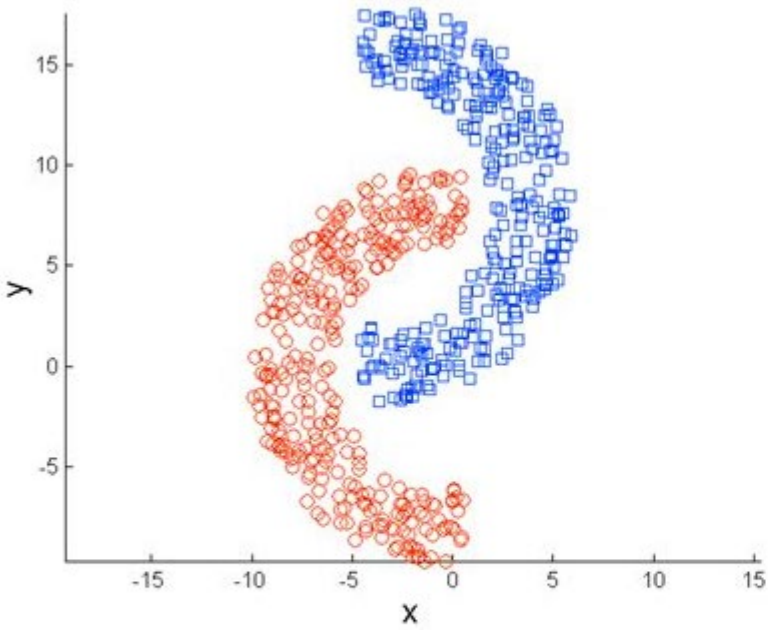
Spectral Clustering

Announcements

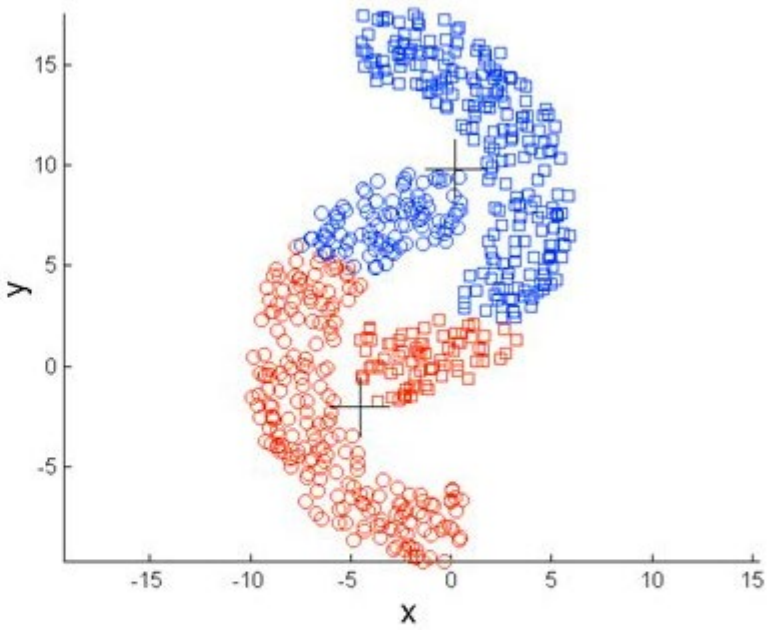
- Exam 2 next Thursday
  - Same format as Exam 1
  - Emphasis on material since Exam 1, plus kernel methods which were not covered on Exam 1
  - 1 cheat sheet allowed as for Exam 1
  - Practice problems will be posted, along with solutions
- New version of HW8 posted.
- Great Lakes accounts available
- Proposal almost graded

Motivation

- k-means and GMMs produce convex clusters
- Spectral clustering is an alternative approach that can find nonconvex clusters



Original Points



K-means (2 Clusters)

Outline

- Similarity graphs
- Graph Laplacian
- Graph cuts
- Relaxation of an optimization problem
- Spectral clustering

Similarity Graphs

- Unlabeled data to be clustered: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$
- An *affinity matrix* is a matrix

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$$

where the weights are nonnegative, symmetric

$\forall i,j \quad w_{ij} \geq 0 \qquad w_{ij} = w_{ji}$

and w_{ij} captures the similarity between x_i and x_j.

