Provable Spectral Clustering Algorithms

"Spectral Clustering" widely used in practice and enjoys much empirical success. Here : in what situations is one able to prove that it succeeds in polynomial time ?

Many alg first do approximate clustering which correctly clusters most points and then clean-up to fix the mis-clustered points. First part usually much simpler - useful for MMDS.

Success (Strictly) : There is one "intended" clustering the data "generator" has in mind and the alg finds precisely that clustering. Definition avoids subjective features.

First Cut : Assume data is generated at random and different pieces of data are statistically independent.

Bopanna; Alon, Kahale; McSherry; Vempala, Wang;....

Ulterior Motive Cannot seem to handle worst-case. So lets try random case.

Not so ulterior a motive Perhaps attacking the random case will give us insights into the general problem.

Will also describe : Moving away from Random Models. Purely Linear Algebra sufficient conditions for success ?

(Known : Spectral works for Planar Graphs with bounded degrees, Spielman, Teng;)

Mixture Models

k probability densities (or discrete distributions) on $\mathbf{R}^n$ with centers $\mu_1, \mu_2, \ldots \mu_k$. "Data Generator" picks given numbers of points according to the k densities. Alg must cluster into the (those) k clusters.

If centers are given !?! and point clouds are well-separated, can cluster according to the nearest center.

Even if centers are not given, but separations are twice as much - pick any point and put points close to it into one cluster. Repeat....

distance-based clustring

Notation (throughout) :

Points (to be clustered) are in n space.

Max. Standard deviation in any direction $\leq \sigma$

Suppose now, the densities are spherical Gaussians. Points are at distance about

$$O(\sqrt{n}\sigma)$$

from center. Inter-center separation of $\Omega(\sqrt{n}\sigma)$ allows distance based clustering. Improvements by Dasgupta, Schulman; Arora, Kannan.

A matrix with given points as rows.

Best Fit $k-$Subspace : Note subspace spanned by top k singular vectors of A is the $k-$dim space minimizing the sum of squared distances to the points.

Vempala, Wong : For spherical Gaussian densities, symmetry implies Best fit k subspace passes through the centers of the k densities.

SVD yields the sub-space spanned by the centers.

Project to that subspace (Principal Component Analysis. Now inter-center separation of $\sqrt{k}\sigma$ would suffice to do distance-based clustering.

A “different” context where Spectral Clustering provably works :

RANDOM matrix A : Assume either

A_{ij} are INDEPENDENT (not identical) random variables.

OR Above-diagonal A_{ij} are independent (not identical) and A is symmetric.

We dub this the FULL-INDEPENDENCE assumption.

Rows of A are the points to be clustered.

Why full independence ?

Random Graph Model : edges are chosen independently to be in/out.

Theoretical model with practical applications.

Ulterior Motive : Much beautiful theory of random symmetric matrices, starting with the Physicist Wigner.

Another mixture model

A is $n \times n$. There is an UNKNOWN partition of $\{1, 2, \ldots n\}$ into a constant number of parts (clusters) - $V_1, V_2, \ldots V_k$ and a probability p_{rt} associated with edges between V_r and V_t.

A is a 0-1 matrix (a graph) generated under FULL INDEPENDENCE with these edge probabilities.

Given just A, find the partition and the probabilities.

4-word Description : Given A, find EA (the entry-wise expectation of A).

Wigner-type Theorem (Füredi, Komlos; Vu) : Suppose A is symmetric, fully independent $n \times n$ matrix with

Max Variance of any entry at most σ AND $|A_{ij}| \leq 1$.

Then with high probability, the largest eigenvalue of $A - EA$ is bounded :

$$||A - EA|| \leq O(\sqrt{n}\sigma).$$

$A \approx EA$ in spectral norm. In fact, Wigner says top eigen-value is O(length of one row) ! Almost no "correlation" among rows.

McSherry : $A \approx EA$ and EA is of rank k imply : the best rank k approx to A gives us approximately EA. Can be used to cluster all but ϵ fraction.

Clean-up Phase : Correct the mis-clustered points. Currently, tends to be technical, hard and often calling for strong added assumptions on the model.

Major Problem Clean up the clean-up phase.

Beyond Full Independence

New Models - no more just Random Graphs.

