Robustness assessment of black-box models

Quantile-constrained Wasserstein projections

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

EDF R&D, Institut de Mathématiques de Toulouse (IMT), Sinclair AI Lab

About the CIFRE PhD *(EDF R&D and IMT)*
Development of interpretability methods for machine learning models applied to critical systems, at the crossroads between sensitivity analysis (SA) and explainable artificial intelligence (XAI). The goal of this PhD is to propose novel post-hoc tools in order to assess the behavior of black-box models.
Contact: marouane.il-idrissi@edf.fr

Inspired by work from both SA [3] and XAI [1] fields, a novel perturbation scheme of black-box models' input distributions is proposed. It is based on probability measure projections under quantile constraints with respect to (w.r.t.) the 2-Wasserstein distance. These perturbations aim to be generic, interpretable, and suitable for both SA and XAI purposes.

Marginal distribution perturbation

Let f be a black-box model, and $X \sim P \in \mathcal{P}(\mathbb{R})$. The optimally perturbed distribution of P is

$$\begin{aligned} Q = \underset{G \in \mathcal{P}(\mathbb{R})}{\text{argmin}} \quad & \mathcal{D}(P, G) \\ \text{s.t.} \quad & G \in \mathcal{C}. \end{aligned} \tag{1}$$

where $\mathcal{D}$ is a discrepancy between probability measures, and $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R})$ is a perturbation class. P can be **an empirical measure** from an observed dataset, or **admit a positive density**. Marginal perturbations are applied using a copula invariant perturbation map

$$T = (F_Q^{\leftarrow} \circ F_P)$$

where F_P is the cdf of P and $F_Q^{\leftarrow}$ the generalized quantile function of Q defined, for $a \in [0,1]$, as

$$F_Q^{\leftarrow}(a) = \sup \{t \in \mathbb{R} \mid F_Q(t) < a\}$$

Main objectives

1. Define a perturbation class $\mathcal{Q}$ using quantile constraints.
2. Solve Eq. 1 with the 2-Wasserstein distance as a discrepancy.
3. Explore the behavior of f subject to marginal perturbations.

Quantile constraints

Quantile constraints are of the form, given a perturbed quantile value $b \in \mathbb{R}$

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

Let $\mathcal{V}$ be part of $\mathcal{F}^{\leftarrow}$, the space of left-continuous, non-decreasing functions on $[0,1]$. The **quantile perturbation class** is defined, for $i = 1 \dots, K$ as

$$\mathcal{Q}_{\mathcal{V}} = \left\{ Q \in \mathcal{P}(\mathbb{R}) \mid F_Q^{\leftarrow} \in \mathcal{V}, F_Q^{\leftarrow}(\alpha_i) \geq b_i \geq F_Q^{\rightarrow}(\alpha_i) \right\}.$$

Different types of perturbations can be defined:

- Perturbations driven by an intensity parameter θ (quantile shift, operating domain dilatation).
- Perturbations for modelling purposes (e.g., expert knowledge).

Quantile-constrained Wasserstein projections

The problem in Eq. 1, can be equivalently written as a projection in $L^2([0,1])$:

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

- $\mathcal{V} = \mathcal{F}^{\leftarrow}$: **analytical solution**.
- $\mathcal{V} =$ monotone piece-wise continuous polynomials: **convex problem**.

