

Density-Based Clustering in Semi-Metric Space for Large-Scale Data

Luiza Krzepkowska

Supervisor: DI Dr.techn. Sebastian Ratzenböck

1. Introduction

Context: The **Gaia mission** collects precise astrometric data for stars.
Challenge: Identify **open clusters** (stars from the same molecular cloud) despite massive background noise.
Issue: **Missing radial velocity** data makes full velocity vector estimation difficult.

2. Goal

Develop a clustering method for large-scale data in semi-metric space, where distances don't follow the triangle inequality.

3. Velocity Space Complexity

Components of velocity:

- Proper motion:** Observable angular movement across sky.
- Radial velocity:** Motion along the line of sight – often unmeasured.

Total velocity of a star: $v = T(x)q$, $q = (\mu_\alpha^*, \mu_\delta, \mu_{rad})^T$ where:

- $T(x)$ is a transformation matrix,
- q is the motion vector,
- x represents the star's spatial coordinates.

Pairwise 3D velocity difference:
 $\|\Delta v\| = \|v_1 - v_2\| = \|T(x_1)q_1 - T(x_2)q_2\|$

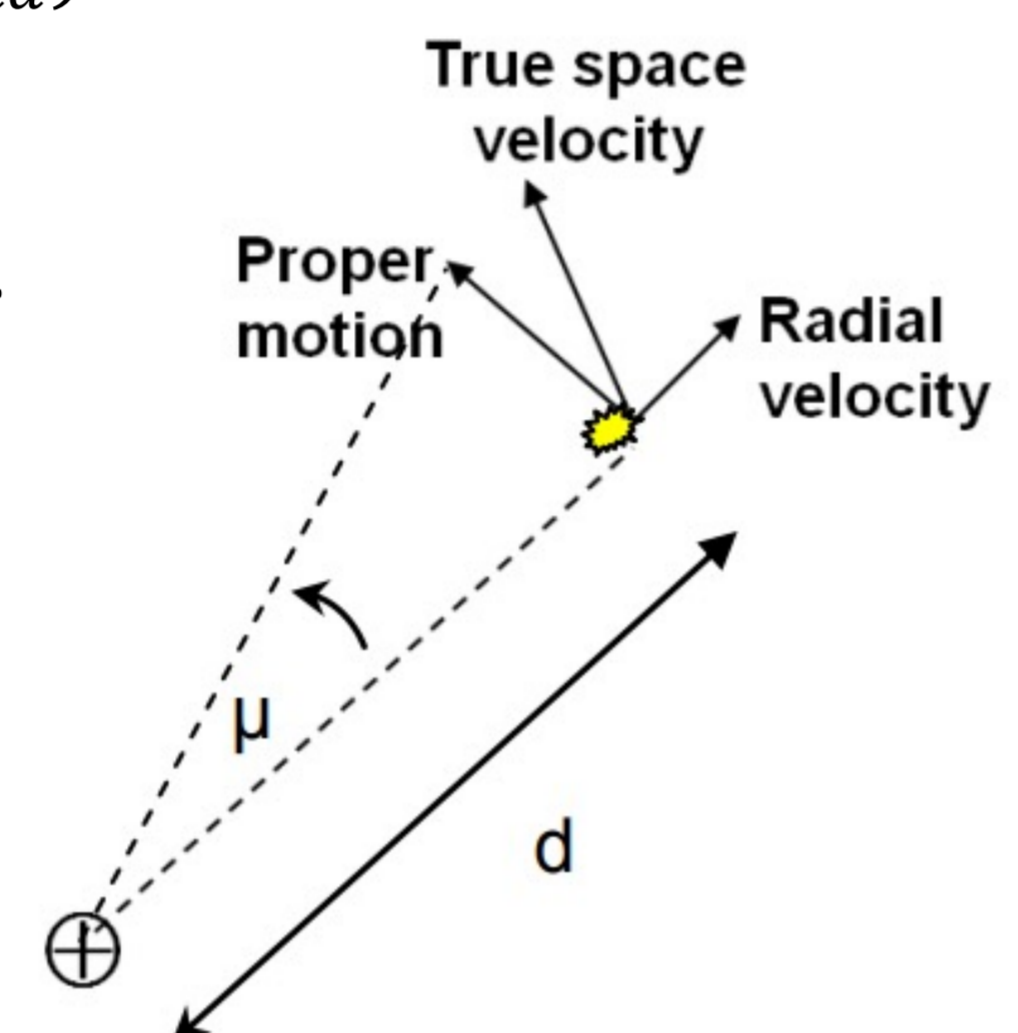


Fig. 1: Components of stellar motion

4. Why Semi-Metric Space?

- Radial velocity is often missing** and must be estimated.
- It is estimated by minimizing the velocity difference: $\frac{\partial \|\Delta v\|}{\partial v_{rad,(1,2)}} = 0$.
- These approximated distances are not guaranteed to follow the triangle inequality.
- Semi-metric space satisfies nonnegativity, null condition, symmetry, but **does not follow triangle inequality**.

