Mining the Minoria: Unknown, Under-represented, and Under-performing Minority Groups

Mohsen Dehghankar, Abolfazl Asudeh

University of Illinois Chicago
{mdehgh2, asudeh}@uic.edu

51st International Conference on Very Large Data Bases
September 1-5, 2025 – London, United Kingdom
Research Track 21 – Specialized and Domain-Specific Data Management

Outline

Motivation Example: A data-sharing platform

- Before sharing their datasets, `Chicago Open Data Portal` would like to specify groups that are *under-represented* & *under-performing*.
- This is to **limit the scope of use** of shared datasets.
- **Challenge:**
  1. The datasets either do not include grouping attributes (such as `race`) or only contain some of those.
  2. Targeting a comprehensive audit, they do not want to limit their scope to a small set of predefined groups.
- **Goal:** To *proactively* detect *any meaningful* "problematic" group.

Outline

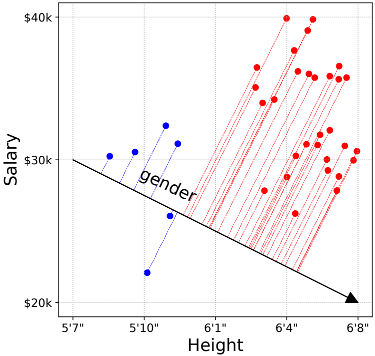
Problem Formulation: **Minoria Mining**

- **Given:** a dataset $\mathcal{D} = \{t_i\}^n$, where $t_i = \langle X = \langle \mathbf{x}_1, \cdots, \mathbf{x}_d \rangle, y \rangle$. $\mathcal{D}$ is used for training a model $h_\theta(X)$ that predicts y.
- **Find:** groupings of $\mathcal{D}$ to $\mathcal{D}^g$ (group g) and $\mathcal{D}^{!g}$ (others), s.t.:
  1. g is *under represented*: $|\mathcal{D}^g| \ll |\mathcal{D}|$
  2. Predictions based on $\mathcal{D}$ are *not accurate* for g:

$$\mathbb{E}[L_{\mathcal{D}^{\mathbf{g}}}(\theta)] - \mathbb{E}[L_{\mathcal{D}}(\theta)] \geq \tau$$

Our Approach: Finding high-skew projections

- Find the top-ℓ directions f that yield the highest skew when projecting points
  - Projection: $\mathcal{D}_f = \{t_i^\top f \mid t_i \in \mathcal{D}\}$
- High skew $\Rightarrow$ Small group in the tail $\Rightarrow$ Potential Minoria



Pearson's median skewness

$$skew(\mathcal{D}_f) = \frac{3(\mu - \nu)}{\sigma}$$

- μ = mean, σ = std. dev. ν = median
- **Idea?:** The weights are continuous ⇒ Formulate the optimization problem as linear programming (LP)?
- **Challenge:** What is the median?!
    - Every projection has its own median!

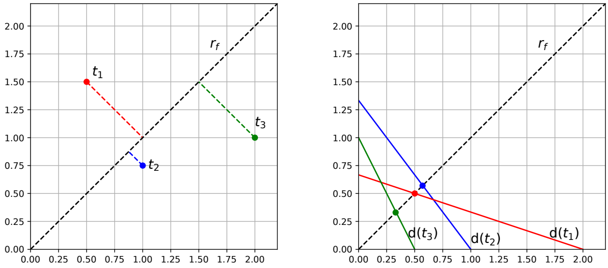
Outline

Dual-space transformation

- **Dual Space:** Tuples $t_i = \langle t_{i_1}, \dots, t_{i_d} \rangle$ represented as *hyperplanes*:

$$\mathsf{d}(t_i):\ t_{i_1}x_1 + \cdots + t_{i_d}x_d = 1$$

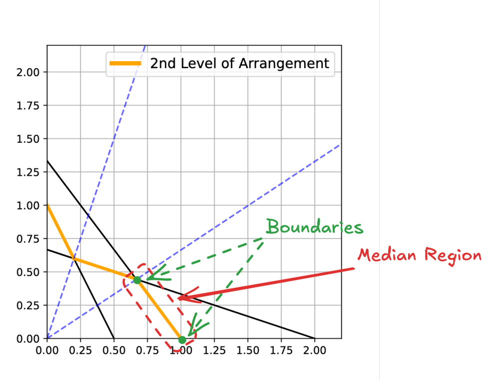
- A projection-direction f in primal $\Rightarrow$ an *origin-anchored ray* r_f in dual.
- The projection order $\mathcal{D}_f = \{t_i^\top f\}$ equals *the order of intersections of* $\mathsf{d}(t_i)$ *with* r_f.
- We use **arrangement of dual hyperplanes**, to track the medians.



Primal space Dual space

Median Regions

- A **Median Region** is a set of directions f that have the same median.
- In dual space, the $\lfloor \frac{n}{2} \rfloor$*-th level of the arrangement* partitions directions into *median regions*.



