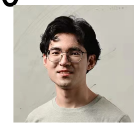
Computing High-dimensional Confidence Sets of Arbitrary Distributions


Chao Gao
U. Chicago
Statistics


Liren Shan
TTIC
Computer Science


Vaidehi Srinivas
Northwestern
Computer Science


Aravindan Vijayaraghavan
Northwestern
Computer Science

PROBLEM SETTING

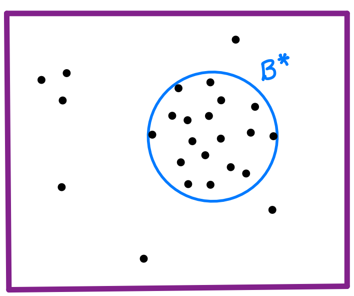
Goal: estimate a high-density region of an arbitrary distribution

With sample access to distribution D over $\mathbb{R}^d$ for coverage factor $\delta \in [0,1]$, set system $\mathcal{C}$ of bounded VC-dimension over $\mathbb{R}^d$ want to approximate

$$\min_{C \in \mathcal{C}} \mathrm{vol}(C) \quad \text{s.t.} \quad \Pr_{x \sim D}[x \in C] \geq \delta$$

→ Volume or Lebesgue measure

▷ interesting even for $\mathcal{C}$ being the set of ℓ_2 balls



Example:

- points drawn from D
- ○ min. volume ball B* containing 70% of D

Applications: Conformal prediction, estimating density level sets, support estimation, goodness-of-fit, robust estimation ... and more!

Informal Theorem [approx. min. vol. confidence set]

For an arbitrary set of points $Y \subseteq \mathbb{R}^d$, coverage factor $\delta \in [0,1]$ and coverage slack $\gamma > 0$, in poly$(|Y|, d)$ time we can find an ellipsoid $\hat{\mathcal{E}}$ s.t.

$$|\hat{\mathcal{E}} \cap Y| \geq (1-\gamma)\delta|Y| \quad \text{and} \quad \mathrm{vol}(\hat{\mathcal{E}}) \leq \mathrm{vol}(B^*)\exp\left(\tilde{O}_{\gamma\delta}\left(d^{2/3}\right)\right)$$

where B^* is the minimum volume ball s.t. $|B^* \cap Y| \geq \delta|Y|$.

MIN. VOL. BALL SUBROUTINE

▷ Want to do "robust center estimation" but center is not a robust quantity because it can depend on only a few points!

can approximate the center with mean, by sacrificing a few points (i.e., γ slack factor)

▷ Produce coarse estimate B' to refine to $\hat{B}$:

① if mean of B' close to center of B^*, we are done!

② if mean of B' far from center of B^*, most points are actually closer to mean

▷ requires some work to bound variance/error