additionally, the perturbations' smoothness is enforced using piece-wise continuous isotonic polynomials of degree up to 12, chosen arbitrarily.

5.2.2. Robustness of the sensitivity analysis

From a global standpoint, one can be interested in the impact of the distributional perturbations on key statistics of the random output of the river water level model. Figure 13 presents estimated values for the mean, standard deviation, 0.025 and 0.975-quantiles (shown by the 95% coverage),

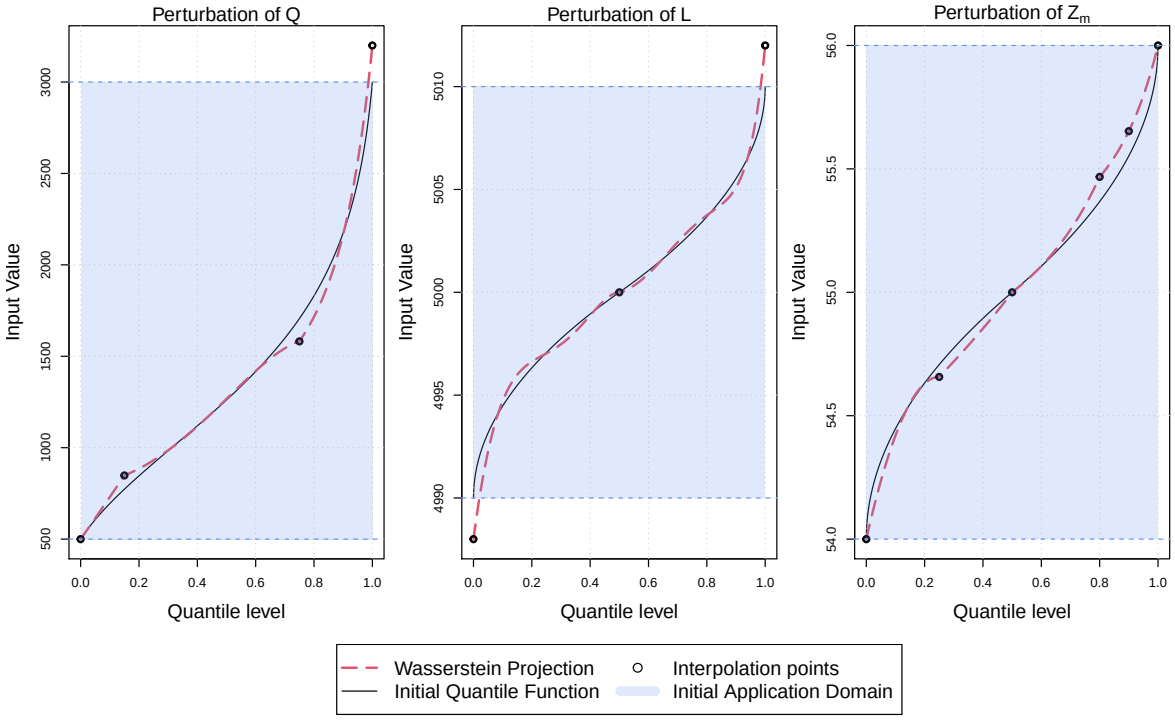

Figure 11: Initial quantile functions, application domains, and corresponding optimally perturbed quantile functions of the Q, L, and Z_m inputs.

and minimum and maximum values of the random output, computed on 10^5 Monte Carlo samples, w.r.t. the dilatation intensity θ. These values are compared to the reference ones according to the initial distribution of the inputs, estimated on a 2×10^5 Monte Carlo sample.

Notice that the expectation, standard deviation, 95% coverage quantiles, and minimum value of the model output remain stable under the distributional perturbations on the application domain of the Strickler coefficient. However, the estimated upper bound of the output support increases exponentially for positive values of θ. Widening the uncertainty on the riverbed type allows for relatively rare events of high river water levels since the 0.975-quantile does not seem dramatically affected by the distributional perturbations.

Figure 14 presents the Shapley effects [65], which are global SA importance measures for real-valued model outputs with dependent inputs. These indices have been computed using a double Monte Carlo scheme as depicted in [83], with fixed simulated sample sizes, for each perturbed distribution Q driven by a value of θ, $N_v = 10^4$ for estimating $\mathrm{Var}_Q(Y)$, as well as $N_o = 10^3$ and $N_i = 100$ to estimate $\mathbb{E}_Q\left[\mathrm{Var}_Q\left(Y \mid X_A\right)\right]$ for every subset $X_A, A \subseteq \{1, \ldots, d\}$ of variables. Additionally, the reference Shapley effects have been computed under the initial distribution with sample sizes $N_v = 10^5$, $N_o = 3 \times 10^3$, and $N_i = 300$.

Note that the distributional perturbations have an impact on the importance measures. More precisely, increasing the range of the uncertainty of the riverbed roughness increases its importance for positive values of θ. Conversely, the importance of both Q and Z_v decreases accordingly. However, the variable importance hierarchy induced by the Shapley effects is preserved. It is also essential to notice that Q and Z_v are considered equally important as θ gets large. Hence, this SA does not seem robust to distributional perturbations and, more precisely, to a widening of the support of the Strickler coefficient in combination with the quantile perturbations put on Q, L, and Z_m.

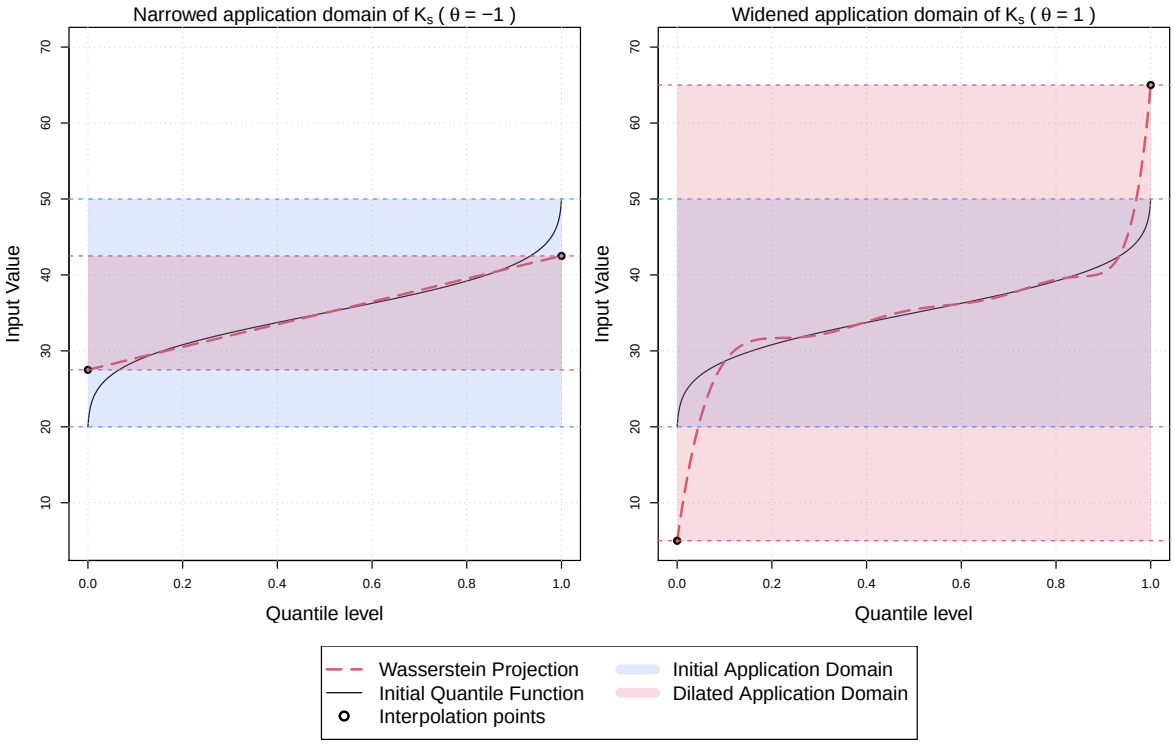

Figure 12: Initial quantile function, application domain and corresponding optimally perturbed quantile functions for K_s, for θ being equal to -1 (left) and 1 (right), for a scaling parameter $\eta = 2$.