- Associated graph: Nodes i and j are connected iff $w_{ij} > 0$.
- A *similarity graph* is a graph associated to an affinity matrix, with edge weights given by the affinity matrix

Similarity Graphs

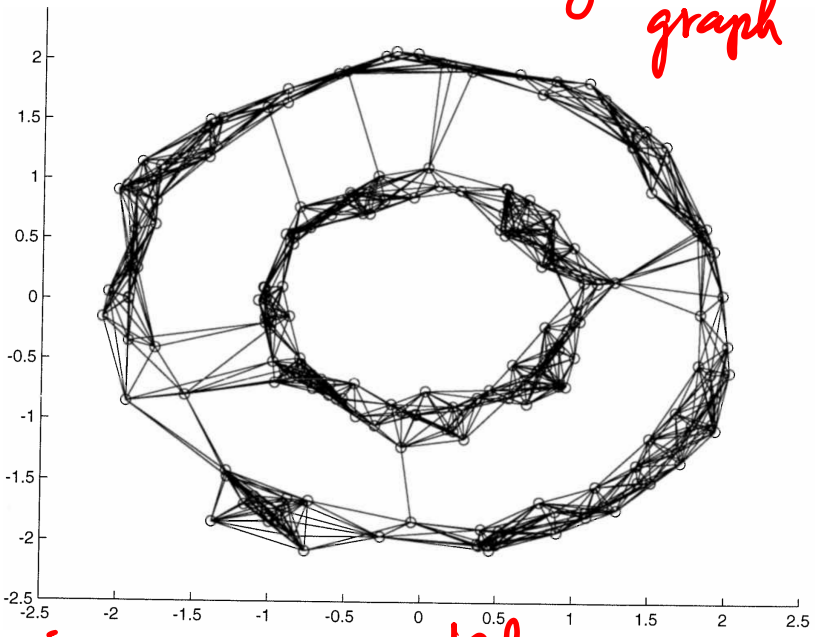
- Similarity graphs are defined by two things.
- **Graph structure:**

11 nearest neighbor graph

k-NN graph

ε-ball graph

complete



- **Weights:**

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{ow} \end{cases}$$

$$w_{ij} = \begin{cases} \exp\left(-\gamma \|x_i - x_j\|^2\right) & \text{if } i \text{ and } j \text{ connected} \\ 0 & \text{o.w.} \end{cases}$$

Graph Laplacian

- The (weighted) *degree* of a node $\boldsymbol{x}_i$ is

$$d_i = \sum_j w_{ij}$$

- The *degree matrix* is the diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix}$$

- The *unnormalized graph Laplacian* is $L = D - W$
- Note that $\boldsymbol{L}$ is independent of the self-similarity weights w_{ii} because

$$L_{ii} = d_i - w_{ii} = \sum_{j \neq i} w_{ij}$$

Poll

True or False: The graph Laplacian L is symmetric

(A) True

(B) False

$L = D - W$

Properties of Graph Laplacian

1. For every $\boldsymbol{f} \in \mathbb{R}^n$

$$\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2, \quad \boldsymbol{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

2. $\boldsymbol{L}$ is symmetric + PSD.

3. The smallest eigenvalue of $\boldsymbol{L}$ is 0, and

$$L\boldsymbol{v} = 0\boldsymbol{v} = \boldsymbol{0}$$

$$\mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

is a corresponding eigenvector.

