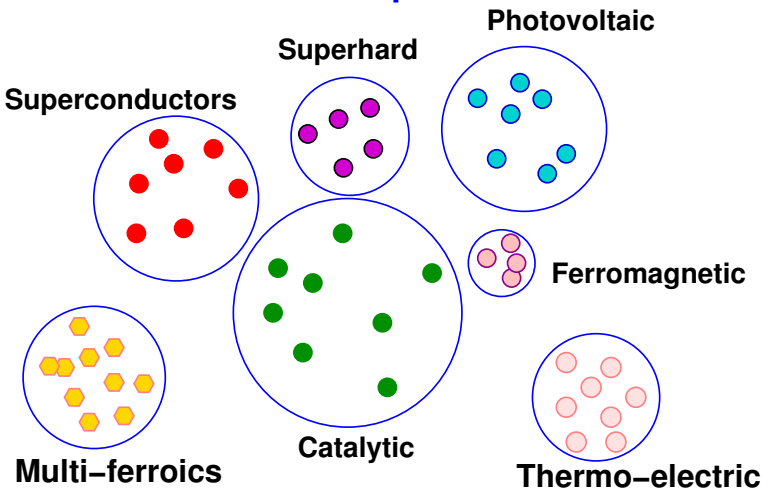
APPLICATIONS OF GRAPH LAPLACEANS: CLUSTERING

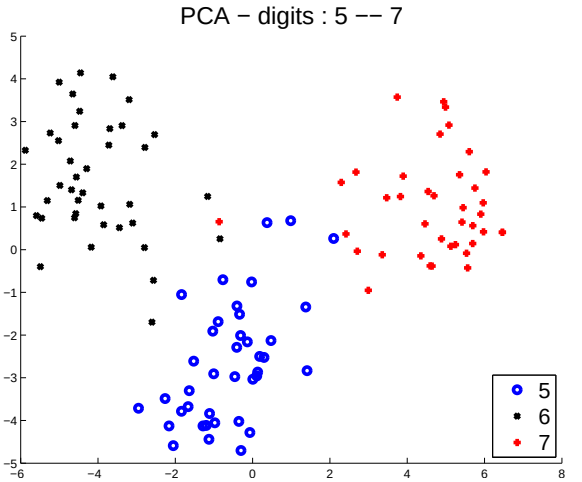
Clustering

➤ Problem: we are given n data items: $x_1, x_2, \cdots, x_n$. Would like to *'cluster'* them, i.e., group them so that each group or cluster contains items that are similar in some sense.

➤ Example: materials



➤ Example: Digits



➤ Each group is a 'cluster' or a 'class'

➤ 'Unsupervised learning'

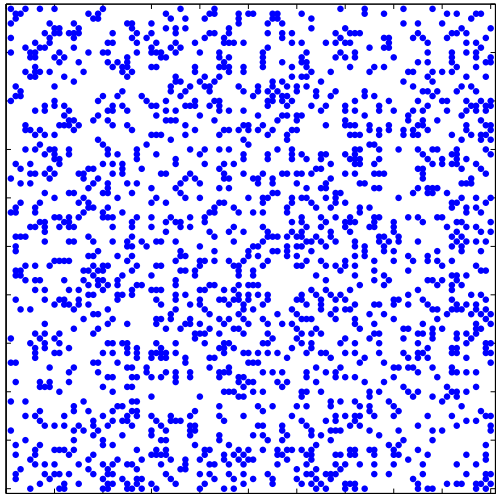
What is 'Unsupervised Learning'?

Ans: Class of methods that do not exploit labeled data

- Example of digits: perform a 2-D projection
- Images of same digit tend to cluster (more or less)
- Such 2-D representations are popular for visualization
- Can also try to find natural clusters in data, e.g., in materials
- Basic clusterning technique: K-means

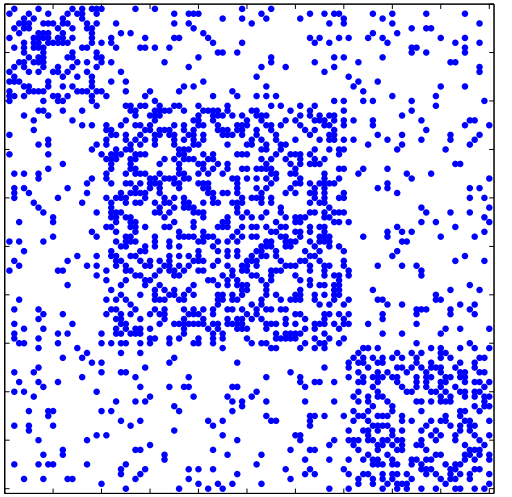
Example: Community Detection

- Communities modeled by an 'affinity' graph [e.g., 'user A sends frequent e-mails to user B'] . [data: www-personal.umich.edu/~mejn/netdata/]



← Original Adj. matrix

Goal: Find ordering so blocks are as dense as possible →



- Use 'blocking' techniques for sparse matrices
- Advantage of this viewpoint: need not know # of clusters.

Example of application Data set from :

http://www-personal.umich.edu/~mejn/netdata/

➤ Network connecting bloggers of different political orientations [2004 US presidentual election]

➤ 'Communities': liberal vs. conservative

➤ Graph: $1,490$ vertices (blogs) : first 758: liberal, rest: conservative.