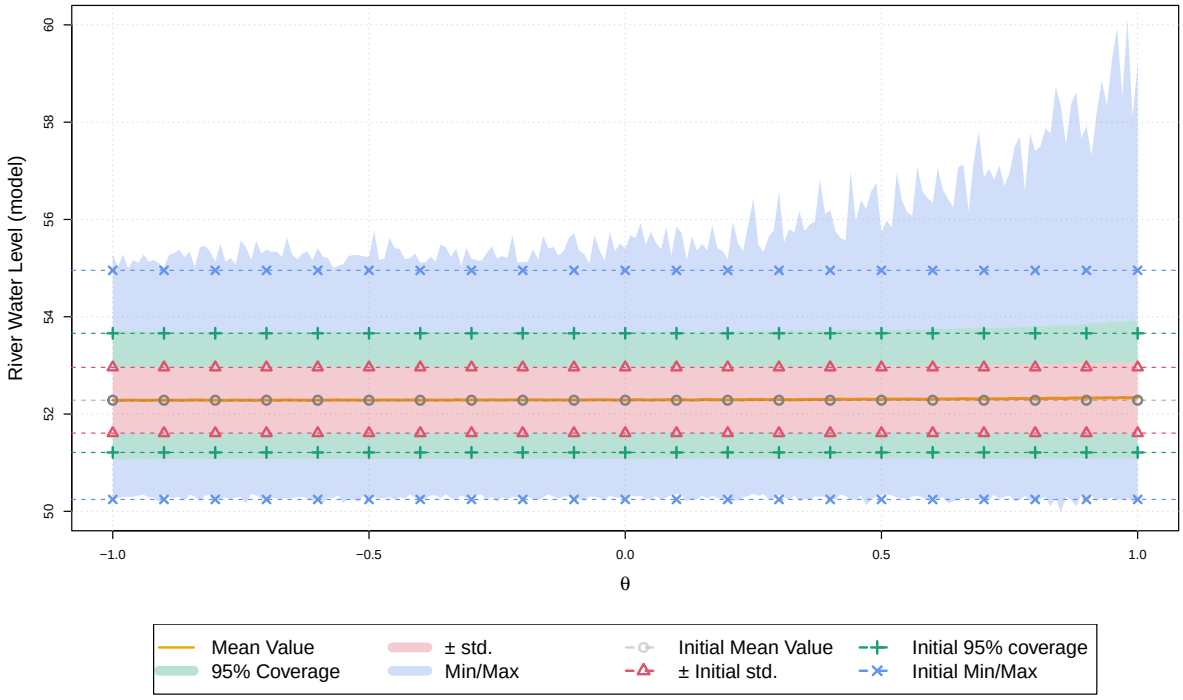

Figure 13: Expectation, standard deviation, 95% coverage, minimum and maximum estimators of the river water level, w.r.t. the application domain dilatation intensity θ.

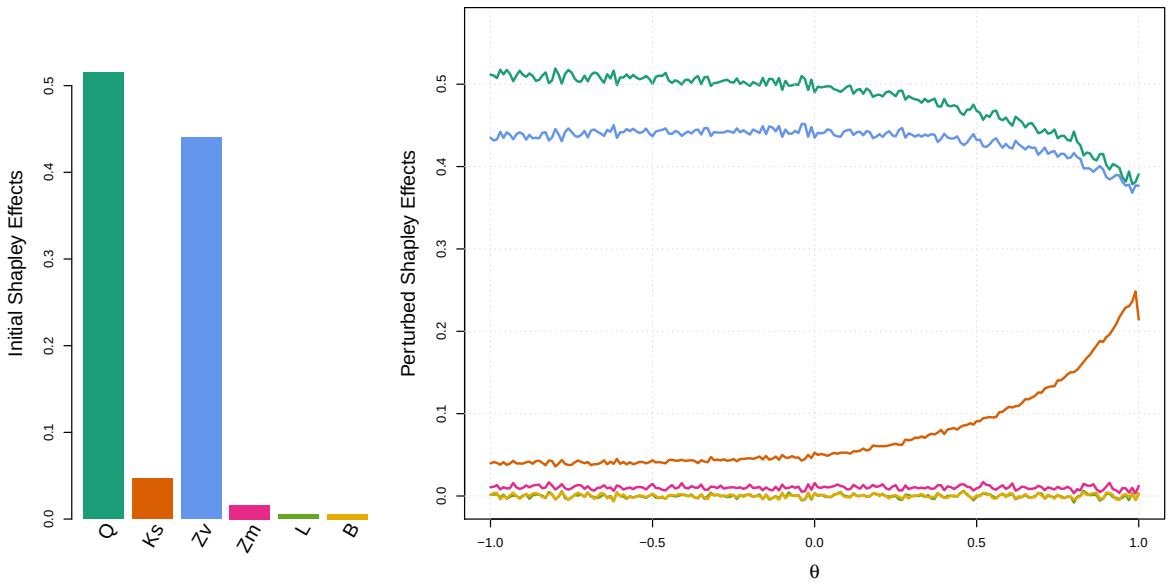

Figure 14: Reference Shapley effects (left) and Shapley effects of the river water level model under optimally dilated application domain w.r.t. θ (right), using the same color panel.

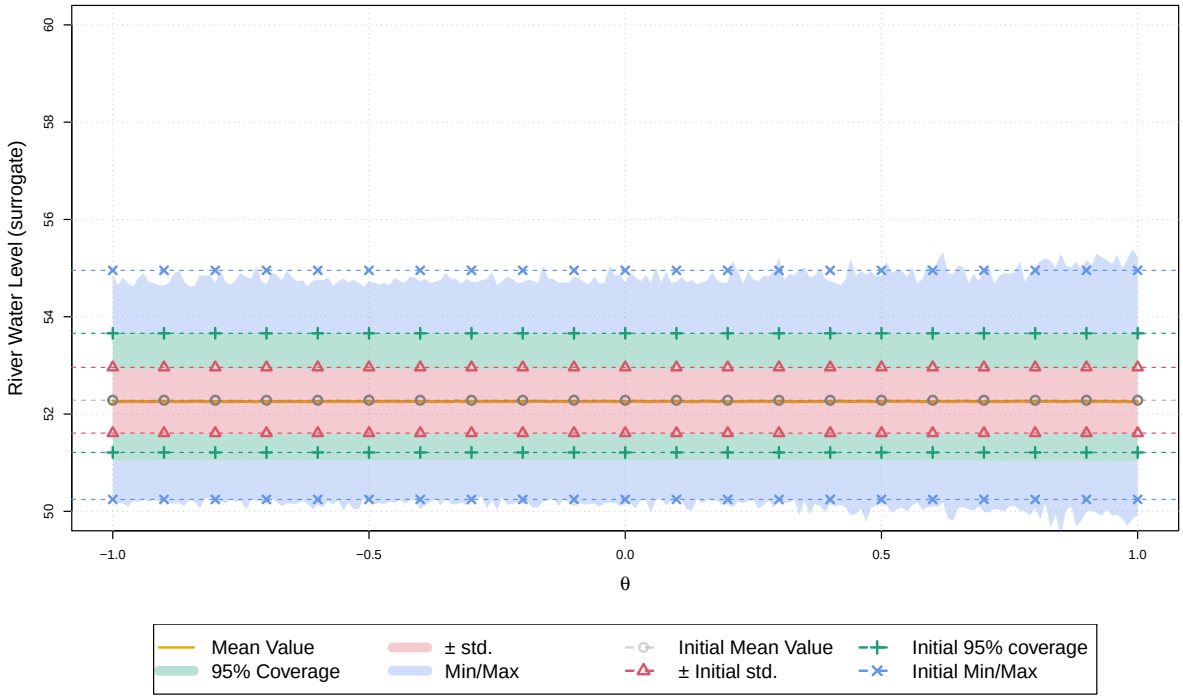

Figure 15: Expectation, standard deviation, 95% coverage, minimum and maximum estimators of the surrogate model, w.r.t. the application domain dilatation intensity θ.