4. The number of connected components of the similarity graph is equal to the dimension of the 0-eigenspace (i.e., the subspace of eigenvectors for the eigenvalue zero, which is also the nullspace of $\boldsymbol{L}$)

Proofs

1.

$$\boldsymbol{f}^T L \boldsymbol{f} = \boldsymbol{f}^T D \boldsymbol{f} - \boldsymbol{f}^T W \boldsymbol{f} = \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j} w_{ij} f_i f_j$$

$$= \frac{1}{2} \left(\sum_{i=1}^{n} d_i f_i^2 - 2 \sum_{i,j} w_{ij} f_i f_j + \sum_{j=1}^{n} d_j f_j^2 \right)$$

$$= \frac{1}{2} \sum_{i,j} w_{ij} (f_i^2 - 2 f_i f_j + f_j^2) = \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2$$

2. Follows from the first property and symmetry of W.

3. $L\mathbb{1} = D\mathbb{1} - W\mathbb{1} = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} - \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = 0\mathbb{1}$

Normalized Graph Laplacian

- The *normalized graph Laplacian* is

$$\tilde{L} = D^{-1}L = I - D^{-1}W$$

- $\tilde{\boldsymbol{L}}$ has n real eigenvalues
- The smallest eigenvalue of $\tilde{\boldsymbol{L}}$ is 0, and

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

  is a corresponding eigenvector.

- The number of connected components of the similarity graph is equal to the dimension of the 0-eigenspace (i.e., the subspace of eigenvectors for the eigenvalue zero, which is also the nullspace of $\tilde{\boldsymbol{L}}$)

Spectral Clustering

- Input: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, desired number of clusters K, parameters of similarity graph
- Construct a similarity graph and form the graph Laplacian $\boldsymbol{L}$ (or $\tilde{\boldsymbol{L}}$)
- Determine K smallest eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K$ of $\boldsymbol{L}$ (or $\tilde{\boldsymbol{L}}$) and corresponding eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_K \in \mathbb{R}^n$
- Set $\boldsymbol{y}_i = [u_{1i}, u_{2i}, \ldots, u_{Ki}]^T$, $i \in \{1, 2, \ldots, n\}$
- Cluster $\{\boldsymbol{y}_i\}_{i=1}^n$ using K-means clustering, and assign $\{\boldsymbol{x}_i\}_{i=1}^n$ to corresponding clusters

$L = U \Lambda U^T$

$U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix}$

$U_{:,1:K} = \begin{bmatrix} u_{11} & u_{21} & \cdots & u_{K1} \\ u_{12} & u_{22} & \cdots & u_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n} & u_{2n} & \cdots & u_{Kn} \end{bmatrix} \begin{matrix} \rightarrow y_1 \\ \rightarrow y_2 \\ \vdots \\ \rightarrow y_n \end{matrix}$