Preliminary idea for finding the high-skew projections

1. Identify the median regions
2. For each region, form an LP and solve it to find the highest skew.

- *Theoretically Polynomial* (in n)
- **Not Practical!** (Needs to solve **many** LPs)
- **Resolution:** *Can we avoid the LP optimizations?*

Key Theorem

- **Theorem:** The *highest skew* happens either in **the boundary of median regions** or

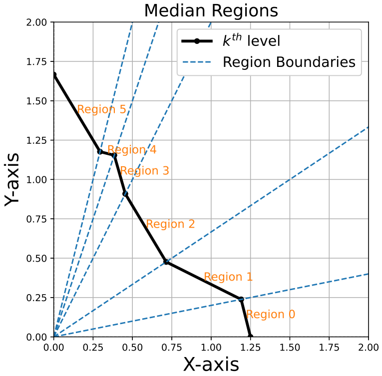
$$f^* = \frac{(QQ^\top)^{-1} q_{m_f}}{\|(QQ^\top)^{-1} q_{m_f}\|}, \quad q_i = t_i - \mu(\mathcal{D})$$

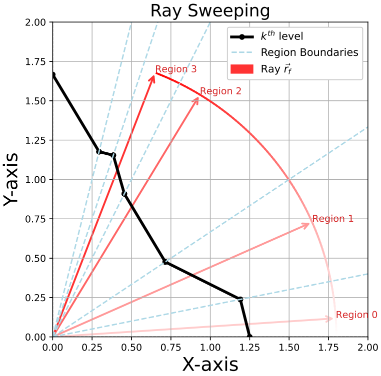
- **Result:** Enough to check **Only a few candidate directions per region**.

Minoria Mining in 2D

- **Overall approach:**
  1. Build the $\frac{n}{2}$-th level arrangement $\mathcal{A}_{\frac{n}{2}}$.
     - ★ Number of regions $= O(n^{4/3})$
  2. Enumerate boundary nodes (and f^* directions) of the median regions.
  3. At each node, compute Pearson's skew of its corresponding direction.
- **Naïve algorithm:** Each skew takes $O(n)$ time.
  - Time complexity: $O(n \cdot n^{4/3}) = O(n^{\frac{7}{3}})$
- **Our algorithm (Ray sweeping):** By updating median, mean, and std incrementally, skew can be computed in **constant time**.
  - **Time complexity:** $O(n^{\frac{4}{3}})$

Ray Sweeping: Example





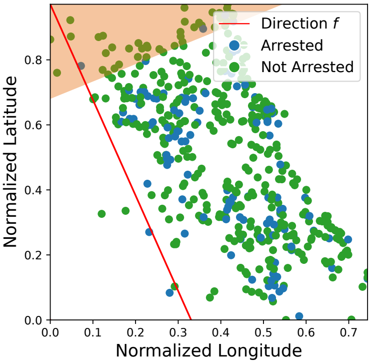
Mining in Higher Dimensions

- **Generalized Ray-Sweeping:** Works for $d > 2$ by traversing the $\frac{n}{2}$-th level arrangement.
  - Complexity: $O(d \cdot n^d)$ (enumerating $\mathcal{A}_{\frac{n}{2}}$ and computing skew).
  - **Curse of dimensionality**: arrangement size grows exponentially with d.
- **Practical heuristics:** To make the method feasible in higher dimensions, we use:
  - **Space discretization:** sample directions via grid partitioning or diverse candidate generation.
  - **Exploration & exploitation:** balance random search with refinement near promising directions.
  - **Focused exploration:** identify error-prone regions with the model and restrict search around them.

Outline

2D Experiments: Chicago Crimes

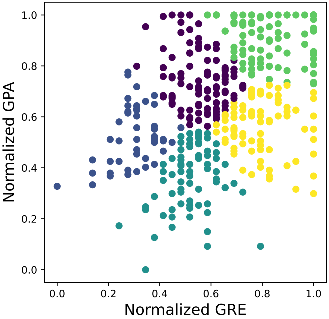
- Dataset: **Chicago Crimes** (2001–2023), projected on *Long* & *Lat*
- Classifier: 1-hidden-layer NN (F1 = 0.72)
- **Finding:** the top skewed direction aligns roughly **North Side**; tail shows **F1-score significantly drops**.



Percentile	F1
1	0.72
0.1	**0.62**
0.01	**0.68**
0.001	**0.40**

Why Not Clustering? (College Admissions)

- k-means clusters have f/m ratios close to the whole data (≈ 1.1)
- Our discovered high-skew tail shows **much higher** female/male ratios (and F1 drops)



k-means (k=5)

Percentile (tail p)	Acc.	F1	Female/Male (tail)
1.00	0.70	0.36	1.10
0.50	0.68	0.42	1.12
0.20	0.67	0.48	0.80
0.10	**0.61**	**0.34**	**2.00**
0.08	**0.64**	0.42	**1.81**

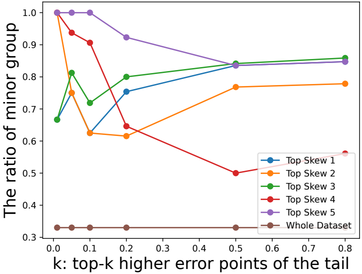
Tail eval on highest-skew direction (skew = 0.07).

Cluster ID	Size	Female/Male
0	92	0.95
1	72	0.94
2	108	1.11
3	45	1.50
4	83	1.24
Total	**400**	**1.10**

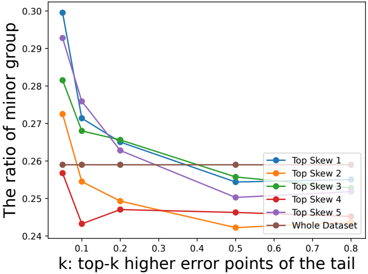
Cluster ratios near dataset baseline.

Experiments in Higher Dimensions: Focused Exploration

- High-skew directions expose **hidden minority groups** with higher model errors.
- Minority ratios **grow in the tails** (left side of plots), showing errors are not uniformly distributed.
- Even subtle groups (ratios < 0.3) are systematically highlighted with Focused Exploration algorithm.



(a) Adults dataset: minority ratio rises in tail directions.



(b) Diabetes dataset: subtle minorities (< 0.3) still detected.

Thank you, Question?