5.2.3. Surrogate model validation

A surrogate model is trained on an input-output sample of size 5×10^5 of the initial probabilistic structure and validated on a validation dataset of size 5×10^4. The surrogate model is a neural network comprised of 3 hidden layers, 64 neurons each, and ReLu as an activation function. The R^2 of the model is 99.5% R^2 on the training data and 99.5% on the validation data. Despite the model's good results on the validation data, it does not behave the same way as the initial model when perturbed similarly. Echoing Figure 13, Figure 15 illustrates the model's behavior when subject to the previously introduced perturbations.

One can notice that the surrogate model does not behave as the numerical model w.r.t. the epistemic uncertainty of the riverbed roughness. Even though, on the surface, the surrogate model generalizes well on validation data, its behavior on the perturbed data differs from the initial numerical model. More precisely, the maximal value of the river water level does not seem to be impacted by the epistemic uncertainty of the riverbed roughness. However, the other statistics (mean, variance, and 95% coverage) align with the numerical model. Hence, despite its good fit, using this surrogate model would not be advised if the goal of the sensitivity analysis is to study rare events.

5.3. Conclusions

These two use cases illustrate the different insights the perturbation methodology can offer in UQ and ML studies. On the ML side, for classification tasks, it allows assessing the global behavior of black-box models under input perturbations. This assessment is quantified either through studying the prediction shifts due to the perturbation or through the behavior of feature importance metrics. Locally, it allows the detection of low-stability regions of interest (regions where small perturbations induce a classification change). In addition to classical accuracy metrics, our method can be used to assess confidence in a predictive model. On the UQ side, it allows for studying the impact of distributional perturbations (whose intensity can be tuned to represent epistemic uncertainties) on

the model output, even in situations where inputs are correlated. Furthermore, in a SA context, the behavior of classical sensitivity indices under those perturbations can also be studied, and their robustness (for instance, the preservation of the input importance hierarchy) w.r.t. the probabilistic modeling on the inputs can be assessed. In both cases, meaningful perturbations allow for a complete picture, beyond classical validation metrics, of a black-box model's behavior outside of the initial distribution.

6. Discussion and perspective

Obtaining robustness diagnoses on the influence of input variables and the behavior of a model considered a black box is essential for its acceptance and use. This paper provides a tool to answer this problem by modifying the distributions of the input variables. These perturbations modify the quantile of marginal distributions while, in some instances, preserving the dependence structure. This method revolves around probability measure projection under a 2-Wasserstein cost, leading to interpretable, generic, and close solutions, allowing for data exploration. Regularity conditions can be enforced, and the case of piece-wise interpolating isotonic polynomials is studied. The robustness analyses conducted on real case studies illustrate its potential flexibility and adequate computational cost, which are essential for high-dimensional cases. These studies highlighted validation insights beyond classical tools, allowing for a more complete understanding of the black-box model's behavior.

Several avenues of improvement can be considered. First, concerning the piece-wise interpolating isotonic polynomials. An enlightened polynomial degree selection is required. An idea would be to use prior information on the order of differentiability of the sought-after perturbed gqf. In an ML framework, nonparametric approaches to isotonic regression of the marginal gqfs of P can provide answers through statistical testing [32, 25] or criteria enforcing a trade-off between approximation error and sparsity (e.g., inspired from AIC or BIC). Moreover, while the proposed methodology allows for continuous results, differentiability is not guaranteed. However, inspiration from the literature on isotonic splines [46, 79, 37, 91] can be leveraged to offer the practitioner a better range of smooth solutions. Additionally, other spaces of functions can also be used for smoothing purposes. Following the work of [7], abstract reproducing kernel Hilbert space of nonnegative functions can be reached through particular kernels. Hence, it would allow accessing different sets of nonnegative functions whose regularities can be assessed through a thorough study of these kernels.

Second, the proposed methodology only focuses on marginal perturbation preserving the dependence structure of the inputs. As pointed out by the reviewers of this article, one may wish to perturb the dependence structure as well. However, it is argued that copula perturbation should be done independently of marginal perturbations for the sake of the final interpretation of the robustness analyses. It allows separating the effects in the marginal perturbation of the effects of the stochastic dependence perturbation. Association and concordance measures appear as the most interpretable tools for copula manipulation (and are frequently used to incorporate expert opinion) [21, 92, 10]. An alternative approach to perturb the stochastic dependence structure and the marginal would be to consider multivariate quantile functions. However, defining multivariate quantile functions is not trivial and not as natural as in the univariate case. Among the many approaches to defining such a notion, the most theoretically accomplished today is the one resulting from the concept of *center-outward distribution function* [19, 45, 11]. Perturbing these quantile contours can be leveraged to go beyond marginal consideration and will be the subject of future work.

Finally, one of the primary motivations for using the 2-Wasserstein distance as a projection metric is that it metricizes weak convergence on a broad set of probability measures. Other distances between probability measures are endowed with similar properties, such as the Prokhorov-Levy distance. Leveraging the different relationships between such distances (see [42]) could be beneficial for generalizing the proposed approach.

Acknowledgements

Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is gratefully acknowledged.

The authors warmly thank the Editor In-Chief and the two anonymous reviewers from their helpful remarks, as well as Jean-Bernard Lasserre (Institut de Mathématique de Toulouse) and Guillaume Dalle (CERMICS) for their help in solving the optimization problem at the heart of this work, Clément Bénesse (Institut de Mathématiques de Toulouse) and Antoine Paolini (UVSQ Université Paris Saclay) for their support on some of the mathematical aspects of this work.

References

[1] A. Alfonsi and B. Jourdain. A remark on the optimal transport between two probability measures sharing the same copula. *Statistics & Probability Letters*, 84:131–134, January 2014.

[2] D.L. Allaire and K. E. Willcox. Distributional sensitivity analysis. *Procedia - Social and Behavioral Sciences*, 2:7595–7596, 2010.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

[4] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML), 10-15, 2018*, volume 80, pages 284–293, 2018.

[5] European Banking Authority. *2021 EU-Wide Stress Test*. European Banking Authority, 2020.

[6] F. Bachoc, F. Gamboa, M. Halford, J-M. Loubes, and L. Risser. Explaining machine learning models using entropic variable projection. *Information and Inference: A Journal of the IMA*, 12(3), 05 2023. iaad010.

[7] J. A. Bagnell and A-M Farahmand. Learning positive functions in a hilbert space. *Preprint*, 2015.

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

[9] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. SHAFF: Fast and consistent SHApley eFfect estimates via random Forests. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 5563–5582, 2022.