➤ Edge: $i \to j$: a citation between blogs i and j

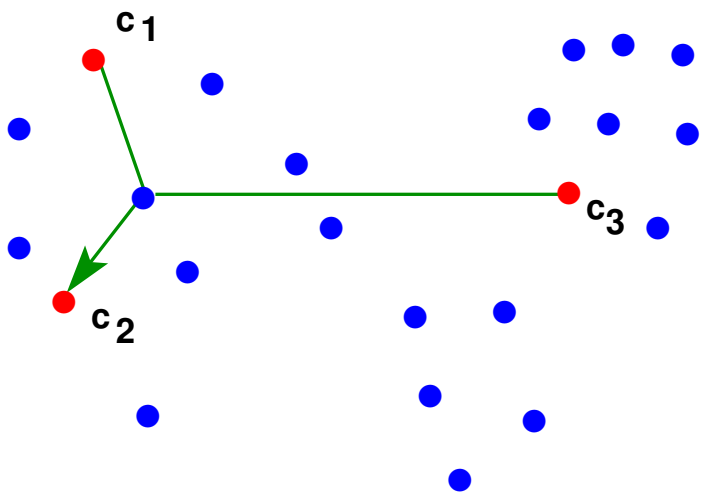
➤ Blocking algorithm (Density theshold=0.4): subgraphs [note: density = $|E|/|V|^2$.]

➤ Smaller subgraph: conservative blogs, larger one: liberals

A basic method: K-means

- A basic algorithm that uses Euclidean distance

1. Select p initial centers: $c_1, c_2, ..., c_p$ for classes $1, 2, \cdots, p$
2. For each x_i do: determine *class* of x_i as $\text{argmin}_k \|x_i - c_k\|$
3. Redefine each c_k to be the centroid of class k
4. Repeat until convergence



- Simple algorithm
- Works well (gives good results) but can be slow
- Performance depends on initialization

Methods based on similarity graphs

➤ Class of Methods that perform clustering by exploiting a graph that describes the similarities between any two items in the data.

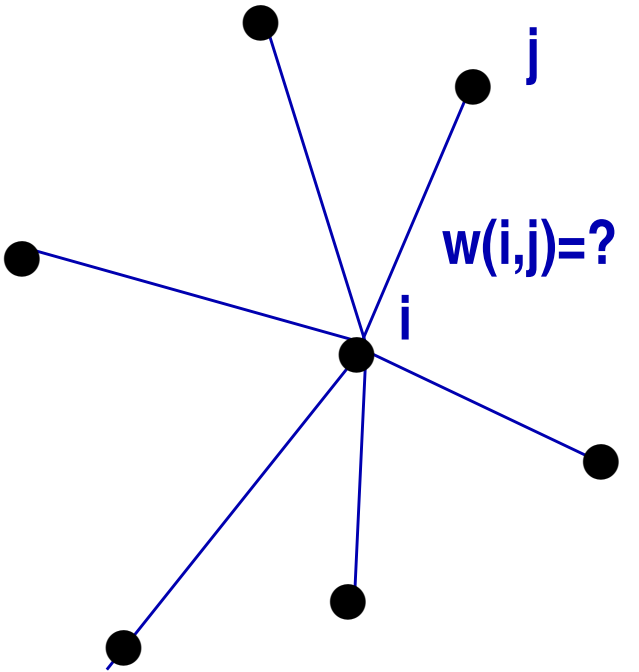
➤ Need to:

1. decide what nodes are in the neighborhood of a given node
2. quantify their similarities - by assigning a weight to any pair of nodes.

Example: For text data: Can decide that any columns i and j with a cosine greater than 0.95 are 'similar' and assign that cosine value to w_{ij}

First task: build a 'similarity' graph

➤ Goal: to build a similarity graph, i.e., a graph that captures similarity between any two items



