# On Spectral Custering: Analysis and an algorithm

Andrew Y. Ng, Michael I. Jordan, Yair Weiss

ENEE698A Recent Advance in Clustering

Presenter: Xavier Gibert

September 27, 2011

# Outline

# Background on Segmentation using Eigenvectors



Several segmentation algorithms rely on spectral spectral clustering methods

- Weiss (1999)
  *Segmentation eigenvectors: a unifying view*
- Meila and Shi (2001)
  *Learning segmentation by random walks*
- Perona and Feeman (1998)
  *A factorization approach to grouping*
- Shi and Malik (1997)
  *Normalized cuts and image segmentation*
- Scott and Longuet-Higgens (1990)
  *Feature grouping by relocalization of eigenvectors of the proximity matrix*

# Motivation



- Several algorithms based on *spectral clustering* use eigenvectors, but there is no agreement on what is the optimal way to use them, Weiss (1999).
- Empirical successes, but no proof of optimality
- *Spectral graph theory* shows that the second eigenvector of a graph's *Laplacian* can be used to define a semi-optimal cut.
- Several algorithms implement *k-way* partitioning by recursively using *2-way* partitions. It is desirable to use a *k-way* partition directly.
- This approach is an improvement over Meila and Shi (2001).

# Algorithm set up



- Input: Set of $n$ points $S = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^l$, $1 \leq i \leq n$

# Algorithm set up



- Input: Set of $n$ points $S = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^l$, $1 \leq i \leq n$
- Parameters:
  - $k$: number of clusters
  - $\sigma$: scaling parameter (can be detected automatically)

# Algorithm set up

- Input: Set of $n$ points $S = \{s_1, \ldots, s_n\}$, $s_i \in \mathbb{R}^l$, $1 \leq i \leq n$
- Parameters:
  - $k$: number of clusters
  - $\sigma$: scaling parameter (can be detected automatically)
- Output: Cluster labels for each of the $n$ points.
  $L = \{l_1, \ldots, l_n\}$, $l_i \in \{1, \ldots, k\}$, $1 \leq i \leq n$

# Algorithm



Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

# Algorithm





Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

- **Step 1:** Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = exp(-||s_i - s_j||^2/2\sigma^2)$ if $i \neq j$ and $A_{ii} = 0$.

$$A = A^T = \begin{pmatrix} 0 & A_{12} & \ldots & A_{1n} \\ A_{12} & 0 & \ldots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \ldots & 0 \end{pmatrix}$$

# Algorithm



Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

- **Step 1:** Form the affinity matrix $A \in \mathbb{R}^{n \times n}$
- **Step 2:** Define $D$ to be the diagonal matrix whose $(i, i)$-element is the sum of $A$'s $i$-th row, and the matrix $L = D^{-1/2} A D^{-1/2} \in \mathbb{R}^{n \times n}$ (Laplacian).

$$D = \operatorname{diag}\left( \sum_{j=1}^{n} A_{1j}, \ldots, \sum_{j=1}^{n} A_{nj} \right)$$

$$L = \begin{pmatrix} D_{11}^{-\frac{1}{2}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & D_{nn}^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} 0 & \ldots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{1n} & \ldots & 0 \end{pmatrix} \begin{pmatrix} D_{11}^{-\frac{1}{2}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & D_{nn}^{-\frac{1}{2}} \end{pmatrix}$$

# Algorithm



Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

- **Step 1:** Form the affinity matrix $A \in \mathbb{R}^{n \times n}$
- **Step 2:** Calculate the Laplacian matrix $L \in \mathbb{R}^{n \times n}$
- **Step 3:** Find $x_1, x_2, \ldots, x_k$ the $k$ largest eigenvectors of $L$, and then form the matrix $X = [x_1, x_2, \ldots, x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.

