Robust mixture learning when outliers overwhelm small groups

ETH AI CENTER
UNIVERSITY OF COPENHAGEN
ETH zürich

Daniil Dmitriev*, Rares-Darius Buhai*, Stefan Tiegel, Alexander Wolters,
Gleb Novikov, Amartya Sanyal, David Steurer, Fanny Yang

*equal contribution

Problem Setting

- Dimension $d \in \mathbb{N}_+$, number of clusters $k \in \mathbb{N}_+$
- **Input distribution:**

$$\mathcal{X} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, I_d) + \varepsilon Q,$$

  where Q is *adversarial*, i.e., can be any distribution.
- Weights $w_1, \ldots, w_k$ and outliers fraction ε, s.t. $\varepsilon + \sum w_i = 1$
- Cluster centers $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$
- We allow *large* ε and assume that $\|\mu_i - \mu_j\|$ is large
- *Lower bound on the mixture weights:* $w_{\min} \leq w_i$ for all $i \in [k]$
- **Goal:**
  - Given i.i.d. samples from $\mathcal{X}$, estimate $\mu_1, \ldots, \mu_k$
  - Weights w_i's are unknown, only $w_{\min}$ is given
  - Output a *small* list with *small* error

Prior works

- We rely on *mean estimation* paradigm, which models data as $\mathcal{X} = \alpha \mathcal{N}(\mu^*, I_d) + (1-\alpha)Q$, where Q is adversarial. This model can be applied directly to our case with $\alpha = w_{\min}$.
- **List-decodable mean estimation**: Applies when $\alpha \leq 1/2$
  - ✗ when $\alpha = w_{\min}$, all points in Q are treated as outliers $\Longrightarrow$ sub-optimal error and list size guarantees.
  - ✓ our work: leverages structure in the data, so that only real outlier points are considered outliers.
- **Robust mean estimation**: achieves optimal error, but requires $\alpha > 1/2$.
  - ✗ cannot out-of-the-box handle cases when $k \geq 2$.
  - ✓ our work: as long as $\varepsilon \ll w_i$, we obtain guarantees from existing robust mean estimation algorithms.
- **Mixture learning**: only applicable when $\varepsilon \leq w_{\min}$.

Take-home message:

We achieve optimal mean recovery guarantees in the presence of large number of adversarial points, with the small list size $k + O\left(\frac{\varepsilon}{w_{\min}}\right)$

Our algorithm: *Outer stage* and *Inner stage*

- *Outer stage* splits the input dataset into (intersecting) subsets, each containing points from at most one inlier cluster
- *Inner stage* runs *base learners* at different scale to identify the correct inlier cluster size
- *Base learners*: (i) List-decodable mean estimation algorithm (*LD-ME(α)*) and (ii) Robust mean estimation (*RME(α)*)
- Any improvement for *base learners* results in improvements for our *mixture learning* task (e.g., better error)

Outer stage

Repeat until only a few points remain:

Step 1. select area with many points, where at most one inlier cluster is present

Step 2. add this set to the collection $\mathscr{T}$

Step 3. remove points from the selected area (in reality, only a subset of points are removed)

