Median-of-Means for Outlier-Robust MMD Estimation*

Matthieu Lerasle[1,3], Zoltán Szabó[2], Timothée Mathieu[3], Guillaume Lecué[4]

[1]CNRS, Université Paris Saclay, France [2]CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris, Palaiseau, France [3]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, France [4]CREST ENSAE ParisTech, France

Quick Summary

- Mean embedding, MMD: information theory on kernel-enriched domains.
- Goal: their outlier-robust estimation.
- Contribution:
  - Optimal sub-Gaussian deviation bound (minimal 2nd order assumption).
  - Practical algorithms.

Target Quantities

- Mean embedding:

$$\mathbb{P} \mapsto \mu_\mathbb{P} = \int_\mathcal{X} \underbrace{\varphi(x)}_{\text{example: } e^{\langle \cdot, x\rangle}} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_K.$$

- Maximum mean discrepancy (MMD):

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_{\mathcal{H}_K} = \sup_{f\in B_K} \underbrace{\langle f, \mu_\mathbb{P} - \mu_\mathbb{Q}\rangle_{\mathcal{H}_K}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}.$$

Notes:
- Large number of applications; review [1].
- Numerous kernel-endowed domains. $K(x,y) = \langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}_K}$, $\varphi(x) = K(\cdot, x)$.

Goal

- Design outlier-robust estimators.
- Interest: unbounded kernels
  - exponential kernel: $K(x,y) = e^{\gamma\langle x,y\rangle}$.
  - polynomial kernel: $K(x,y) = (\langle x,y\rangle + \gamma)^p$.
  - string, time series or graph kernels.

- Issue with average: A single outlier can ruin it.

Estimator

- Idea (MOM):
  1. Partition: $\underbrace{x_1,\ldots,x_{N/Q}}_{S_1}, \ldots, \underbrace{x_{N-N/Q+1},\ldots,x_N}_{S_Q}$.
  2. Compute average in each block:

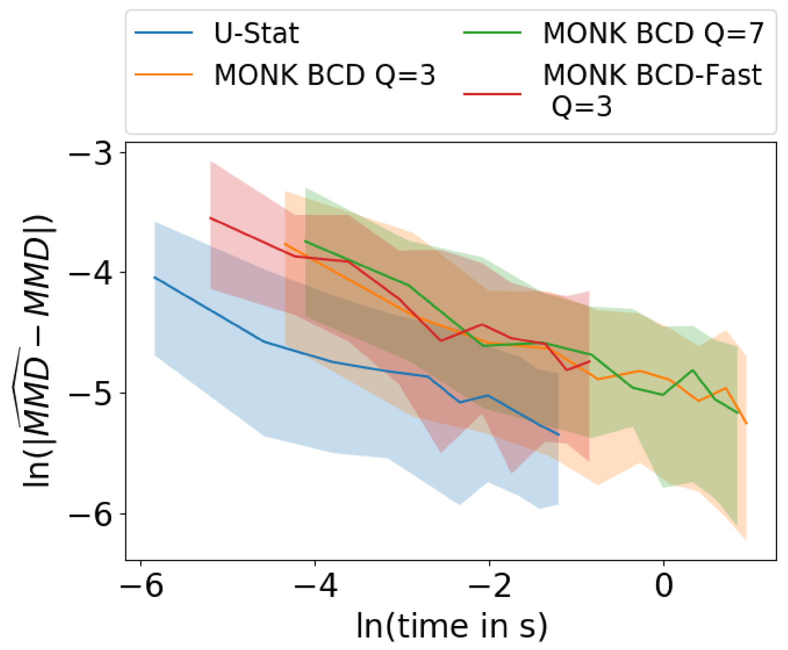
$$a_1 = \frac{1}{|S_1|}\sum_{i\in S_1} x_i, \quad \ldots, \quad a_Q = \frac{1}{|S_Q|}\sum_{i\in S_Q} x_i.$$