**Note:** Since $L$ is real and symmetric its eigenvectors are orthonormal:

$$L = Q \Lambda Q^T$$

$$L \approx X \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) X^T$$

# Algorithm



Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

- **Step 1:** Form the affinity matrix $A \in \mathbb{R}^{n \times n}$
- **Step 2:** Calculate the Laplacian matrix $L \in \mathbb{R}^{n \times n}$
- **Step 3:** Find $k$ largest eigenvectors of $L$, $X \in \mathbb{R}^{n \times k}$
- **Step 4:** Form the matrix $Y \in \mathbb{R}^{n \times k}$ from $X$ by renormalizing each of $X$'s rows to have unit length

$$Y_{ij} = \frac{X_{ij}}{\left(\sum_{j=1}^{k} X_{ij}^2\right)^{1/2}}$$

# Algorithm



Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

- **Step 1:** Form the affinity matrix $A \in \mathbb{R}^{n \times n}$
- **Step 2:** Calculate the Laplacian matrix $L \in \mathbb{R}^{n \times n}$
- **Step 3:** Find $k$ largest eigenvectors of $L$, $X \in \mathbb{R}^{n \times k}$
- **Step 4:** Form the matrix $Y \in \mathbb{R}^{n \times k}$ from $X$ by renormalizing each of $X$'s rows to have unit length
- **Step 5:** Treating each row of $Y$ as a point in $\mathbb{R}^k$, cluster them into $k$ clusters via K-means (or any other algorithm that attempts to minimize distortion)

# Algorithm



Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster in $k$ subsets:

- **Step 1:** Form the affinity matrix $A \in \mathbb{R}^{n \times n}$
- **Step 2:** Calculate the Laplacian matrix $L \in \mathbb{R}^{n \times n}$
- **Step 3:** Find $k$ largest eigenvectors of $L$, $X \in \mathbb{R}^{n \times k}$
- **Step 4:** Form the matrix $Y \in \mathbb{R}^{n \times k}$ from $X$ by renormalizing each of $X$'s rows to have unit length
- **Step 5:** Treating each row of $Y$ as a point in $\mathbb{R}^k$, cluster them into $k$ clusters via K-means (or any other algorithm that attempts to minimize distortion)
- **Step 6:** Finally, assign the original point $s_i$ to cluster $j$ if and only if row $i$ of the matrix $Y$ was assigned to cluster $j$.

# Analysis - Notation

- $n_i$, the number of samples in cluster $i$. Therefore, $n = \sum_{i=1}^{k} n_i$.
- $S_i = \{s \in S : l_i = i\}$., the samples in cluster $i$. We assume $S_i \neq \emptyset \; \forall i$, and $S = \bigcup_{i=1}^{j} S_i$.
- $A^{(ii)} \in \mathbb{R}^{n_i \times n_i}$, the matrix of "intra-cluster" affinities for cluster $i$.
- $d^{(i)} \in \mathbb{R}^{n_i}$, the vector containing $D^{(ii)}$'s diagonal elements.
- $d \in \mathbb{R}^{n}$, the vector containing $D$'s diagonal elements.

# Analysis - The "ideal" case

First, consider the "ideal" case in which points in different clusters are *infinitely far apart*.

- Let $k = 3$
- Assume $S = \{s_1, \ldots, s_n\} = \{\{S_1\}, \ldots, \{S_k\}\}$ (points are ordered based on cluster labels).
- Define $\widehat{A}$ with $\widehat{A}_{ij} = 0$ if $s_i$ and $s_j$ belong to different clusters, $\widehat{A}_{ij} = A_{ij}$ otherwise.
- Define $\widehat{L}$, $\widehat{D}$, $\widehat{X}$, and $\widehat{Y}$, be defined as in the previous algorithm.

# Analysis - The "ideal" case



First, consider the "ideal" case in which points in different clusters are *infinitely far apart*.