[10] N. Benoumechiara, N. Bousquet, B. Michel, and P. Saint-Pierre. Detecting and modeling critical dependence structures between random inputs of computer models. *Dependence Modeling*, 8(1):263–297, 2020.

[11] Bernard Bercu, Jeremie Bigot, and Gauthier Thurin. Monge-kantorovich superquantiles and expected shortfalls with applications to multivariate risk measurements, 2023.

[12] D. P. Bertsekas. *Nonlinear programming*. Athena scientific, Belmont, Mass, 3rd ed edition, 2016.

[13] N. Bloom. The impact of uncertainty shocks. *Econometrica*, 77(3):623–685, 2009.

[14] E. Borgonovo, A. Figalli, E. Plischke, and G. Savare. Probabilistic Sensitivity with Optimal Transport. *Preprint*, 2022.

[15] B. Broto, F. Bachoc, and M. Depecker. Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716, 2020.

[16] L. Bruzzone and M. Marconcini. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

[17] C. Bénard, S. Da Veiga, and E. Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*, 109(4):881–900, 02 2022.

[18] G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - Application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012.

[19] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge-Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017.

[20] Y. Chung, W. Neiswanger, I. Char, and J. Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10971–10984, 2021.

[21] R. T. Clemen and T. Reilly. Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224, 1999.

[22] I. Covert, S. Lundberg, and S.-I. Lee. Understanding Global Feature Contributions With Additive Importance Measures. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223, 2020.

[23] I. Csiszár. I-Divergence Geometry of Probability Distributions and Minization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[24] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. *Basics and Trends in Sensitivity Analysis. Theory and Practice in R*. SIAM. Computational Science and Engineering, 2021.

[25] S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. *Annales de la Faculté des Sciences de Toulouse*, 3:529–555, 2012.

[26] L. De Lara, A. González-Sanz, N. Asher, and J-M Loubes. Transport-based counterfactual models. *arXiv preprint arXiv:2108.13025*, 2021.

[27] Lucas De Lara, Alberto González-Sanz, and Jean-Michel Loubes. Diffeomorphic registration using sinkhorn divergences. *SIAM Journal on Imaging Sciences*, 16(1):250–279, 2023.

[28] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in Industrial Practice*. John Wiley and Sons, Ltd, Chichester, UK, April 2008.

[29] H. Dette and W. J. Studden. *The theory of canonical moments with applications in statistics, probability, and analysis*. Wiley series in probability and statistics. Wiley, New York, 1997.

[30] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 43:835–1234, 2021.

[31] J.-M. Dufour. Distribution and quantile functions. *McGill University Report*, 1995.

[32] C. Durot and A.-S. Tocquet. Goodness of fit test for isotonic regression. *ESAIM:P&S*, 5:119–140, 2001.

[33] G. Ecoto, A. Bibault, and A. Chambaz. One-step ahead Super Learning from short time series of many slightly dependent data, and anticipating the cost of natural disasters. *arXiv:2107:13291*, 2021.

[34] Gal Elidan. Copulas in machine learning. In Piotr Jaworski, Fabrizio Durante, and Wolfgang Karl Härdle, editors, *Copulae in Mathematical and Quantitative Finance*, pages 39–60, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[35] T. Fel, R. Cadene, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. In *Advances in Neural Information Processing Systems*, volume 34, pages 26005–26014, 2021.

[36] J-C Fort, T. Klein, and A. Lagnoux. Global Sensitivity Analysis and Wasserstein Spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021.

[37] S. Fredenhagen, H. J. Oberle, and G. Opfer. On the Construction of Optimal Monotone Cubic Spline Interpolations. *Journal of Approximation Theory*, 96(2):182–201, 1999.

[38] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T.A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[39] A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020.

[40] S. Fu, M. Couplet, and N. Bousquet. An adaptive kriging method for solving nonlinear inverse statistical problems. *Environmetrics*, 28(4):e2439, 2017.

[41] C. Gauchy, J. Stenger, R. Sueur, and B. Iooss. An information geometry approach to robustness analysis for the uncertainty quantification of computer codes. *Technometrics*, 64:80–91, 2022.

[42] A.L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

[43] U. Grömping. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7:137–152, 2015.

[44] Shimodaira; H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[45] M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, April 2021. Publisher: Institute of Mathematical Statistics.

[46] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer: New York, 2009.

[47] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J-M Loubes. On the coalitional decomposition of parameters of interest, 2023.

[48] M. Il Idrissi, V. Chabridon, and B. Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling and Software*, 143:105115, 2021.

[49] B. Iooss, V. Chabridon, and V. Thouvenot. Variance-based importance measures for machine learning model interpretability. In *Actes du 23ème Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement ($\lambda\mu$23)*, Saclay, France, october 2022.

[50] B. Iooss, R. Kennet, and P. Secchi. Different views of interpretability. In A. Lepore, B. Palumbo, and J-M. Poggi, editors, *Interpretability for Industry 4.0: Statistical and Machine Learning Approaches*. Springer, 2022.

[51] B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In G. Dellino and C. Meloni, editors, *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, pages 101–122. Springer US, 2015.

[52] O. Kallenberg. *Foundations of modern probability*. Probability theory and stochastic modelling. Springer, Cham, Switzerland, 2021.

[53] M. Koklu and Y. S. Taspinar. Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods. *IEEE Access*, 9:86207–86216, 2021.

[54] J-B. Lasserre. *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2015.

[55] P. Lemaître. *Analyse de sensibilité en fiabilité des structures*. PhD thesis, Université de Bordeaux, Bordeaux, 2014.

[56] P. Lemaître, E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. Density modification-based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85(6):1200–1223, 2015.

[57] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18:92–106, 2006.

[58] S. M. Lundberg and S-I. Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[59] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. leanpub.com, 1 edition, 2021.

[60] S-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[61] K. Murray, S. Müller, and B. A. Turlach. Fast and flexible methods for monotone polynomial fitting. *Journal of Statistical Computation and Simulation*, 86(15):2946–2966, 2016.

[62] A. Narayan and D. Xiu. Distributional sensitivity for uncertainty quantification. *Communications in Computational Physics*, 10(1):140–160, 2011.

[63] R. B. Nelsen. *An introduction to copulas*. Springer series in statistics (2nd edition). Springer, New York, 2006.

[64] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[65] A. B. Owen. Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

[66] T. Paananen, J. Piironen, M. Riis Andersen, and A. Vehtari. Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1743–1752, 2019.

[67] P. A. Parrilo. Algebraic Optimization and Semidefinite Optimization. *MIT Lectures Notes (EIDMA Minicourse)*, 2010.

[68] P. A. Parrilo. Polynomial optimization, sums of squares, and applications. In *Semidefinite Optimization and Convex Algebraic Geometry*, pages 47–157. SIAM, 2012.

[69] M.K. Paul, M.R. Islam, and Sarowar Sattar A.H.M. An efficient perturbation approach for multivariate data in sensitive and reliable data mining. *Journal of Information Security and Applications*, 62:102954, 2021.

[70] S.M. Pesenti. Reverse Sensitivity Analysis for Risk Modelling. *Risks*, 10:141, 2022.

[71] E. Plischke and E. Borgonovo. Copula theory and probabilistic sensitivity analysis: Is there a connection? *European Journal of Operational Research*, 277(3):1046–1059, 2019.

[72] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J.H.A. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H.R. Maier. The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling and Software*, 137:104954, 2021.

[73] S. I. Resnick. Preliminaries. In S. I. Resnick, editor, *Extreme Values, Regular Variation and Point Processes*, Springer Series in Operations Research and Financial Engineering, pages 1–37. Springer, New York, NY, 1987.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[75] C.J. Roy and W.L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011.

