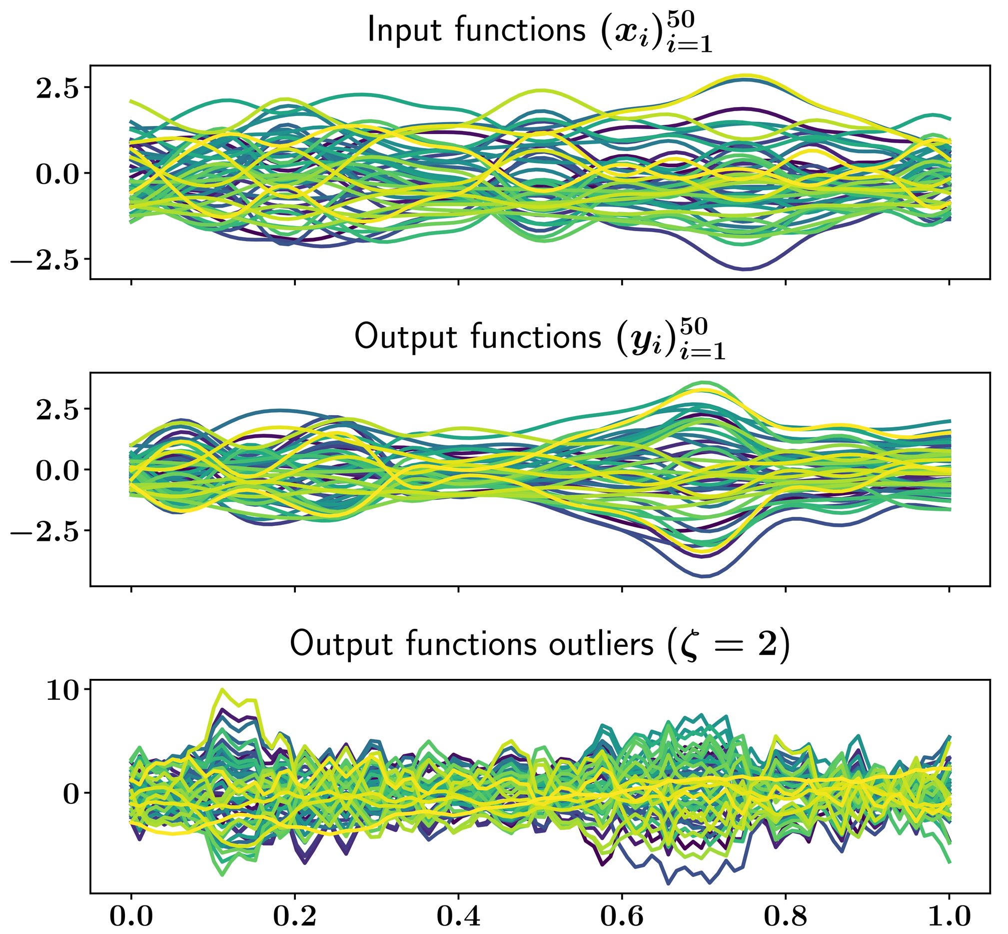
Functional Output Regression with Infimal Convolution: Exploring the Huber and ϵ-insensitive Losses

A. Lambert[†⋆], D. Bouche[†], Z. Szabó[♣], F. d'Alché-Buc[†]

† LTCI, Télécom Paris. ⋆ ESAT, KU Leuven. ♣ Department of Statistics, London School of Economics.

Goal

Go beyond the square loss in functional output regression to better handle outliers and sparsity



Input space $\mathcal{X}$, output space $\mathcal{Y} := L^2[\Theta, \mu]$ where $\Theta \subset \mathbb{R}$ compact. Build

$$h : \mathcal{X} \to \mathcal{Y}$$

Proposed loss functions

Typical loss function: square loss $L(f) = \frac{1}{2}\|f\|_{\mathcal{Y}}^2 = \frac{1}{2}\int_\Theta f(\theta)^2 \mathrm{d}\mu(\theta)$. [2]

- Sensible to outliers, no sparsity

Key idea: use a loss obtained with infimal convolution

$$L = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, g,$$

where g is a well-chosen function that enforces robustness or sparsity. Suited to dual approaches as Fenchel-Legendre conjugate is

$$\left(\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, g\right)^\star = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 + g^\star.$$

