UNIVERSITY OF WARSAW

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Munich Center for Machine Learning

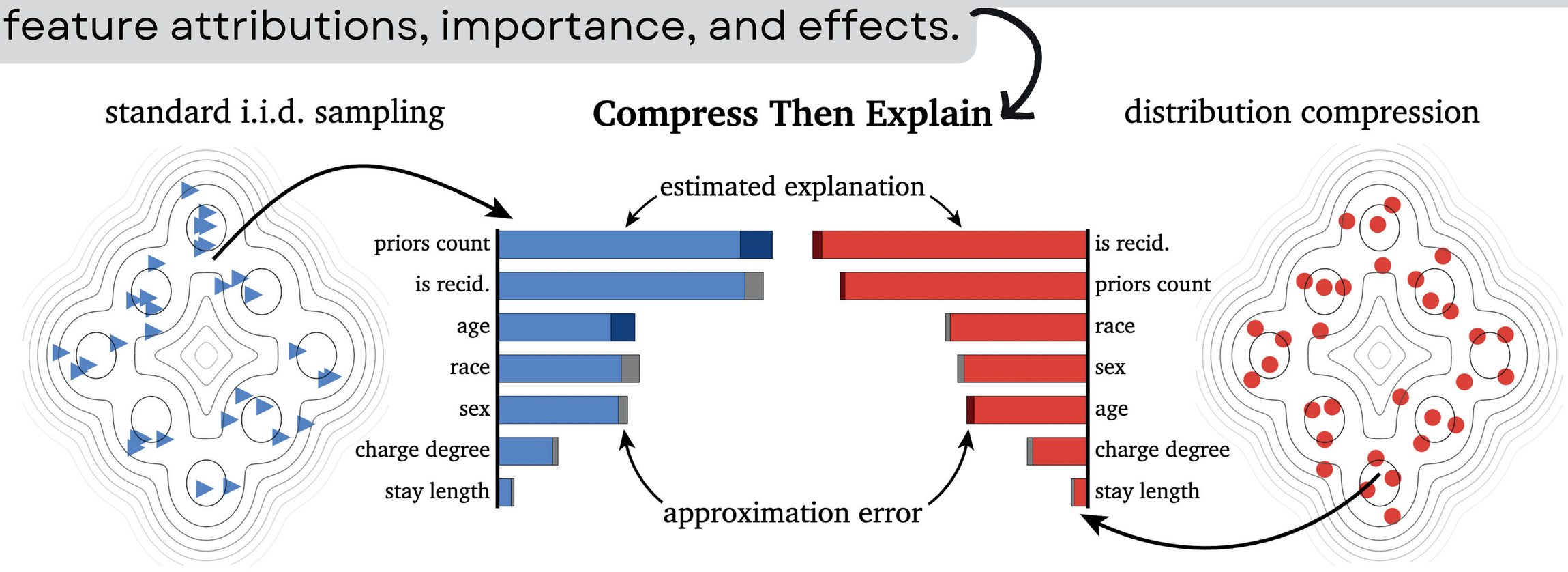
ICML International Conference On Machine Learning

Efficient and accurate explanation estimation with distribution compression

#data-efficient-XAI

Hubert Baniecki, Giuseppe Casalicchio, Bernd Bischl, Przemyslaw Biecek

TL;DR: Distribution compression improves the explanation approximation error of feature attributions, importance, and effects.



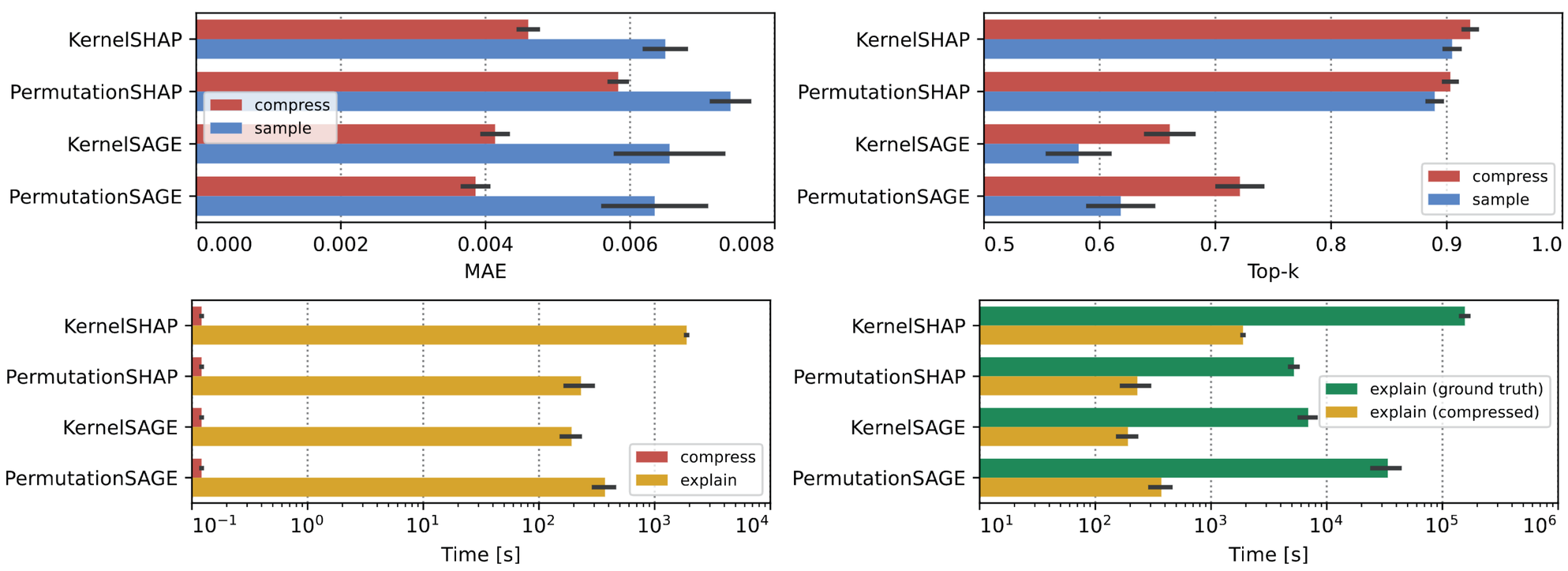
1. Motivation: Sampling from the dataset is prevalent in various post-hoc explanation algorithms. Sampling decreases the **computational cost** of estimation but introduces an **approximation error**.

