Supplementary Material: Density kernel depth for outlier detection in functional data

Nicolás Hernández, Alberto Muñoz and Gabriel Martos

1 Proof of Proposition 1

Proof The proof for properties P1 and P2 are proven straightforward by Eq. (6) and (7) given that the Density kernel depth is an f-monotonous function.

To prove P3 consider $\|\boldsymbol{\lambda}\| \to \infty$ then it holds that $f(\boldsymbol{\lambda}) \to 0$, which implies that $g(\boldsymbol{\lambda}, f)g(\mathbf{m}, f) \to 0$ and therefore $\mathrm{DKD}(\tilde{x}(t), f, K_f) \to 0$.

To prove P4, consider the affine map $\tau \in \mathcal{T}$ of the form $\tau \circ \tilde{x}(t) = a + b\tilde{x}(t) = \tilde{x}'(t)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ the representation coefficients of $\tilde{x}(t)$. Then by Eq (5) it holds that $\boldsymbol{\lambda}' = (a + b\lambda_1, \ldots, a + b\lambda_d)$ are the representation coefficients of $\tilde{x}'(t)$. Since τ is an affine map, by Eq. (6) and (7) it holds that:

$$\mathrm{DKD}(\tilde{x}(t), f, K_f) = a + b\mathrm{DKD}(\tilde{x}(t), f, K_f).$$

□

2 Proposition 2

The order induced by the DKD is invariant under RKHS representation. Consider:

i) The ordering map $I : \mathbb{R} \to \{1, \ldots, n\}$. I is a monotonous function that ranks the sample of functions from 1 to the sample size n.

ii) Consider two systems of orthonormal basis functions $B = \{\phi_i\}_{i \geq 1}$ with projection coefficients $\boldsymbol{\lambda}$, and $G = \{\psi_i\}_{i \geq 1}$ with projection coefficients $\boldsymbol{\kappa}$. Such that for a fixed d, it holds that $\mathrm{Span}\{B\} = \mathrm{Span}\{G\}$.

Then the order induced by the DKD remains invariant under the RKHS representation:

$$I\big(g(\boldsymbol{\lambda}, f_B)g(\mathbf{m}_B, f_B)\big) = I\big(g(\boldsymbol{\kappa}, f_G)g(\mathbf{m}_G, f_G)\big)$$

where $\mathbf{m}_B = \mathrm{argmax}_{\boldsymbol{\lambda} \in \mathbb{R}^d} f_B(\boldsymbol{\lambda})$, and $\mathbf{m}_G = \mathrm{argmax}_{\boldsymbol{\kappa} \in \mathbb{R}^d} f_G(\boldsymbol{\kappa})$.

Proof Using Eq. (5) $\tilde{x}(t) \in Span\{B\}$, can be approximated by two basis systems:

$$\tilde{x}(t) \approx \sum_{j=1}^{d} \lambda_j \phi_j(t), \text{ and} \tag{1a}$$

$$\tilde{x}(t) \approx \sum_{j=1}^{d} \kappa_j \psi_j(t), \tag{1b}$$

then each ϕ_j can be expressed in terms of $\{\psi_l\}_{l=1}^{d}$–and conversely–, as follows

$$\phi_j(t) \approx \sum_{l=1}^{d} \gamma_l \psi_l(t) \tag{2}$$

Then combining Eq 1a with 2, we have

$$\tilde{x}(t) \approx \sum_{j=1}^{d} \lambda_j \sum_{l=1}^{d} \gamma_l \psi_l(t). \tag{3}$$

Therefore we can represent $\tilde{x}(t)$ in the subspace generated by the Span$\{G\}$ by means of a change of basis -an affine transformation-, hence by P4 it holds that: $I\big(g(\boldsymbol{\lambda}, f_B)g(\mathbf{m}_B, f_B)\big) = I\big(g(\boldsymbol{\kappa}, f_G)g(\mathbf{m}_G, f_G)\big)$ which concludes the proof.

$\square$

3 Outliergram and Functional Boxplot results

In the following figure are shown the results of the Outliergram [1] and the Functional Boxplot [2] for the Australian Mortality data. These two devices were designed for functional outlier detection.

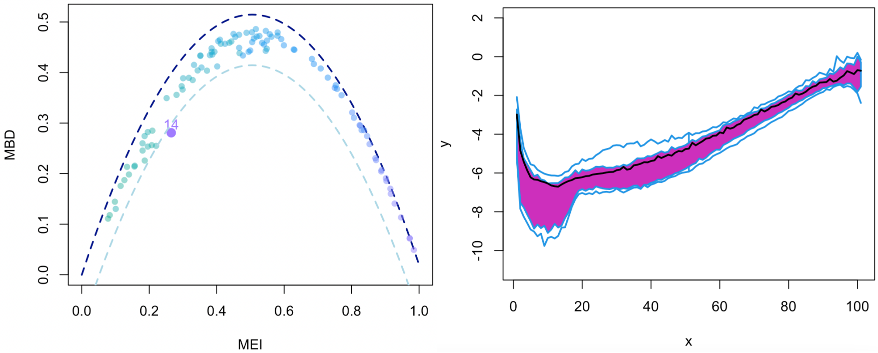

Fig. 1: Left: Outliergram highlighting year 1914 as the identified outlier curve. Right: Functional Boxplot of the data highlighting in black the deepest curve.

4 Monte Carlo experiment for sample size

In this section we present an additional example to illustrate the evolution of the detection rate of true positive outliers as the sample size increases. We perform this experiment for Scenarios A, B and C (see the experimental section of the paper). The Monte Carlo experiment was carried out with 1000 samples for different sample size $N = 100, 200, 300, 500, 1000$, for a contamination level of $q = 10\%$. In the table we below we report the average True Positive Rate. The corresponding standard-error are reported in parenthesis.

Table 1: Monte Carlo experiment: Avg. TPR for different Scenarios and Sample sizes, with a contamination level of $q = 10\%$. The corresponding standard-error are reported in parenthesis.

Scenario	Sample Size				
	100	200	300	500	1000
Symmetric	0.86 (0.096)	0.879 (0.065)	0.891 (0.052)	0.903 (0.042)	0.91 (0.032)
Asymmetric	0.857 (0.082)	0.855 (0.057)	0.853 (0.045)	0.851 (0.035)	0.854 (0.026)
Bi-modal	0.573 (0.128)	0.618 (0.095)	0.674 (0.071)	0.769 (0.063)	0.821 (0.042)

References

[1] Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. Biostatistics **15**(4), 603–619 (2014)

[2] Sun, Y., Genton, M.G.: Functional boxplots. Journal of Computational and Graphical Statistics **20**(2), 316–334 (2011)