Leverage p-norms for flexible choice of g, where $p \in [1, +\infty]$. Denoting q the conjugate exponent ($\frac{1}{p} + \frac{1}{q} = 1$), $\iota_{\mathcal{C}}(\cdot)$ the indicator function of a convex set $\mathcal{C}$, and $\mathcal{B}_\kappa^p$ the p-ball of radius κ in $\mathcal{Y}$,

Extended Huber loss ($\kappa \geq 0$):

$$H_\kappa^p := \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, \kappa\|\cdot\|_p, \qquad (H_\kappa^p)^\star = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 + \iota_{\mathcal{B}_\kappa^q}(\cdot).$$

Extended ϵ-insensitive loss ($\epsilon \geq 0$):

$$\ell_\epsilon^p := \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, \iota_{\mathcal{B}_\epsilon^p}(\cdot), \qquad (\ell_\epsilon^p)^\star = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 + \epsilon\|\cdot\|_q.$$

Dual Formulation in vv-RKHSs

Extension of kernel methods to handle vector-valued outputs. [1]

- $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ two scalar-valued kernels
- $T_{k_\Theta} \in \mathcal{L}(\mathcal{Y})$ the integral operator associated to k_Θ
- $K = k_{\mathcal{X}} \cdot T_{k_\Theta}$ with vv-RKHS $\mathcal{H}_K$

$$\inf_{h \in \mathcal{H}_K} \frac{1}{n}\sum_{i \in [n]} L(y_i - h(x_i)) + \frac{\lambda}{2}\|h\|_{\mathcal{H}_K}^2, \qquad \lambda > 0$$

Dual problem for $L = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, g$ reads [3]

$$\inf_{(\alpha_i)_{i\in[n]} \in \mathcal{Y}^n} \sum_{i \in [n]} \left[\frac{1}{2}\|\alpha_i\|_{\mathcal{Y}}^2 - \langle \alpha_i, y_i\rangle_{\mathcal{Y}} + g^\star(\alpha_i)\right] + \frac{1}{2\lambda n}\sum_{i,j \in [n]} k_{\mathcal{X}}(x_i, x_j)\,\langle \alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}}.$$

Challenges: $(\alpha_i)_{i=1}^n$ are functions and need suitable representation that ensures computability of proximal operator of $g^\star$ and all other quantities.

Optimization

Representing the dual variables: we choose a linear splines representation for the $(\alpha_i)_{i=1}^n$ based on some fixed anchors $(\theta_{ij})_{i,j \in [n]\times[m]}$ distributed i.i.d. as μ. This allows for a finite dimensional encoding of the dual variables in a matrix $\mathbf{A}$ of size $n \times m$ with $a_{ij} = \alpha_i(\theta_{ij})$.

Computing the objective function: The different terms are computed using Monte-Carlo approximation with the anchors $(\theta_{ij})_{i,j \in [n]\times[m]}$.