2. Problem: How large is the error? Is it significant? **Can we reduce it?**

3. Spoiler: Yes we can! Error is sometimes large, e.g. it can impact the feature importance ranking.

4. Method: Distribution compression, e.g. the kernel thinning algorithm *(Dwivedi & Mackey, COLT 2021)*

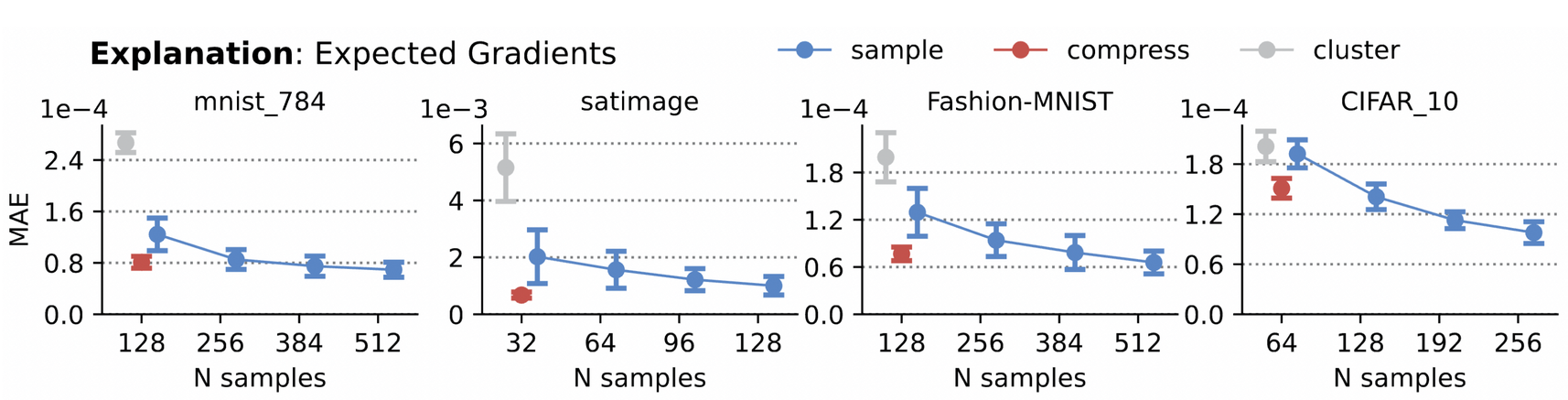
5. Results:



6. Why? Intuition:

Proposition 3.4 (Feature marginalization is bounded by the total variation distance between data samples)**.** *For two distributions $q_{\mathbb{X}}, q_{\widetilde{\mathbb{X}}}$, we have $\left|f(\mathbf{x}_s; q_{\mathbb{X}}) - f(\mathbf{x}_s; q_{\widetilde{\mathbb{X}}})\right| \leq C_f \cdot \mathrm{TV}(q_{\mathbb{X}}, q_{\widetilde{\mathbb{X}}})$, where C_f denotes a constant that bounds the model function f.*

Proposition 3.5 (Global explanation is bounded by total variation distance between data samples)**.** *For two distributions $q_{\mathbb{X}}, q_{\widetilde{\mathbb{X}}}$, we have $\left\|G(q_{\mathbb{X}}; f, g) - G(q_{\widetilde{\mathbb{X}}}; f, g)\right\|_1 \leq C_g \cdot \mathrm{TV}(q_{\mathbb{X}}, q_{\widetilde{\mathbb{X}}})$, where C_g denotes a constant that bounds the local explanation function g.*



Compress Then Explain (CTE): not only 2–3x more efficient, but also more stable!

Critical **d**ifference **d**iagrams of average ranks (**lower is better**) aggregated over **6 explanation estimators and 48 dataset–model pairs**. **Left:** MAE averaged over repeats. **Right:** Standard deviation of MAE over repeats that corresponds to the stability of estimation.