Image Segmentation

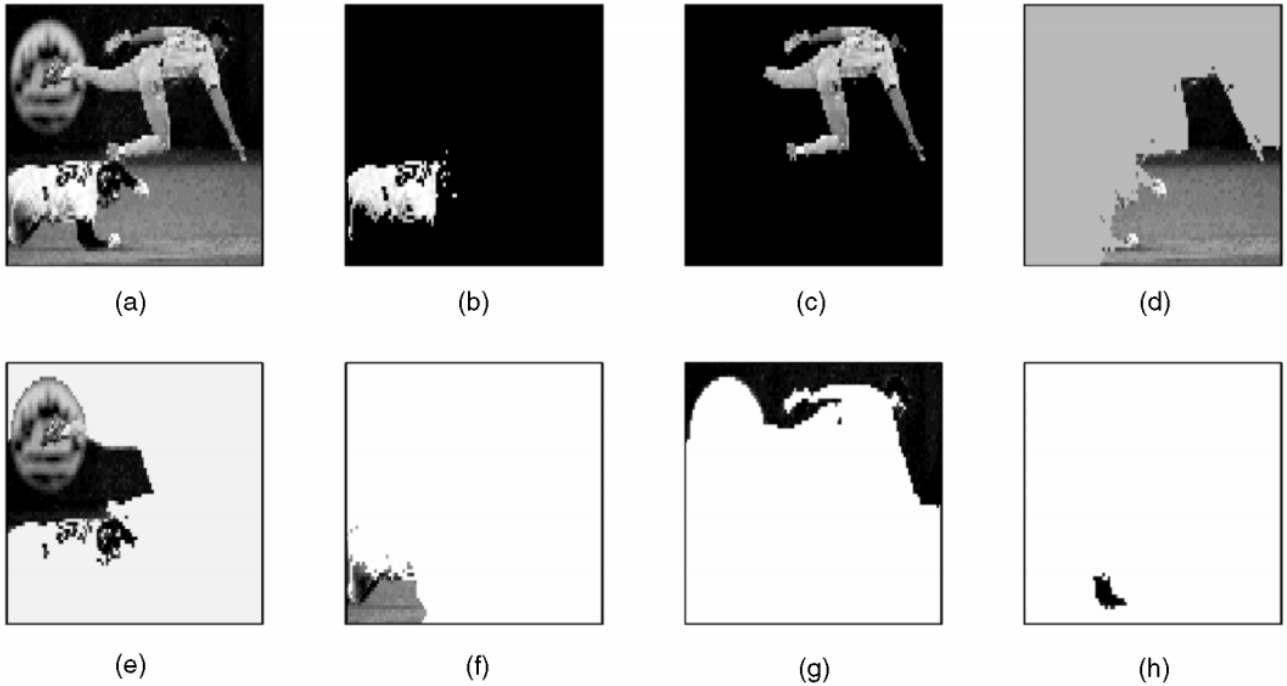


Fig. 4. (a) shows the original image of size 80×100. Image intensity is normalized to lie within 0 and 1. Subplots (b)-(h) show the components of the partition with $Ncut$ value less than 0.04. Parameter setting: $\sigma_I = 0.1$, $\sigma_X = 4.0$, $r = 5$.

$$w_{ij} = e^{\frac{-\|\boldsymbol{F}(i)-\boldsymbol{F}(j)\|_2^2}{\sigma_I}} * \begin{cases} e^{\frac{-\|\boldsymbol{X}(i)-\boldsymbol{X}(j)\|_2^2}{\sigma_X}} & \text{if } \|\boldsymbol{X}(i)-\boldsymbol{X}(j)\|_2 < r \\ 0 & \text{otherwise,} \end{cases}$$

From Shi and Malik (2000)

Big Picture

- Spectral clustering determines clusters from the eigenvalue (i.e., spectral) decomposition of L or $\tilde{L}$.
- There are many ways to derive spectral clustering.
- We will focus on one particular way based on graph cuts.

Graph Cuts

- A graph cut is a partition of a graph
- Assume there are two clusters
- Idea: Determine clusters by finding a good graph cut
- In particular: Find a cut such that
  - w_{ij} is large if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are in the same cluster
  - w_{ij} is small if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are in different clusters
- One way to quantify this:

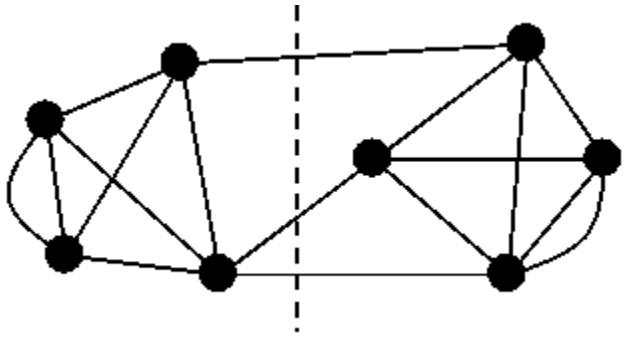
$\bar{A}$ = complement of A

$A \subseteq \{1, \ldots, n\}$

$$\min_A \quad C(A, \overline{A})$$

where

$$C(A,B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$$



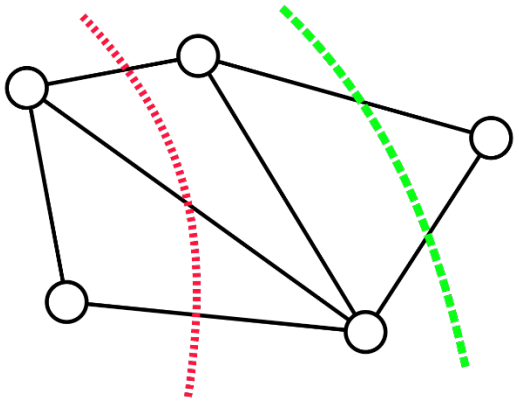
Mincut Problem

- Now assume K clusters
- Given a similarity graph/affinity matrix, we would like to find a partition $A_1, \ldots, A_K$ of $\{1, 2, \ldots, n\}$ such that:
  - w_{ij} is large if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are in the same cluster
  - w_{ij} is small if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are in different clusters
- The *mincut* problem aims to minimize

$$\text{cut}(A_1, \ldots, A_K) = \frac{1}{2} \sum_{k=1}^{K} C(A_k, \bar{A}_k)$$

with respect to the partition $A_1, \ldots, A_K$.

