Clustering using Max-Norm Constrained Optimization

Ali Jalali[†] and Nathan Srebro[‡]

[†]University of Texas at Austin and [‡]Toyota Technological Institute at Chicago

Problem Definition

Correlation Clustering [1]: Given a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ on n nodes and affinity matrix A (for all $u,v\in\mathcal{V}: 0\le A_{uv}\le 1$ and $A_{uu}=1$), find the clustering $\mathcal{C}=\{C_1,\ldots,C_k\}$ that minimizes the disagreement

$$\mathcal{D}(\mathcal{C}) = \underbrace{\sum_{i=1}^{k}\sum_{u,v\in C_i}(1-A_{uv})}_{\text{Internal Disagreement}} + \underbrace{\sum_{i\neq j=1}^{k}\sum_{u\in C_i;v\in C_j}A_{uv}}_{\text{External Disagreement}}.$$

- Represent $\mathcal{C}$ using incident matrix $K(\mathcal{C})\in\mathbb{R}^{n\times n}$ such that $K(\mathcal{C})_{uv}=1$ if u and v belong to same cluster and zero otherwise.
- Equivalent disagreement definition:

  $\mathcal{D}(\mathcal{C}) = \|A-K(\mathcal{C})\|_1 = \sum_{u,v}|A_{uv}-K(\mathcal{C})_{uv}|$ or $\mathcal{D}(\mathcal{C}) = \sum_{u,v}K(\mathcal{C})_{uv}(1-2A_{uv}) + \sum_{u,v}A_{uv}$
- Correlation Clustering:

$$\min_K \ \mathcal{D}(K) \quad \text{s.t.} \quad K \text{ is a valid clustering matrix.} \tag{1}$$

  This constraint is NOT convex.
- A valid clustering matrix is a low-rank matrix; trace-norm is a proxy to rank [2]

$$\min_K \ \|A-K\|_1 \quad \text{s.t.} \quad \|K\|_* = n.$$

Is this the tightest convex relaxation of the set of valid clustering matrices?

Max-Norm Relaxation

- Max-Norm of a matrix K is $\|K\|_{\max} = \max_{K=RL^T}\|R\|_{\infty,2}\|L\|_{\infty,2}$, where, $\|\cdot\|_{\infty,2}$ is the maximum of the 2-norm of the rows.
- Max-norm is a tighter relaxation:

$$\{K: K \text{ is a valid clustering}\} \subset \{K: \|K\|_{\max}\le 1\} \subsetneq \{K: \|K\|_* \le n\}$$

Hence, better to solve $\min_K \ \|A-K\|_1 \quad \text{s.t.} \quad \|K\|_{\max}=1.$

◇ Fundamental Noise Bound

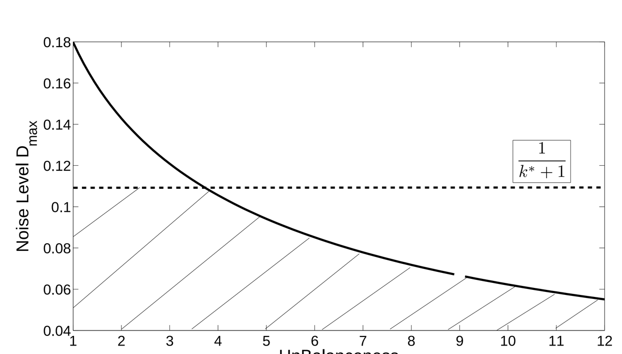
- Suppose the optimal clustering is $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$
- For any node u and any cluster C_i^*, let $d_{u,i} = \frac{1}{|C_i^*|}\sum_{v\in C_i^*}A_{u,v}$ if $u\notin C_i^*$ and $d_{u,i} = 1-\frac{1}{|C_i^*|}\sum_{v\in C_i^*}A_{u,v}$ if $u\in C_i^*$. Let $D_{\max}(A,K(\mathcal{C}^*)) = \max_{u,i} d_{u,i}$.

