# Robust Kernel Density Estimation with Median-of-Means principle

ICML 2022

Pierre Humbert, Batiste Le Bars, Ludovic Minvielle

June 24, 2022

# Introduction

**Data:** $X_1, \dots, X_n$ i.i.d. with density $f(\cdot)$
**Objective:** Estimate $f$ from the sample

**Applications:**

- Data visualization, clustering, classification
- Outlier detection

$\longrightarrow$ **One possibility:** Kernel Density Estimation (KDE)

**Problem:** The KDE is not robust to outliers

## KDE with outliers: Toy example

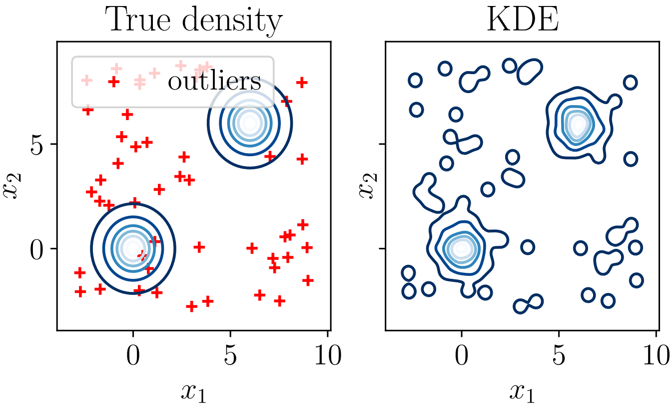


Figure: True density and outliers. Estimation with KDE.

# Outlier-robust KDE

**Objective:** Propose a KDE robust to outliers

$\longrightarrow$ Combination of KDE and Median-of-Means (MoM)

**Why?** KDE can be seen as a mean. MoM is known to robustify mean estimators.

## Outlier setup

$\mathcal{O} \cup \mathcal{I}$ **framework:**

- $\{X_i \mid i \in \mathcal{I}\}$ with i.i.d. inliers with density $f$
- $\{X_i \mid i \in \mathcal{O}\}$ with outliers.

  $\longrightarrow$ No assumption on them (can be adversarial)

# Median-of-Means KDE

## How to compute the MoM-KDE?

1. Randomly split the dataset in $S$ blocks, $[\![1, n]\!] = \sqcup_{s=1}^{S} B_s$
2. At a given $x_0$, compute a standard KDE $\hat{f}_{n_s}(x_0)$ over $B_s$ for each $s$
3. Compute the MoM-KDE:

$$\hat{f}_{MoM}(x_0) \propto \text{Median}\left(\hat{f}_{n_1}(x_0), \cdots, \hat{f}_{n_S}(x_0)\right)$$

Recall the standard KDE: $\quad \hat{f}_{n_s}(x_0) = \frac{1}{|B_s| h^d} \sum_{i \in B_s} K\left(\frac{X_i - x_0}{h}\right)$

## Back to our toy example

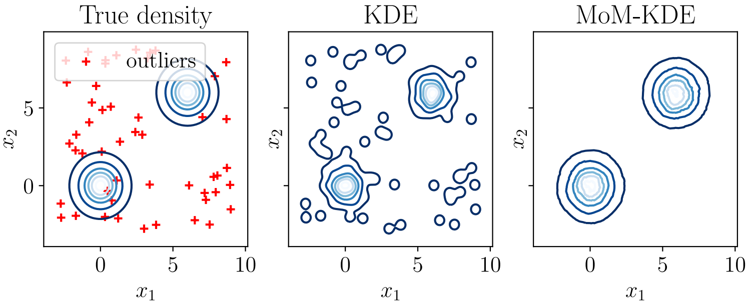


Figure: True density and outliers. Estimation with KDE. Estimation with MoM-KDE

# Theoretical contributions

1. **Consistency results** – Under mild assumptions:
   - With high probability,

$$\|\hat{f}_{MoM} - f\|_\infty \leq C_1 \sqrt{\frac{S(\log(S) + \gamma + \log(1/h))}{nh^d}} + C_2 h^\alpha \ ,$$

   - With probability higher than $1 - \frac{1}{n}$, we have

$$\|\hat{f}_{MoM} - f\|_\infty \quad \lesssim \left(\frac{|\mathcal{O}|}{n}\log(n)\right)^{\alpha/(2\alpha+d)} + \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)} \ .$$

   - $\|\hat{f}_{MoM} - f\|_1 \xrightarrow[n\to\infty]{\mathcal{P}} 0 \ .$

2. **Influence Function (IF)**
   - Introduction of an IF adapted to the $\mathcal{O} \cup \mathcal{I}$ framework
   - IF lower for the MoM-KDE than for the standard KDE

# Empirical contributions

1. Extensive results on synthetic data: Density estimation
   - 3 metrics
   - 4 type of outliers including adversarial
   - Comparison with 5 methods
2. Extensive results on real dataset: Outlier detection
   - 6 different real datasets
   - Several amount of outliers
3. Empirical experiments on a bootstrap version of the MoM-KDE

## Conclusion

- We propose MoM-KDE a robust estimator for density estimation by combining KDE and MoM principle
- We prove its $L_\infty$ and $L_1$ convergence under mild assumptions
- We introduce an influence function adapted to the $\mathcal{O} \cup \mathcal{I}$ framework
- We show the robustness of the MoM-KDE on synthetic and real data
- We perform additional empirical experiments on a bootstrap version of the MoM-KDE

Thank you !