$$C(A, B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$$



Other Graph Cut Problems

- Mincut unfortunately leads to small and often singleton clusters. Therefore some modifications have been proposed:

- *RatioCut* (Hagen and Kahng, 1992)

$$\text{RatioCut}(A_1, \ldots, A_K) = \frac{1}{2} \sum_{k=1}^{K} \frac{C(A_k, \bar{A}_k)}{|A_k|}$$

where $|A| = \#$ of nodes in A.

- *Normalized Cut (Ncut)* (Shi and Malik, 2000)

$$\text{Ncut}(A_1, \ldots, A_K) = \frac{1}{2} \sum_{k=1}^{K} \frac{C(A_k, \bar{A}_k)}{\text{vol}(A_k)}$$

where

$$\text{vol}(A) = \sum_{i \in A} \sum_{j \in V} w_{ij}$$

and V is the set of all nodes (vertices) in the graph.

Relaxations

- Unfortunately, introducing these "balancing" terms causes these problems to be NP-hard.
- Therefore we will consider *relaxed* versions of these optimization problems
- A *relaxation* of a constrained optimization problem is obtained by enlarging the feasible set so as to make the problem (more) tractable.
- We will see that relaxations of RatioCut and NCut can be solved in terms of spectral clustering with $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$, respectively.

Relaxation of RatioCut

$n=5$, $A=\{3,4\}$,

$$f_A = \begin{bmatrix} -\sqrt{2/3} \\ -\sqrt{2/3} \\ \sqrt{3/2} \\ \sqrt{3/2} \\ -\sqrt{2/3} \end{bmatrix}$$

- Assume $K = 2$ (the argument generalizes but this case is simpler)
- We wish to minimize

$$\text{RatioCut}(A, \overline{A}) = \frac{1}{2}\left[\frac{C(A,\overline{A})}{|A|} + \frac{C(A,\overline{A})}{|\overline{A}|}\right]$$

- Given $A \subseteq \{1, 2, \ldots, n\}$, define $\boldsymbol{f}_A = (f_{A_1}, \ldots, f_{A_n})^T \in \mathbb{R}^n$ by

$$(f_A)_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & \text{if } i \in A \\ -\sqrt{|A|/|\bar{A}|} & \text{if } i \notin A \end{cases}$$

- Claim: For all $A \subseteq \{1, 2, \ldots, n\}$,

$$f_A^T L f_A = n \,\text{RatioCut}(A, \bar{A})$$

Relaxation of RatioCut

- *Proof of Claim*:

$$\begin{aligned}
\boldsymbol{f}_A^T L \boldsymbol{f}_A &= \frac{1}{2}\sum_{i,j} w_{ij}\left(f_{A_i} - f_{A_j}\right)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in\overline{A}} w_{ij}\left(\sqrt{\frac{|\overline{A}|}{|A|}} + \sqrt{\frac{|A|}{|\overline{A}|}}\right)^2 + \sum_{i\in\overline{A}, j\in A} w_{ij}\left(-\sqrt{\frac{|A|}{|\overline{A}|}} - \sqrt{\frac{|\overline{A}|}{|A|}}\right)^2 \\
&= \frac{1}{2}C(A,\overline{A})\left(\frac{|\overline{A}|}{|A|} + \frac{|A|}{|\overline{A}|} + 2\right) + \frac{1}{2}C(A,\overline{A})\left(\frac{|A|}{|\overline{A}|} + \frac{|\overline{A}|}{|A|} + 2\right) \\
&= C(A,\overline{A})\left(\frac{|\overline{A}|}{|A|} + \frac{|A|}{|\overline{A}|} + 2\right) \\
&= C(A,\overline{A})\left(\frac{|A| + |\overline{A}|}{|\overline{A}|} + \frac{|\overline{A}| + |A|}{|A|}\right) \\
&= n\left(\frac{C(A,\overline{A})}{|\overline{A}|} + \frac{C(A,\overline{A})}{|A|}\right) = n \cdot \mathrm{RatioCut}(A,\overline{A})
\end{aligned}$$