Lemma 1. *Given any clustering $\mathcal{C}$ and $\gamma > \frac{5}{2+n^2/\sum|C_i|^2} \approx \frac{5}{2+k}$, there exists A s.t. $D_{\max}(A,K(\mathcal{C}))=\gamma$ and combinatorial program (1) does NOT output $\mathcal{C}$.*

◇ Guarantee

- Let $\hat{K}_\mu = \arg\min_K \frac{1-\mu}{n^2}\|A-K\|_1 + \mu\|K\|_{\max}$

Theorem 1. *For binary A, if $D_{\max}\le\frac{1}{k+1}$ is small enough s.t. Unbalanceness $=\frac{1}{k}\sum_i\left(\frac{|C_i^*|}{|C_{\min}|}\right)^2 \le \frac{(1-3D_{\max})^2}{(1+D_{\max})D_{\max}}$, then for any μ in a known range, $\hat{K}_\mu = K^*$.*



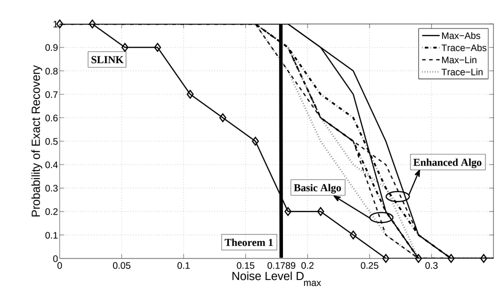
Comparison

◇ Guarantee Comparison

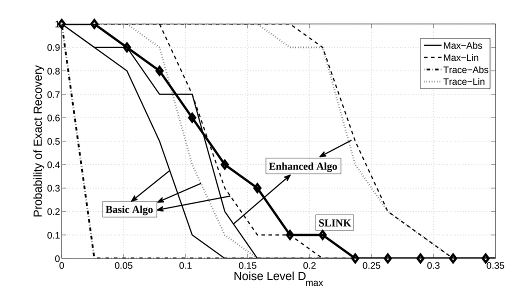
- Balanced Clustering
  - Trace-Norm: $D_{\max}\le\frac{1}{4k}$
  - Max-Norm: $D_{\max}\le\min(\frac{1}{k+1}, 0.1789)$
- Unbalanced Clustering
  - Trace-Norm: $D_{\max}=o(\frac{1}{n})$
  - Max-Norm: $D_{\max}=o(\frac{k}{n})$

◇ Numerical Comparison

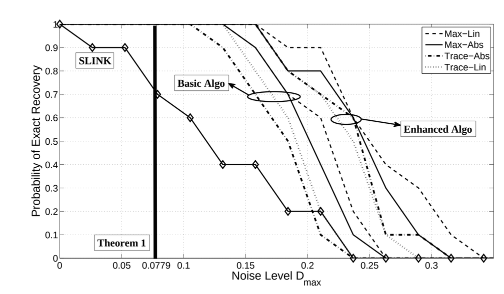
- Balanced: 4 clusters of size 25
- Unbalanced: 3 clusters of size 30 and a cluster of size 10
- Absolute loss function $\|A-K\|_1$
- Linear loss $\sum_{u,v}K_{u,v}(1-2A_{u,v})$



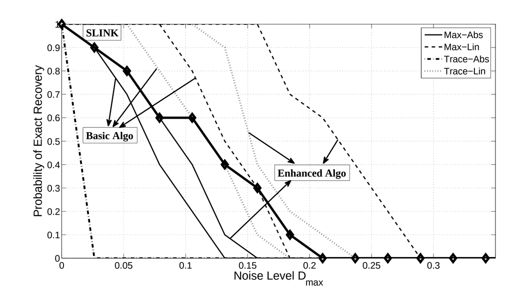
Balanced; Binary



Balanced; Fractional



Unbalanced; Binary



Unbalanced; Fractional

Max-Norm + ℓ_1 Optimization

- **Semi-Definite Programming:** Solve

$$\widehat{K} = \arg\min_{K,L,R} \|A-K\|_1 \quad \text{s.t.} \quad \begin{bmatrix} L & K \\ K^T & R\end{bmatrix}\succeq 0 \quad\text{and}\quad L_{ii},R_{ii}\le 1$$

- **Factorization Method:** Solve $\widehat{K}=\widehat{L}\widehat{R}^T = \arg\min_{L,R}\|A-LR^T\|_1$ s.t. $\|L\|_{\infty,2},\|R\|_{\infty,2}\le 1$, with updates

$$\begin{bmatrix}L\\R\end{bmatrix}_{k+1} = \mathcal{P}_{\max}\left(\begin{bmatrix}L\\R\end{bmatrix}_k + \frac{\tau}{\sqrt{k}}\begin{bmatrix}\mathbf{Sign}(A-LR^T)\ R\\ \mathbf{Sign}(A-LR^T)^T\ L\end{bmatrix}_k\right),$$

  where, $\mathcal{P}_{\max}(\cdot)$ scales the rows with euclidean norm greater than one by their norm.