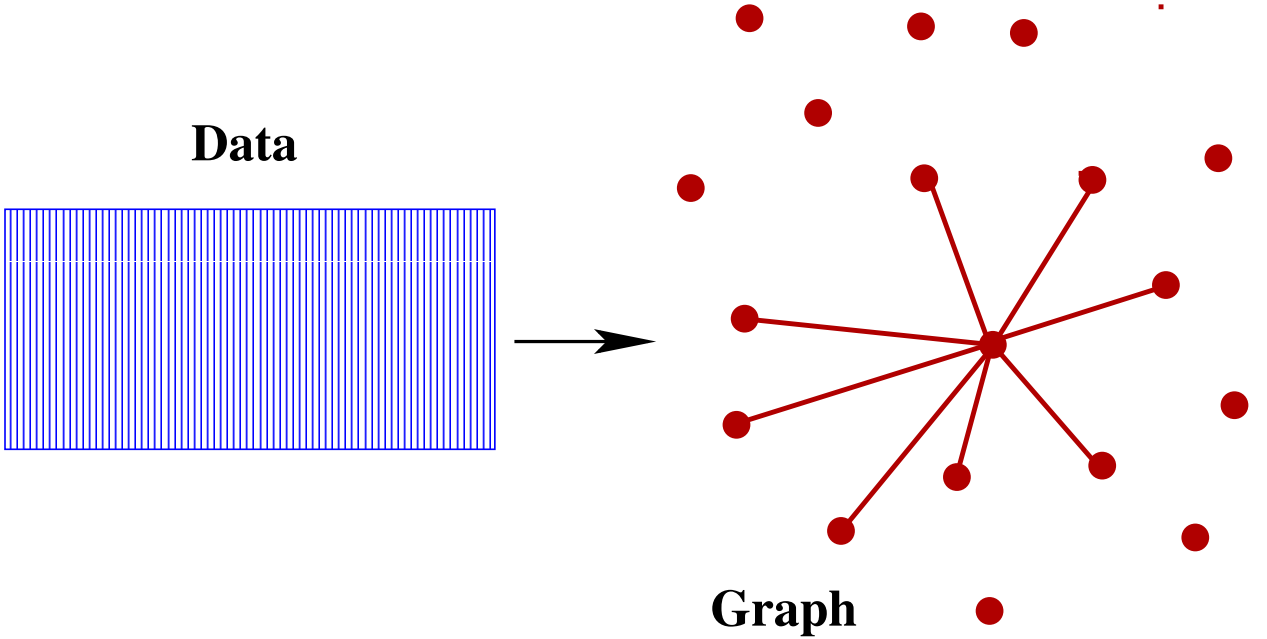
➤ Two methods: K-nearest Neighbor graphs or use Gaussian ('heat') kernel

K-nearest neighbor graphs

- Given: a set of n data points $X = \{x_1, \ldots, x_n\} \rightarrow$ vertices
- Given: a proximity measure between two data points x_i and x_j – as measured by a quantity $dist(x_i, x_j)$
- Want: For each point x_i a list of the 'nearest neighbors' of x_i (edges between x_i and these nodes).
- Note: graph will usually be directed $\rightarrow$ need to symmetrize

Nearest neighbor graphs

- For each node, get a few of the nearest neighbors → Graph



- Problem: How to build a nearest-neighbor graph from given data
- We will revisit this later.

Two types of nearest neighbor graph often used:

ϵ-graph: Edges consist of pairs (x_i, x_j) such that $\rho(x_i, x_j) \leq \epsilon$

kNN graph: Nodes adjacent to x_i are those nodes x_ℓ with the k with smallest distances $\rho(x_i, x_\ell)$.

➤ ϵ-graph is undirected and is geometrically motivated. Issues: 1) may result in disconnected components 2) what ϵ?

➤ kNN graphs are directed in general (can be trivially fixed).

➤ kNN graphs especially useful in practice.

Similarity graphs: Using 'heat-kernels'

Define weight between i and j as:

$$w_{ij} = f_{ij} \times \begin{cases} e^{\frac{-\|x_i - x_j\|^2}{\sigma_X^2}} & \text{if } \|x_i - x_j\| < r \\ 0 & \text{if not} \end{cases}$$

- Note $\|x_i - x_j\|$ could be any measure of distance...
- f_{ij} = optional = some measure of similarity - other than distance
- Only nearby points kept.
- Sparsity depends on parameters

Edge cuts, ratio cuts, normalized cuts, ...

- Assume now that we have built a 'similarity graph'
- Setting is identical with that of graph partitioning.
- Need a Graph Laplacean: $L = D - W$ with $w_{ii} = 0, w_{ij} \geq 0$ and $D = diag(W * ones(n, 1))$ [in matlab notation]
- Partition vertex set V in two sets A and B with

$$A \cup B = V, \quad A \cap B = \emptyset$$

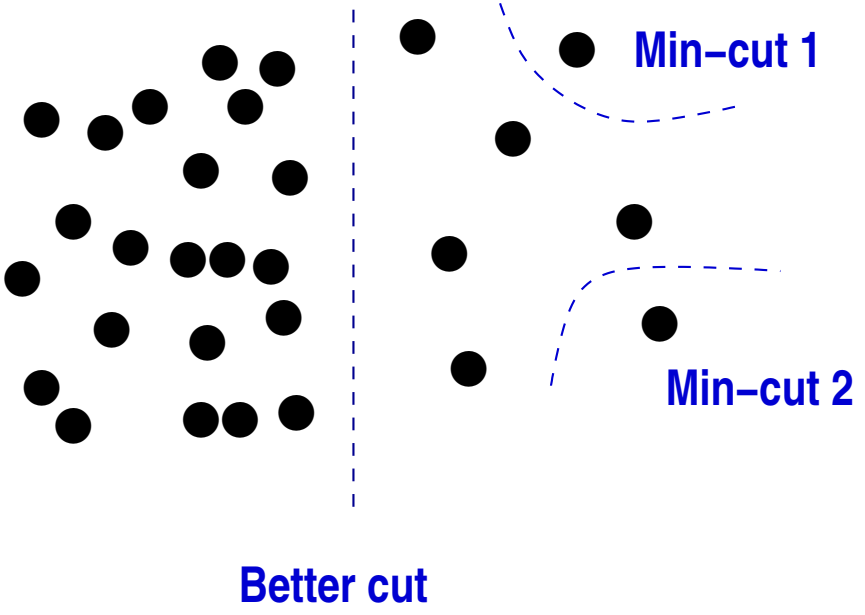
- Define

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v)$$

- First (naive) approach: use this measure to partition graph, i.e.,

... Find A and B that minimize $cut(A, B)$.

