Robustness assessment of black-box models to feature perturbations

Finding optimal probability measures under distributional constraints

[1]**Université du Québec à Montréal** (UQÀM)
[2]**Institut Intelligence et Données** (IID) - Université Laval

Montréal - Guanajuato Workshop on Probability and Machine Learning
Centro de Investigación en Matemáticas - Guanajuato, Gto, México
February 25-27, 2025

Marouane Il Idrissi[12]
Joint work with: Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, Jean-Michel Loubes

Robustness to input perturbations

How does an ML model behave when the features are perturbed?

Robustness to input perturbations

How does an ML model behave when the features are perturbed?

Why is this question important?

- **Prospective studies** to anticipate risks
- **Exploratory studies** to assess a model's generalization capabilities
- **Expertise injection** to the feature's probabilistic modeling
- **Enhance the overall confidence** in the predictive model

How does an ML model behave when the features are perturbed?

Why is this question important?

- **Prospective studies** to anticipate risks
- **Exploratory studies** to assess a model's generalization capabilities
- **Expertise injection** to the feature's probabilistic modeling
- **Enhance the overall confidence** in the predictive model

Illustrative example: A predictive model of a river's **water level**

☞ The **range of values** of the **riverbed roughness coefficient** K_S is expected to change
Due to reasons (e.g., climate change) its **support** changes

How does an ML model behave when the features are perturbed?

Why is this question important?

- **Prospective studies** to anticipate risks
- **Exploratory studies** to assess a model's generalization capabilities
- **Expertise injection** to the feature's probabilistic modeling
- **Enhance the overall confidence** in the predictive model

Illustrative example: A predictive model of a river's **water level**

☞ The **range of values** of the **riverbed roughness coefficient** K_S is expected to change
Due to reasons (e.g., climate change) its **support** changes

Question: **What are the consequences on the predictions of the river's water level?**
e.g., how is the predicted flood risk going to change?

How does an ML model behave when the features are perturbed?

Why is this question important?

- **Prospective studies** to anticipate risks
- **Exploratory studies** to assess a model's generalization capabilities
- **Expertise injection** to the feature's probabilistic modeling
- **Enhance the overall confidence** in the predictive model

Illustrative example: A predictive model of a river's **water level**

☞ The **range of values** of the **riverbed roughness coefficient** K_S is expected to change
Due to reasons (e.g., climate change) its **support** changes

Question: **What are the consequences on the predictions of the river's water level?**
e.g., how is the predicted flood risk going to change?

Intuition: Pass a **perturbed version** $\widetilde{K_S}$ **of** K_S through the model and see what changes

Ingredients for feature perturbations

☞ Let $X = (X_1, \ldots, X_d)^\top$ be an ($\mathbb{R}^d$-valued) random vector
i.e., the initial features

Ingredients for feature perturbations

☞ Let $X = (X_1, \ldots, X_d)^\top$ be an ($\mathbb{R}^d$-valued) random vector
i.e., the initial features

☞ Let $\widetilde{X} = (\widetilde{X_1}, \ldots, \widetilde{X_d})^\top$ be another ($\mathbb{R}^d$-valued) random vector
i.e., the perturbed features

Ingredients for feature perturbations

☞ Let $X = (X_1, \dots, X_d)^\top$ be an ($\mathbb{R}^d$-valued) random vector
i.e., the initial features

☞ Let $\widetilde{X} = (\widetilde{X_1}, \dots, \widetilde{X_d})^\top$ be another ($\mathbb{R}^d$-valued) random vector
i.e., the perturbed features

☞ Suppose that $X \sim P$, and $\widetilde{X} \sim Q$
P is the law of/probability measure induced by X

Ingredients for feature perturbations

☞ Let $X = (X_1, \dots, X_d)^\top$ be an ($\mathbb{R}^d$-valued) random vector
i.e., the initial features

☞ Let $\widetilde{X} = (\widetilde{X_1}, \dots, \widetilde{X_d})^\top$ be another ($\mathbb{R}^d$-valued) random vector
i.e., the perturbed features

☞ Suppose that $X \sim P$, and $\widetilde{X} \sim Q$
P is the law of/probability measure induced by X

☞ Let $\mathcal{P}(\mathbb{R}^d)$ be the set of (joint) probability measures on $\mathbb{R}^d$

Ingredients for feature perturbations

☞ Let $X = (X_1, \dots, X_d)^\top$ be an ($\mathbb{R}^d$-valued) random vector
i.e., the initial features

☞ Let $\widetilde{X} = (\widetilde{X_1}, \dots, \widetilde{X_d})^\top$ be another ($\mathbb{R}^d$-valued) random vector
i.e., the perturbed features

☞ Suppose that $X \sim P$, and $\widetilde{X} \sim Q$
P is the law of/probability measure induced by X

☞ Let $\mathcal{P}(\mathbb{R}^d)$ be the set of (joint) probability measures on $\mathbb{R}^d$

☞ Let $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R}^d)$ be the set of probability measures **that respect a certain set of constraints**
e.g., for $Q \in \mathcal{C}$, if $Y \sim Q$, then $\mathbb{E}[Y] = \eta \in \mathbb{R}$

Ingredients for feature perturbations

☞ Let $X = (X_1, \ldots, X_d)^\top$ be an ($\mathbb{R}^d$-valued) random vector
i.e., the initial features

