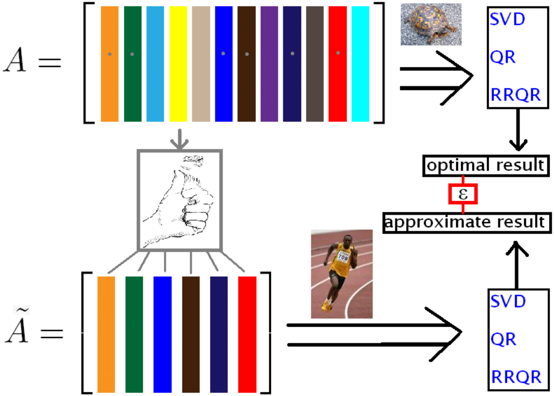
Approximate Spectral Clustering via Randomized Sketching

Christos Boutsidis
Yahoo! Labs, New York

Joint work with Alex Gittens (Ebay), Anju Kambadur (IBM)

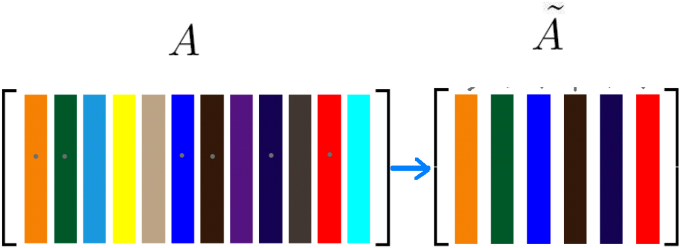
The big picture: “sketch” and solve



Tradeoff: Speed (depends on the size of Ã) with accuracy (quantified by the parameter $\varepsilon > 0$).

Sketching techniques (high level)

1. **Sampling:** $A \rightarrow \tilde{A}$ by picking a subset of the columns of A



2. **Linear sketching:** $A \rightarrow \tilde{A} = AR$ for some matrix R.
3. **Non-linear sketching:** $A \rightarrow \tilde{A}$ (no linear relationship).

Sketching techniques (low level)

1. **Sampling:**
   - Importance sampling: randomized sampling with probabilities proportional to the norms of the columns of A [Frieze, Kannan, Vempala, FOCS 1998], [Drineas, Kannan, Mahoney, SISC 2006].
   - Subspace sampling: randomized sampling with probabilities proportional to the norms of the rows of the matrix V_k containing the top k right singular vectors of A (leverage-scores sampling) [Drineas, Mahoney, Muthukrishnan, SODA 2006].
   - Deterministic sampling: Deterministically selecting rows from V_k - equivalently columns from A [Batson, Spielman, Srivastava, STOC 2009], [Boutsidis, Drineas, Magdon-Ismail, FOCS 2011].
2. **Linear sketching:**
   - Random Projections: Post-multiply A with a random gaussian matrix [Johnson,Lindenstrauss 1982].
   - Fast Random Projections: Post-multiply A with an FFT-type random matrix [Ailon, Chazelle 2006].
   - Sparse Random Projections: Post-multiply A with a sparse matrix [Clarkson, Woodruff STOC 2013].
3. **Non-linear sketching:**
   - Frequent Directions: SVD-type transform. [Liberty, KDD '13], [Ghashami, Phillips, SODA '14].
   - Other non-linear dimensionality reduction methods such as LLE , ISOMAP etc.

Problems

Linear Algebra:

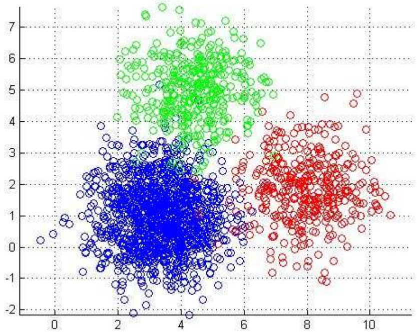
1. **Matrix Multiplication** [Drineas, Kannan, Rudelson, Virshynin, Woodruff, Ipsen, Liberty, and others]
2. **Low-rank Matrix Approx.** [Tygert, Tropp, Clarkson, Candes, B., Despande, Vempala, and others]
3. **Element-wise Sparsification** [Achlioptas, McSherry, Kale, Drineas, Zouzias, Liberty, Karnin, and others]
4. **Least-squares** [Mahoney, Muthukrishnan, Dasgupta, Kumar, Sarlos, Roklhin, Boutsidis, Avron, and others]
5. **Linear Equations with SDD matrices** [Spielman, Teng, Koutis, Miller, Peng, Orecchia, Kelner, and others]
6. **Determinant of SPSD matrices** [Barry, Pace, B., Zouzias and others]
7. **Trace of SPSD matrices** [Avron, Toledo, Bekas, Roosta-Khorasani, Uri Ascher, and others]

Machine Learning:

1. **Canonical Correlation Analysis** [Avron, B., Toledo, Zouzias]
2. **Kernel Learning** [Rahimi, Recht, Smola, Sindhwani and others]
3. ***k*-means Clustering** [B., Zouzias, Drineas, Magdon-Ismail, Mahoney, Feldman, and others]
4. **Spectral Clustering** [Gittens, Kambadur, Boutsidis, Strohmer and others]
5. **Spectral Graph Sparsification** [Batson, Spielman, Srivastava, Koutis, Miller, Peng, Kelner, and others]
6. **Support Vector Machines** [Paul, B., Drineas, Magdon-Ismail and others]
7. **Regularized least-squares classification** [Dasgupta, Drineas, Harb, Josifovski, Mahoney]

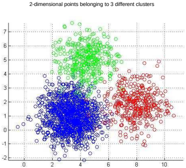
What approach should we use to cluster these data?

2-dimensional points belonging to 3 different clusters



Answer: *k*-means clustering

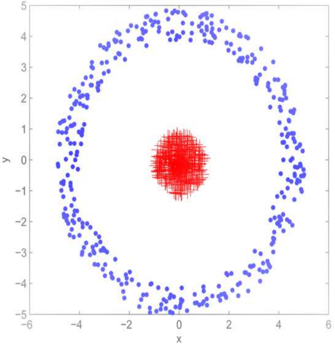
k-means optimizes the "right" metric over this space

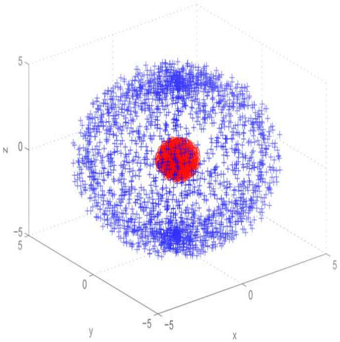


- $\mathcal{P} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \in \mathbb{R}^d$. number of clusters k.
- k-partition of $\mathcal{P}$: a collection $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_k\}$ of sets of points.
- For each set S_j, let $\mu_j \in \mathbb{R}^d$ be its centroid.
- k-means objective function: $\mathcal{F}(\mathcal{P}, \mathcal{S}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mu(\mathbf{x}_i)\|_2^2$
- Find the best partition:

$$\mathcal{S}_{opt} = \arg\min_{\mathcal{S}} \mathcal{F}(\mathcal{P}, \mathcal{S}).$$

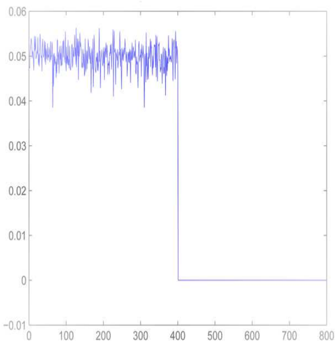
What approach should we use to cluster these data?





Answer: *k*-means will fail miserably. What else?

Spectral Clustering: Transform the data into a space where *k*-means would be useful



1-d representation of points from the first dataset in previous picture (this is an eigenvector from an appropriate graph).