5. Softmax Clustering in Semi-Metric Space

Cannot use standard algorithms: Traditional clustering requires metric distance that follow the triangular inequality.

Softmax Clustering solution: A probabilistic algorithm that operates on semi-cohesion matrix – a similarity measure derived from pairwise distances [1].

- Algorithm steps:**
- Initialize soft membership probabilities for each point across all clusters.
 - Compute expected cohesion between each point and cluster.
 - Update memberships using a softmax function, which favours higher cohesion.
 - Gradually sharpen assignments by increasing the confidence.

Semi-cohesion measure:

$$\gamma(a, b) = \frac{1}{n} \sum_{z_2 \in \Omega} d(z_2, b) + \frac{1}{n} \sum_{z_1 \in \Omega} d(a, z_1) - \frac{1}{n^2} \sum_{z_2 \in \Omega} \sum_{z_1 \in \Omega} d(z_2, z_1) - d(a, b)$$

[1] Chia-Tai Chang and Cheng-Shang Chang "A Unified Framework for Sampling, Clustering and Embedding Data Points in Semi-Metric Spaces." 2017

6. Final Clustering Algorithm for Large-Scale Data

- Grid Partitioning:**
 - Divide space into square cells.
- Local Clustering:**
 - Run the Softmax Clustering Algorithm independently in each grid cell.
- Grid Transformation:**
 - Shift the grid.
 - Rotate the dataset.
- Consensus Clustering:**
 - Merge clustering outputs from all grid versions.