- Issue: Small sets, isolated nodes, big imbalances,



Ratio-cuts

- Standard Graph Partitioning approach: Find A, B by solving

$$\text{Minimize} \quad cut(A, B), \text{ subject to } |A| = |B|$$

- Condition $|A| = |B|$ not too meaningful in some applications - too restrictive in others.

- Minimum Ratio Cut approach. Find A, B by solving:

$$\text{Minimize} \quad \frac{cut(A,B)}{|A|.|B|}$$

- Difficult to find solution (original paper [Wei-Cheng '91] proposes several heuristics)

- Approximate solution : spectral .

Theorem [Hagen-Kahng, 91] If λ_2 is the 2nd smallest eigenvalue of L, then a lower bound for the cost c of the optimal ratio cut partition, is:

$$c \geq \frac{\lambda_2}{n}.$$

Proof: Consider an optimal partition A, B and let $p = |A|/n, q = |B|/n$. Note that $p + q = 1$. Let x be the vector with coordinates

$$x_i = \begin{cases} q & \text{if } i \in A \\ -p & \text{if } i \in B \end{cases}$$

Note that $x \perp \mathbb{1}$. Also if (i, j) == an edge-cut then $|x_i - x_j| = |q - (-p)| = |q + p| = 1$, otherwise $x_i - x_j = 0$. Therefore, $x^T L x = \sum_{(i,j)\in E}(x_i - x_j)^2 = w(A, B)$. In addition:
$\|x\|^2 = pq^2n + qp^2n = pq(p + q)n = pqn = \frac{|A|.|B|}{n}$.

Therefore, by the Courant-Fischer theorem:

$$\lambda_2 \leq \frac{(Lx, x)}{(x, x)} = n \times \frac{w(A, B)}{|A|.|B|} = n \times c.$$

Hence result. ■

➤ Idea is to use eigenvector associated with λ_2 to determine partition, e.g., based on sign of entries. Use the ratio-cut measure to actually determine where to split.

Normalized cuts [Shi-Malik,2000]

➤ Recall notation $w(X, Y) = \sum_{x \in X, y \in Y} w(x, y)$ - then define:

$$\text{ncut}(A, B) = \frac{cut(A,B)}{w(A,V)} + \frac{cut(A,B)}{w(B,V)}$$

➤ Goal is to avoid small sets A, B

✍ What is $w(A, V)$ in the case when $w_{ij} == 1$?

➤ Let x be an indicator vector:

$$x_i = \begin{cases} 1 \ if \ i \in A \\ 0 \ if \ i \in B \end{cases}$$

➤ Recall that: $x^T L x = \sum_{(i,j) \in E} w_{ij} |x_i - x_j|^2$ (note: each edge counted once)

- Therefore:

$$cut(A,B) = \sum_{x_i=1, x_j=0} w_{ij} = x^T L x$$

$$w(A,V) = \sum_{x_i=1} d_i = x^T W \mathbb{1} = x^T D \mathbb{1}$$

$$w(B,V) = \sum_{x_j=0} d_j = (\mathbb{1} - x)^T W \mathbb{1} = (\mathbb{1} - x)^T D \mathbb{1}$$

- Goal now: to minimize ncut

$$\min_{A,B} \text{ncut}(A,B) = \min_{x_i \in \{0,1\}} \frac{x^T L x}{x^T D x} + \frac{x^T L x}{(\mathbb{1} - x)^T D x}$$

- Let

$$\beta = \frac{w(A,V)}{w(B,V)} = \frac{x^T D\,\mathbb{1}}{(\mathbb{1}-x)^T D\,\mathbb{1}}$$
$$y = x - \beta(\mathbb{1}-x)$$

- Then we need to solve:

$$\min_{y_i \in \{0,-\beta\}} \frac{y^T L y}{y^T D y} \quad \text{Subject to} \quad y^T D\,\mathbb{1} = 0$$

- + Relax $\rightarrow$ need to solve Generalized eigenvalue problem

$$Ly = \lambda Dy$$

- $y_1 = \mathbb{1}$ is eigenvector associated with eigenvalue $\lambda_1 = 0$
- y_2 associated with second eigenvalue solves problem.

A few properties

✍ Show that

$$ncut(A, B) = \sigma \times \frac{cut(A, B)}{w(A, V) \times w(B, V)}$$

where σ is a constant

✍ How do ratio-cuts and normalized cuts compare when the graph is d-regular (same degree for each node).

Extension to more than 2 clusters

- Just like graph partitioning we can:

1. Apply the method recursively [Repeat clustering on the resulted parts]
2. or compute a few eigenvectors and run K-means clustering on these eigenvectors to get the clustering.

Application: Image segmentation

- First task: obtain a graph from pixels.
- Common idea: use “Heat kernels”
- Let $F_j =$ feature value (e.g., brightness), and Let $X_j =$ spatial position.