Relaxation of RatioCut

- Furthermore, $\boldsymbol{f}_A$ satisfies the following two properties:
  1. $\mathbf{1}^T \boldsymbol{f}_A = 0$
  2. $\|\boldsymbol{f}_A\|^2 = n$

$$|A|\sqrt{\frac{|\bar{A}|}{|A|}} - |\bar{A}|\sqrt{\frac{|A|}{|\bar{A}|}} = 0$$

$$|A| \cdot \frac{|\bar{A}|}{|A|} + |\bar{A}|\frac{|A|}{|\bar{A}|} = |A| + |\bar{A}| = n$$

- Therefore, RatioCut can be written as the following optimization problem:

$$\begin{aligned} \min_{A \subset \{1,\ldots,n\}} \quad & \boldsymbol{f}_A^T L \boldsymbol{f}_A \\ s.t. \quad & \mathbf{1}^T \boldsymbol{f}_A = 0 \\ & \|\boldsymbol{f}_A\| = \sqrt{n} \end{aligned}$$

- Note that it would still be RatioCut without the above two constraints, but we include them to keep the relaxation closer to the original problem.

Relaxation of RatioCut

- We can now state the relaxation.
- For reference, we have written RatioCut as

$$\begin{aligned} \min_{A \subset \{1,\dots,n\}} \quad & \boldsymbol{f}_A^T \boldsymbol{L} \boldsymbol{f}_A \\ s.t. \quad & \mathbf{1}^T \boldsymbol{f}_A = 0 \\ & \|\boldsymbol{f}_A\| = \sqrt{n} \end{aligned}$$

- A relaxation of RatioCut is

$$\min_{f \in \mathbb{R}^n} \; f^T L f \quad s.t. \quad \mathbb{1}^T f = 0, \quad \|f\| = \sqrt{n}$$

f^* ⟵

Try to find A such that f_A is close to f^*.

Poll

The solution to the previous optimization problem is

(A) An eigenvector associated to the smallest eigenvalue of $\boldsymbol{L}$

(B) An eigenvector associated to the second smallest eigenvalue of $\boldsymbol{L}$

(C) An eigenvector associated to the largest eigenvalue of $\boldsymbol{L}$

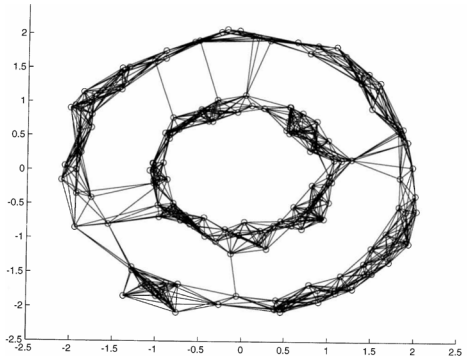
(D) An eigenvector associated to the second largest eigenvalue of $\boldsymbol{L}$