Rows of A can represent objects.

Columns of A are features.

Example : Document-Term Matrix : A_{ij} is the number of occurances of j th term in i th document.

Example : Consumer-Product Matrix : A_{ij} the preference the i th consumer has for the j th product.

Again, model is EA. Given A. Infer model.

Azar, Karlin, Fiat, McSherry, Saia.

Can we assume Full Independence ?

Consumers may be reasonably assumed to function independently of each other. But does one consumer choose products to buy independent;y of each other.

At least budget constraints ?

Documents in a collection may be independent random variables. But one cannot assume a particular document chooses independently whether to include/exclude each word.

: Partial Independence Rows of A are independent vector-valued random variables. Columns possibly correlated.

Joint with : A. Dasgupta, J. Hopcroft, and P. Mitra.

Can we carry out these algorithms under partial independence ??

Theorem Suppose A is an $m \times n$ matrix with independent rows (vector-valued random variables). Suppose
(i) the maximum variance of any row in any direction is at most σ^2 and
(ii) $|$ row $- E($row$)| \leq M$. Then, whp,

$$||A - EA|| \leq O((\log n)^c)(M + \sqrt{n}\sigma).$$

Remark Relaxing full indpendence to partial independence costs only polylog factors.

Proof based on Functional Analysis work of Rudelson, Lust-Picard, Milman,....

What are minimal Linear Algebra sufficient conditions (no randomness) under which spectral clustering works ?? Again, say points are the rows of A. There is a positive real σ (playing the role of std. dev. if the data were random) such that

(i) Spectral Norm bound There exists a matrix C (of cluster centers) with each row one of k distinct vectors such that

$$||A - C|| \leq c\sqrt{n}\sigma.$$

(ii) Separation Each pair of distinct centers (rows of C) are separated by at least $O(\text{poly}(k)\sigma)$.

(iii) Let $\hat{A}$ denote projection to the row space of C and $_{(i)}$ denote i th row. Assume (each point is nearer its own cluster center than other cluster centers in the projection) :

$$|\hat{A}_{(i)} - C_{(i)}| \leq |\hat{A}_{(i)} - C_{(j)}| - \text{poly}(k)\sigma \ , \ \forall i, \forall j : C_{(j)} \neq C_{(}$$

Conjecture : Under the assumptions, spectral clustering can be used to cluster perfectly.

Planted Clique Problem

Random Graph $G(V, E)$ with edge probabilities 1/2. ($|V| = n$).

Hidden Clique of size p. Find it.

If $p \geq c\sqrt{n \log n}$, then the clique vertices have highest degrees and this gives the clique away.

If $p \geq c\sqrt{n}$, can still find them using "Spectral Methods" (eigenvalues). (Alon, Krievelevich, Sudukoff).

Major open question : Can we find the hidden clique assuming only $p > \Omega(n^{1/3}), \Omega(n^{(1/2)-\epsilon})$??

Define an $n \times n \times n$ array A by :

$$A_{ijk} = \pm 1$$

if the number of edges of the graph among $\{i, j, k\}$ is odd or even respectively.

A has a solid block of $p \times p \times p$ 1's on $P \times P \times P$ (P is the hidden clique). But elsewhere, the entries of A should be random ± 1 and so in other blocks, the total will be close to zero because of cancellations.

So, if we can find this large block of "highly correlated" entries, we should be done.

Theorem (A. Frieze, K.) If A is an $n \times n \times n$ array constructed as above from a purely random graph (no hidden clique), Whp, for any unit length vector x,

$$|\sum_{ijk} A_{ijk} x_i x_j x_k| \leq c\sqrt{n}(\log n)^2.$$

[A generalization of Wigner-type Theorem to 3-dimensional arrays.]

The unit length vector u which puts $1/\sqrt{p}$ on each $i \in P$ gives for our A - constructed from random graph + hidden clique:

$$\sum_{ijk} A_{ijk} u_i u_j u_k = p^{3/2}.$$

If $p > cn^{1/3}(\log n)^2$,, then $p^{3/2} > c\sqrt{n}(\log n)^2$.

Can we find maxima of such cubic forms - $\sum_{ijk} A_{ijk} x_i x_j x_k$??