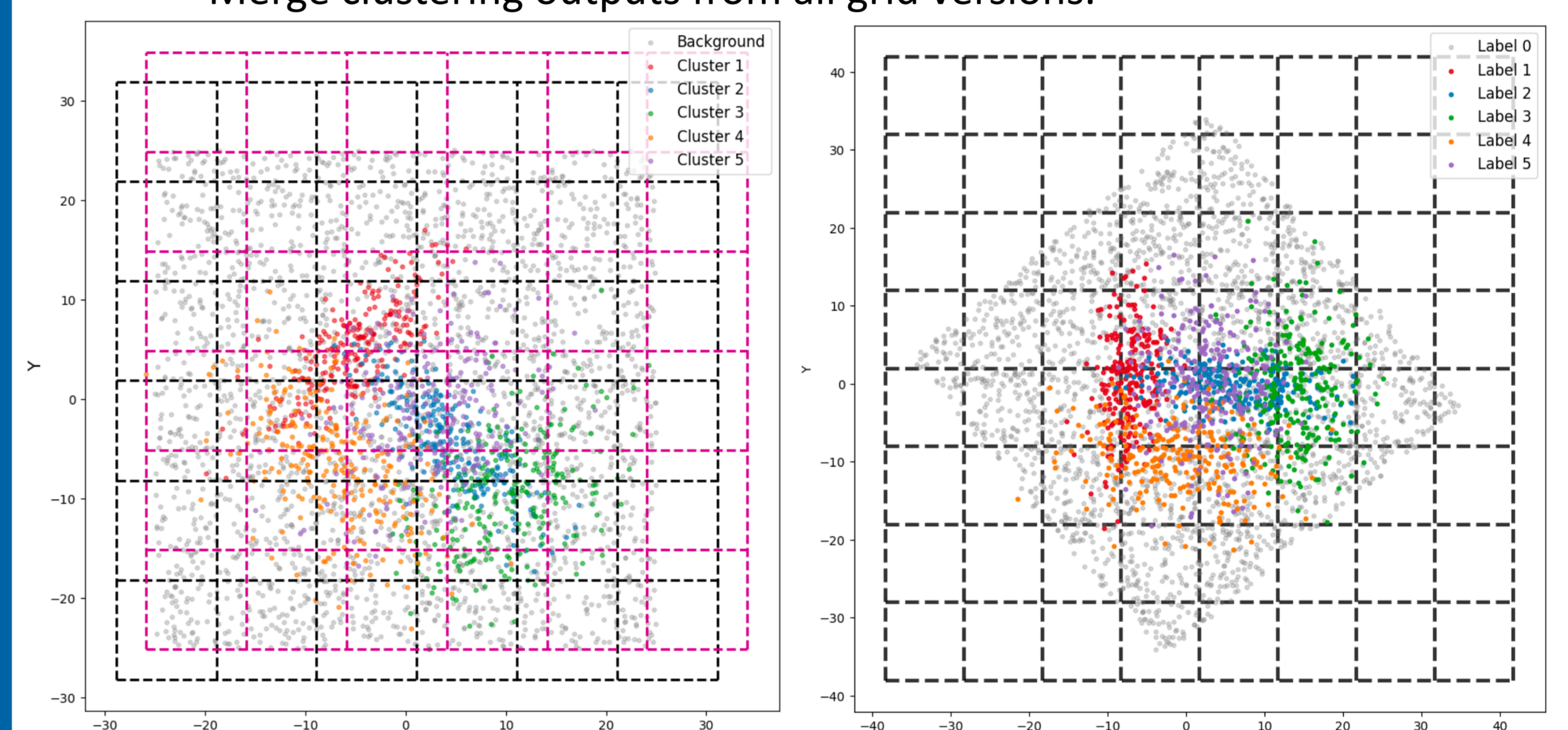


Fig. 2: Grid partitioning with a shift (left) and on a rotated dataset (right)

7. Results on Synthetic Data

- Best clustering results achieved NMI = 0.90.**
- Grid size, rotation, and consensus clustering solver affect quality.

Table 1: Effect of grid and solver parameters

Grid Size	Rotation	Shift Vector	Solver	NMI	Runtime (s)
20	30°	[3, 3, 3]	cspa	0.90	10.29
20	30°	[3, 3, 3]	hbgf	0.41	25.23
10	30°	[3, 3, 3]	cspa	0.57	06.64