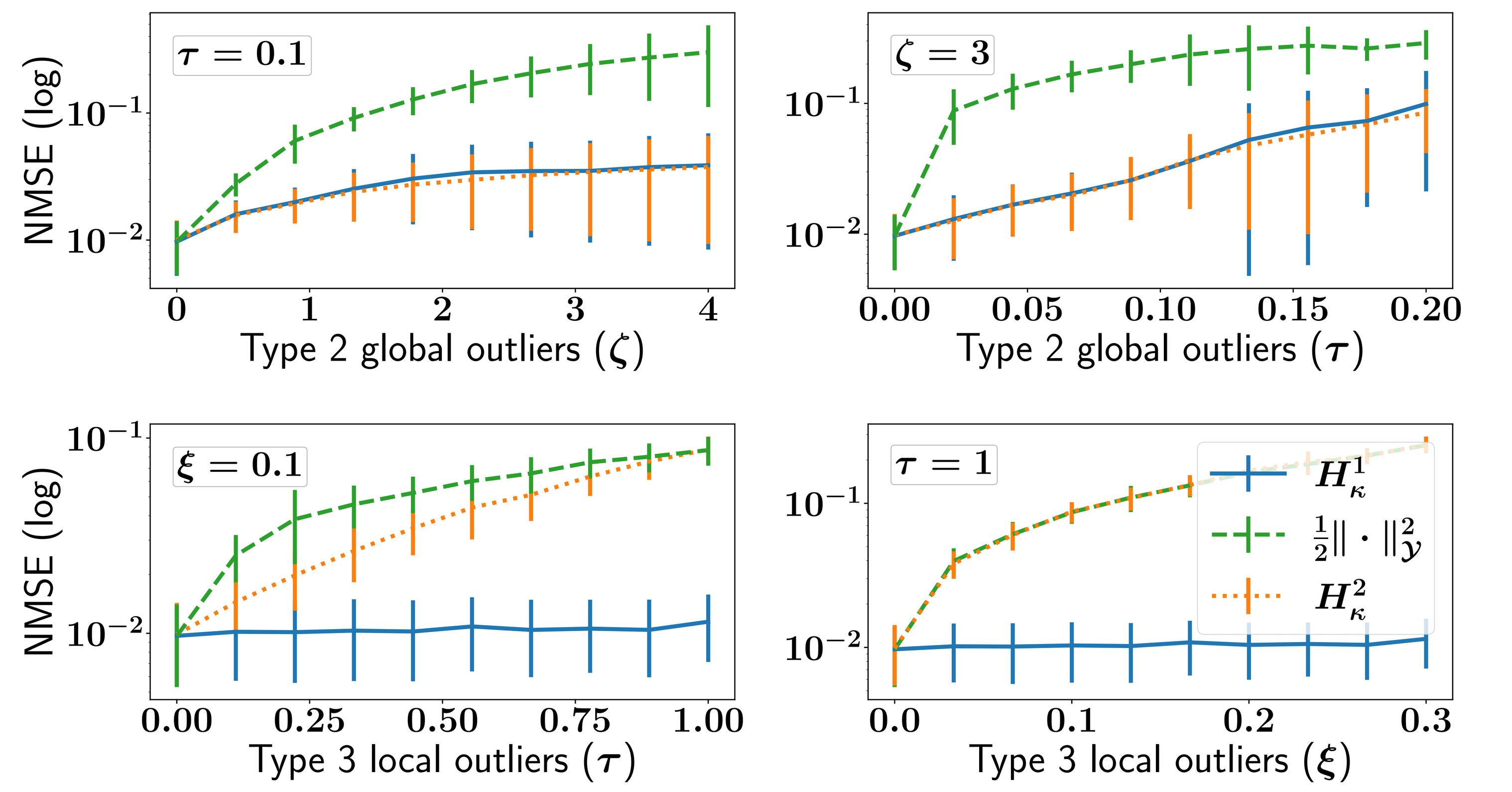
Composite optimization problem: Because $g^\star$ is non-smooth, we consider accelerated proximal gradient descent. For the Huber loss, the proximal step amounts to projecting on some q-ball which is tractable when $q \in \{2, +\infty\}$. For the ϵ-insensitive loss, it corresponds to a soft thresholding operator when $q = 1$ and a block soft thresholding operator when $q = 2$.

Overall estimator: Once the matrix $\mathbf{A}$ is known, the estimator reduces to

$$h(x)(\theta) = \frac{1}{\lambda nm}\sum_{i \in [n]} k_{\mathcal{X}}(x, x_i)\sum_{j \in [m]} a_{ij} k_\Theta(\theta, \theta_j).$$