- Note that $\widehat{A}$ and $\widehat{L}$ are block-diagonal:

$$\widehat{A} = \begin{bmatrix} A^{(11)} & 0 & 0 \\ 0 & A^{(22)} & 0 \\ 0 & 0 & A^{(33)} \end{bmatrix}$$

$$\widehat{L} = \begin{bmatrix} \widehat{L}^{(11)} & 0 & 0 \\ 0 & \widehat{L}^{(22)} & 0 \\ 0 & 0 & \widehat{L}^{(33)} \end{bmatrix} \tag{1}$$

where $\widehat{L}^{(ii)} = \left(\widehat{D}^{(ii)}\right)^{-1/2} A^{(ii)} \left(\widehat{D}^{(ii)}\right)^{-1/2}$.

# Analysis - The "ideal" case



First, consider the "ideal" case in which points in different clusters are *infinitely far apart*.

- Since $\widehat{L}$ is block-diagonal its eigenvalues and eigenvectors are the union of those of its blocks.

# Analysis - The "ideal" case

First, consider the "ideal" case in which points in different clusters are *infinitely far apart*.

- Since $\widehat{L}$ is block-diagonal its eigenvalues and eigenvectors are the union of those of its blocks.
- Each $\widehat{L}$ has a strictly positive principal eigenvector $x_1^{(i)} \in \mathbb{R}^{n_i}$ with $\lambda_1^{(i)} = 1$.

# Analysis - The "ideal" case





First, consider the "ideal" case in which points in different clusters are *infinitely far apart*.

- Since $\widehat{L}$ is block-diagonal its eigenvalues and eigenvectors are the union of those of its blocks.
- Each $\widehat{L}$ has a strictly positive principal eigenvector $x_1^{(i)} \in \mathbb{R}^{n_i}$ with $\lambda_1^{(i)} = 1$.
- Also, since $A_{jk}^{(ii)} > 0$ $(j \neq k)$, the next eigenvalue $\lambda_2^{(i)} < 1$

# Analysis - The "ideal" case



First, consider the "ideal" case in which points in different clusters are *infinitely far apart*.

- Since $\widehat{L}$ is block-diagonal its eigenvalues and eigenvectors are the union of those of its blocks.
- Each $\widehat{L}$ has a strictly positive principal eigenvector $x_1^{(i)} \in \mathbb{R}^{n_i}$ with $\lambda_1^{(i)} = 1$.
- Also, since $A_{jk}^{(ii)} > 0$ $(j \neq k)$, the next eigenvalue $\lambda_2^{(i)} < 1$
- Thus, stacking $\widehat{L}$'s eigenvectors in columns we have

$$\widehat{X} = \begin{bmatrix} x_1^{(1)} & \vec{0} & \vec{0} \\ \vec{0} & x_1^{(2)} & \vec{0} \\ \vec{0} & \vec{0} & x_1^{(3)} \end{bmatrix} \in \mathbb{R}^3 \quad (2)$$

# Analysis - The "ideal" case





- After normalizing $\widehat{X}$ we obtain

$$\widehat{Y} = \begin{bmatrix} \widehat{Y}^{(1)} \\ \widehat{Y}^{(2)} \\ \widehat{Y}^{(3)} \end{bmatrix} = \begin{bmatrix} \vec{1} & \vec{0} & \vec{0} \\ \vec{0} & \vec{1} & \vec{0} \\ \vec{0} & \vec{0} & \vec{1} \end{bmatrix} R \tag{3}$$

$$R \in \mathbb{R}^{3\times 3}$$

$$R^T R = RR^T = I$$

# Analysis - Proposition



**Proposition 1** *Let $\widehat{A}$'s off-diagonal blocks $\widehat{A}^{(ij)}, i \neq j$, be zero. Also assume that each cluster $S_i$ is connected. Then there exist $k$ orthonormal vectors $r_1, \ldots, r_k$, $r_i^T r_j = \delta[i-j]$ so that $\widehat{Y}$'s rows satisfy*

$$\hat{y}_j^{(i)} = r_i \tag{4}$$

*for all $i = 1, \ldots, k$, $j = 1, \ldots, n_i$.*

# Analysis - The general case



- In the general case we consider $A = \widehat{A} + E$ to be a perturbed version of $\widehat{A}$.

