CLUSTERING OF DATA WITH MISSING ENTRIES

Computational Biomedical Imaging Group (CBIG)

Sunrita Poddar and Mathews Jacob

The University of Iowa, Iowa City, IA, USA

The University of Iowa

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

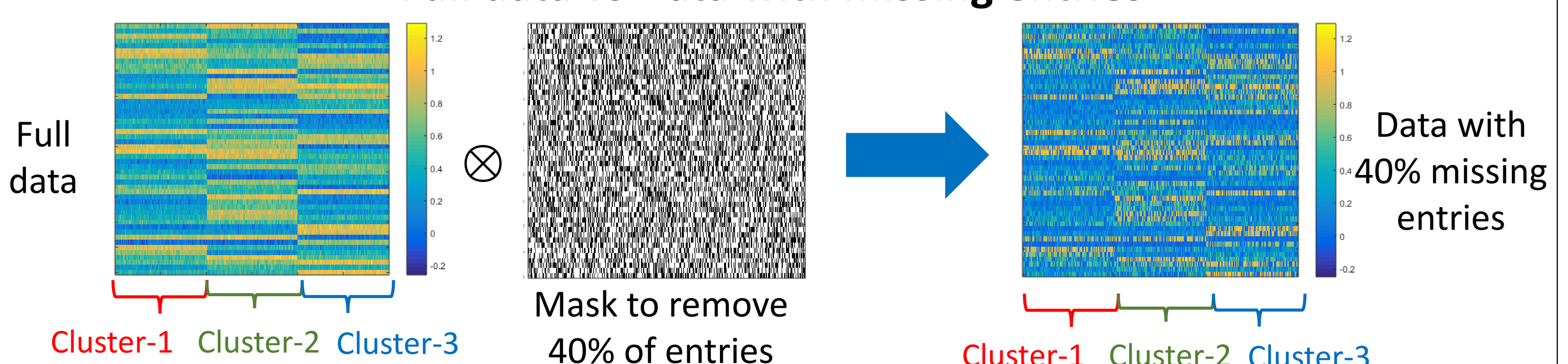
When does data have missing entries? In most practical situations!

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



Our aim: Design algorithm that finds the same clusters for both datasets

PROPOSED SCHEME

l_0 penalty based optimization problem

$$\{\mathbf{u}_i^*\} = \min_{\{\mathbf{u}_i\}} \sum_{i,j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_{2,0} \text{ s.t } \|\mathbf{S}_i\,(\mathbf{x}_i - \mathbf{u}_i)\|_\infty \le \frac{\epsilon}{2}, i \in \{1 \dots KM\}$$

Estimated centres — Selects sampled entries

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

Relaxed optimization problem

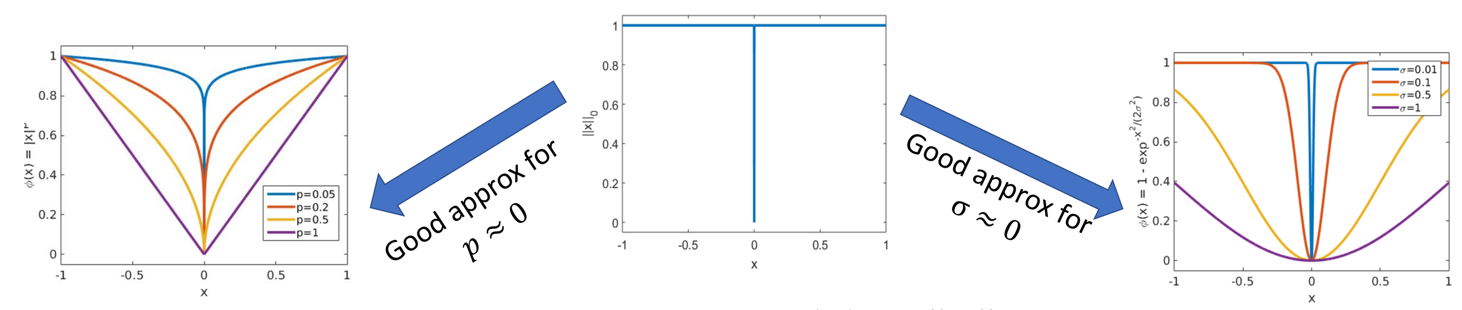
$$\{\mathbf{u}_i^*\} = \arg\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 + \lambda \sum_{i,j=1}^{KM} \phi(\|\mathbf{u}_i - \mathbf{u}_j\|_2)$$