☞ Let $\widetilde{X} = (\widetilde{X_1}, \ldots, \widetilde{X_d})^\top$ be another ($\mathbb{R}^d$-valued) random vector
i.e., the perturbed features

☞ Suppose that $X \sim P$, and $\widetilde{X} \sim Q$
P is the law of/probability measure induced by X

☞ Let $\mathcal{P}(\mathbb{R}^d)$ be the set of (joint) probability measures on $\mathbb{R}^d$

☞ Let $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R}^d)$ be the set of probability measures **that respect a certain set of constraints**
e.g., for $Q \in \mathcal{C}$, if $Y \sim Q$, then $\mathbb{E}[Y] = \eta \in \mathbb{R}$

☞ Let $\mathcal{D}$ be a discrepancy between probability measures
e.g., f-divergences, integral probability metrics

Feature perturbation problem

We define Q as the **solution of a constrained optimization problem** over **probability measures**.

Feature perturbation problem

We define Q as the **solution of a constrained optimization problem** over **probability measures**.

Formally, in general,

$$Q \in \underset{P' \in \mathcal{P}(\mathbb{R}^d)}{\operatorname{argmin}} \quad \mathcal{D}\left(P, P'\right) \qquad \text{s.t.} \quad P' \in \mathcal{C}$$

Feature perturbation problem

We define Q as the **solution of a constrained optimization problem** over **probability measures**.

Formally, in general,

$$
\begin{aligned}
Q \in \operatorname*{argmin}_{P' \in \mathcal{P}(\mathbb{R}^d)} \quad & \mathcal{D}\left(P, P'\right) \\
\text{s.t.} \quad & P' \in \mathcal{C}
\end{aligned}
$$

Interpretation: Q is the probability measure **closest** to the initial measure P, that **respect the constraints characterizing** $\mathcal{C}$.

Feature perturbation problem

We define Q as the **solution of a constrained optimization problem** over **probability measures**.

Formally, in general,

$$Q \in \underset{P' \in \mathcal{P}(\mathbb{R}^d)}{\operatorname{argmin}} \quad \mathcal{D}\left(P, P'\right)$$
$$\text{s.t.} \quad P' \in \mathcal{C}$$

Interpretation: Q is the probability measure **closest** to the initial measure P, that **respect the constraints characterizing** $\mathcal{C}$.

Different choices of $\mathcal{D}$ and $\mathcal{C}$ lead to different Qs with different properties

Kullback-Leibler divergence, and constraints on generalized moments

In Lemaitre et al. (2015), they studied the case where:

- P is univariate
- $\mathcal{D}(\cdot,\cdot)$ is the Kullback-Leibler divergence, i.e.,

$$\mathrm{KL}(P,P') = \int_{\mathbb{R}} \log\left(\frac{dP(x)}{dP'(x)}\right) dx, \quad \text{provided the integral exists and } P \ll P'$$

- $\mathcal{C}$ is characterized using **constraints on generalized moments**, i.e., of the form

$$\int_{\mathbb{R}} g_k(x) dP'(x) = \eta_k, \quad k = 1, \dots, K$$

  for some functions g_k such that the above integral exists, and $\eta_k \in \mathbb{R}$.

Theorem *(Lemaitre et al. 2015)*. If P has density f_P and Q is further restricted to $P \ll Q$ with density f_Q, then Q is the unique solution of the feature perturbation optimization problem, and it is of the form:

$$f_Q(x) = f_P(x) \exp\left[\sum_{i=1}^{K} \lambda_k^* g_k(x) - \log\left(\int_{\mathbb{R}} f_P(x) \exp\left\{\sum_{i=1}^{K} \lambda_k^* g_k(x)\right\}\right)\right] = f_P(x)\alpha^*(x, \eta_1, \dots, \eta_K) \qquad (1)$$

Kullback-Leibler divergence, and constraints on generalized moments

In Lemaitre et al. (2015), they studied the case where:

- P is univariate
- $\mathcal{D}(\cdot,\cdot)$ is the Kullback-Leibler divergence, i.e.,

$$\mathrm{KL}(P,P') = \int_{\mathbb{R}} \log\left(\frac{dP(x)}{dP'(x)}\right) dx, \quad \text{provided the integral exists and } P \ll P'$$

- $\mathcal{C}$ is characterized using **constraints on generalized moments**, i.e., of the form

$$\int_{\mathbb{R}} g_k(x) dP'(x) = \eta_k, \quad k = 1, \ldots, K$$

  for some functions g_k such that the above integral exists, and $\eta_k \in \mathbb{R}$.

Theorem *(Lemaitre et al. 2015)*. If P has density f_P and Q is further restricted to $P \ll Q$ with density f_Q, then Q is the unique solution of the feature perturbation optimization problem, and it is of the form:

$$f_Q(x) = f_P(x) \exp\left[\sum_{i=1}^{K} \lambda_k^* g_k(x) - \log\left(\int_{\mathbb{R}} f_P(x) \exp\left\{\sum_{i=1}^{K} \lambda_k^* g_k(x)\right\}\right)\right] = f_P(x)\alpha^*(x, \eta_1, \ldots, \eta_K) \qquad (1)$$

The perturbed density is a reweighing the initial density

Kullback-Leibler divergence, and constraints on generalized moments

In Bachoc et al. (2023), they studied the same problem, but when, P is **a purely-atomic empirical measure** supported on n datapoints $(x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^d$), i.e.,

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

where δ is the Kronecker delta.