[76] R.Y. Rubinstein. Sensitivity analysis and performance extrapolation for computer simulation models. *Operation Research*, 37(1):72–81, 1989.

[77] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K-R. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2019.

[78] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015.

[79] J. W. Schmidt and W. Heß. Positivity of cubic polynomials on intervals and positive spline interpolation. *BIT Numerical Mathematics*, 28(2):340–352, 1988.

[80] R. C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Computational Science & Engineering. SIAM, 2014.

[81] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001.

[82] O. Sobrie, N. Gillis, V. Mousseau, and M. Pirlot. UTA-poly and UTA-splines: Additive value functions with polynomial marginals. *European Journal of Operational Research*, 264(2):405–418, 2018.

[83] E. Song, B. L. Nelson, and J. Staum. Shapley Effects for Global Sensitivity Analysis: Theory and Computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016.

[84] A. Stevens, P. Deruyck, Z. Van Veldhoven, and J. Vanthienen. Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1241–1248, 2020.

[85] T. Sullivan. *Introduction to Uncertainty Quantification*. Springer, 2017.

[86] Y. S. Taspinar, M. Koklu, and M. Altin. Classification of flame extinction based on acoustic oscillations using artificial intelligence methods. *Case Studies in Thermal Engineering*, 28:101561, December 2021.

[87] Y. S. Taspinar, M. Koklu, and M. Altin. Acoustic-Driven Airflow Flame Extinguishing System Design and Analysis of Capabilities of Low Frequency in Different Fuels. *Fire Technology*, 58(3):1579–1597, May 2022.

[88] N. Tripuraneni, B. Adlam, and J. Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[89] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, March 2003.

[90] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

[91] X. Wang and F. Li. Isotonic Smoothing Spline Regression. *Journal of Computational and Graphical Statistics*, 17(1):21–37, 2008.

[92] M. Zondervan-Zwijnenburg, W. van de Schoot-Hubeek, K. Lek, H. Hoijtink, and R. van de Schoot. Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations. *Frontiers in Psychology*, 8:90, 2017.

Appendix A. Proofs

Proof of Lemma 1. Notice that if (9) is respected, then the constraints are non-decreasing. Then, there exists at least a function $F^{\leftarrow}$ in $\mathcal{F}^{\leftarrow}$ such that the constraints are respected (e.g., the linear interpolant of the constraints). So, there exists a probability measure with $F^{\leftarrow}$ as a generalized quantile function. □

Proof of Lemma 2. Since $[\eta_0, \eta_1]$ is bounded, one can define a standardized intensity parameter $\theta \in \Theta = [-1, 1]$ as:

$$\theta(b) = \frac{p_\alpha - b}{p_\alpha - \eta_1} \mathbb{1}_{\{b > p_\alpha\}}(b) + \frac{b - p_\alpha}{p_\alpha - \eta_0} \mathbb{1}_{\{b < p_\alpha\}}(b).$$

Equivalently, one can express b in terms of θ, which directly provides the expression of $b_\alpha(\boldsymbol{\eta}, \theta)$. □

Proof of Lemma 3. Preserving the midpoint of Ω_X while perturbing its width requires that, for any couple $(b_0, b_1) \in \mathbb{R}^2$, that

$$\begin{cases} \dfrac{b_0 + b_1}{2} = \dfrac{\omega_0 + \omega_1}{2} \\ b_1 - b_0 = \kappa(\omega_1 - \omega_0) \end{cases} \iff \begin{cases} b_1 = \dfrac{\omega_1(\kappa + 1) - \omega_0(\kappa - 1)}{2} \\ b_0 = \dfrac{\omega_0(\kappa + 1) - \omega_1(\kappa - 1)}{2} \end{cases}$$

where $\kappa \in [\frac{1}{\eta}, \eta]$. Using the transformation

$$\theta(\kappa) = \begin{cases} -\dfrac{\kappa - 1}{\frac{1}{\eta} - 1} & \text{if} \quad \dfrac{1}{\eta} \leq \kappa < 1 \\ 0 & \text{if} \quad \kappa = 1 \\ \dfrac{\kappa - 1}{\eta - 1} & \text{if} \quad 1 < \kappa < \eta \end{cases}$$

allows defining the formulas for b_0 and b_1 provided in the result's statement. □

Proof of Lemma 4. **(i)** Suppose that P is empirical. Notice that the empirical copula (see Section 2.2.2) only depends on the *ranks* of the observed data points. Since each $F_i^{\leftarrow}$ is strictly monotone increasing, the ranks between the initial and perturbed data points are preserved. Hence, the empirical copula between X and $\widetilde{X}$ is the same.
(ii) Let $F \in \mathcal{F}$, and recall that if $F^{\leftarrow}$ is strictly increasing then from [31], for all $u \in [0, 1]$

$$(F \circ F^{\leftarrow})(u) = u$$

Now let $F_1, \ldots, F_d \in \mathcal{F}$, such that $F_i^{\leftarrow}$ is strictly increasing, and denote:

$$\begin{array}{rcl} F : \mathbb{R} & \to & [0,1]^d \\ (u_1, \ldots, u_d)^\top & \mapsto & (F_1(u_1), \ldots, F_d(u_d))^\top \end{array}$$

One then has that:

$$F(T(X)) = F_P(X) \text{ a.s.}$$

and hence, X and $T(X)$ have the same copula. □

Proof of Lemma 5. Notice that, from Lemma 4, every probability measure in $\widetilde{Z}(P, \theta)$ has the same copula as P. Leveraging the work in [1] (Proposition 1.1), if P and Q share the same copula, one can rewrite their 2-Wasserstein distance as:

$$W_2^2(P, Q) = \sum_{i=1}^{d} W_2^2(P_i, Q_i) = \sum_{i=1}^{d} \int_0^1 \left(F_{P_i}^{\rightarrow}(x) - F_{Q_i}^{\rightarrow}(x)\right)^2 dx \tag{A.1}$$

Moreover, noticing that each marginal perturbation class $\mathcal{Q}_i(\theta)$ can be written as constraints on the generalized inverses of the cdf of Q_i. Hence, minimizing (A.1) entails minimizing each univariate transportation problem under marginal constraints. Finally, the perturbation map T is thus optimal between P and Q. □

Proof of Proposition 1. First, note that the intervals $A_i, i = 1, \ldots, K$ are disjoint. Moreover for any $i = 1, \ldots, K-1$, consider the four cases:

1. If $\alpha_i < \beta_i < \alpha_{i+1}$ and, then $A_i = (\alpha_i, \beta_i]$;
2. If $\beta_i < \alpha_i < \beta_{i+1}$ and, then $A_i = (\beta_i, \alpha_i]$;
3. If $\alpha_i < \beta_i$ and assume that $\alpha_{i+j} < \beta_{i+j-1}$ for $j = 1, \ldots, m$ where $m \leq K - i$ is some non-negative integer, then $A_i = (\alpha_i, \alpha_{i+1}]$, additionally for $j = i+1, \ldots, i+m-1$, $A_j = (\alpha_j, \alpha_{j+1}]$ and finally $A_{i+m} = (\alpha_{i+m}, \beta_{i+m}]$;
4. If $\beta_i < \alpha_i$ and assume that $\alpha_{i+j} < \beta_{i+j+1}$ for $j = 1, \ldots, m$ where $m \leq K - i - 1$ is some non-negative integer, then $A_i = (\beta_i, \alpha_i]$ and for $j = i+1, \ldots, i+m$, $A_j = (\alpha_{j-1}, \alpha_j]$.