  3. Estimate $\mathbb{E}X$: $\mathrm{med}_{q\in[Q]} a_q$.
- On MMD: replace the expectation with MON

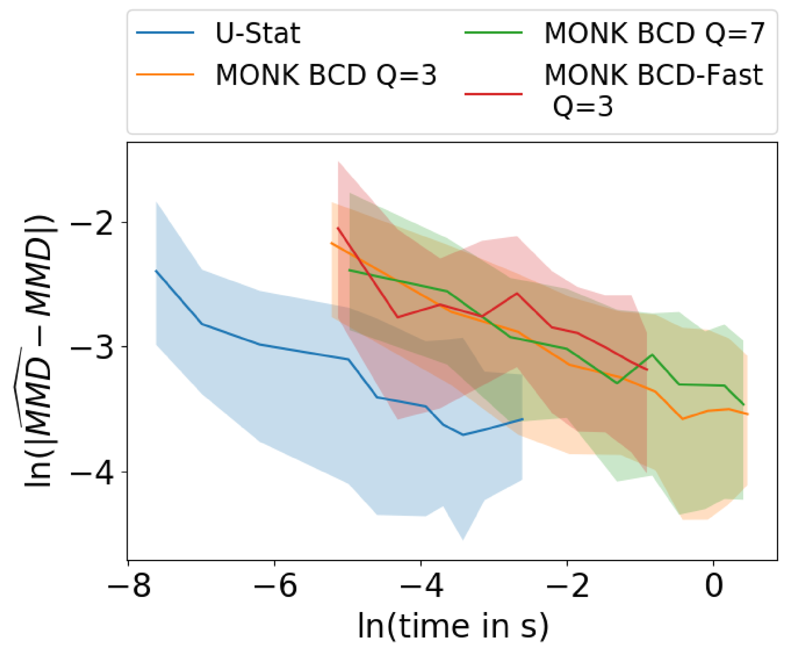
$$\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]}\left\{\frac{1}{|S_q|}\sum_{j\in S_q} f(x_j) - \frac{1}{|S_q|}\sum_{j\in S_q} f(y_j)\right\}.$$