☞ In this case, Q must also be purely atomic and supported on the same datapoints

In fact,

Theorem *(Bachoc et al. 2023)*. The unique solution of the feature perturbation optimization problem Q is of the form

$$Q(x) = \frac{1}{n} \sum_{i=1}^{n} \alpha_i^*(x, \eta_1, \ldots \eta_k) \delta_{x_i}(x)$$

Kullback-Leibler divergence, and constraints on generalized moments

In Bachoc et al. (2023), they studied the same problem, but when, P is **a purely-atomic empirical measure** supported on n datapoints $(x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^d$), i.e.,

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

where δ is the Kronecker delta.

☞ In this case, Q must also be purely atomic and supported on the same datapoints

In fact,

Theorem *(Bachoc et al. 2023)*. The unique solution of the feature perturbation optimization problem Q is of the form

$$Q(x) = \frac{1}{n} \sum_{i=1}^{n} \alpha_i^*(x, \eta_1, \ldots \eta_k) \delta_{x_i}(x)$$

The perturbed datapoints are a reweighting the initial datapoints

These two results are based on the pioneering work of Csiszár (1975) on the geometry of probability measures under the KL divergence

From KL to Wasserstein

However, there are a couple of (practical) drawbacks to the KL-version of the perturbation:

- **The "nature" (continuous, atomic...) of Q depends on the one of P**
  Not possible to have a continuous Q for an empirical measure P

From KL to Wasserstein

However, there are a couple of (practical) drawbacks to the KL-version of the perturbation:

- **The "nature" (continuous, atomic...) of Q depends on the one of P**
  Not possible to have a continuous Q for an empirical measure P
- **Q is a reweighting of P: the evaluation points do not change (only their probability)**
  We're not going to evaluate the model on new datapoints

From KL to Wasserstein

However, there are a couple of (practical) drawbacks to the KL-version of the perturbation:

- **The "nature" (continuous, atomic...) of Q depends on the one of P**
  Not possible to have a continuous Q for an empirical measure P
- **Q is a reweighting of P: the evaluation points do not change (only their probability)**
  We're not going to evaluate the model on new datapoints
- **Generalized moments can be limited in practice**
  Key quantities (e.g., quantiles) cannot be expressed as generalized moments

From KL to Wasserstein

However, there are a couple of (practical) drawbacks to the KL-version of the perturbation:

- **The "nature" (continuous, atomic...) of Q depends on the one of P**
  Not possible to have a continuous Q for an empirical measure P
- **Q is a reweighting of P: the evaluation points do not change (only their probability)**
  We're not going to evaluate the model on new datapoints
- **Generalized moments can be limited in practice**
  Key quantities (e.g., quantiles) cannot be expressed as generalized moments

To alleviate these drawbacks, we propose:

- **Using the 2-Wasserstein distance**
  New **unobserved** datapoints, only condition on Q and P is finite variance

From KL to Wasserstein

However, there are a couple of (practical) drawbacks to the KL-version of the perturbation:

- **The "nature" (continuous, atomic...) of Q depends on the one of P**
  Not possible to have a continuous Q for an empirical measure P
- **Q is a reweighting of P: the evaluation points do not change (only their probability)**
  We're not going to evaluate the model on new datapoints
- **Generalized moments can be limited in practice**
  Key quantities (e.g., quantiles) cannot be expressed as generalized moments

To alleviate these drawbacks, we propose:

- **Using the 2-Wasserstein distance**
  New **unobserved** datapoints, only condition on Q and P is finite variance
- **Preserving the copula of X**
  Ceteris paribus interpretation of the model's change in behavior due to marginal perturbations

From KL to Wasserstein

However, there are a couple of (practical) drawbacks to the KL-version of the perturbation:

- **The "nature" (continuous, atomic...) of Q depends on the one of P**
  Not possible to have a continuous Q for an empirical measure P
- **Q is a reweighting of P: the evaluation points do not change (only their probability)**
  We're not going to evaluate the model on new datapoints
- **Generalized moments can be limited in practice**
  Key quantities (e.g., quantiles) cannot be expressed as generalized moments

To alleviate these drawbacks, we propose:

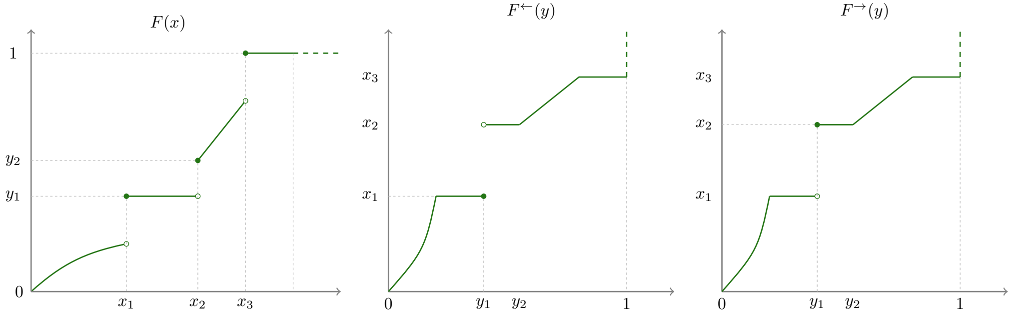
- **Using the 2-Wasserstein distance**
  New **unobserved** datapoints, only condition on Q and P is finite variance
- **Preserving the copula of X**
  Ceteris paribus interpretation of the model's change in behavior due to marginal perturbations
- **To put constraints on the quantiles of the marginal distributions**
  They always exist, and they're pretty meaningful in practice (e.g., risk measures)