Then define

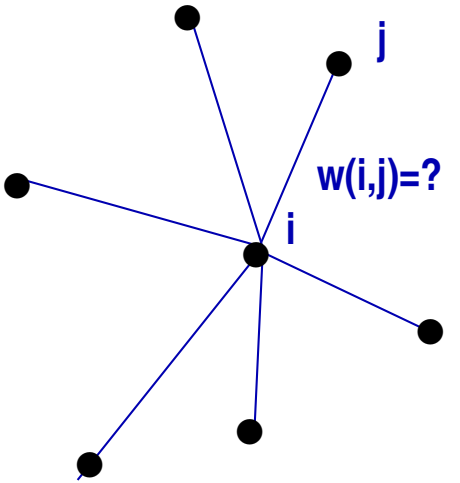
$$w_{ij} = e^{\frac{-\|F_i - F_j\|^2}{\sigma_I^2}} \times \begin{cases} e^{\frac{-\|X_i - X_j\|^2}{\sigma_X^2}} & \text{if} \|X_i - X_j\| < r \\ 0 & \text{else} \end{cases}$$

- Sparsity depends on parameters

Spectral clustering: General approach

1 Given: Collection of data samples $\{x_1, x_2, \cdots, x_n\}$

2 Build a similarity graph between items

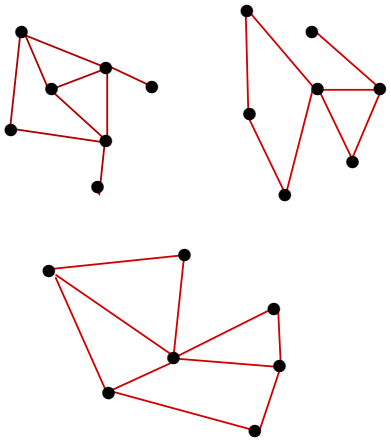


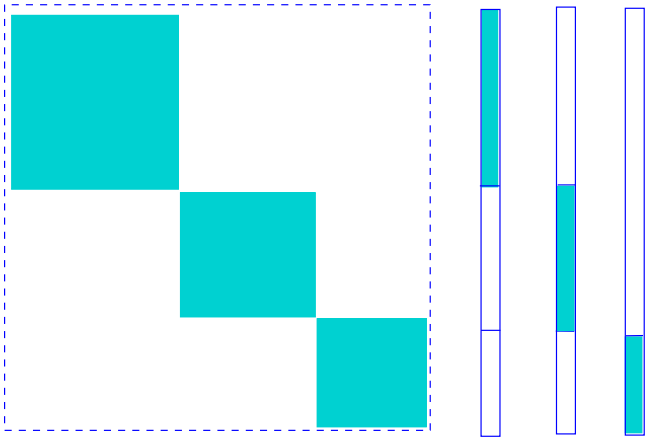
3 Compute (smallest) eigenvector (s) of resulting graph Laplacean

4 Use k-means on eigenvector (s) of Laplacean

➤ For Normalized cuts solve generalized eigen problem.

- Recall observation made earlier:

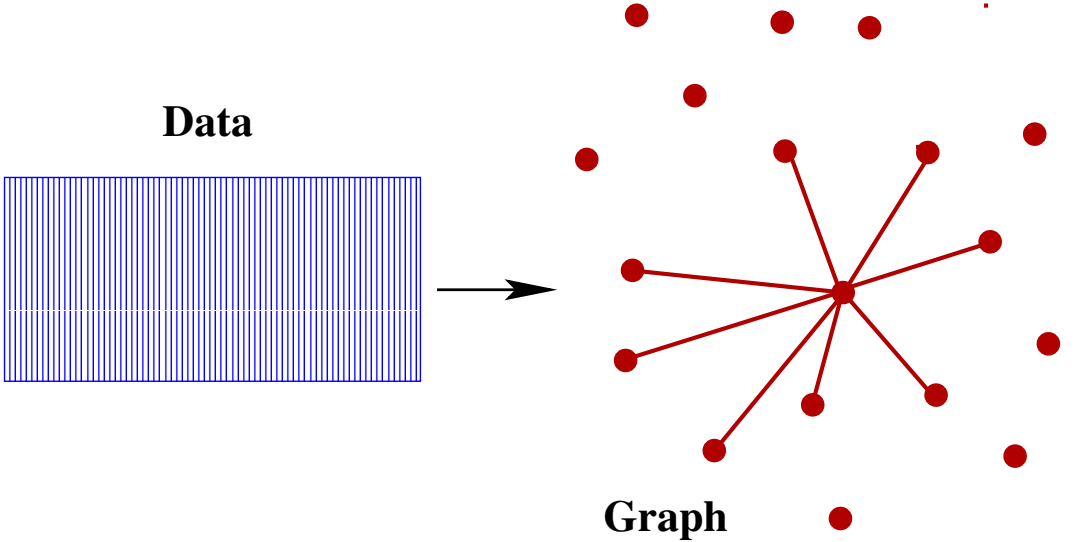




- Alg. Multiplicity of eigenvalue zero = # connected components.

Building a nearest neighbor graph

➤ Question: How to build a nearest-neighbor graph from given data?



➤ Will demonstrate the power of a divide a conquer approach combined with the Lanczos algorithm.