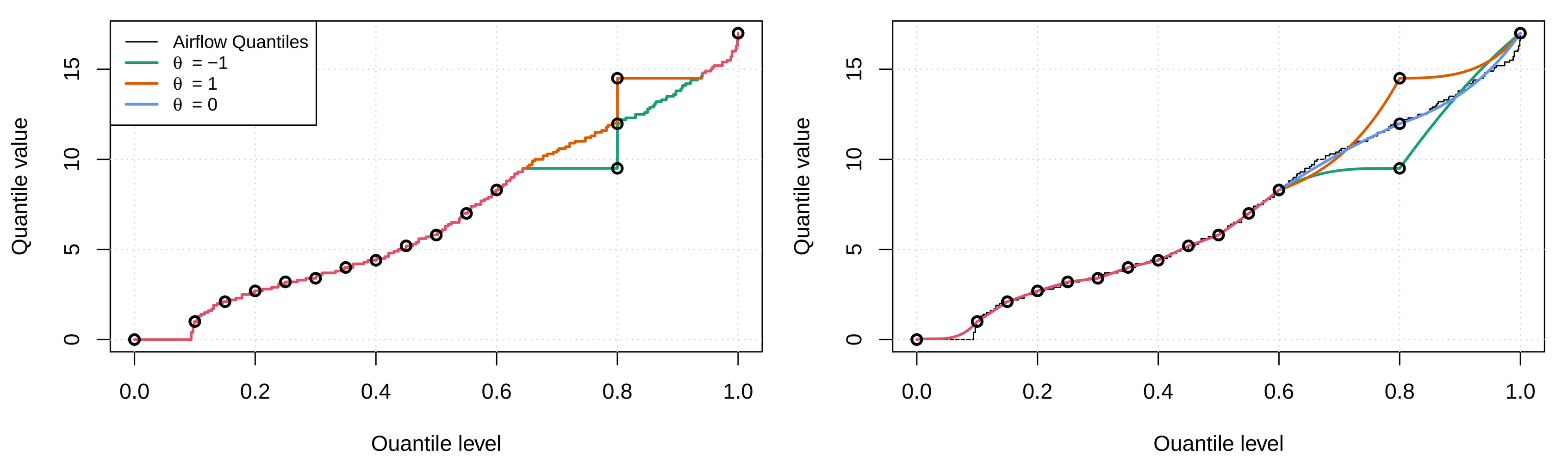

Figure 1: Quantile shifting perturbations. Analytical solution when $\mathcal{V} = \mathcal{F}^{\leftarrow}$ (left), and perturbation using isotonic polynomials of degree 9 (right).

Acoustic fire extinguisher

15390 experiments of sound wave fire extinguishing. Classification task on 6 variables measured during the experiments.

Black-box model: 1-layer neural network [2] trained with an accuracy of 95.15% (validation accuracy of 94.26%).

Perturbation scheme: shift of the Airflow 0.8-quantile: initial value at 12, shift between 9.5 ($\theta = -1$) and 14.5 ($\theta = 1$) by polynomial perturbation approximation of degree 9 (see, Fig. 1).