Generalized quantile functions

Generalized quantile functions are the generalized inverses of a cdf (de la Fortelle 2015).

$$F_P^{\leftarrow}(a) = \sup\{t \in \mathbb{R} \mid F_P(t) < a\} \\ = \inf\{t \in \mathbb{R} \mid F_P(t) \geq a\}.$$

$$F_P^{\rightarrow}(a) = \sup\{t \in \mathbb{R} \mid F_P(t) \leq a\} \\ = \inf\{t \in \mathbb{R} \mid F_P(t) > a\},$$



☞ They **characterize** probability measures (Dufour 1995)

Quantile perturbation class

Quantile perturbation classes are defined using constraints of the form

$$F_{Q}^{\leftarrow}(\alpha) \geq b \geq F_{Q}^{\rightarrow}(\alpha)$$

i.e., the α-quantile of Q must be equal to some $b \in \mathbb{R}$.

Quantile perturbation class

Quantile perturbation classes are defined using constraints of the form

$$F_Q^{\leftarrow}(\alpha) \geq b \geq F_Q^{\rightarrow}(\alpha)$$

i.e., the α-quantile of Q must be equal to some $b \in \mathbb{R}$.

We can define the subset of $\mathcal{P}(\mathbb{R})$

$$\mathcal{Q}_{\mathcal{V}} = \{Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow} \in \mathcal{V}, \quad F_Q^{\leftarrow}(\alpha_i) \geq b_k \geq F_Q^{\rightarrow}(\alpha_i), \quad k = 1, \ldots, K\}.$$

where $\mathcal{V} \subseteq \mathcal{F}^{\leftarrow}$ is a subset of the **space of quantile functions**
i.e., a subset of the functions from $[0, 1]$ to $\mathbb{R}$ that are non-decreasing cadlàg

Wasserstein distance

For two multivariate probability measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$ **having the same copula** (Alfonsi and Jourdain 2014):

$$W_p^p(P, Q) = \sum_{i=1}^{d} W_p^p(P_i, Q_i). \tag{2}$$

where the $P_i, Q_i \in \mathcal{P}(\mathbb{R})$ are the marginal distributions of P and Q

Wasserstein distance

For two multivariate probability measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$ **having the same copula** (Alfonsi and Jourdain 2014):

$$W_p^p(P, Q) = \sum_{i=1}^{d} W_p^p(P_i, Q_i). \tag{2}$$

where the $P_i, Q_i \in \mathcal{P}(\mathbb{R})$ are the marginal distributions of P and Q

Each element of the sum reduces to (Santambrogio 2015):

$$W_p^p(P_i, Q_i) = \int_0^1 \left| F_{P_i}^{\rightarrow}(x) - F_{Q_i}^{\rightarrow}(x) \right|^p dx$$

Wasserstein distance

For two multivariate probability measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$ **having the same copula** (Alfonsi and Jourdain 2014):

$$W_p^p(P, Q) = \sum_{i=1}^{d} W_p^p(P_i, Q_i). \tag{2}$$

where the $P_i, Q_i \in \mathcal{P}(\mathbb{R})$ are the marginal distributions of P and Q

Each element of the sum reduces to (Santambrogio 2015):

$$W_p^p(P_i, Q_i) = \int_0^1 \left| F_{P_i}^{\rightarrow}(x) - F_{Q_i}^{\rightarrow}(x) \right|^p dx$$

Copula preservation:

Transportation problem in $\mathbb{R}^d$ $\iff$ d transportation problems in $\mathbb{R}$

Perturbation problem

Thus, the **marginal perturbation** problem can be expressed as:

$$
\begin{aligned}
Q = \underset{G \in \mathcal{P}(\mathbb{R})}{\operatorname{argmin}} \quad & W_2\left(P, G\right) \\
\text{s.t.} \quad & G \in \mathcal{Q}_{\mathcal{V}}
\end{aligned} \tag{3}
$$

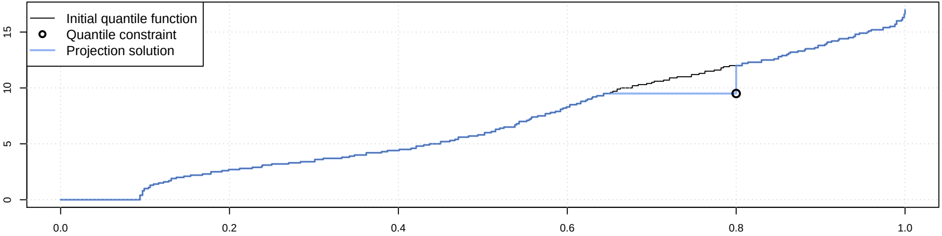
Proposition . The solution Q of the problem in Eq. (3) the unique probability measure with quantile function solution of