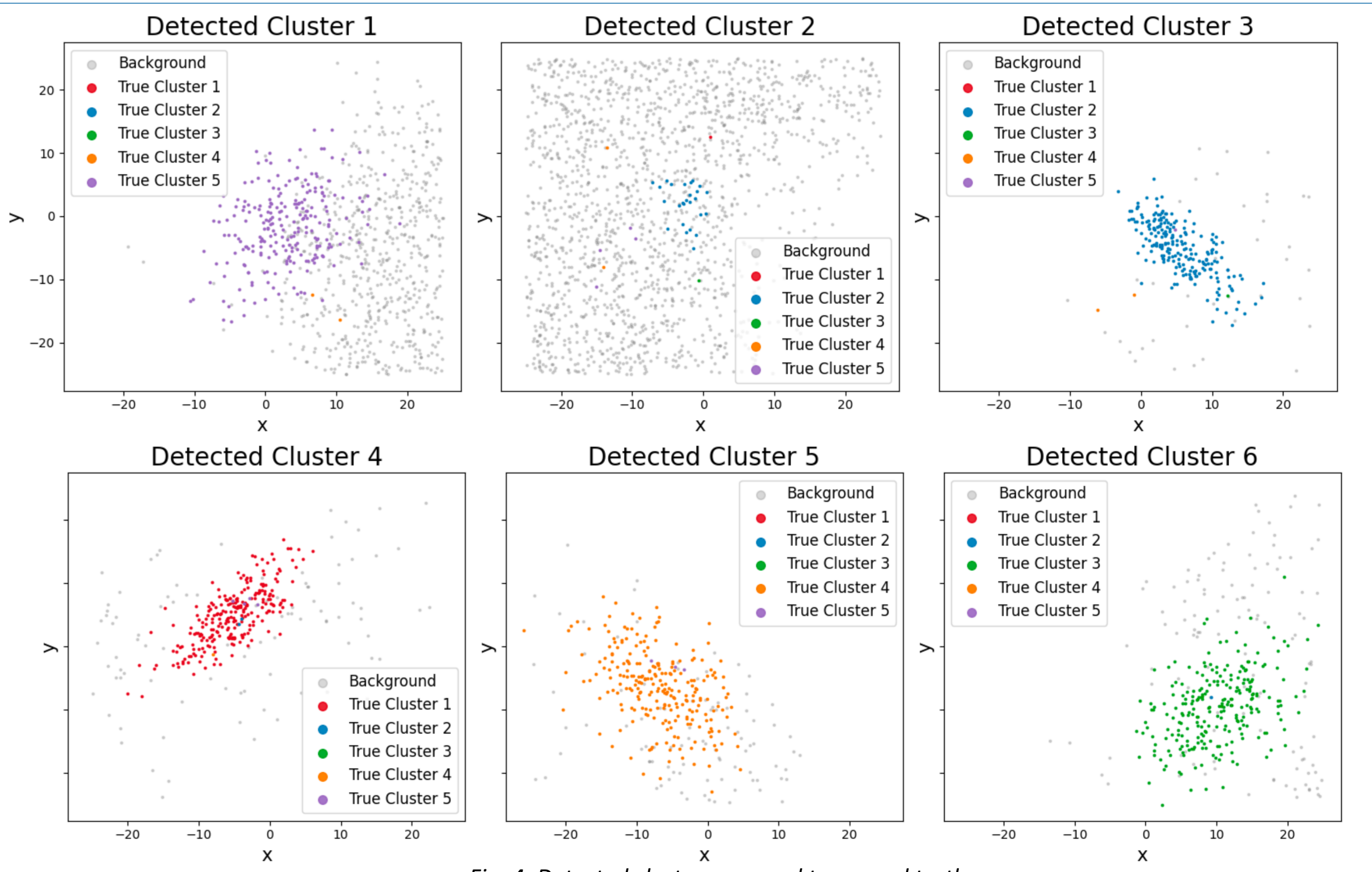


Fig. 4: Detected clusters mapped to ground truth

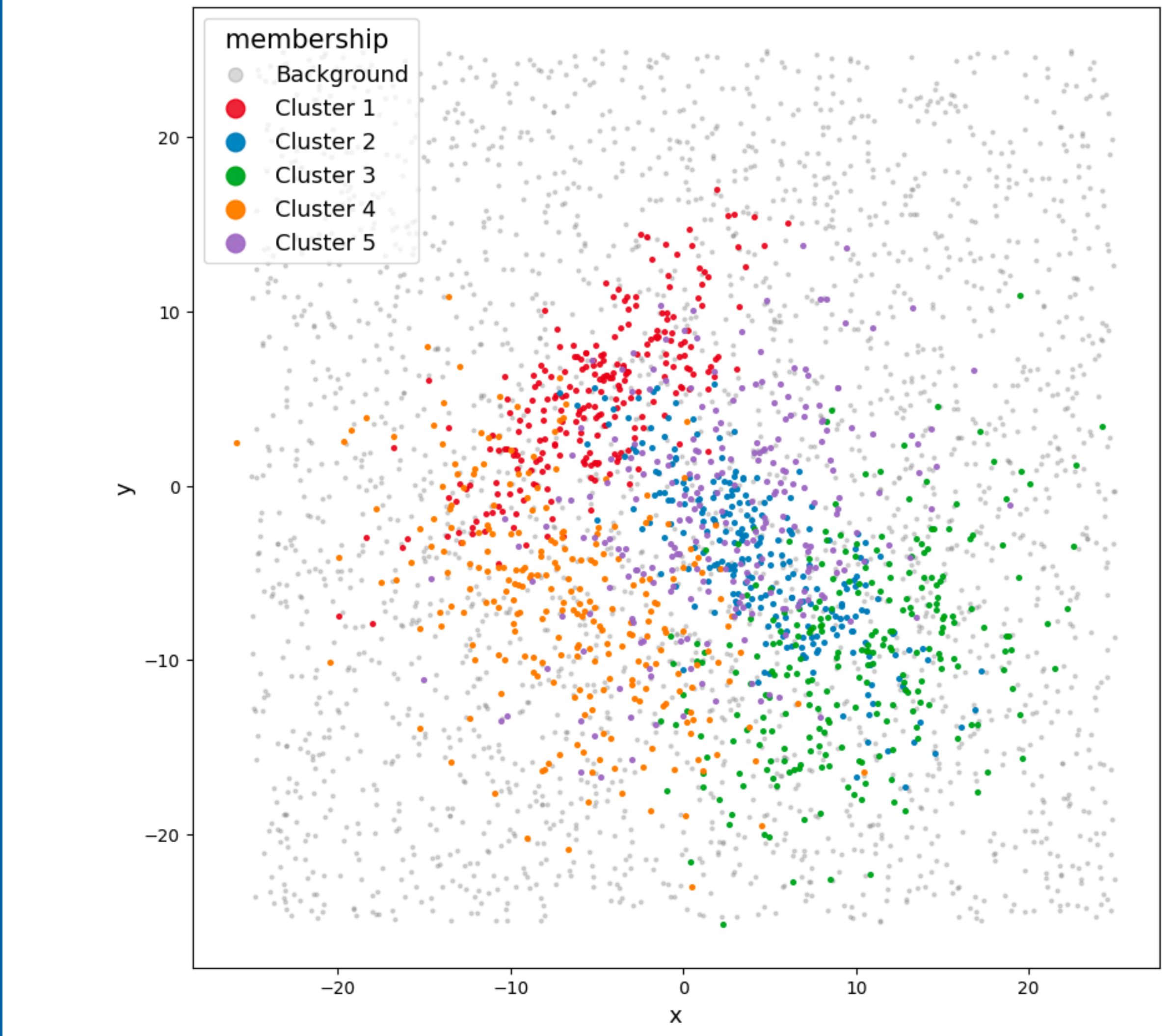


Fig. 3: Synthetic dataset before clustering

Table 2: True label distribution in detected clusters

Cluster	Dominant True Label	Points from Label	Background	Other Labels
1	5	238	602	2
2	Background	1557	1557	34
3	2	221	38	4
4	1	249	94	7
5	4	243	91	3
6	3	248	118	1