


## ABSTRACT

- We propose a method to perform *clustering of data with missing entries*.
- **The technique is able to recover the original clusters.**
- Useful for *analyzing and visualizing patterns in large datasets*.


## MOTIVATION

When does data have missing entries? In most practical situations!




**Netflix**

- Each user rates a small fraction of available movies
- Most ratings are **missing**



**Surveys**

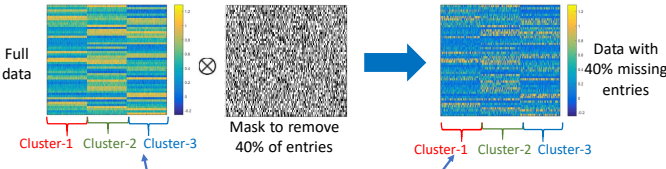
- Many respondents leave some questions unanswered
- These form **missing entries**



**Medical Records**

- All information is not available for each patient
- These form **missing entries**

Full data vs Data with missing entries



## PROPOSED SCHEME

### l<sub>0</sub> penalty based optimization problem

$$\{u_i^*\} = \min_{\{u_i\}} \sum_{i,j=1}^{KM} \|u_i - u_j\|_{2,0} \text{ s.t. } \|S_i(x_i - u_i)\|_\infty \leq \frac{\epsilon}{2}, i \in \{1 \dots KM\}$$

↑ Estimated centres      ↑ Selects sampled entries

Solving this problem is computationally intensive  
Hence, we solve a relaxation of this problem

### Relaxed optimization problem

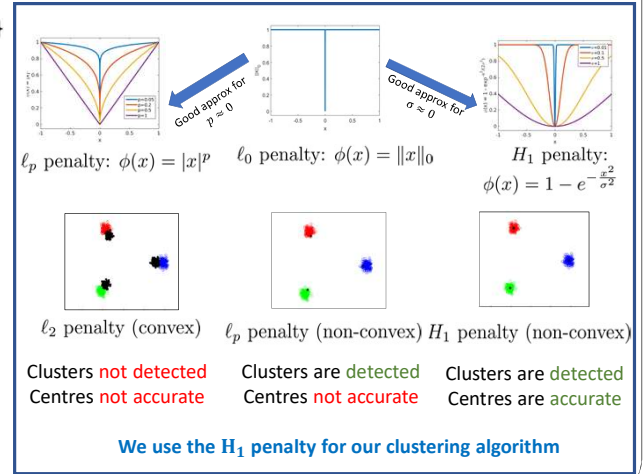
$$\{u_i^*\} = \arg \min_{\{u_i\}} \sum_{i=1}^{KM} \|S_i(u_i - x_i)\|_2^2 + \lambda \sum_{i,j=1}^{KM} \phi(\|u_i - u_j\|_2)$$

### Solve using majorize-minimize formulation

$$\{u_i^{(n)}\} = \arg \min_{\{u_i\}} \sum_{i=1}^{KM} \|S_i(u_i - x_i)\|_2^2 + \lambda \sum_{i,j=1}^{KM} w_{i,j}^{(n-1)} \|u_i - u_j\|_2^2$$

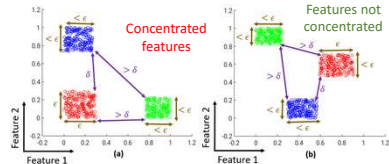
$$w_{i,j}^{(n)} = \frac{\phi'(\|u_i^{(n)} - u_j^{(n)}\|)}{2\|u_i^{(n)} - u_j^{(n)}\|}$$

## Effect of different penalties on clustering



## Definitions and Assumptions

- $K$  clusters
- $M$  points in each cluster
- $p_0$ : probability that a feature is measured



- **Cluster Separation:**  $\geq \delta$
- **Cluster size:**  $\leq \epsilon$
- **Feature concentration:** Coherence of difference between points in different clusters is  $\mu$