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

Analytical solution

The problem can be solved analytically if $\mathcal{V} = \mathcal{F}^{\leftarrow}$

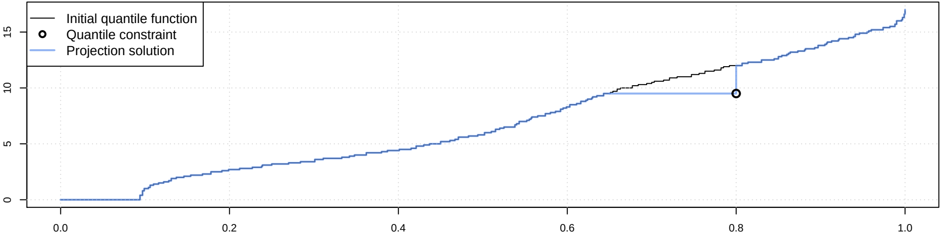
☞ Add atoms to P with sufficient mass to satisfy the constraint



Analytical solution

The problem can be solved analytically if $\mathcal{V} = \mathcal{F}^{\leftarrow}$

☞ Add atoms to P with sufficient mass to satisfy the constraint



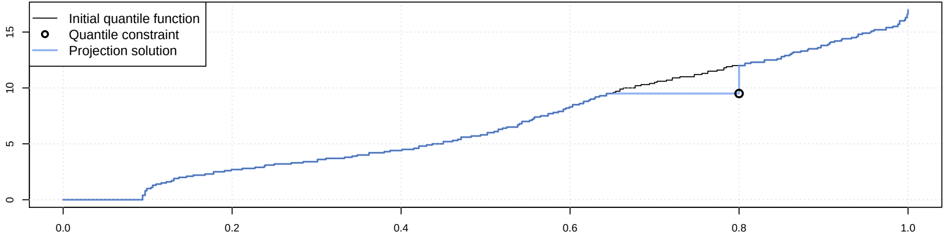
Not useful in practice: parts of the initial support cannot be observed anymore

We don't evaluate the model on these points anymore

Analytical solution

The problem can be solved analytically if $\mathcal{V} = \mathcal{F}^{\leftarrow}$

☞ Add atoms to P with sufficient mass to satisfy the constraint



Not useful in practice: parts of the initial support cannot be observed anymore

We don't evaluate the model on these points anymore

Idea: Set $\mathcal{V}$ to a subset of $\mathcal{F}^{\leftarrow}$ containing "smoother" quantile functions

i.e., at least continuous

Isotonic piece-wise continous polynomials

Our contribution: Non-decreasing polynomials that interpolate the constraints

Isotonic piece-wise continous polynomials

Our contribution: Non-decreasing polynomials that interpolate the constraints

We constraint the **derivative of the polynomials to be positive**

This can be achieved using SOS polynomials

Isotonic piece-wise continous polynomials

Our contribution: Non-decreasing polynomials that interpolate the constraints

We constraint the **derivative of the polynomials to be positive**

This can be achieved using SOS polynomials

☞ The resulting quantile function is **non-decreasing**

Because its derivative is positive

☞ The resulting quantile function is **continuous**

Because we interpolate the constraints

Isotonic piece-wise continous polynomials

Our contribution: Non-decreasing polynomials that interpolate the constraints

We constraint the **derivative of the polynomials to be positive**

This can be achieved using SOS polynomials

☞ The resulting quantile function is **non-decreasing**

Because its derivative is positive

☞ The resulting quantile function is **continuous**

Because we interpolate the constraints

And, it's a

Convex constrained quadratic optimization problem

Existence of **optimal solutions**, and it can be solved numerically!

Transportation map

Once we have access to the optimally perturbed marginal quantile functions $\{F_{Q_i}^{\leftarrow}\}_{i=1}^{d}$, we can define

$$\widetilde{X} = \begin{pmatrix} F_{Q_1}^{\leftarrow} \circ F_{P_1}(X_1) \\ \vdots \\ F_{Q_d}^{\leftarrow} \circ F_{P_d}(X_d) \end{pmatrix} \sim Q \tag{4}$$

Transportation map

Once we have access to the optimally perturbed marginal quantile functions $\left\{F^{\leftarrow}_{Q_i}\right\}_{i=1}^{d}$, we can define

$$\widetilde{X} = \begin{pmatrix} F^{\leftarrow}_{Q_1} \circ F_{P_1}(X_1) \\ \vdots \\ F^{\leftarrow}_{Q_d} \circ F_{P_d}(X_d) \end{pmatrix} \sim Q \tag{4}$$

Proposition . If each $F^{\leftarrow}_{Q_i}$ is strictly increasing, then this transportation map in Eq. (4) is optimal transportation map for the **joint feature perturbation problem** such that Q and P have the same copula.

Transportation map

Once we have access to the optimally perturbed marginal quantile functions $\{F^{\leftarrow}_{Q_i}\}_{i=1}^{d}$, we can define

$$\widetilde{X} = \begin{pmatrix} F^{\leftarrow}_{Q_1} \circ F_{P_1}(X_1) \\ \vdots \\ F^{\leftarrow}_{Q_d} \circ F_{P_d}(X_d) \end{pmatrix} \sim Q \tag{4}$$

Proposition . If each $F^{\leftarrow}_{Q_i}$ is strictly increasing, then this transportation map in Eq. (4) is optimal transportation map for the **joint feature perturbation problem** such that Q and P have the same copula.