The integral can be decomposed as follows:

$$\int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx = \int_{\overline{A}} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx + \sum_{i=1}^{K} \int_{A_i} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx$$

where

$$\int_{\overline{A}} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \geq 0.$$

Since the quantile constraints are of the form:

$$L(\alpha_i) \leq b_i \leq L\left(\alpha_i^+\right).$$

one can always write $L(y) = b_i + h(y)$ for $y \in A_i$, and where h is an non-decreasing, left-continuous function. Moreover, note that:

- $h(y)$ is non-negative, and $F_P^{\rightarrow}(y) - b_i \leq 0$ if A_i falls in cases 2. and 4.
- $h(y)$ is non-positive, and $F_P^{\rightarrow}(y) - b_i \geq 0$ if A_i falls in cases 1. and 3.

Then one has:

$$\begin{aligned}
\int_{A_i} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx &= \int_{A_i} \left(L(x) - b_i - h(y)\right)^2 dx \\
&= \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx + \int_{A_i} h(x)^2 dx \\
&\quad - 2\int_{A_i} h(x)\left(F_P^{\rightarrow}(x) - b_i\right) dx \\
&\geq \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx
\end{aligned}$$

since $h(x)$ and $F_P^{\rightarrow}(x) - b_i$ have different sign. Due to the constraint and the left-continuous non-decreasing nature of L, this bound is tight and is attained if and only if $h(y) = 0$ for all $y \in A_i$. Globally, this entails that

$$\int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \geq \sum_{i=1}^{K} \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx$$

and this tight bound is uniquely attained by the left-continuous non-decreasing function defined as

$$F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A} \\ b_i & \text{if } y \in A_i, \quad i = 1, \ldots, K. \end{cases}$$

□

Proof of Theorem 1 (ingredients). This proof relies on the following results from [67, 68, 54], and further recalled in [82]. They involve sum-of-squares (SOS) polynomials, which can be defined as follows.

Definition 5 (SOS polynomials)**.** *A polynomial S of even degree p is said to be a SOS polynomial if, for $m \in \mathbb{N}^*$, there exists $s_1, \ldots, s_m$ polynomials of degree at most equal to $\frac{d}{2}$, and such that, $\forall x \in \mathbb{R}$:*

$$S(x) = \sum_{i=1}^{m} \big(s_i(x)\big)^2.$$

Theorem 2. *Let $t_0, t_1 \in \mathbb{R}$ such that $t_0 < t_1$, and let $p \in \mathbb{N}^*$.*

(i) *A univariate polynomial S of even degree $d = 2p$ is non-negative on $[t_0, t_1]$ if and only if it can be written as, $\forall x \in [t_0, t_1]$*

$$S(x) = Z(x) + (x - t_0)(t_1 - x)W(x)$$

where Z is a SOS polynomial of degree at most equal to d, and W is an SOS polynomial of degree at most equal to $d - 2$.

(i) *An univariate polynomial S of odd degree $d = 2p+1$ is non-negative on $[t_0, t_1]$ if and only if it can be written as, $\forall x \in [t_0, t_1]$*

$$S(x) = (x - t_0)Z(x) + (t_1 - x)W(x)$$

where Z, W are SOS polynomials of degree at most equal to d.

It is important to note that Theorem 2 is quite general in the sense that it allows for extensions to multivariate polynomials (i.e., polynomials taking values from $\mathbb{R}^d$). As pointed out in [29] (Thm. 1.4.2), nonnegative polynomials on compact intervals can also be defined as a linear combination of squared polynomials. It may facilitate the identification of the nonnegative polynomials' coefficients, as done in [61] in the context of statistical learning. However, for the sake of potential future genericity, the direct powerful link between SOS polynomials and semi-definite positive matrices is leveraged, as expressed in the following theorem.

Theorem 3. *Let S be a univariate polynomial of even degree $d = 2p$, with coefficients $s = (s_0, \dots, s_d)$, and denote x_p the usual monomial basis of polynomials of degree at most equal to p, i.e., $x_p = (1, x, x^2, \dots, x^{p-1}, x^p)^\top$. S is an SOS polynomial if and only if there exists a $(p \times p)$ symmetric semi-definite positive (SDP) matrix*

$$\Gamma = \left[\Gamma_{ij}\right]_{i,j=1,\dots,p}$$

that satisfies, $\forall x \in \mathbb{R}$,

$$S(x) = x_p^\top \Gamma x_p.$$

Moreover, for $k = 0, \dots, d$, let $\mathbb{I}_k^p$ be the $(p \times p)$ matrix defined by, for $i, j = 1, \dots, p$:

$$\left[\mathbb{I}_k^p\right]_{i,j} = \mathbb{1}_{\{i+j=k+2\}}(i, j).$$

Then one additionally has that, for $i = 0, \dots, d$

$$s_i = \langle \mathbb{I}_i^p, \Gamma \rangle_F = \sum_{j+k=i+2} \Gamma_{j,k} \tag{A.2}$$

where, $\langle ., . \rangle_F$ denotes the Frobenius norm on matrices.

Theorem 4. *Let $\mathbb{S}_n$ the subspace of real-valued symmetric matrices, in the vector space of square matrices. The set of symmetric SDP matrices Σ_N is a proper cone in $\mathbb{S}_n$, and thus is a closed convex set.*

A few results on the preservation of convexity of sets under transformations are also required. These lemmas can be found in [12].

Lemma 6 (Linear maps preserve convexity)**.** *Let V, W be two vector spaces over the same field F. Let $T : V \to W$ be a linear map, and let $\mathcal{C} \subset V$ be a convex set. Then the image of $\mathcal{C}$ under T, i.e., :*

$$T(C) = \{T(x) \in W \mid x \in C \subset V\}$$

is also a convex set.

Lemma 7 (Cartesian product of convex sets is a convex set)**.** *Let C_1 be a subset of $\mathbb{R}^m$ and C_2 be a convex subset of $\mathbb{R}^n$. Then, the Cartesian product $C_1 \times C_2$ is a convex subset of $\mathbb{R}^m \times \mathbb{R}^n$.*

Two additional results, proven beneath, are required before proceeding to the proof of Theorem 1.

Lemma 8. *The mapping in (A.2), $V : \mathbb{S}_p \to \mathbb{R}^{2p}$, defined, for any $\Gamma \in \mathbb{S}_p$, as:*

$$V(\Gamma) = \left(\sum_{j+k=i+2} \Gamma_{j,k} \right)_{i=0,\dots,2p}$$

is linear.

Proof of Lemma 8. We need to show that:

- For $A, B \in \mathbb{S}_p$, $T(A+B) = T(A) + T(B)$;
- For $\alpha \in \mathbb{R}$, $\Gamma \in \mathbb{S}_p$, $T(\alpha\Gamma) = \alpha T(\Gamma)$.

