Convex relaxation for the planted cluster problem

Brendan Ames[1] Stephen Vavasis[1]

[1]Department of Combinatorics and Optimization
University of Waterloo

2010 SIAM Annual Meeting
Minisymposium on Matrix Rank Minimization
July 14, 2010

Clustering

- Clustering: Given data points with known pairwise distances, group them into clusters such that points in each cluster are closer to each other than points in other clusters.

- Combinatorial clustering: Given a graph G with N data points where any two data points are adjacent if and only if they are compatible. Want to group the data into disjoint clusters of compatible data points (i.e. cliques).

The maximum node k-disjoint-clique problem

- Given integer $k \in [1, N]$ a *k-disjoint-clique subgraph* is a subgraph of G composed of k disjoint cliques.
- Maximum node k-disjoint-clique problem: Given graph G find a k-disjoint-clique subgraph covering the maximum number of nodes.
- Each clique corresponds to a rank one matrix in $\mathbf{R}^{N \times N}$.
  - $N \times N$ node-node adjacency matrix of the clique, plus 1's on the diagonal corresponding to nodes in clique.
- Want to find a rank k matrix corresponding to k disjoint cliques with maximum number of nonzero entries

The maximum clique problem

- Maximum clique problem: Given a graph G, want to find the largest clique of G.
- Well-known to be NP-hard.
- Problem is equivalent to finding the largest rank one matrix corresponding to a subgraph of G.

Formulation as rank minimization

- A clique of $G = (V, E)$ containing at least n nodes can be found (if one exists) by solving the rank minimization problem:

$$\begin{array}{ll} \min & \operatorname{rank}(X) \\ \text{s.t.} & \sum\sum X_{ij} \geq n^2 \\ & X_{ij} = 0 \qquad\qquad \forall\, (i,j) \notin E,\ i \neq j \\ & X \in \{0,1\}^{V \times V} \end{array}$$

- Convexify by replacing $\operatorname{rank}(X)$ with $\|X\|_*$, removing binary constraints.

Results

- Construct graph $G = (V, E)$ with $|V| = N$ nodes as follows:
  - Start with clique K with n nodes.
  - Add noise edges: either by an *adversary* or *randomly*.

- [A-V] If n is sufficiently large and not too many noise edges are added then can find K by solving the nuclear norm min relaxation.

Relaxation as SDP

- The maximum node k-disjoint-clique problem can be relaxed as

$$
\begin{array}{rll}
 & \max & \sum\sum X_{ij} \\
 & \text{s.t.} & X\mathbf{e} \leq \mathbf{e}, \\
(SDR) & & X_{ij} = 0 \qquad \forall (i,j) \notin E,\ i \neq j \\
 & & \operatorname{trace}(X) = k, \\
 & & X \succeq 0
\end{array}
$$

- Every feasible X is positive semidefinite $\Rightarrow \|X\|_* = \operatorname{trace}(X)$.
- The k-disjoint-clique subgraph K composed of the disjoint cliques $C_1, \ldots, C_k$ induces the feasible solution

$$X^* = \sum_{i=1}^{k} (1/|C_i|)\mathbf{v}_i\mathbf{v}_i^T.$$

where $\mathbf{v}_i \in \mathbf{R}^N$ is the characteristic vector of the index set C_i.

Planted k-disjoint-clique construction

- We consider the graphs G constructed as follows:
- A k-disjoint-clique subgraph K consisting of disjoint cliques of size $r_1, \ldots, r_k$ respectively.
- A set C_{k+1} of r_{k+1} additional nodes and a number of additional edges are added to the graph either deterministically by an adversary or at random independently with fixed probability p.
- Let X^* be the feasible solution of (SDR) corresponding to K.

Notation

- Let $\hat{r} = \min\{r_1, \ldots, r_k\}$.
- Let n_v^q denote the # of neighbours of the node v in the vertex set C_q for all $v \in V$, $q \in \{1, \ldots, k+1\}$.
- Let $cl(v)$ denote the vertex set containing v for all $v \in V$.

Results: Adversary chosen noise

Theorem

- *Suppose that the adversary adds at most* $r_{k+1} = O(\hat{r}^2)$ *additional nodes and* $O(\hat{r}^2)$ *edges such that*

$$n_v^q \leq \delta \min\{r_q, r_{cl(v)}\}$$

for all $v \in V$, $q \in \{1, \ldots, k+1\} - cl(v)$ *for some scalar* δ *satisfying*

$$0 < \delta < (1-\delta)^2.$$

- *Then* X^* *is the unique optimal solution of (SDR) and* K *is the unique maximum node* k*-disjoint-clique subgraph of* G.

Results: Random noise

Theorem

- *Suppose that a small number ($<< k$) of cliques have size $O(\hat{r}^\alpha)$ for some scalar $\alpha \in (1, 3/2)$.*
- *Remaining cliques have size $O(\hat{r})$.*
- *If K consists of*

$$k \leq O(\min\{\hat{r}^{1/2}, \hat{r}^{3-2\alpha}\})$$

disjoint cliques and G contains at most $r_{k+1} \leq O(\hat{r}^2)$ nodes not in K then X^* is the unique optimal solution of (SDR) and K is the unique maximum node k-disjoint-clique subgraph of G.