# Analysis - The general case

- In the general case we consider $A = \widehat{A} + E$ to be a perturbed version of $\widehat{A}$.
- When can we expect the resulting rows of $Y$ to cluster similarly to the rows of $\widehat{Y}$?

## Analysis - The general case





- In the general case we consider $A = \widehat{A} + E$ to be a perturbed version of $\widehat{A}$.
- When can we expect the resulting rows of $Y$ to cluster similarly to the rows of $\widehat{Y}$?
- Under what conditions will the eigenvectors of $L$ be "close" to those of $\widehat{L}$?

# Analysis - The general case



- In the general case we consider $A = \widehat{A} + E$ to be a perturbed version of $\widehat{A}$.
- When can we expect the resulting rows of $Y$ to cluster similarly to the rows of $\widehat{Y}$?
- Under what conditions will the eigenvectors of $L$ be "close" to those of $\widehat{L}$?
- Matrix perturbation theory indicates that stability of eigenvectors are determined by the *eigengap*.

$$\max_i \lambda_2^{(i)} << \min_i \lambda_1^{(i)} \approx 1$$

# Analysis - General Assumptions



**Assumption A1.** There exists $\delta > 0$ so that, for all $i = 1, \ldots, k$, $\lambda_2^{(i)} \leq 1 - \delta$

# Analysis - General Assumptions



**Assumption A1.** There exists $\delta > 0$ so that, for all $i = 1, \dots, k$, $\lambda_2^{(i)} \leq 1 - \delta$

**Assumption A1.1.** There exists $\delta > 0$ so that $(h(S_i))^2/2 \geq \delta$ where the *Cheeger constant* of cluster is defined as

$$h(S_i) = \min_{\mathcal{I}} \frac{\sum_{j \in \mathcal{I}, k \notin \mathcal{I}} A_{jk}^{(ii)}}{\min\{\sum_{j \in \mathcal{I}} \hat{d}_j^{(i)}, \sum_{k \notin \mathcal{I}} \hat{d}_k^{(i)}\}} \tag{5}$$

where the outer minimum is over all index subsets $\mathcal{I} \subseteq \{1, \dots, n_i\}$.

# Analysis - General Assumptions

**Assumption A1.** There exists $\delta > 0$ so that, for all $i = 1, \ldots, k$, $\lambda_2^{(i)} \leq 1 - \delta$

**Assumption A1.1.** There exists $\delta > 0$ so that $(h(S_i))^2/2 \geq \delta$ where the *Cheeger constant* of cluster is defined as

$$h(S_i) = \min_{\mathcal{I}} \frac{\sum_{j \in \mathcal{I}, k \notin \mathcal{I}} A_{jk}^{(ii)}}{\min\{\sum_{j \in \mathcal{I}} \hat{d}_j^{(i)}, \sum_{k \notin \mathcal{I}} \hat{d}_k^{(i)}\}} \tag{5}$$

where the outer minimum is over all index subsets $\mathcal{I} \subseteq \{1, \ldots, n_i\}$.

**Assumption A1.1** $\Longrightarrow$ **Assumption A1**

# Analysis - General Assumptions



**Assumption A2.** There is some fixed $\epsilon_1 > 0$, so that for every $i_1, i_2 \in \{1, \ldots, k\}$, $i_1 \neq i_2$, we have that

$$\sum_{j \in S_{i_1}} \sum_{j \in S_{i_2}} \frac{A_{jk}^2}{\hat{d}_j \hat{d}_k} \leq \epsilon_1 \tag{6}$$

# Analysis - General Assumptions



**Assumption A2.** There is some fixed $\epsilon_1 > 0$, so that for every $i_1, i_2 \in \{1, \ldots, k\}$, $i_1 \neq i_2$, we have that

$$\sum_{j \in S_{i_1}} \sum_{j \in S_{i_2}} \frac{A_{jk}^2}{\hat{d}_j \hat{d}_k} \leq \epsilon_1 \tag{6}$$