- **Loss Function Method:** Solve $\widehat{K} = A-\widehat{Z} = \arg\min_{Z,L,R}\|Z\|_1 + \lambda\|A-Z-LR^T\|_2^2$ s.t. $\|L\|_{\infty,2},\|R\|_{\infty,2}\le 1$, using updates

$$Z_{k+1} = \mathcal{P}_{\ell_1}\left(Z_k + \frac{\tau\lambda}{\sqrt{k}}(A-Z-LR^T)_k\right) \quad \& \quad \begin{bmatrix}L\\R\end{bmatrix}_{k+1} = \mathcal{P}_{\max}\left(\begin{bmatrix}L\\R\end{bmatrix}_k + \frac{\tau\lambda}{\sqrt{k}}\begin{bmatrix}(A-Z-LR^T)\ R\\ (A-Z-LR^T)^T\ L\end{bmatrix}_k\right),$$

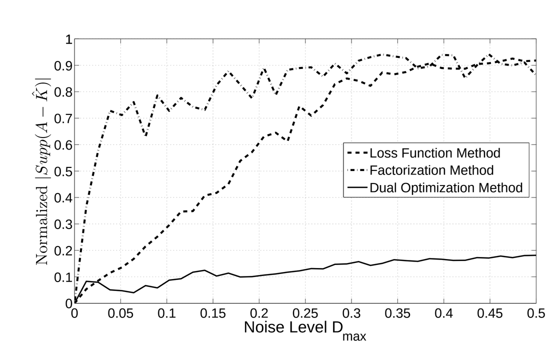
  where, $\mathcal{P}_{\ell_1}(\cdot)$ set the entries who change their sign during the update to zero.

- **Dual Method:** Solve $\widehat{K} = \arg\max_\Lambda\min_{Z,K}\ \|A-K\|_1 + \langle\langle\Lambda,\ K-Z\rangle\rangle$ s.t. $\|Z\|_{\max}\le 1$ by solving separate problems

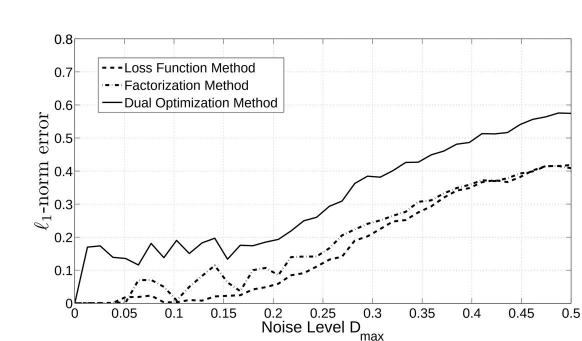
$$\widehat{K}(\Lambda) = \arg\min_K \|A-K\|_1 + \langle\langle\Lambda,\ K\rangle\rangle \quad \& \quad \widehat{Z}(\lambda) = \arg\min_Z -\langle\langle\Lambda,\ Z\rangle\rangle \quad \text{s.t.} \quad \|Z\|_{\max}\le 1.$$

  and update $\Lambda_{k+1} = \Lambda_k - \frac{\tau}{\sqrt{k}}(\widehat{K}(\Lambda_k) - \widehat{Z}(\Lambda_k))$.

  - **Comparison:** 2000 iterations on three clusters of size 20 for different values of noise