In practice: Solve a **relaxed problem** where we require each $F^{\leftarrow}_{Q_i}$ to only be increasing

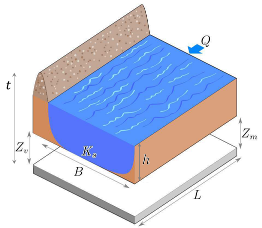
Computationally easier, and most of the time it still returns strictly increasing polynomials

Illustration: River Water Level

Simplified **numerical model of a river water level** (Iooss and Lemaître 2015).

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

- Q: River maximum annual water flow rate.
- K_s: **Strickler riverbed roughness coefficient.**
- Z_v: Downstream river level.
- Z_m: Upstream river level.
- L: River length.
- B: River width.



Structure probabiliste :

Input	Distribution	Support
Q	$\mathcal{G}(1013, 558)$ trunc.	$[500, 3000]$
K_s	$\mathcal{N}(30, 7)$ **trunc.**	$[20, 50]$
Z_v	$\mathcal{T}(49, 50, 51)$	$[49, 51]$
Z_m	$\mathcal{T}(54, 55, 56)$	$[54, 56]$
L	$\mathcal{T}(4990, 5000, 5010)$	$[4990, 5010]$
B	$\mathcal{T}(295, 300, 305)$	$[295, 305]$

The **inputs are correlated** by means of a Gaussian copula: $\rho(Q, K_s) = 0.5$ and $\rho(Z_v, Z_m) = \rho(L, B) = 0.3$.

Perturbation strategy

Prospective study:
What does a wider/narrower range of value for K_s entail on the river water level?

Strategy:
Perturb the **support** of K_s, i.e., the 0 and 1-quantiles.

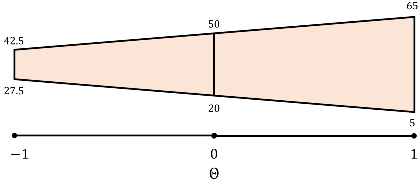
Perturbation strategy

Prospective study:
What does a wider/narrower range of value for K_s entail on the river water level?

Strategy:
Perturb the **support** of K_s, i.e., the 0 and 1-quantiles.

Intensity coefficient $\theta \in [-1, 1]$:

- $\theta = -1$: support's width is **halved**.
- $\theta = 0$: no change.
- $\theta = 1$: support's width is **doubled**.



Perturbation strategy

Prospective study:

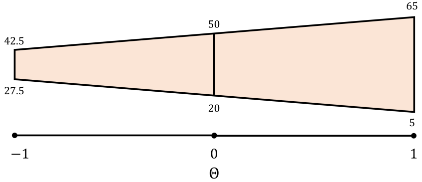
What does a wider/narrower range of value for K_s entail on the river water level?

Strategy:

Perturb the **support** of K_s, i.e., the 0 and 1-quantiles.

Intensity coefficient $\theta \in [-1, 1]$:

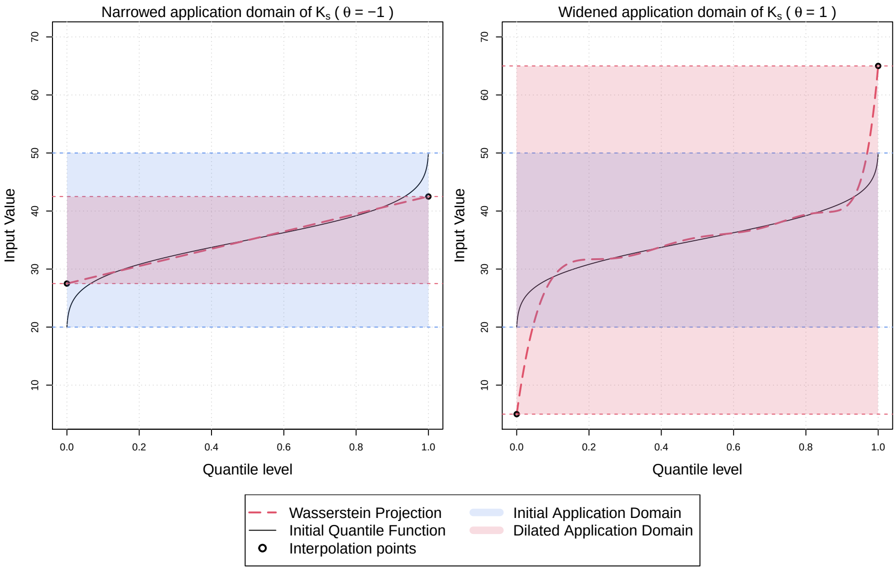
- $\theta = -1$: support's width is **halved**.
- $\theta = 0$: no change.
- $\theta = 1$: support's width is **doubled**.