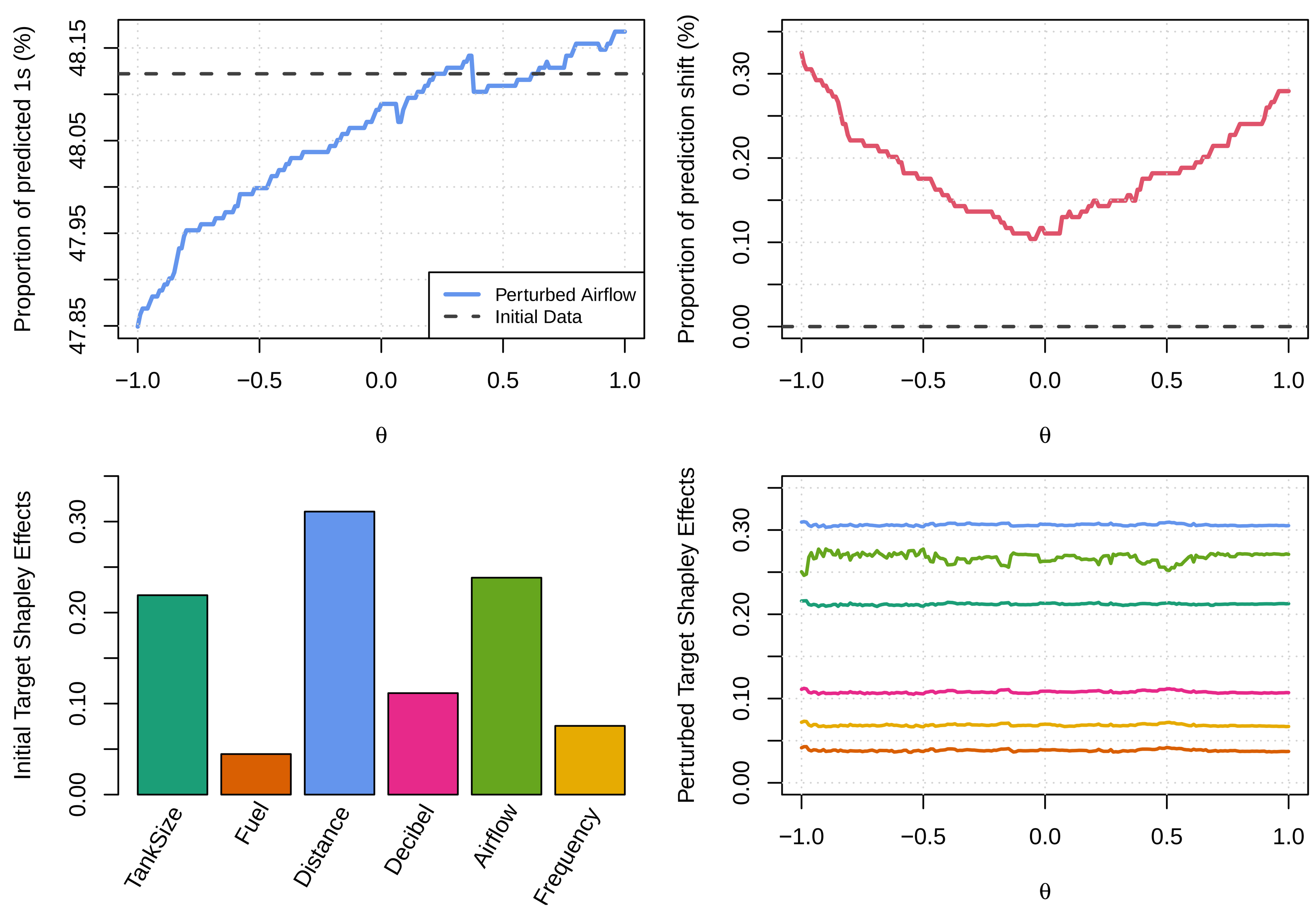

Figure 2: Global metrics under airflow quantile perturbations. Top row are the proportion of predicted put-out fire and prediction changes w.r.t. the initial data. Bottom row is the sensitivity of target Shapley importance metrics w.r.t. the perturbations.

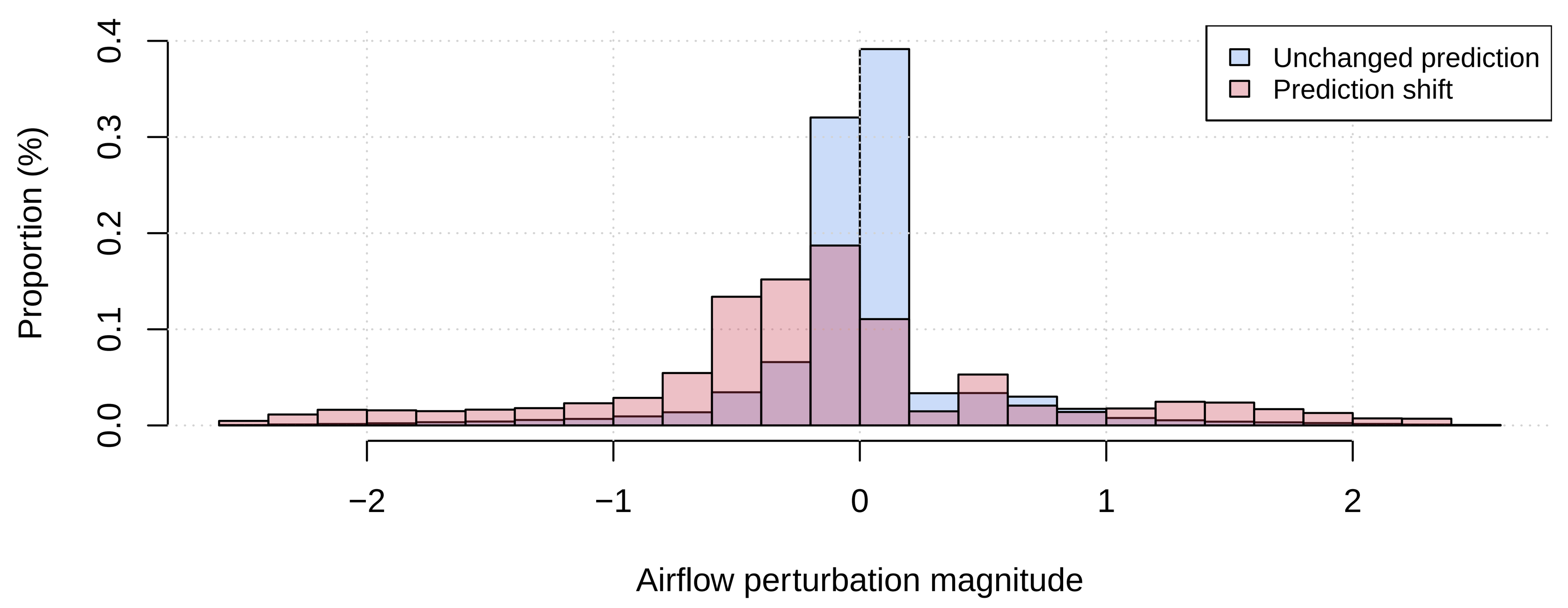

Figure 3: Signed airflow perturbation magnitude of instances inducing either a prediction shift, or no prediction change.

Conclusion and perspectives

Generic, interpretable and easy to compute perturbation scheme, leading to robustness to input perturbation diagnostics for SA and ML black-box models.

Future work:

- Parallel and efficient implementation in R (soon).
- Polynomial optimal degree-selection scheme and isotonic splines.
- Multivariate (copula) perturbations, and other discrepancies (Prokhorov).
- Other general smoothing spaces $\mathcal{V}$ (Sobolev, RKHS).
- Super-quantile perturbations.

References

[1] F. Bachoc, F. Gamboa, M. Halford, J-M. Loubes, and L. Risser. Explaining Machine Learning Models using Entropic Variable Projection. *arXiv:1810.07924 [cs, stat]*, December 2020. arXiv: 1810.07924.

[2] 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. Conference Name: IEEE Access.

[3] P. Lemaître et al. Density modification-based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85(6):1200–1223, 2015.

SINCLAIR