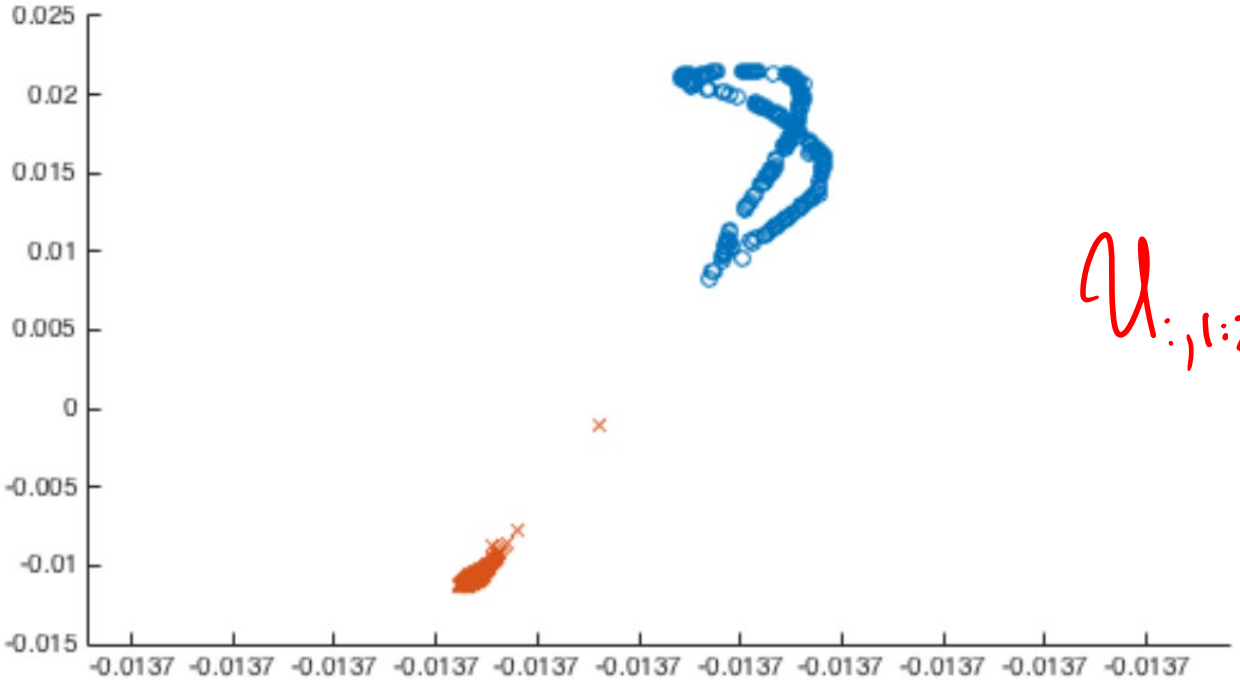
$$\min_{f \in \mathbb{R}^n} f^T L f \quad \text{s.t.} \quad \mathbb{1}^T f = 0, \quad \|f\| = \sqrt{n}$$

$$\max a^T S a \quad \text{s.t.} \quad a^T u_1 = 0, \quad \|a\| = 1$$

Example

- This figure shows $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ for the circular clusters dataset shown earlier.
- How should we recover the clusters?





$$U_{:,1:2} = \begin{bmatrix} 1 & f_1 \\ 1 & f_2 \\ \vdots & \vdots \\ 1 & f_2 \end{bmatrix} \begin{matrix} \rightarrow y_1 \\ \rightarrow y_2 \\ \vdots \\ \rightarrow y_n \end{matrix}$$

Remarks

- In practice the similarity graph is chosen to have only one connected component (if there is more then one connected component, just cluster each one separately). So the 0-eigenspace has dimension 1, and there is no ambiguity in the choice of $\mathbf{1}$ as the smallest eigenvector.
- A similar analysis applies to $K > 2$ and NCut.
- The gap between the optimal value of RatioCut/NCut and the optimal value of its relaxation can be arbitrarily large.

Final Remarks

- The mapping $\boldsymbol{x} \mapsto \boldsymbol{y}$ is actually a form of nonlinear dimensionality reduction (called Laplacian eigenmaps) that transforms the data to a space where k-means can be successfully applied.
- Model selection: Choose K such that λ_{K+1} is the first "large" eigenvalue.
- In practice $\tilde{\boldsymbol{L}}$ is preferred to $\boldsymbol{L}$.
- There is yet a third graph Laplacian defined as

$$\tilde{\tilde{\boldsymbol{L}}} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{L} \boldsymbol{D}^{\frac{1}{2}}$$

  It has properties similar to $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$, but the spectral clustering algorithm needs to be tweaked.
- For a very thorough tutorial on spectral clustering see Ulrike von Luxburg, "A Tutorial on Spectral Clustering", 2007.