Spectral Clustering: the graph theoretic perspective

- n points $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ in d-dimensional space.
- $G(V, E)$ is the corresponding graph with n nodes.
- Similarity matrix $\mathrm{W} \in \mathbb{R}^{n\times n}$ $\mathrm{W}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}}$ (for $i \neq j$); $\mathrm{W}_{ii} = 0$.
- Let k be the number of clusters.

Definition

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and $k = 2$ are given. Find subgraphs of G, denoted as A and B, to minimize:

$$\mathrm{Ncut}(A, B) = \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(A, V)} + \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(B, V)},$$

where $\mathrm{cut}(A, B) = \sum_{\mathbf{x}_i \in A, \mathbf{x}_j \in B} \mathrm{W}_{ij}$; and

$$\mathrm{assoc}(A, V) = \sum_{\mathbf{x}_i \in A, \mathbf{x}_j \in V} \mathrm{W}_{ij}; \mathrm{assoc}(B, V) = \sum_{\mathbf{x}_i \in B, \mathbf{x}_j \in V} \mathrm{W}_{ij}.$$

Spectral Clustering: the linear algebraic perspective

For *any* G, A, B and partition vector $\mathbf{y} \in \mathbb{R}^n$ with $+1$ to the entries corresponding to A and -1 to the entries corresponding to B it is:

$$4 \cdot \mathrm{Ncut}(A, B) = \mathbf{y}^T(\mathrm{D} - \mathrm{W})\mathbf{y}/(\mathbf{y}^T \mathrm{D}\mathbf{y}).$$

Here, $\mathrm{D} \in \mathbb{R}^{n \times n}$ is the diagonal matrix of degree nodes: $\mathrm{D}_{ii} = \sum_j \mathrm{W}_{ij}$.

Definition

Given graph G with n nodes, adjacency matrix W, and degrees matrix D find $\mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{y} = \underset{\mathbf{y} \in \mathbb{R}^n, \mathbf{y}^T \mathrm{D} \mathbf{1}_n}{\operatorname{argmin}} \frac{\mathbf{y}^T(\mathrm{D} - \mathrm{W})\mathbf{y}}{\mathbf{y}^T \mathrm{D}\mathbf{y}}.$$

Spectral Clustering: Algorithm for k-partitioning

Cluster n points $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ into k clusters

1. Construct the similarity matrix $\mathrm{W} \in \mathbb{R}^{n \times n}$ as $\mathrm{W}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}}$ (for $i \neq j$) and $\mathrm{W}_{ii} = 0$.
2. Construct $\mathrm{D} \in \mathbb{R}^{n \times n}$ as the diagonal matrix of degree nodes: $\mathrm{D}_{ii} = \sum_j \mathrm{W}_{ij}$.
3. Construct $\tilde{\mathrm{W}} = \mathrm{D}^{-\frac{1}{2}} \mathrm{W} \mathrm{D}^{-\frac{1}{2}} \in \mathbb{R}^{n \times n}$.
4. Find the largest k eigenvectors of $\tilde{\mathrm{W}}$ and assign them as columns to a matrix $\mathrm{Y} \in \mathbb{R}^{n \times k}$.
5. Apply k-means clustering on the rows of Y, and cluster the original points accordingly.

In a nutshell, compute the top k eigenvectors of $\tilde{\mathrm{W}}$ and then apply k-means on the rows of the matrix containing those eigenvectors.

Spectral Clustering via Randomized Sketching

Cluster n points $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ into k clusters

1. Construct the similarity matrix $\mathrm{W} \in \mathbb{R}^{n \times n}$ as $\mathrm{W}_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}}$ (for $i \neq j$) and $\mathrm{W}_{ii} = 0$.
2. Construct $\mathrm{D} \in \mathbb{R}^{n \times n}$ as the diagonal matrix of degree nodes: $\mathrm{D}_{ii} = \sum_j \mathrm{W}_{ij}$.
3. Construct $\tilde{\mathrm{W}} = \mathrm{D}^{-\frac{1}{2}} \mathrm{W} \mathrm{D}^{-\frac{1}{2}} \in \mathbb{R}^{n \times n}$.
4. Let $\tilde{\mathrm{Y}} \in \mathbb{R}^{n \times k}$ contain the left singular vectors of
$$\mathrm{B} = (\tilde{\mathrm{W}}\tilde{\mathrm{W}}^T)^p \tilde{\mathrm{W}} \mathrm{S},$$
with $p \geq 0$, and $\mathrm{S} \in \mathbb{R}^{n \times k}$ being a matrix with *i.i.d* random Gaussian variables.
5. Apply k-means clustering on the rows of $\tilde{\mathrm{Y}}$, and cluster the original data points accordingly.