Solve using majorize-minimize formulation

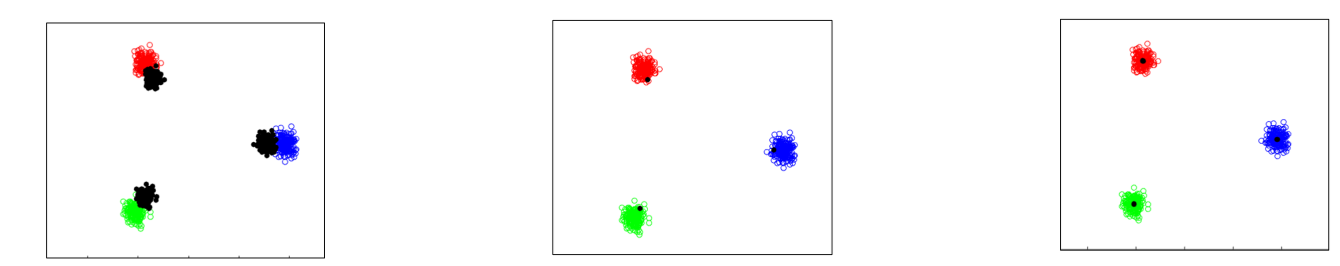
$$\{\mathbf{u}_i^{(n)}\} = \arg\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 + \lambda \sum_{i,j=1}^{KM} w_{ij}^{(n-1)} \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$$

$$w_{i,j}^{(n)} = \frac{\phi'(\|\mathbf{u}_i^{(n)} - \mathbf{u}_j^{(n)}\|)}{2\|\mathbf{u}_i^{(n)} - \mathbf{u}_j^{(n)}\|}$$

Effect of different penalties on clustering



ℓ_p penalty: $\phi(x) = |x|^p$ — ℓ_0 penalty: $\phi(x) = \|x\|_0$ — H_1 penalty: $\phi(x) = 1 - e^{-\frac{x^2}{\sigma^2}}$



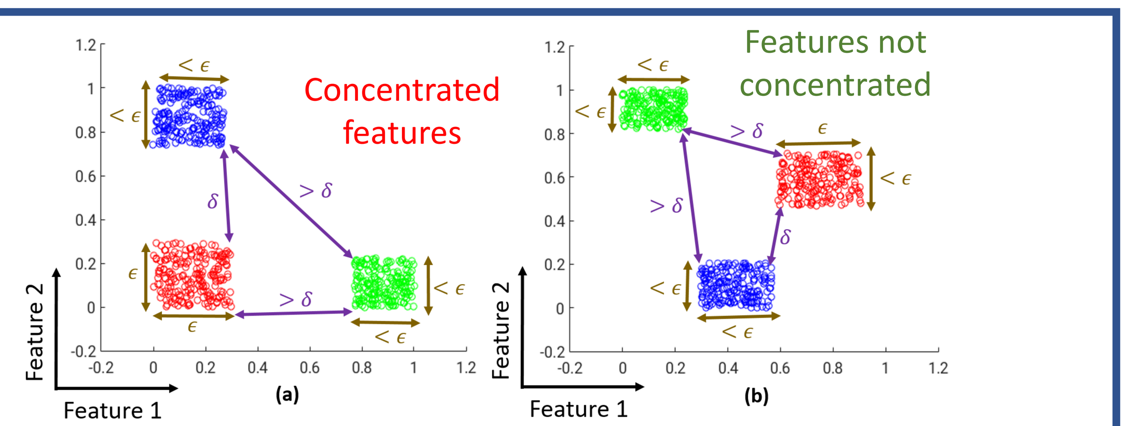
ℓ_2 penalty (convex)	ℓ_p penalty (non-convex)	H_1 penalty (non-convex)
Clusters not detected Centres not accurate	Clusters are detected Centres not accurate	Clusters are detected Centres are accurate