## THEORETICAL GUARANTEES

### Clustering using l<sub>0</sub> penalty

$$\text{Let } \kappa = \frac{\epsilon\sqrt{P}}{\delta}$$

- $P$ : Dimensionality
- $\epsilon$ : Intra-cluster separation
- $\delta$ : Inter-cluster separation

### Result with missing entries:

If  $\kappa < 1$ , correct clustering with probability  $> 1 - \eta_0$

### Probability of success is higher for:

- More points ( $M$ )
- Few missing entries
- Few clusters ( $K$ )
- Well separated clusters

### Result with no missing entries:

If  $\kappa < 1$ , correct clustering is guaranteed

## Computing probability of success

- Probability of 2 points from different clusters sharing a centre  $< \beta_0$
- For 2 clusters, probability of clustering failure:

$$\eta_0 = \sum_{i=1}^{M-1} \left[ \beta_0^{i(M-i)} \binom{M}{i}^2 \right] \leq M^3 \beta_0^{M-1}$$

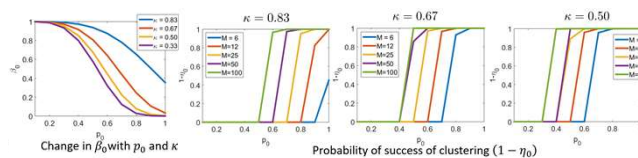
- Generalized to  $K$  clusters:

$$\eta_0 = \sum_{\{m_j\} \in \mathcal{S}} \left[ \beta_0^{\frac{1}{2}(M^2 - \sum_j m_j^2)} \prod_j \binom{M}{m_j} \right]$$

where  $\mathcal{S}$  is the set of all sets with  $\leq K$  non-zero positive integers with sum  $M$

## RESULTS

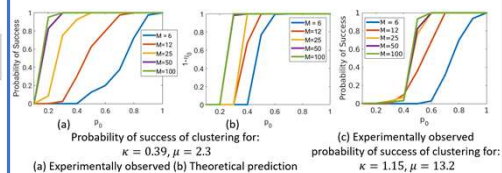
### Study of theoretical guarantees



Probability of correct clustering increases with:

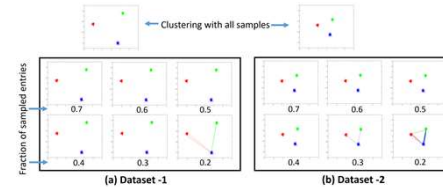
- More points:  $M \uparrow$
- More measured features:  $p_0 \uparrow$
- $\kappa \downarrow$

### Theoretical guarantees vs experimental results



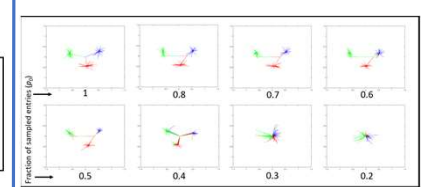
Comparison on a simulated dataset with 2 clusters using 20 experimental trials

### Clustering of simulated data



- 2 simulated datasets: 3 clusters, 200 points in each
- Successful clustering for 70% missing entries in data-1 and 60% missing entries in data-2

### Clustering of wine data



- 3 classes of Wine, 40 samples in each
- Successful clustering for 50% missing entries

## CONCLUSION

- Proposed algorithm can **reliably cluster datasets with large fractions of missing entries**.
- Performance degrades with: (1) More number of missing entries (2) Outliers (3) Less separation between clusters (4) High variance within clusters (5) High feature concentration.

## REFERENCES

1. T. D. Hocking et al, "Clusterpath an algorithm for clustering using convex fusion penalties", ICML 2011.
2. B. Eriksson et al, "High-Rank Matrix Completion and Subspace Clustering with Missing Data", arXiv: 1112.5629.
3. C. Zhu et al, "Convex optimization procedure for clustering: Theoretical revisit", NIPS 2014.