In a nutshell, "approximate" the top k eigenvectors of $\tilde{\mathrm{W}}$ and then apply k-means on the rows of the matrix containing those eigenvectors.

Related work

The Nystrom method: Uniform random sampling of the similarity matrix W and then compute the eigenvectors. [Fowlkes et al. 2004]

The Spielman-Teng iterative algorithm: Very strong theoretical result based on their fast solvers for SDD systems of linear equations. Complex algorithm to implement. [2009]

Spectral clustering via random projections: Reduce the dimensions of the data points before forming the similarity matrix W. No theoretical results are reported for this method. [Sakai and Imiya, 2009].

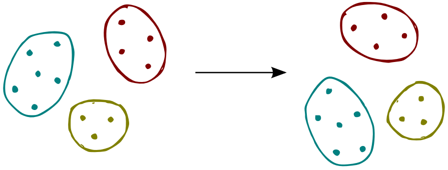
Power iteration clustering: Like our idea but for the $k = 2$ case. No theoretical results reported. [Lin, Cohen, ICML 2010]

Other approximation algorithms: [Yen et al. KDD 2009]; [Shamir and Tishby, AISTATS 2011]; [Wang et al. KDD 2009]

Approximation Framework for Spectral Clustering

- Assume that $\|\mathrm{Y} - \tilde{\mathrm{Y}}\|_2 \leq \varepsilon$.
- For all $i = 1:n$, let $\mathbf{y}_i^T, \tilde{\mathbf{y}}_i^T \in \mathbb{R}^{1\times k}$ be the ith rows of $\mathrm{Y}, \tilde{\mathrm{Y}}$.
- Then,
$$\|\mathbf{y}_i - \tilde{\mathbf{y}}_i\|_2 \leq \|\mathrm{Y} - \tilde{\mathrm{Y}}\|_2 \leq \varepsilon.$$
- Clustering the rows of Y and the rows of $\tilde{\mathrm{Y}}$ with the same method should result to the same clustering.
- A distance-based algorithm such as k-means would lead to the same clustering as $\varepsilon \to 0$.
- This is equivalent to saying that k-means is robust to small perturbations to the input.

Approximation Framework for Spectral Clustering



- The rows of $\tilde{Y}$ and $\tilde{Y}Q$, where Q is some square orthonormal matrix, are clustered identically.

Definition (Closeness of Approximation)

Y and $\tilde{Y}$ are close for "clustering purposes" if there exists a square orthonormal Q such that

$$\|Y - \tilde{Y}Q\|_2 \leq \varepsilon.$$

This is really a problem of bounding subspaces

Lemma

There is an orthonormal matrix $Q \in \mathbb{R}^{n \times k}$ *(*$Q^T Q = I_k$*) such that:*

$$\|Y - \tilde{Y}Q\|_2^2 \leq 2k\|YY^T - \tilde{Y}\tilde{Y}^T\|_2^2.$$

- $\|YY^T - \tilde{Y}\tilde{Y}^T\|_2^2$ corresponds to the cosine of the principal angle between *span*(Y) and *span*($\tilde{Y}$).
- Q is the solution of the following "Procrustes Problem":

$$\min_{Q} \|Y - \tilde{Y}Q\|_F$$

The Singular Value Decomposition (SVD)

- Let A be an $m \times n$ matrix with $rank(A) = \rho$ and $k \leq \rho$.

$$A = U_A \Sigma_A V_A^T = \underbrace{\begin{pmatrix} U_k & U_{\rho-k} \end{pmatrix}}_{m \times \rho} \underbrace{\begin{pmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{\rho-k} \end{pmatrix}}_{\rho \times \rho} \underbrace{\begin{pmatrix} V_k^T \\ V_{\rho-k}^T \end{pmatrix}}_{\rho \times n}.$$

- U_k: $m \times k$ matrix of the top-k left singular vectors of A.
- V_k: $n \times k$ matrix of the top-k right singular vectors of A.
- Σ_k: $k \times k$ diagonal matrix of the top-k singular values of A.