➤ Note: The Lanczos algortithm will be covered in detail later

Recall: Two common types of nearest neighbor graphs

ϵ-graph: Edges consist of pairs (x_i, x_j) such that $\rho(x_i, x_j) \leq \epsilon$

kNN graph: Nodes adjacent to x_i are those nodes x_ℓ with the k with smallest distances $\rho(x_i, x_\ell)$.

➤ ϵ-graph is undirected and is geometrically motivated. Issues: 1) may result in disconnected components 2) what ϵ?

➤ kNN graphs are directed in general (can be trivially fixed).

➤ kNN graphs especially useful in practice.

Divide and conquer KNN: key ingredient

- Key ingredient is *Spectral bisection*
- Let the data matrix $\boldsymbol{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$
- Each column == a data point.
- Center the data: $\hat{\boldsymbol{X}} = [\hat{x}_1, \ldots, \hat{x}_n] = \boldsymbol{X} - ce^T$
  where c == centroid; $e = ones(d, 1)$ (matlab)

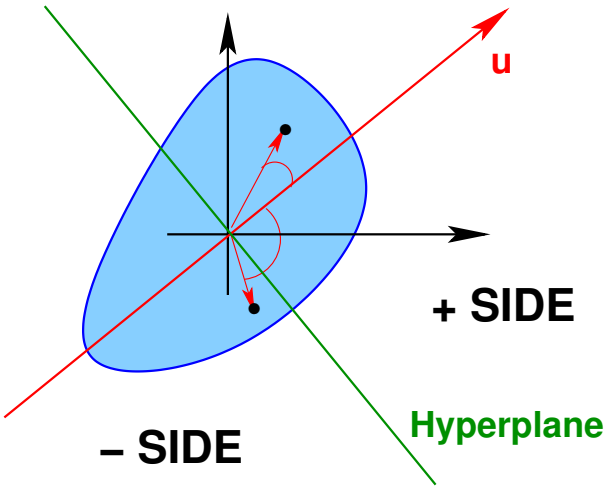
Goal: Split $\hat{X}$ into halves using a hyperplane.

Method: Principal Direction Divisive Partitioning D. Boley, '98.

Idea: Use the (σ, u, v) = largest singular triplet of $\hat{X}$ with: $u^T \hat{X} = \sigma v^T$.

➤ Hyperplane is defined as $\langle u, x \rangle = 0$, i.e., it splits the set of data points into two subsets:

$$X_+ = \{x_i \mid u^T \hat{x}_i \geq 0\} \quad \text{and} \quad X_- = \{x_i \mid u^T \hat{x}_i < 0\}.$$



➤ Note that $u^T \hat{x}_i = u^T \hat{X} e_i = \sigma v^T e_i \rightarrow$

$$X_+ = \{x_i \mid v_i \geq 0\} \quad \text{and} \quad X_- = \{x_i \mid v_i < 0\},$$

where v_i is the i-th entry of v.

- In practice: replace above criterion by

$$X_+ = \{x_i \mid v_i \geq \text{med}(v)\} \;\&\; X_- = \{x_i \mid v_i < \text{med}(v)\}$$

where *med*(v) == *median of the entries of* v.

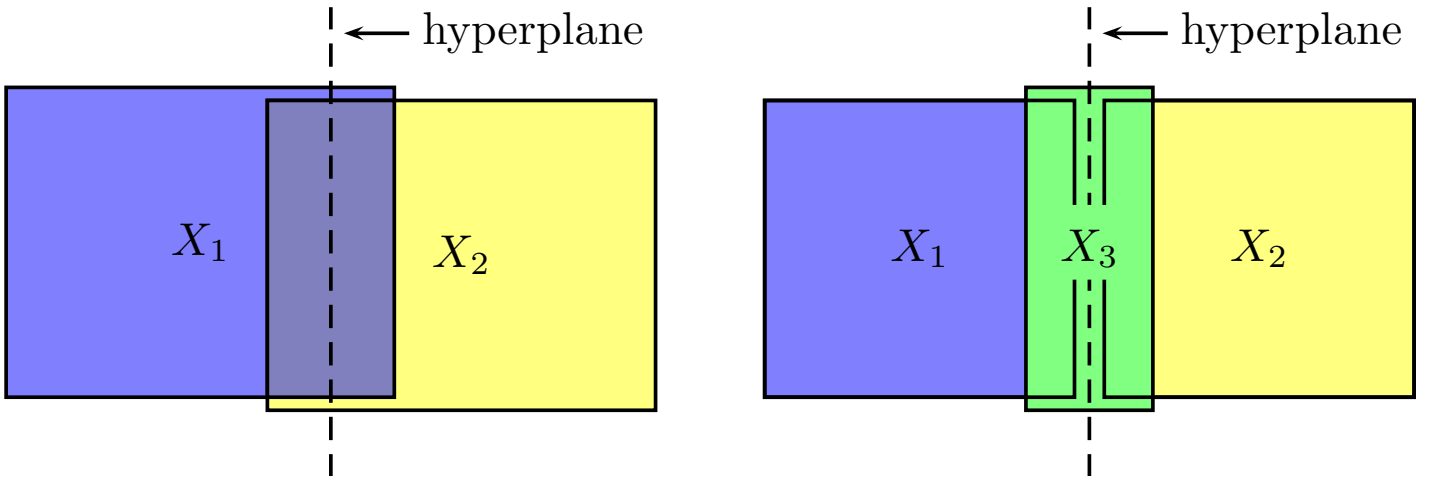
- For largest singular triplet (σ, u, v) of $\hat{X}$: use Golub-Kahan-Lanczos algorithm or Lanczos applied to $\hat{X}\hat{X}^T$ or $\hat{X}^T\hat{X}$

- Cost (assuming s Lanczos steps) : $O(n \times d \times s)$; Usually: d very small

Two divide and conquer algorithms

Overlap method: divide current set into two overlapping subsets X_1, X_2

Glue method: divide current set into two disjoint subsets X_1, X_2 plus a third set X_3 called gluing set.