**Assumption A3.** For some fixed $\epsilon_2 > 0$, for every $i = 1, \ldots, j$, $j \in S_i$, we have

$$\frac{\sum_{k : k \notin S_i} A_{jk}}{\hat{d}_j} \leq \epsilon_2 \left( \sum_{k,l \in S_i} \frac{A_{kl}^2}{\hat{d}_k \hat{d}_l} \right)^{-1/2} \tag{7}$$

# Analysis - General Assumptions

**Assumption A4.** There is some constant $C > 0$ so that for every $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, we have

$$\hat{d}_j^{(i)} \geq \frac{\sum_{k=1}^{n_i} \hat{d}_k^{(i)}}{C n_i}$$

# Analysis - General Assumptions Summary



**Assumption A1.** Clusters are tight

**Assumption A1.1.** Clusters are compact

**Assumption A2.** There is low connectivity across clusters

**Assumption A3.** Connectivity across clusters relative to within clusters is small

**Assumption A4.** No points in a cluster are "too much less" connected than other points in the same cluster

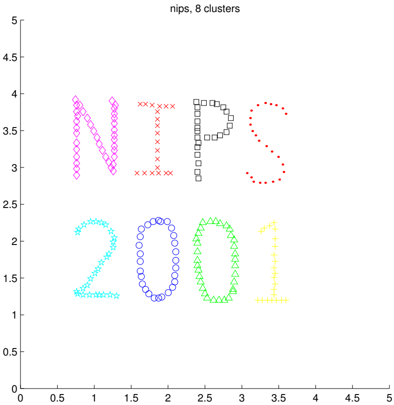
# Analysis - Theorem

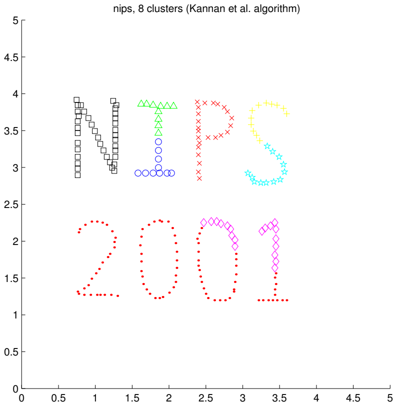


**Theorem 2** *Let assumptions A1, A2, A3, and A4 hold. Set* $\epsilon = \sqrt{k(k-1)\epsilon_1 + k\epsilon_2^2}$. *If* $\delta > (2+\sqrt{2})\epsilon$, *then there exist* $k$ *orthogonal vectors* $r_1, \ldots, r_k$ ($r_i^T r_j = 1$ *if* $i = j$, $0$ *otherwise) so that* $Y$*'s rows satisfy*

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \|y_j^{(i)} - r_i\|_2^2 \leq 4C\left(4 + 2\sqrt{k}\right)^2 \frac{\epsilon^2}{(\delta - \sqrt{2}\epsilon)} \qquad (8)$$