Results: Random noise (a special case)

Theorem

- *Suppose all cliques roughly the same size ($\sim \hat{r}$).*
- *Then if K can contains*

$$k \leq O(\hat{r}^{1/2})$$

and at most $O(\hat{r}^2)$ diversionary nodes and X^* is the unique optimal solution of (SDR) and K is the unique maximum node k-disjoint-clique subgraph of G.

Optimality conditions

- Suppose that X is feasible for (SDR) and there exists $\lambda \in \mathbf{R}^N_+, \mu \in \mathbf{R}, \eta \in \mathbf{R}^{N \times N}$ and $S \in \Sigma^N_+$ such that

$$-\mathbf{e}\mathbf{e}^T + \lambda \mathbf{e}^T + \mathbf{e}\lambda^T + \mu I + \sum_{\substack{(i,j) \notin E \\ i \neq j}} \eta_{ij} \mathbf{e}_i \mathbf{e}_j^T = S,$$

$$\lambda^T (X\mathbf{e} - \mathbf{e}) = 0,$$

$$\langle S, X \rangle = 0.$$

- Then X is optimal for (SDR).

Proof outline

- *Proof idea*: For a particular choice of λ, μ show how to construct a S that satisfies KKT conditions in the case that G satisfies the hypothesis of either theorem.
- KKT conditions give explicit formulas for entries of S corresponding to edges but remaining entries of S are unknown.
- Establishing that S is PSD involves norm bounds for its off-diagonal blocks.

Parametrization of S

- For each $i \in C_q$, $j \in C_s$ such that, $q, s \in \{1, \ldots, k\}$, $q \neq s$, and $(i, j) \notin E$ we take

$$S_{ij} = \mathbf{y}_i + \mathbf{z}_j$$

where $(\mathbf{y}, \mathbf{z})$ is a solution of the system

$$\begin{pmatrix} \mathrm{Diag}(\mathbf{d}) + \theta \mathbf{e}\mathbf{e}^T & H - \theta \mathbf{e}\mathbf{e}^T \\ H^T - \theta \mathbf{e}\mathbf{e}^T & \mathrm{Diag}(\mathbf{f}) + \theta \mathbf{e}\mathbf{e}^T \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{b}.$$

  - $\mathbf{d}_i = r_s - n_i^q$: # of nonneighbours of $i \in C_q$ in C_s
  - $\mathbf{f}_i = r_q - n_i^s$: # of nonneighbours of $i \in C_s$ in C_q
  - H : (C_q, C_s) block of the adjacency matrix of the complement of G.
  - $\mathbf{b}$: chosen so that rows/columns of the (C_q, C_s) block of S sum to 0.
  - θ : equal to 1 in adv. case, $1 - p$ in random case.

Optimality conditions, v.2

- Let $\tilde{S} \in \mathbf{R}^{N \times N}$ be such that

$$\tilde{S}_{C_q, C_s} = \begin{cases} 0 & \text{if } q = s, q \in \{1, \ldots, k\} \\ S_{C_q, C_s} & \text{if } q \neq s \\ S_{C_{k+1}, C_{k+1}} - \hat{r} I & \text{if } q = s = k+1 \end{cases}$$

- Then X^* is optimal for (SDR) and K is the maximum node k-disjoint-clique subgraph of G if $\|\tilde{S}\| \leq \hat{r} - 1$.
- If $\|\tilde{S}\| < \hat{r} - 1$ and $n_v^q < r_q$ for all $v \in V$ and $q \in \{1, \ldots, k+1\} - cl(v)$ then X^* is the unique optimal solution of (SDR) and K is the unique maximum node k-disjoint-clique subgraph of G.

Bounds on $\|\tilde{S}\|$

- *Adversarial case*: using the weak bound $\|\tilde{S}_{C_q,C_s}\| \leq \|\tilde{S}_{C_q,C_s}\|_F$ can show that

$$\sum_{q=1}^{k+1}\sum_{s=1}^{k+1} \|\tilde{S}_{C_q,C_s}\|^2 \leq O(\#\text{ noise edges} + r_{k+1}).$$

- *Random case*: use combination of norm bounds on random matrices (Geman; Füredi and Komlós), Chernoff bounds, and Bernstein analysis to show that

$$\|\tilde{S}_{C_q,C_s}\| \leq O\left(\frac{\max\{r_q, r_s\}}{\sqrt{\min\{r_q, r_s\}}}\right) \quad q, s \in \{1, \ldots, k\},\ q \neq s,$$

and the remaining blocks of $\tilde{S}$ contribute at most $O(\sqrt{N})$ to $\|\tilde{S}\|$.

Conclusions and open questions

- Convex relaxation can find a k-disjoint-clique subgraph in a graph that contains the k-disjoint-clique subgraph plus many diversionary edges.
- Results for the maximum clique problem match the previous results in the literature.
- Would be interesting to extend the technique to other information retrieval problems, e.g., nonnegative matrix factorization.