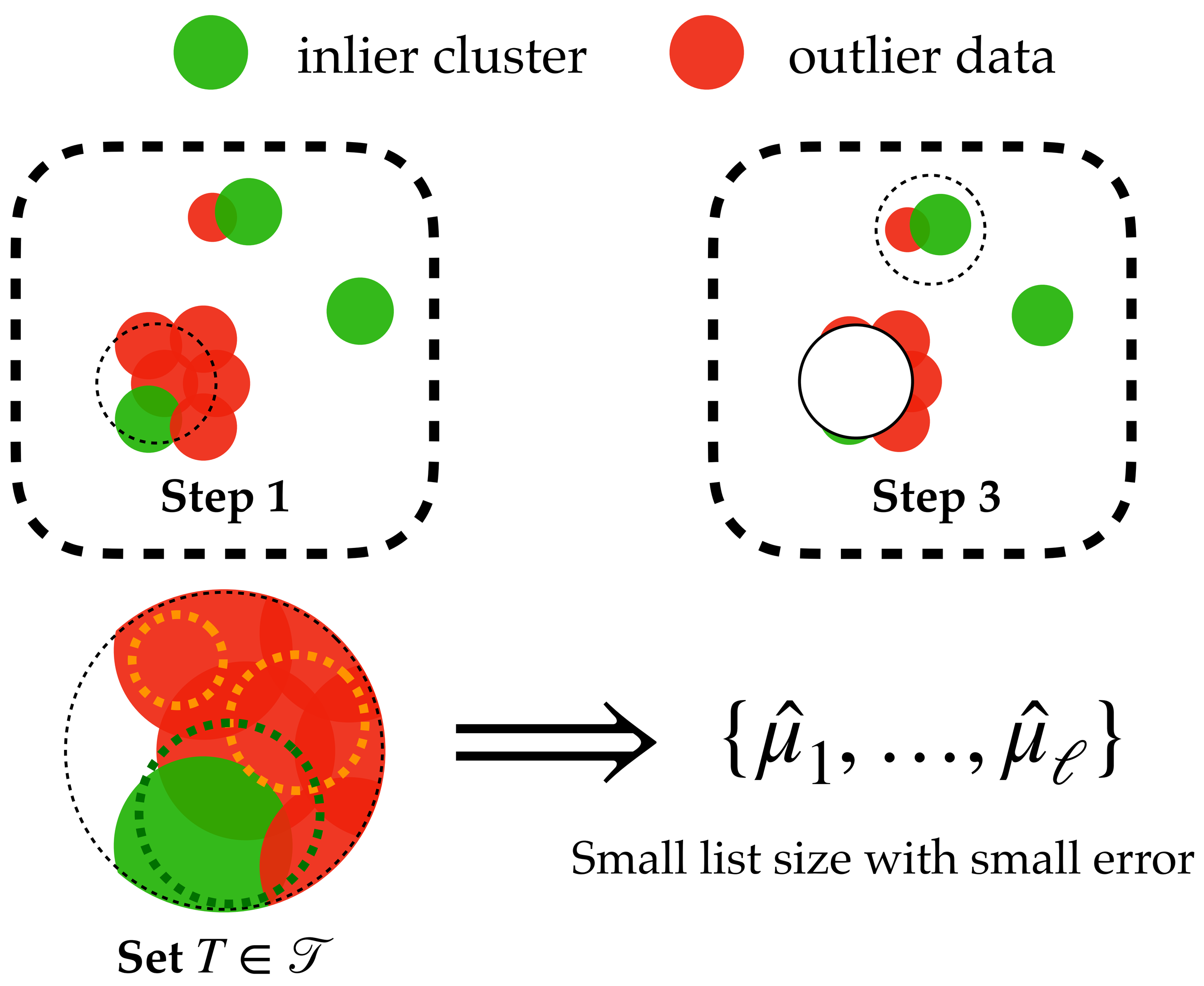
Inner stage (run on each set $T \in \mathscr{T}$)

Initialize $\Delta = \{\alpha_{\min}, 2\alpha_{\min}, \ldots, 1/2\}$

Step 1. run *LD-ME(α)* base learner with $\alpha \in \Delta$

Step 2. concatenate and filter outputs $(\hat{\mu}_i)_i$

Step 3. try to improve errors with *RME(α)*



Theoretical and experimental results

Type of inlier mixture	Best prior work	Ours	IT lower bound
Large ($\forall j: \varepsilon \leq w_j$), sep. groups	$\tilde{O}(\varepsilon/w_i)$	$\tilde{O}(\varepsilon/w_i)$	$\Omega(\varepsilon/w_i)$
Small ($\exists j: \varepsilon \geq w_j$), sep. groups	$O\left(\sqrt{\log \frac{1}{w_{\min}}}\right)$	$O\left(\sqrt{\log \frac{\varepsilon+w_i}{w_i}}\right)$	$\Omega\left(\sqrt{\log \frac{\varepsilon+w_i}{w_i}}\right)$
Non-separated groups	$O\left(\sqrt{\log \frac{1}{w_{\min}}}\right)$	$O\left(\sqrt{\log \frac{1}{w_i}}\right)$	$\Omega\left(\sqrt{\log \frac{1}{w_i}}\right)$