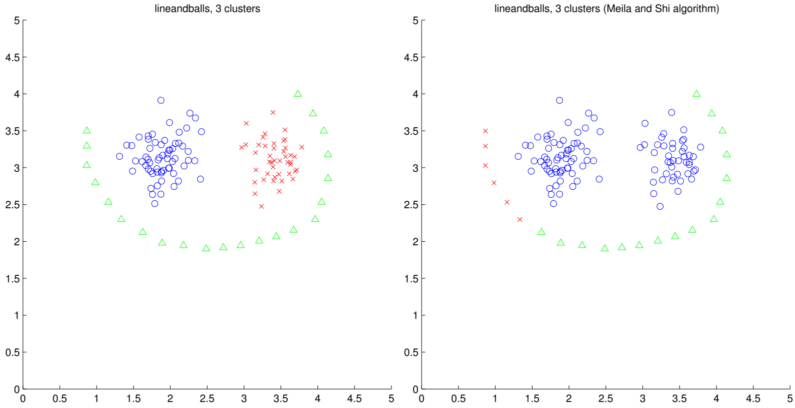
# Experimental Results I



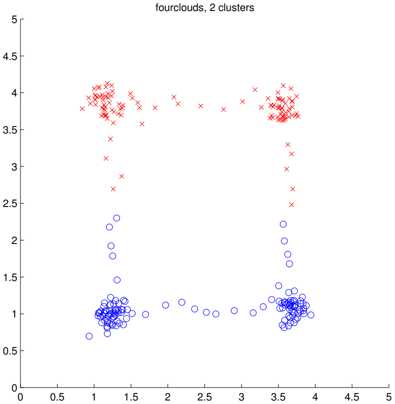
# Experimental Results II

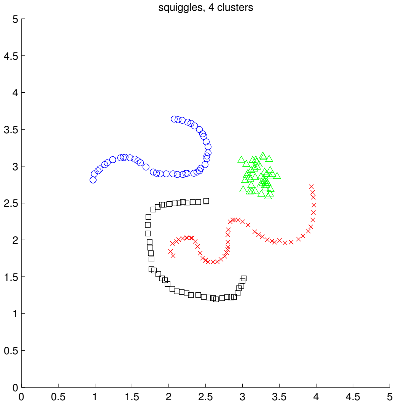




# Experimental Results III

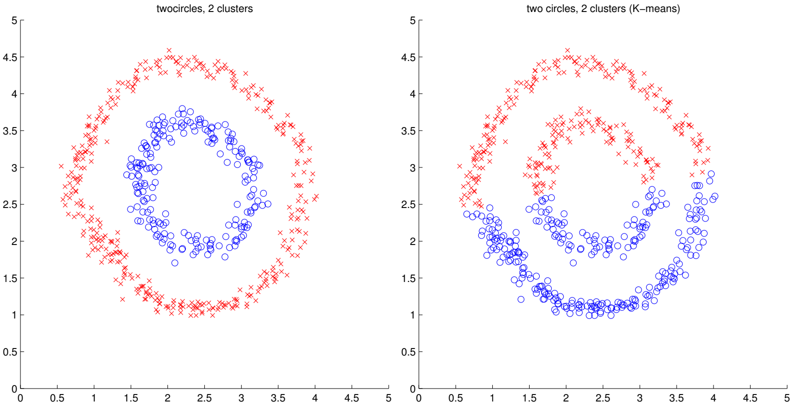




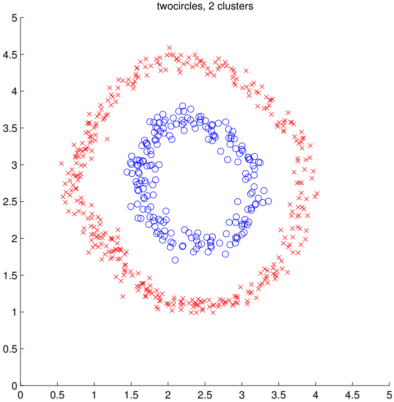
# Experimental Results IV

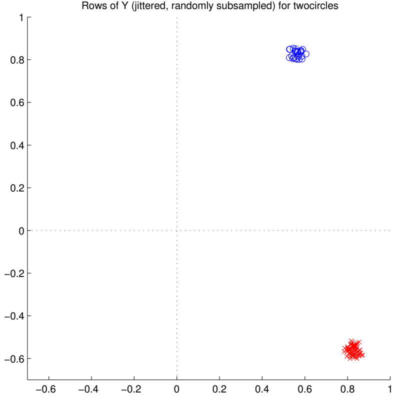




# Experimental Results IVb

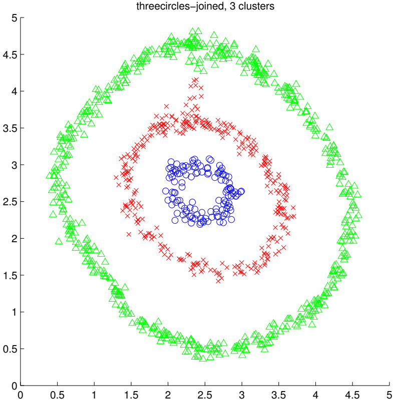


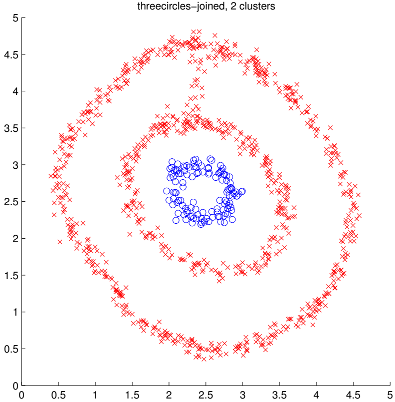


# Experimental Results V

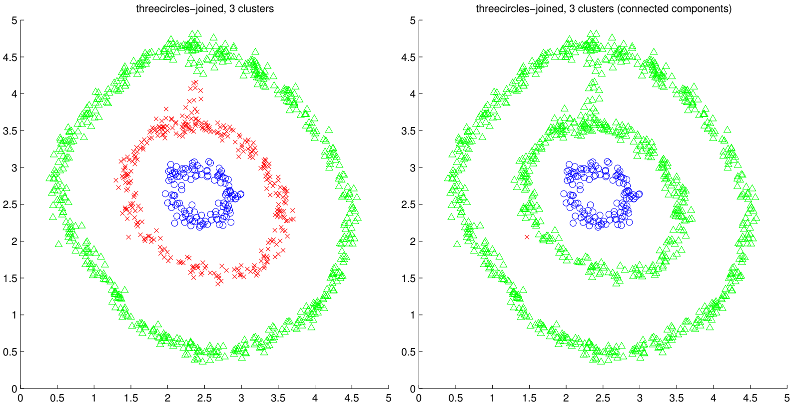




# Experimental Results Vb





# Discussion I



- A simple clustering algorithm that solves non-trivial clustering problems has been presented.
- The algorithm has been derived using *spectral graph theory*.
- An analysis has been done in two steps:
  - **Proposition 1** shows that in the "ideal" case the algorithm always provides the correct solution.
  - **Theorem 2** shows that it works when it does not deviate too much from the "ideal" case.
- The authors observe that this algorithm is similar to *Kernel PCA* with a Gaussian Kernel.

# Discussion II



- In real-world problems, assumptions A1-A4 may not hold with reasonably large $\delta$ and small $\epsilon_1$, $\epsilon_2$, and $C$.
- Knowledge of $k$ is required. Incorrect choice of $k$ causes assumptions A1-A4 to fail.