➤ Exploit recursivity

The Overlap Method

➤ Divide current set X into two overlapping subsets:

$$X_1 = \{x_i \mid v_i \geq -h_\alpha(S_v)\} \quad \text{and} \quad X_2 = \{x_i \mid v_i < h_\alpha(S_v)\},$$

- where $S_v = \{|v_i| \mid i = 1, 2, \ldots, n\}$.
- and $h_\alpha(\cdot)$ is a function that returns an element larger than $(100\alpha)\%$ of those in S_v.

➤ Rationale: to ensure that the two subsets overlap $(100\alpha)\%$ of the data, i.e.,

$$|X_1 \cap X_2| = \lceil \alpha |X| \rceil \, .$$

The Glue Method

Divide the set X into two disjoint subsets X_1 and X_2 with a gluing subset X_3:

$$X_1 \cup X_2 = X, \quad X_1 \cap X_2 = \emptyset, \quad X_1 \cap X_3 \neq \emptyset, \quad X_2 \cap X_3 \neq \emptyset.$$

Criterion used for splitting:

$$X_1 = \{x_i \mid v_i \geq 0\}, \quad X_2 = \{x_i \mid v_i < 0\},$$
$$X_3 = \{x_i \mid -h_\alpha(S_v) \leq v_i < h_\alpha(S_v)\}.$$

Note: gluing subset X_3 here is just the intersection of the sets X_1, X_2 of the overlap method.

Theorem The time complexity for the overlap method is

$$T_{\mathsf{o}}(n) = \Theta(dn^{t_{\mathsf{o}}}), \qquad \text{where:} \qquad t_{\mathsf{o}} = \log_{2/(1+\alpha)} 2 = \frac{1}{1 - \log_2(1+\alpha)}.$$

Theorem The time complexity for the glue method is

$$T_{\mathsf{g}}(n) = \Theta(dn^{t_{\mathsf{g}}}/\alpha), \quad \text{where} \quad t_{\mathsf{g}} \equiv \text{sol. to the equ.:} \quad \frac{2}{2^t} + \alpha^t = 1.$$

Example: When $\alpha = 0.1$, then $t_{\mathsf{o}} = 1.16$ while $t_{\mathsf{g}} = 1.12$.

Reference:

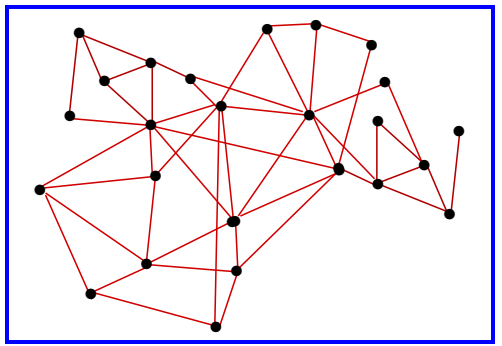
Jie Chen, Haw-Ren Fang and YS, "Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection" JMLR, vol. 10, pp. 1989-2012 (2009).

APPLICATIONS OF GRAPH LAPLACEANS: GRAPH EMBEDDINGS

Graph embeddings

- We have seen how to build a graph to represent data
- *Graph embedding* does the opposite: maps a graph to data

Given: a graph that models some data (e.g., a kNN graph)

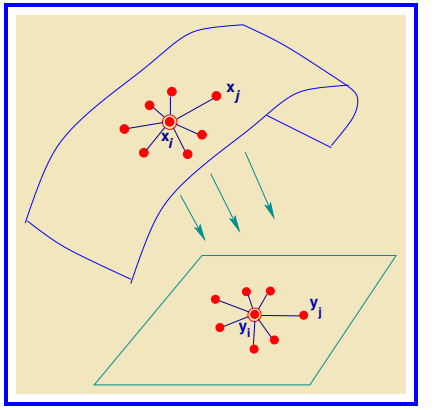


$\longrightarrow$ Data: $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]$ in $\mathbb{R}^d$

- Trivial use: visualize a graph ($d = 2$)
- Wish: mapping should preserve *similarities* in graph.

Vertex embedding: map every vertex x_i to a vector $y_i \in \mathbb{R}^d$

➤ Many applications [clustering, finding missing link, semi-supervised learning, community detection, ...]