- Code: `https://bitbucket.org/TimotheeMathieu/monk-mmd`

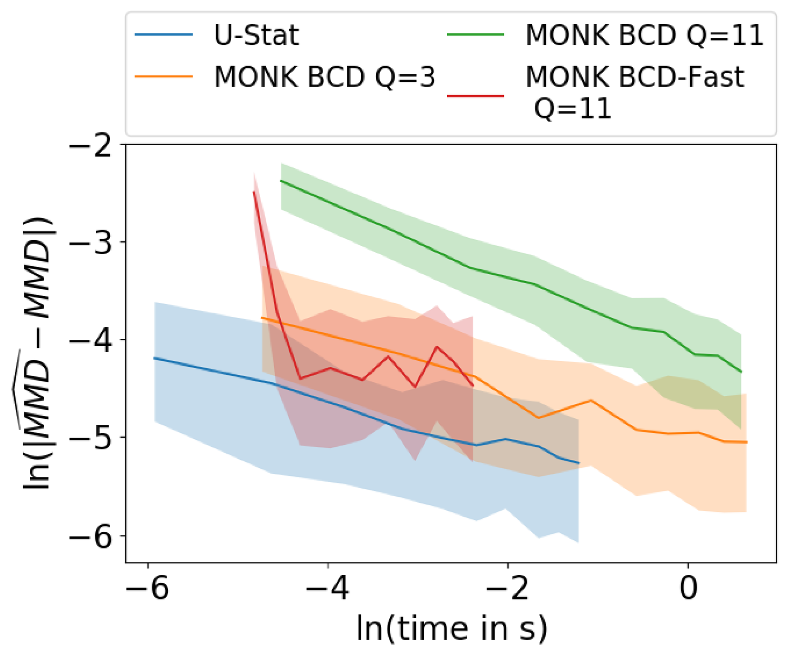
Numerical Illustrations



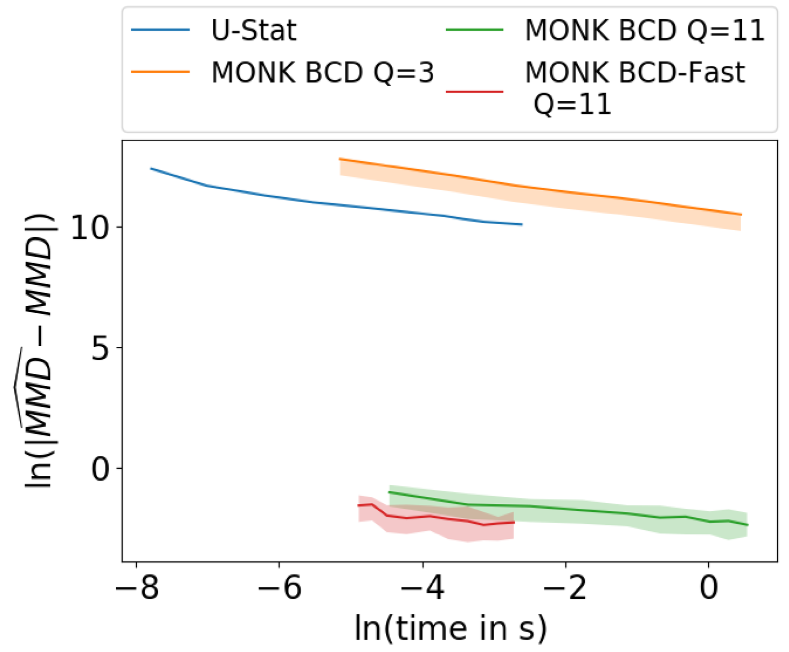
(a) Gaussian distribution, $N_c = 0$ (no outlier), RBF kernel.



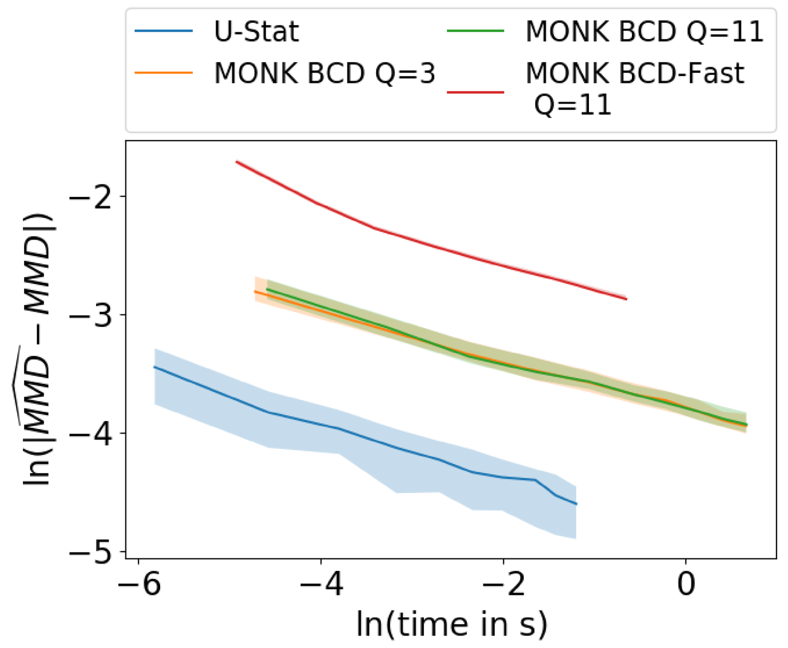
(b) Gaussian distribution, $N_c = 0$ (no outlier), quadratic kernel.