# Potential Improvements





Manifold clustering

- All points in matrix $Y$ lie in a unit hypersphere in $\mathbb{R}^k$, a $(k-1)$-dimensional Riemmanian manifold $\mathcal{M}$.
- Running K-means in $\mathbb{R}^k$ results in cluster centers off the manifold $\mathcal{M}$.
- Can we improve clustering in $Y$ by taking advantage of its manifold structure?
  - Karcher mean
  - Geodesic distance

# Other questions



- How does this algorithm handle outliers?
- How does it handle combinations of dense and sparse clusters?
- How does performance degrade when some assumptions are not met?
- What is the relationship between the upper bound in **Theorem 2** and the probability of convergence?

# References I


A. Ngm M. Jordan, and Y. Weiss.
On Spectral Clustering: Analysis and an algorithm.
*Neural Information Processing Systems, 2001.*

F. Chung.
*Spectral Graph Theory.*
Number 92 in CBMS Regional Conference Series in Mathematics.
American Mathematical Society, 1997.

G. W. Stewart and J.-G. Sun
*Matrix Perturbation Theory.*
Academic Press, 1990.

# References II



Y. Weiss.
Segmentation using eigenvectors: A unifying view.
*International Conference on Computer Vision, 1999.*

M. Meila and J. Shi 2001
Learning segmentation by random walks.
*Neural Information Processing Systems 13, 2001.*