We use the $\mathbf{H_1}$ penalty for our clustering algorithm

THEORETICAL GUARANTEES

Definitions and Assumptions

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



- **Cluster Separation:** $\ge \delta$
- **Cluster size:** $\le \epsilon$
- **Feature concentration:** Coherence of difference between points in different clusters is μ

Clustering using l_0 penalty

Let $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$
- P: Dimensionality
- ϵ : Intra-cluster separation
- δ : 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:

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

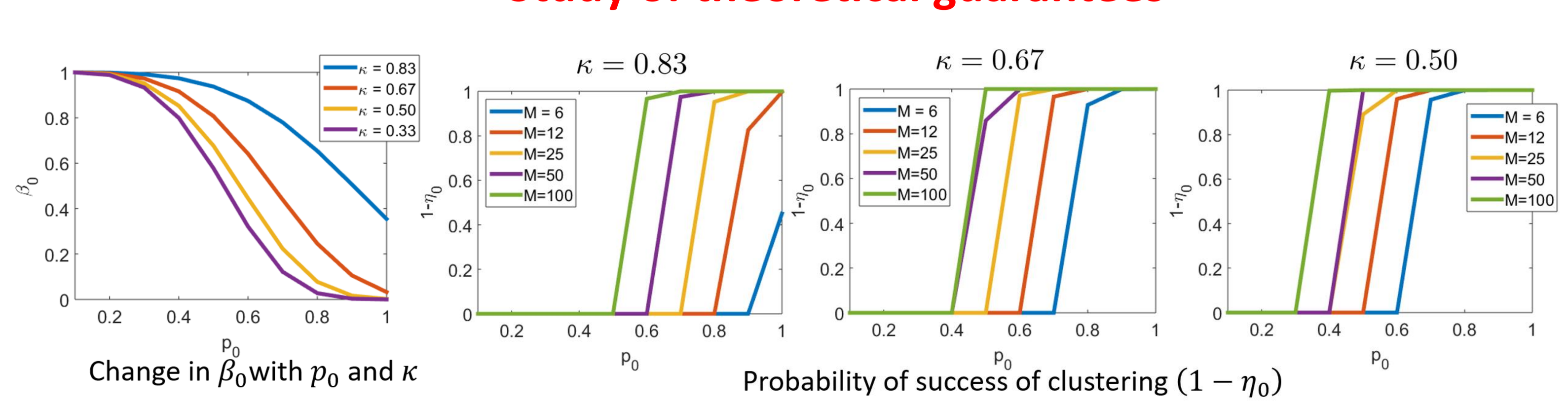
- Generalized to K clusters:

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

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

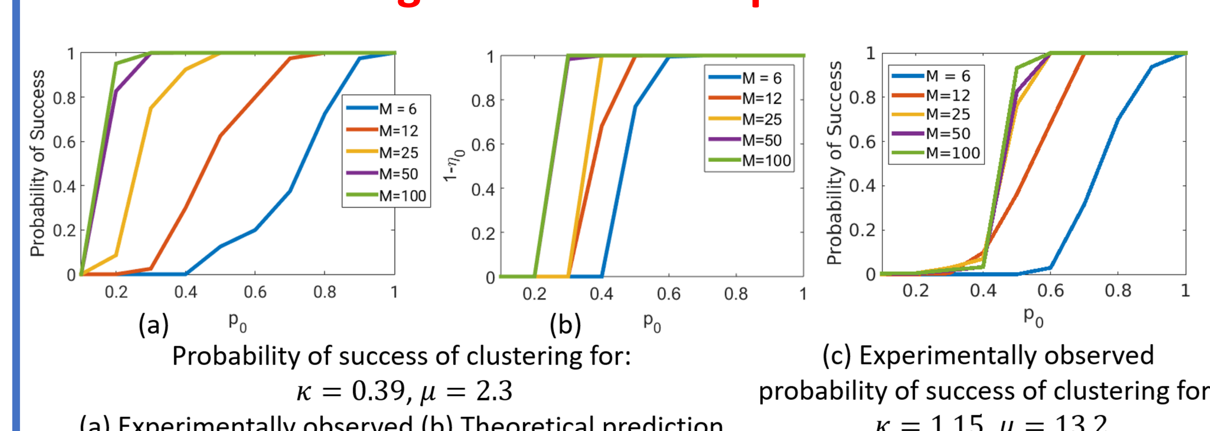
RESULTS

Study of theoretical guarantees



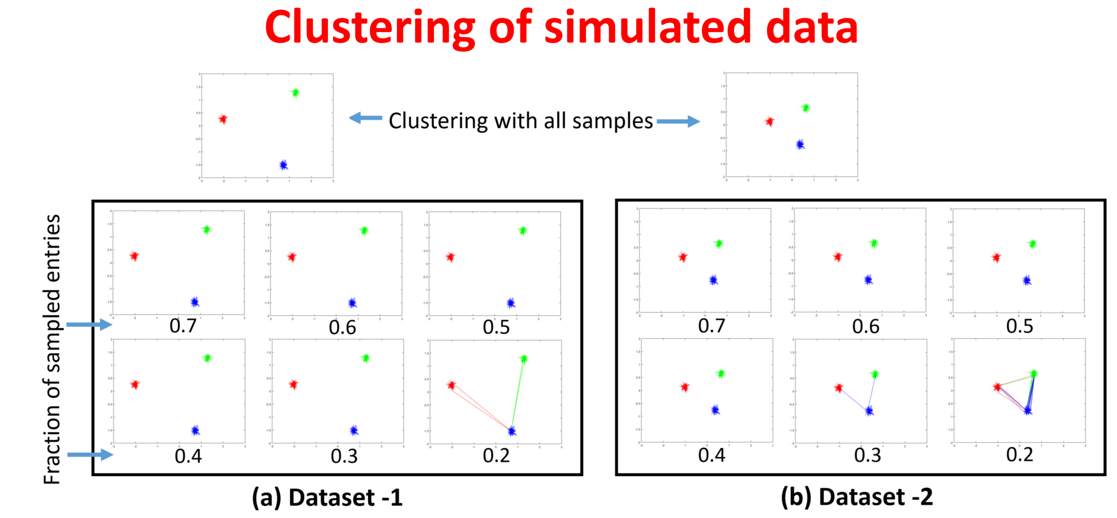
Probability of correct clustering increases with:
- More points: $M \uparrow$
- More measured features: $p_0 \uparrow$
- $\kappa \downarrow$

Theoretical guarantees vs experimental results



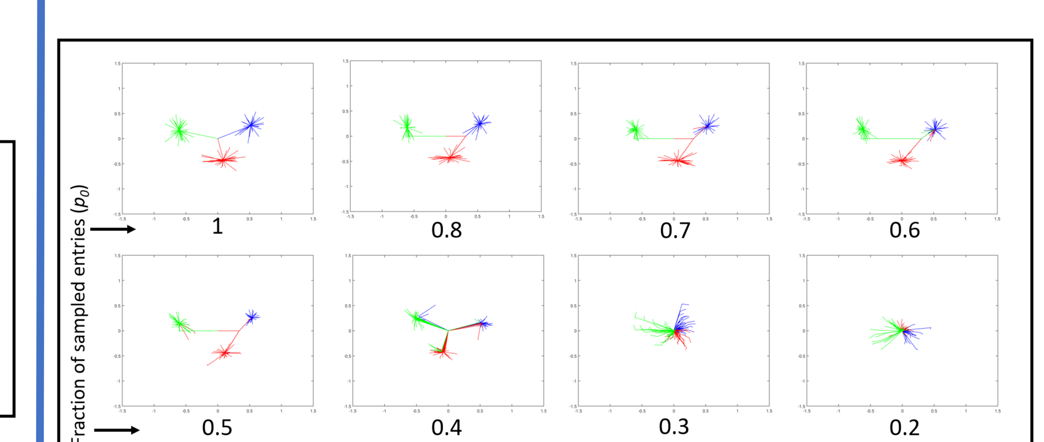
Comparison on a simulated dataset with 2 clusters using 20 experimental trials

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.