First, one has, for $i = 0, \dots, 2p$:

$$\begin{aligned}
\Big[T(A+B)\Big]_i &= \sum_{j+k=2p-i} \Big[A+B\Big]_{jk} \\
&= \sum_{j+k=i+2} A_{jk} + B_{jk} \\
&= \sum_{j+k=i+2} A_{jk} + \sum_{j+k=i+2} B_{jk} \\
&= \Big[T(A)\Big]_i + \Big[T(B)\Big]_i
\end{aligned}$$

since it holds for $i = 0, \dots, 2p$, it entails:

$$T(A+B) = T(A) + T(B).$$

Moreover, one has, for $i = 0, \dots, 2p$:

$$\begin{aligned}
\Big[T(\alpha\Gamma)\Big]_i &= \sum_{j+k=i+2} \alpha\Gamma_{jk} \\
&= \alpha \Big[T(\Gamma)\Big]_i
\end{aligned}$$

and since it holds for $i = 0, \dots, 2p$, it entails:

$$T(\alpha\Gamma) = \alpha T(\Gamma).$$

Hence T is a linear map between $\mathbb{S}_p$ and $\mathbb{R}^{2p}$. □

Lemma 9. *Let S be a univariate polynomial of degree d and $s = (s_0, \dots, s_d)^\top \in \mathbb{R}^{d+1}$ its coefficients. Let S' be its derivative, i.e., a polynomial of degree $d-1$, with coefficients $\breve{s} = (s_1, \dots, s_d)^\top \in \mathbb{R}^d$. Let Z and W be SOS polynomials, with coefficients z and w, and assume that S' is non-negative on $[t_0, t_1]$ as a combination of Z and W as in Theorem 2. Moreover, let*

$$D = \mathrm{diag}(1, 2, \dots, d)$$

be the $(d \times d)$ diagonal matrix with $(1, \dots, d)$ as a diagonal elements and denote the bloc-matrices

$$\overline{\mathcal{I}}_{i,d} = \begin{pmatrix} I_d \\ \mathbf{0}_{i,d} \end{pmatrix}, \quad \underline{\mathcal{I}}_{i,d} = \begin{pmatrix} \mathbf{0}_{i,d} \\ I_d \end{pmatrix}, \quad \overline{\underline{\mathcal{I}}}_{i,d} = \begin{pmatrix} \mathbf{0}_{i,d} \\ I_d \\ \mathbf{0}_{i,d} \end{pmatrix}$$

where $\mathbf{0}_{i,d}$ *denotes the* $(i \times d)$ *matrix of zeros, and* I_d *be the* $(d \times d)$ *identity matrix. If* d *is odd, then* $z \in \mathbb{R}^d$ *and* $w \in \mathbb{R}^{d-2}$ *and furthermore*

$$\breve{s} = Az + Bw$$

where A *and* B *are* $(d \times d)$ *and* $(d \times d-2)$ *matrices, respectively. If the degree* d *of* S *is even, one has that* $z, w \in \mathbb{R}^{d-1}$ *and furthermore:*

$$\breve{s} = Cz + Dw.$$

where C *and* D *are* $(d \times d-1)$ *matrices. More specifically,*

$$A = \mathcal{D}_d^{-1}, \qquad B = \mathcal{D}_d^{-1}\left((t_0+t_1)\overline{\mathcal{I}}_{1,d-2} - \underline{\mathcal{I}}_{2,d-2} - t_0t_1\overline{\mathcal{I}}_{2,d-2}\right),$$
$$C = \mathcal{D}_d^{-1}\left(\underline{\mathcal{I}}_{1,d-1} - t_0\overline{\mathcal{I}}_{1,d-1}\right), \qquad D = \mathcal{D}_d^{-1}\left(t_1\overline{\mathcal{I}}_{1,d-1} - \underline{\mathcal{I}}_{1,d-1}\right).$$

Proof of Lemma 9. First, assume that S is a polynomial of odd degree $d = 2p+1$, meaning that its derivative, S', is a polynomial of even degree $2p$. From Theorem 2, one has that $S'(x)$ is positive on an interval $[t_0, t_1]$ if and only if it can be expressed as :

$$S'(x) = Z(x) + (x - t_0)(t_1 - x)W(x)$$

where Z is an SOS polynomial of degree at most equal to $d-1$ and W is an SOS polynomial of degree at most equal to $d-3$. Denote $\breve{s} = (s_1, \ldots, s_d) \in R^d$ the coefficients of S' and $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$ and $w = (w_1, \ldots, w_{d-2)} \in \mathbb{R}^{d-2}$ the coefficients of Z and W respectively. One has that :

$$\begin{aligned} S'(x) &= \sum_{i=1}^{d} i s_i x^{i-1} \\ &= \sum_{j=0}^{d-1} (j+1) s_{i+1} x^i \end{aligned}$$

and if S' is assumed to be non-negative on $[t_0, t_1]$

$$\begin{aligned} S'(x) &= Z(x) + (x - t_0)(t_1 - x)W(x) \\ &= \sum_{j=0}^{d-1} z_{j+1} x^j + (-x^2 + (t_0+t_1)x - t_0t_1) \sum_{j=0}^{d-3} w_{j+1} x^j \end{aligned}$$

leading to the following identification :

$$\begin{cases} s_1 = z_1 - t_0t_1w_1 \\ s_2 = \frac{1}{2}\left(z_2 - t_0t_1w_2 + (t_0+t_1)w_1\right) \\ s_i = \frac{1}{i}\left(z_i - t_0t_1w_i + (t_0+t_1)w_{i-1} - w_{i-2}\right), \quad \text{for } i = 3, \ldots, d-2 \\ s_{d-1} = \frac{1}{d-1}\left(z_{d-1} + (t_0+t_1)w_{d-2} - w_{d-3}\right) \\ s_d = \frac{1}{d}\left(z_{d-1} - w_{d-2}\right), \end{cases}$$

or, written in a matrix form:

$$\breve{s} = \mathcal{D}_d^{-1}\left(z + \left((t_0+t_1)\overline{\mathcal{I}}_{1,d-2} - \underline{\mathcal{I}}_{2,d-2} - t_0t_1\overline{\mathcal{I}}_{2,d-2}\right)w\right).$$

If S is assumed to be a polynomial of even degree $d = 2p$, S' is necessarily odd degree. From Theorem 2, one has that $S'(x)$ is positive on an interval $[t_0, t_1]$ if and only if it can be expressed as :

$$S'(x) = (x - t_0)Z(x) + (t_1 - x)W(x)$$

where Z and W are SOS polynomials of degree at most equal to $d-2$ with $z=(z_1,\dots,z_{d-1})\in\mathbb{R}^{d-1}$ and $w=(w_1,\dots,w_{d-1)}\in\mathbb{R}^{d-1}$ as coefficients, respectively. It leads to the following identification:

$$\begin{cases} s_1 = -t_0 z_1 + t_1 w_1 \\ s_i = \frac{1}{i}\left(z_{i-1} - t_0 z_i + t_1 w_i - w_{i-1}\right) \quad \text{for } i=2,\dots,d-1 \\ s_d = \frac{1}{d}\left(z_{d-1} - w_{d-1}\right), \end{cases}$$

which can be written in matrix form as

$$\breve{s} = \mathcal{D}_d^{-1}\left(\left(\underline{\mathcal{I}}_{1,d-1} - t_0\overline{\mathcal{I}}_{1,d-1}\right)z + \left(t_1\overline{\mathcal{I}}_{1,d-1} - \underline{\mathcal{I}}_{1,d-1}\right)w\right).$$

□

We can now proceed to prove Theorem 1.

Proof of Theorem 1 (rationale). This rationale can be broken down in two steps: **(a)** proving that the objective function (23) can indeed be written in a quadratic form, and:**(b)** proving that the problem constraints form a feasible set in $\mathbb{R}^{d+1}$ which is closed and convex.