A "structural" result

Theorem

Given $A \in \mathbb{R}^{m \times n}$*, let* $S \in \mathbb{R}^{n \times k}$ *be such that*

$$\mathrm{rank}(A_k S) = k$$

and

$$\mathrm{rank}(V_k^T S) = k.$$

Let $p \geq 0$ *be an integer and let*

$$\gamma_p = \|\Sigma_{\rho-k}^{2p+1} V_{\rho-k}^T S (V_k^T S)^{-1} \Sigma_k^{-(2p+1)}\|_2.$$

Then, for $\Omega_1 = (AA^T)^p AS$*, and* $\Omega_2 = A_k$*, we obtain*

$$\|\Omega_1 \Omega_1^+ - \Omega_2 \Omega_2^+\|_2^2 = \frac{\gamma_p^2}{1 + \gamma_p^2}.$$

Some derivations lead to final result

$$
\begin{aligned}
\gamma_p &\leq \left\|\Sigma_{\rho-k}^{2p+1}\right\|_2 \left\|\mathrm{V}_{\rho-k}^T \mathrm{S}\right\|_2 \left\|(\mathrm{V}_k^T \mathrm{S})^{-1}\right\|_2 \left\|\Sigma_k^{-(2p+1)}\right\|_2 \\
&= \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{2p+1} \frac{\sigma_{\max}\left(\mathrm{V}_{\rho-k}^T \mathrm{S}\right)}{\sigma_{\min}\left(\mathrm{V}_k^T \mathrm{S}\right)} \\
&\leq \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{2p+1} \frac{\sigma_{\max}\left(\mathrm{V}^T \mathrm{S}\right)}{\sigma_{\min}\left(\mathrm{V}_k^T \mathrm{S}\right)} \\
&\leq \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{2p+1} \frac{4\sqrt{n-k}}{\delta/\sqrt{k}} \\
&= \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^{2p+1} \cdot 4\delta^{-1}\sqrt{k(n-k)}.
\end{aligned}
$$

Random Matrix Theory

Lemma (The norm of a random Gaussian Matrix)

Let $A \in \mathbb{R}^{n \times m}$ *be a matrix with i.i.d. standard Gaussian random variables, where* $n \geq m$*. Then, for every* $t \geq 4$,

$$\mathbb{P}\{\sigma_1(A) \geq tn^{\frac{1}{2}}\} \geq e^{-nt^2/8}.$$

Lemma (Invertibility of a random Gaussian Matrix)

Let $A \in \mathbb{R}^{n \times n}$ *be a matrix with i.i.d. standard Gaussian random variables. Then, for any* $\delta > 0$:

$$\mathbb{P}\{\sigma_n(A) \leq \delta n^{-\frac{1}{2}}\} \leq 2.35\delta.$$

Main Theorem

Theorem

If for some $\varepsilon > 0$ *and* $\delta > 0$ *we choose*

$$p \geq \frac{1}{2} \ln(4\varepsilon^{-1}\delta^{-1}\sqrt{k(n-k)}) \ln^{-1}\left(\frac{\sigma_k\left(\tilde{\mathrm{W}}\right)}{\sigma_{k+1}\left(\tilde{\mathrm{W}}\right)}\right),$$

then with probability at least $1 - e^{-n} - 2.35\delta$,

$$\|\mathrm{Y} - \tilde{\mathrm{Y}}\mathrm{Q}\|_2^2 \leq \frac{\varepsilon^2}{1+\varepsilon^2} = O(\varepsilon^2).$$