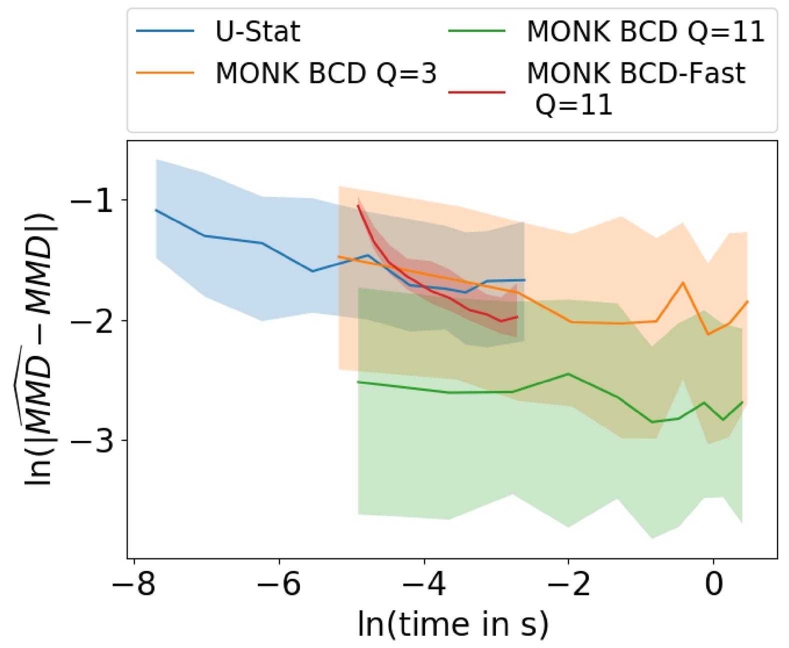
(c) Gaussian distribution, $N_c = 5$ outliers, RBF kernel.



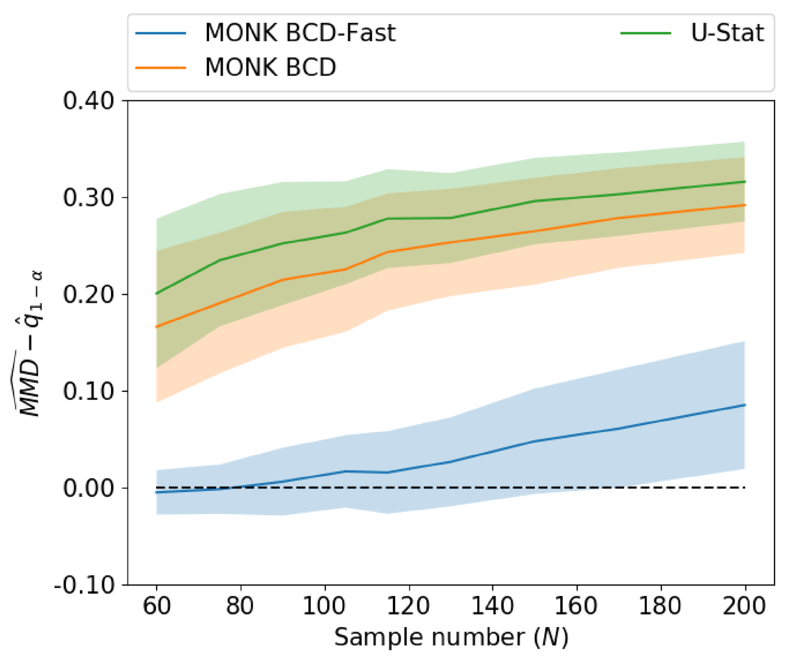
(d) Gaussian distribution, $N_c = 5$ outliers, quadratic kernel.