Robustness experiments

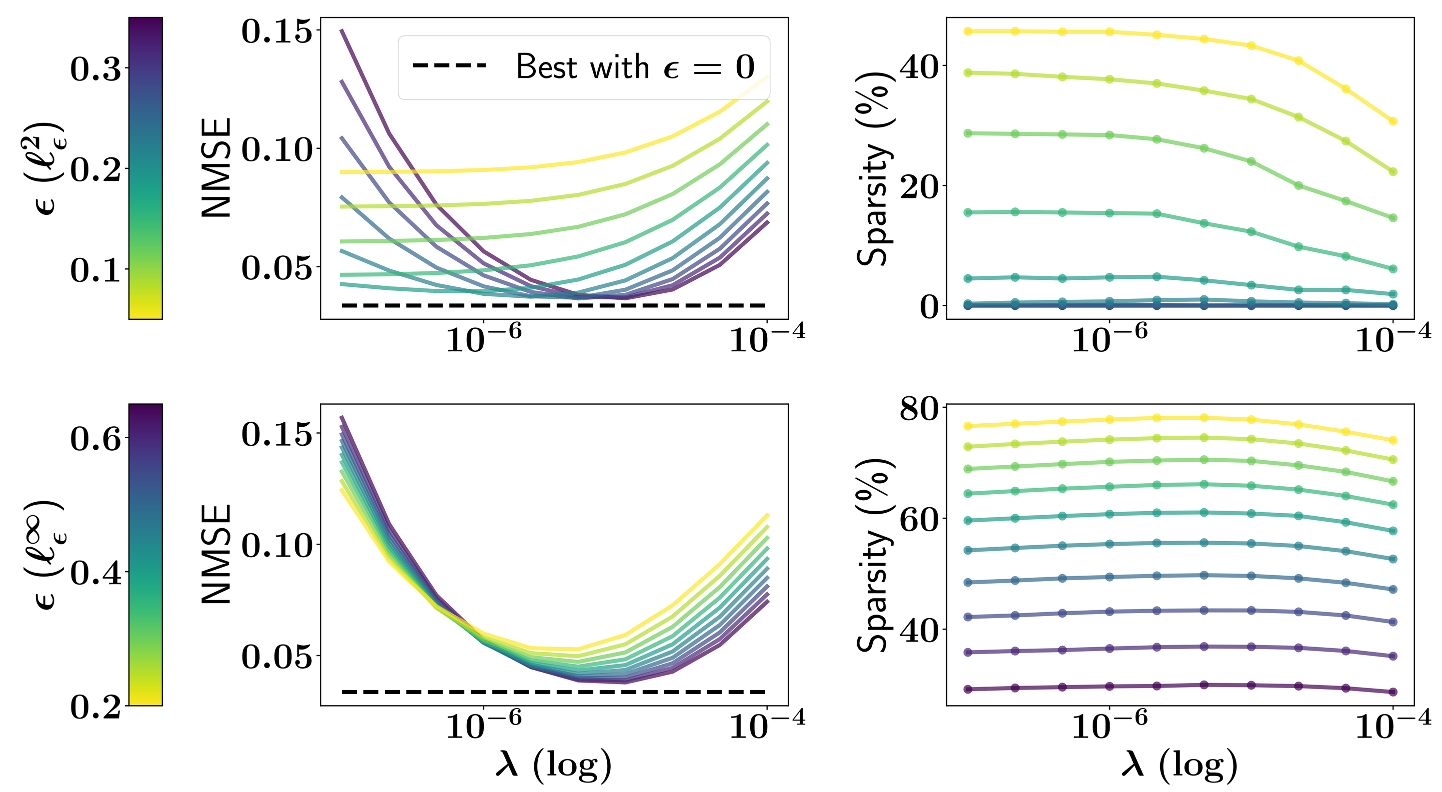
Experimental setup: $k_{\mathcal{X}}$, k_Θ are Gaussian kernels, we contaminate a synthetic dataset using two kind of outliers: local (only a few measurements of the function are corrupted) or global (the function is entirely replaced).



We can see that H_κ^2 struggles against local outliers, whereas H_κ^1 shows good robustness properties.

Sparsity experiments

We show that a compromise can be made between the two parameters λ and ϵ to get increased sparsity with little degradation of the performance.



Code Available

https://github.com/allambert/foreg

A.L. D.B. and F.d.B were funded by the research chair Data Science & Artificial Intelligence for Digitalized Industry and Services at Télécom Paris.

References

[1] Carmeli, Claudio and De Vito, Ernesto and Toigo, Alessandro and Umanitá, Veronica Vector valued reproducing kernel Hilbert spaces and universality. In *Analysis and Applications*, vol 8 pp 19–61, 2010.

[2] Hachem Kadri and Emmanuel Duflos and Philippe Preux and Stéphane Canu and Alain Rakotomamonjy and Julien Audiffren Operator-valued Kernels for Learning from Functional Response Data In *Journal of Machine Learning Research*, pp. 1–54, 2016.

[3] Laforgue, Pierre and Lambert, Alex and Brogat-Motte, Luc and d'Alché-Buc, Florence. Duality in RKHSs with Infinite Dimensional Outputs: Application to Robust Losses. In *International Conference on Machine Learning (ICML)*, pp. 5598–5607, 2020.