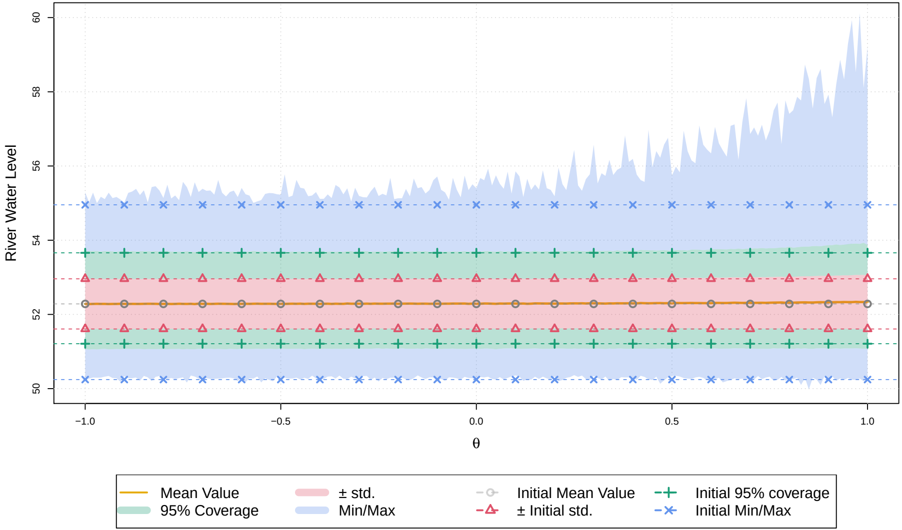
Smoothing constraint using isotonic piece-wise interpolating polynomials

Polynomial degree up to 12.

Perturbed Strickler coefficient



Effects of the perturbation on the numerical model



Surrogate model validation

Surrogate: 3 layer neural network, trained on **the initial (unperturbed) data**.

Training and validation data:

- **Training:** 500.000
- **Validation:** 50.000
- **Loss:** Mean squared error (MSE)

Data type	R^2	Loss
Training	99.5%	0.0119
Validation	99.5%	0.0120

Surrogate model validation

Surrogate: 3 layer neural network, trained on **the initial (unperturbed) data**.

Training and validation data:

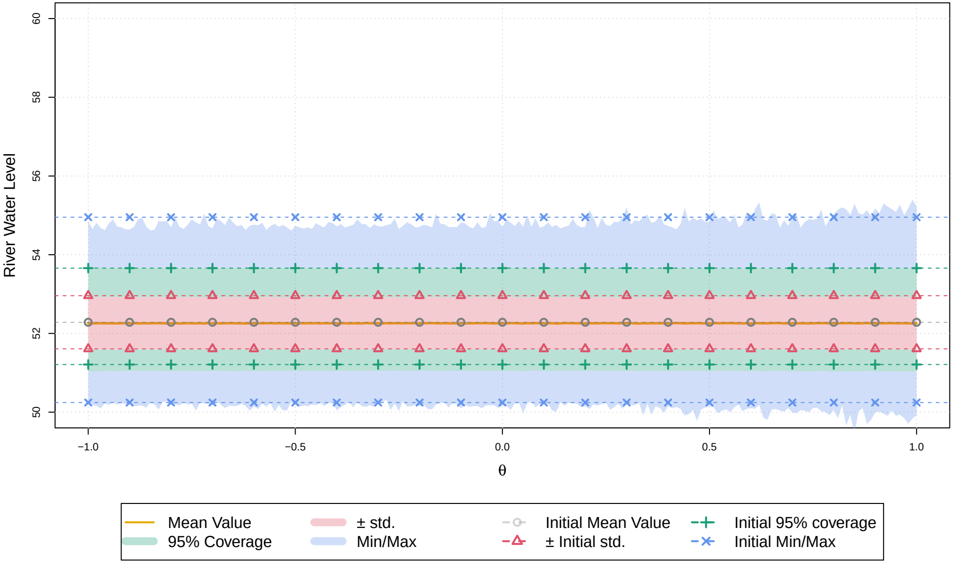
- **Training:** 500.000
- **Validation:** 50.000
- **Loss:** Mean squared error (MSE)

Data type	R^2	Loss
Training	99.5%	0.0119
Validation	99.5%	0.0120

Does the surrogate and the numerical model show the same behavior?

Did the neural network “learn” the numerical model?

Effects of the perturbations on the neural network



Conclusion and perspective

Take away messages:

☞ **Feature perturbations to assess the robustness of models**

Conclusion and perspective

Take away messages:

☞ **Feature perturbations to assess the robustness of models**
☞ **It amounts to solving an constrained optimization problem on probability measures**

Conclusion and perspective

Take away messages:

☞ **Feature perturbations to assess the robustness of models**
☞ **It amounts to solving an constrained optimization problem on probability measures**
☞ **There are many different ways to solve it**

Conclusion and perspective

Take away messages:

- ☞ **Feature perturbations to assess the robustness of models**
- ☞ **It amounts to solving an constrained optimization problem on probability measures**
- ☞ **There are many different ways to solve it**
- ☞ **Wasserstein distance + quantile constraints work well together**

Conclusion and perspective

Take away messages:

☞ **Feature perturbations to assess the robustness of models**
☞ **It amounts to solving an constrained optimization problem on probability measures**
☞ **There are many different ways to solve it**
☞ **Wasserstein distance + quantile constraints work well together**

Most important point:

Performance metrics do not tell the whole story

Conclusion and perspective

Take away messages:

☞ **Feature perturbations to assess the robustness of models**
☞ **It amounts to solving an constrained optimization problem on probability measures**
☞ **There are many different ways to solve it**
☞ **Wasserstein distance + quantile constraints work well together**

Most important point:

Performance metrics do not tell the whole story

Some perspectives:

☞ **Perturbations on the dependence structure**
☞ **Other discrepancies**, and constraints on **other statistics**
☞ **Smoothing using other families of non-decreasing functions** (splines, kernel methods...)