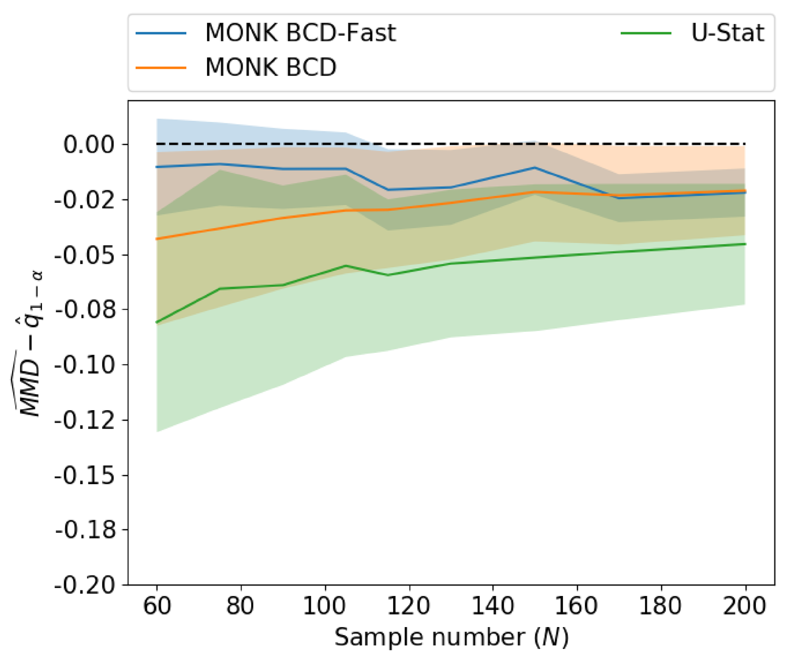
(e) Pareto distribution, RBF kernel.



(f) Pareto distribution, quadratic kernel.



(g) Inter-class: EI-IE



(h) Intra-class: EI-EI

Finite-Sample Bound for $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$ ($\hat{\mu}_\mathbb{P}$: Similar)

Assume:
- Contamination: $\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}$, $N_c \le Q(1/2-\delta)$, $\delta \in (0,1/2]$.
- Mild 2nd-order assumption: $\exists\, \mathrm{Tr}(\Sigma_\mathbb{P}), \mathrm{Tr}(\Sigma_\mathbb{Q})$.

Then, for any $\eta \in (0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q \in \left(N_c/\left(\frac{1}{2}-\delta\right), N/2\right)$, with probability at least $1-\eta$

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \le \frac{12\max\left(\sqrt{\frac{(\|\Sigma_\mathbb{P}\|+\|\Sigma_\mathbb{Q}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}(\Sigma_\mathbb{P})+\mathrm{Tr}(\Sigma_\mathbb{Q})}{N}}\right)}{\delta}.$$

Discussion

(i) N-dependence: $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ is optimal for MMD estimation [2].

(ii) Σ-dependence:
- Optimal sub-Gaussian deviation bound for mean estimation under minimal 2nd-order condition even on $\mathbb{R}^d$ [3] – long-lasting open question.
- They rely on tournament procedure: numerically hard.
- Most practical convex relaxation [4]: $\mathcal{O}(N^{24})$.
- After submission: [5]: $\mathcal{O}(N^4 + dN^2)$, $d < \infty$.

(iii) δ-dependence:
- Larger δ means less outliers,
  - the bound becomes tighter,
  - one needs less blocks.
- optimal?

(iv) Breakdown point – asymptotic concept:
- median $\Rightarrow$ Using Q blocks is resistant to $Q/2$ outliers.
- Q can grow with N, as (almost) $N/2$.
- Breakdown point can be 25%.

(v) Unknown Q:
- One choose Q adaptively by the Lepski method.
- Same guarantee but with increased computional cost.

Acknowledgements

Guillaume Lecué is supported by a grant of the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047).

References

[1] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[2] Ilya Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximal mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1930–1938, 2016.

[3] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019.

[4] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. Technical report, 2018. (`https://arxiv.org/abs/1809.07425`).

[5] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-Gaussian rates. Technical report, 2019. (`https://arxiv.org/abs/1902.01998`).

*Data, Learning and Inference (DALI), San Sebastian, Spain, September 2-6 2019.