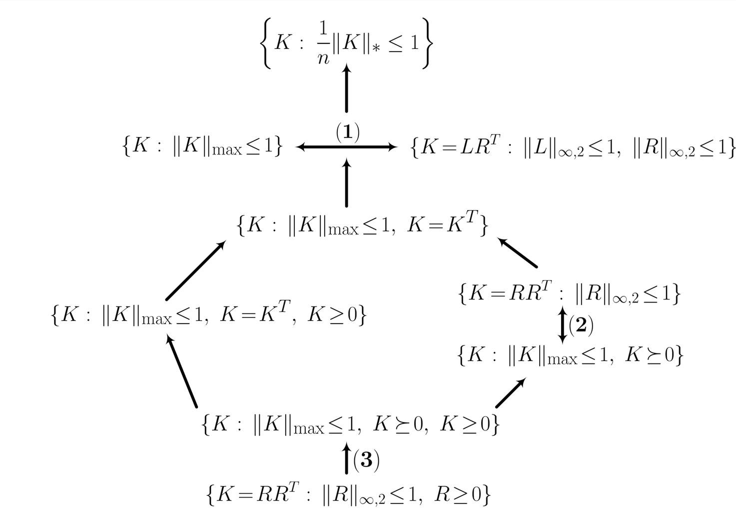


$|\mathrm{Supp}(A-\widehat{K})|$



$\|K^*-\widehat{K}\|_1$

Tighter Relaxations

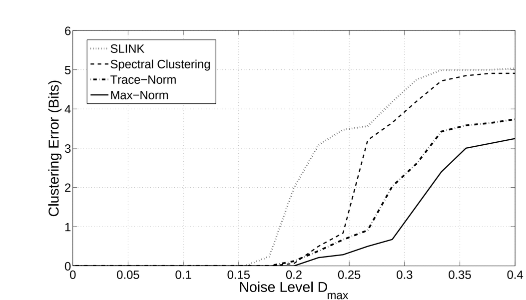


◇ Enhanced Algorithm

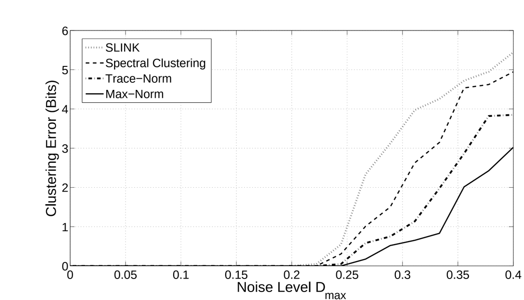
1. Solve $\widehat{R} = \arg\min_R \|A-RR^T\|_1$ s.t. $\|R\|_{\infty,2}\le 1$ & $R\ge 0$, and let $\tilde{K}=\widehat{R}\widehat{R}^T$.
2. Run SLINK on $\tilde{K}$ to get the sequence $K_1,K_2,\ldots,K_n$ and let $\widehat{K} = \arg\min_i\|A-K_i\|_1$.

◇ Comparison

- Comparing the enhanced algorithm (linear objective) with other methods using Variation of Information [3]
- Feeding the correct number of clusters to spectral clustering
- Adding SLINK as a rounding scheme to all methods



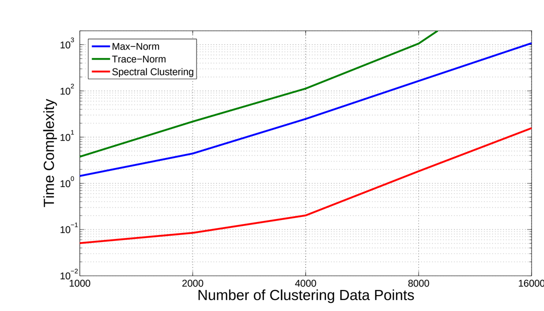
Balanced; Fractional



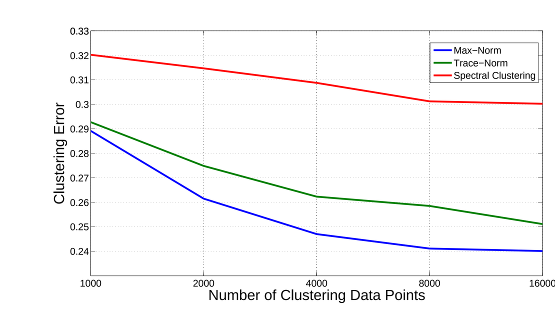
Unbalanced; Fractional

◇ MNIST Dataset

- Picking n data from each class at random
- Measuring the run-time and error as n increases



Time Complexity



Clustering Error

- Max-norm constrained clustering provides stronger guarantees than Trace-norm
- Max-norm constrained clustering outperforms the Trace-norm in practice
- Our optimization methods are scalable to very large size problems

References

[1] N. Bansal, A. Blum and S. Chawla. "Correlation Clustering," In Proceedings of the 43rd Symposium on Foundations of Computer Science, 2002.

[2] A. Jalali, Y. Chen, S. Sanghavi and H. Xu. "Clustering Partially Observed Graphs via Convex Optimization," ICML, 2011.

[3] M. Meila, "Comparing clusterings: an information based distance," Journal of Multivariate Analysis, 98:873-895, 2007.