Paper and more details

Feel free to check out our paper (more theory and illustrations):

M. I., N. Bousquet, F. Gamboa, B. Iooss, and J-M. Loubes. 2024. "Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models." *Electronic Journal of Statistics* 18 (2): 2721–2770

Alfonsi, A., and B. Jourdain. 2014. "A remark on the optimal transport between two probability measures sharing the same copula" [in en]. *Statistics & Probability Letters* 84 (January): 131–134. ISSN: 0167-7152. https://doi.org/10.1016/j.spl.2013.09.035. https://www.sciencedirect.com/science/article/pii/S0167715213003337.

Bachoc, F., F. Gamboa, M. Halford, J. M. Loubes, and L. Risser. 2023. "Explaining machine learning models using entropic variable projection." *Information and Inference: A Journal of the IMA* 12 (3). ISSN: 2049-8772. https://doi.org/10.1093/imaiai/iaad010. eprint: https://academic.oup.com/imaiai/article-pdf/12/3/iaad010/50514888/iaad010.pdf. https://doi.org/10.1093/imaiai/iaad010.

Csiszár, I. 1975. "I-Divergence Geometry of Probability Distributions and Minization problems." *The Annals of Probability* 3 (1): 146–158. https://doi.org/10.1214/aop/1176996454. http://doi.org/10.1214/aop/1176996454.

de la Fortelle, A. 2015. "A study on generalized inverses and increasing functions Part I: generalized inverses" [in en], 14. https://hal-mines-paristech.archives-ouvertes.fr/hal-01255512.

Dufour, J-M. 1995. *Distribution and quantile functions* [in en]. https://jeanmariedufour.github.io/ResE/Dufour_1995_C_Distribution_Quantile_W.pdf.

I., M., N. Bousquet, F. Gamboa, B. Iooss, and J-M. Loubes. 2024. "Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models." *Electronic Journal of Statistics* 18 (2): 2721–2770.

References ii

Iooss, B., and P. Lemaître. 2015. "A Review on Global Sensitivity Analysis Methods." In *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications,* edited by G. Dellino and C. Meloni, 101–122. Springer US. https://doi.org/10.1007/978-1-4899-7547-8_5. https://doi.org/10.1007/978-1-4899-7547-8_5.

Lemaitre, P., E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. 2015. "Density modification-based reliability sensitivity analysis." *Journal of Statistical Computation and Simulation* 85 (6): 1200–1223. https://doi.org/10.1080/00949655.2013.873039. eprint: https://doi.org/10.1080/00949655.2013.873039. https://doi.org/10.1080/00949655.2013.873039.

Santambrogio, F. 2015. *Optimal Transport for Applied Mathematicians.* Vol. 87. Progress in Nonlinear Differential Equations and Their Applications. Cham: Springer International Publishing. ISBN: 978-3-319-20827-5 978-3-319-20828-2. https://doi.org/10.1007/978-3-319-20828-2. http://link.springer.com/10.1007/978-3-319-20828-2.

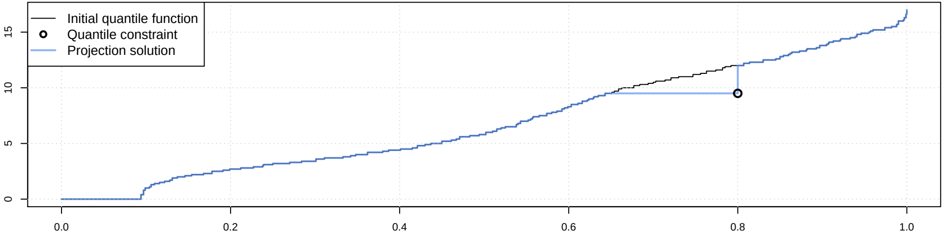
Thank you for your attention!

Any questions?

Analytical solution

If $\mathcal{V}$ is unrestricted, there exists a **unique analytical solution** Q to the problem:

Q is the same as P, except on the intervals between $F_P^{\leftarrow}(\alpha_i)$ and b_i which have no mass, and an atom is added at b_i, taking the initial mass of the interval.



How to explicitly add smoothness to the resulting perturbed quantile function ?

Analytical solution

Proposition . 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 $\alpha_1 < \cdots < \alpha_K$ and $b_1 < \cdots < b_K$, and $\mathcal{Q}(\alpha, b)$ the associated quantile perturbation class. For $i = 1, \ldots, K$, let $\beta_i = F_P(b_i)$. Define the intervals $A_i = (c_i, d_i]$ for $i = 1, \ldots, K$, such that:

$$c_1 = \min(\beta_1, \alpha_1), \quad c_i = \min\left[\max(\alpha_{i-1}, \beta_i), \alpha_i\right], i = 2, \ldots, K,$$

$$d_K = \max(\beta_K, \alpha_K), \quad d_j = \max\left[\min(\beta_j, \alpha_{j+1}), \alpha_j\right], j = 1, \ldots, K-1.$$

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

$$F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A}, \\ b_i & \text{if } y \in A_i, \quad i = 1, \ldots, K. \end{cases} \tag{5}$$