(a) Notice first that the initial objective function

$$\int_{t_0}^{t_1}(L(x)-F_P^{\rightarrow}(x))^2dx$$

where $L\in\mathbb{R}[x]_{\leq d}$ with coefficients $s\in\mathbb{R}^{d+1}$, can be rewritten as:

$$\begin{aligned}\int_{t_0}^{t_1}(F_P^{\rightarrow}(x)-L(x))^2dx &= \int_{t_0}^{t_1}(\sum_{i=0}^{d}s_ix^i - F_P^{\rightarrow}(x))^2dx \\ &= \int_{t_0}^{t_1}\left(\left(\sum_{i=0}^{d}s_ix^i\right)^2 + (F_P^{\rightarrow}(x))^2 - 2\sum_{i=0}^{d}s_ix^iF_P^{\rightarrow}(x)\right)dx \\ &= \int_{t_0}^{t_1}\left(\sum_{i=0}^{d}s_ix^i\right)^2dx - 2\sum_{i=0}^{d}s_i\int_{t_0}^{t_1}x^iF_P^{\rightarrow}(x)dx \\ &\quad + \int_{t_0}^{t_1}\left(F_P^{\rightarrow}(x)\right)^2dx.\end{aligned}$$

Note that

$$\begin{aligned}\int_{t_0}^{t_1}\left(\sum_{i=0}^{d}s_ix^i\right)^2dx &= \sum_{i=0}^{d}\sum_{j=0}^{d}s_is_j\int_{t_0}^{t_1}x^{i+j}dx \\ &= s^\top Ms\end{aligned}$$

where M is the moment matrix of the Lebesgue measure on $[t_0,t_1]$, i.e., defined entry-wise, for $i,j=1,\dots,d+1$ as

$$M_{ij}=\int_{t_0}^{t_1}x^{i+j-2}dx=\frac{(t_1)^{i+j-1}-(t_0)^{i+j-1}}{i+j-1}.$$

and further notice that M is thus positive definite since, for any $u\in\mathbb{R}^{d+1}$,

$$u^\top Mu=\int_{t_0}^{t_1}\left(\sum_{i=0}^{d}u_{i+1}x^i\right)^2dx\geq 0$$

is always non-negative, and equal to 0 if and only if $u_i = 0, i = 1, \dots, d+1$. Moreover, note that:

$$\sum_{i=0}^{d} s_i \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx = s^\top r$$

where $r \in \mathbb{R}^{d+1}$ is the moment vector of G with respect to the Lebesgue measure on $[t_0, t_1]$, defined for $i = 0, \dots, d$ as:

$$r_i = \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx$$

Since a polynomial is completely characterized by its coefficients, searching for:

$$S^* = \underset{L \in \mathbb{R}[x]_{\leq d}}{\operatorname{argmin}} \int_{t_0}^{t_1} (L(x) - F_P^{\rightarrow}(x))^2 dx$$

is equivalent to finding the coefficients s^* of S^*, i.e.,

$$s^* = \underset{s \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \; s^\top M s - 2 s^\top r$$

and thus proving the first part of the proposition.

(b) Notice that the interpolation constraints

$$\begin{cases} S(t_0) = b_0 \\ S(t_1) = b_1 \end{cases}$$

can be written as

$$\begin{cases} s^\top \mathbf{t_0}^d = b_0 \\ s^\top \mathbf{t_1}^d = b_1 \end{cases}$$

where, for $a \in \mathbb{R}$, one denote $\mathbf{a}^d$ the vector of powers of a up to d, i.e., $\mathbf{a}^d = (1, a, \dots, a^{d-1}, a^d) \in \mathbb{R}^{d+1}$. Moreover, by letting:

$$\mathbf{T} = \begin{pmatrix} \mathbf{t_0}^d \\ \mathbf{t_1}^d \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix},$$

where $\mathbf{T}$ is a $(2 \times d + 1)$ bloc-matrix, the constraint can be written as:

$$Ts = b.$$

Furthermore, notice that

$$\mathcal{C}_0 = \{s \in \mathbb{R}^{d+1} \mid Ts = b\}$$

is a convex subset of $\mathbb{R}^{d+1}$, since the equality constraints are linear. Concerning the monotonicity constraint

$$S'(x) \geq 0, \quad \forall x \in [t_0, t_1],$$

from Lemma 9 one can quite generically write

$$\begin{pmatrix} s_d \\ \vdots \\ s_1 \end{pmatrix} = T_0(z, w) := Az + Bw$$

where z and w are the coefficient of SOS polynomials of degrees depending on d. Additionally, notice that the mapping $T_0 : \mathbb{R}^d \times \mathbb{R}^{d-2} \to \mathbb{R}^d$ is linear. Next, let $V_1 : \mathbb{S}_p \to \mathbb{R}^{2p}$, and $V_2 : \mathbb{S}_q \to \mathbb{R}^{2q}$ be

defined as in (A.2), where $p = d - 1/2$ and $q = d - 3/2$ if d is odd, or $p = d - 2/2$ and $q = d - 2/2$ if d is even, and note that both mappings are linear thanks to Lemma 8.

Moreover, denote the following sets:

$$\mathcal{Z} = \{V_1(E) \mid E \in \Sigma_p\}, \quad \mathcal{W} = \{V_2(E) \mid E \in \Sigma_{p-1}\}$$

and notice the polynomial Z (resp. W) is SOS if and only its coefficients z (resp. w) are in $\mathcal{Z}$ (resp. $\mathcal{W}$) thanks to Theorem 4. In addition again, notice that, thanks to Lemma 6, and due to the fact that Σ_p is a closed convex set in $\mathbb{S}_p$ as per Theorem 4, both $\mathcal{Z}$ and $\mathcal{W}$ are convex subsets of $\mathbb{R}^{2p}$ and $\mathbb{R}^{2q}$ respectively. Besides, thanks to Lemma 7, the set $\mathcal{Z} \times \mathcal{W}$ is a convex subset of $\mathbb{R}^{2p} \times \mathbb{R}^{2q}$ as well. Moreover, let

$$\mathcal{C}_1 = \left\{ \begin{pmatrix} T_0(w,z) \\ x \end{pmatrix} \in \mathbb{R}^{d+1} \mid x \in \mathbb{R}, \quad (z,w) \in \mathcal{Z} \times \mathcal{W} \right\}$$

and note that it is a convex subset of $\mathbb{R}^{d+1}$ due to the fact that T_0 is a linear map.

Finally, since both $\mathcal{C}_0$ and $\mathcal{C}_1$ are convex sets, their intersection:

$$\mathcal{K} = \mathcal{C}_0 \cap \mathcal{C}_1$$

is as well, and note that any element $s \in \mathcal{K}$ are the coefficients of a polynomial respecting both equality and monotonicity constraints. In other words, $\mathcal{K}$ is the feasible set of coefficients of the initial optimization problem. □

Appendix B. Computing moment vector of arbitrary quantile functions

One wishes here at computing the vector described in (25). In the case where P is an empirical measure built on a n-sample, one has that for $[t_0, t_1] \in [0,1]$, $i = 0, \dots, p$:

$$\begin{aligned} r_i = & \frac{1}{i+1} \left[\sum_{j \in J} \frac{X_{(j)}}{n^{i+1}} \left((j+1)^{i+1} - j^{i+1} \right) \right. \\ & + X_{(\bar{j})} \left(t_1^{i+1} - \left(\frac{\bar{j}}{n} \right)^{i+1} \right) \\ & \left. + X_{(\underline{j}-1)} \left(\left(\frac{\underline{j}}{n} \right)^{i+1} - t_0^{i+1} \right) \right] \end{aligned}$$

where $J = \{i \in \mathbb{N} \mid \lfloor nt_0 \rfloor < i < \lfloor nt_1 \rfloor\}$, $\bar{j} = \lfloor t_1 n \rfloor$, $\underline{j} = \lfloor t_0 n \rfloor + 1$, and where $X_{(j)}$ denotes the j-th order statistic of the observe sample. In cases where $F_P^{\leftarrow}$ is continuous, it is possible to use numerical quadrature methods in order to evaluate each integral composing the elements r_i of r.