➤ Graph captures similarities, closeness, ..., in data

Objective: Build a mapping of each vertex i to a data point $y_i \in \mathbb{R}^d$

➤ Many methods do this

➤ Eigenmaps and LLE are two of the best known

- Eigenmaps uses the *graph Laplacean*
- Recall: Graph Laplacean is a matrix defined by :

$$L = D - W$$

$$\begin{cases} w_{ij} \geq 0 \text{ if } j \in Adj(i) \\ w_{ij} = 0 \quad \text{else} \end{cases} \qquad D = \text{diag}\left[d_{ii} = \sum_{j \neq i} w_{ij} \right]$$

with $Adj(i)$ = neighborhood of i (excludes i)

- Remember that vertex i represents data item x_i. We will use i or x_i to refer to the vertex.
- We will find the y_i's by solving an optimization problem.

The Laplacean eigenmaps approach

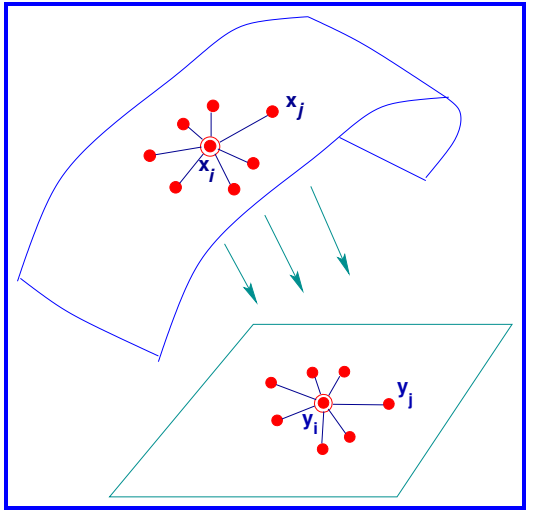
Laplacean Eigenmaps [Belkin-Niyogi '01] *minimizes*

$$\mathcal{F}(Y) = \sum_{i,j=1}^{n} w_{ij} \|y_i - y_j\|^2 \quad \text{subject to} \quad YDY^\top = I$$

Motivation: if $\|x_i - x_j\|$ is small (orig. data), we want $\|y_i - y_j\|$ to be also small (low-Dim. data)

- Original data used indirectly through its graph
- Objective function can be translated to a trace

(see Property 3 in Lecture notes 9) and will yield a sparse eigenvalue problem



- Problem translates to:

$$\begin{cases} \min_{\boldsymbol{Y} \in \mathbb{R}^{d \times n}} & \text{Tr}\left[\boldsymbol{Y}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{Y}^\top\right] . \\ \boldsymbol{Y}\boldsymbol{D}\,\boldsymbol{Y}^\top = \boldsymbol{I} \end{cases}$$

- Solution (sort eigenvalues increasingly):

$$(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{u}_i = \boldsymbol{\lambda}_i \boldsymbol{D}\boldsymbol{u}_i \; ; \quad \boldsymbol{y}_i = \boldsymbol{u}_i^\top ; \quad \boldsymbol{i} = \boldsymbol{1}, \cdots, \boldsymbol{d}$$

- An $n \times n$ sparse eigenvalue problem [In 'sample' space]

- Note: can assume $D = I$. Amounts to rescaling data. Problem becomes

$$(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{u}_i = \boldsymbol{\lambda}_i \boldsymbol{u}_i \; ; \quad \boldsymbol{y}_i = \boldsymbol{u}_i^\top ; \quad \boldsymbol{i} = \boldsymbol{1}, \cdots, \boldsymbol{d}$$

Locally Linear Embedding (Roweis-Saul-00)

➤ LLE is very similar to Eigenmaps. Main differences:

1) Graph Laplacean matrix is replaced by an 'affinity' graph

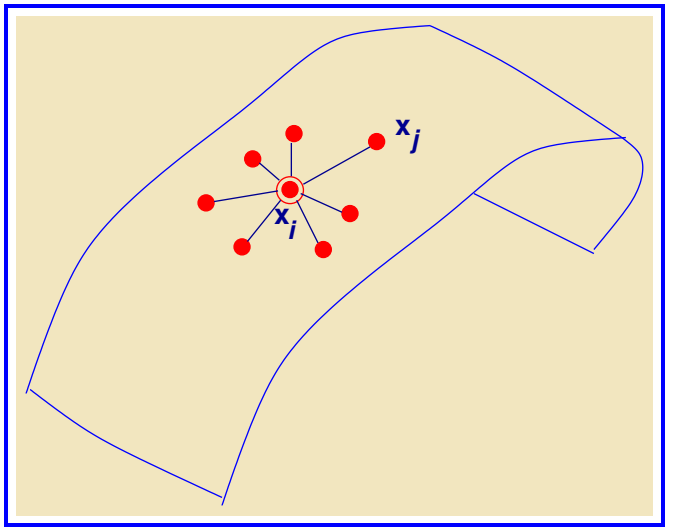
2) Objective function is changed: want to preserve graph

1. Graph: Each x_i is written as a convex combination of its k nearest neighbors:

$x_i \approx \Sigma w_{ij} x_j, \quad \sum_{j \in N_i} w_{ij} = 1$

➤ Optimal weights computed ('local calculation') by minimizing

$$\|x_i - \Sigma w_{ij} x_j\| \quad \text{for} \quad i = 1, \cdots, n$$



2. Mapping:

The y_i's should obey the same 'affinity' as x_i's $\rightsquigarrow$

Minimize:

$$\sum_i \left\| y_i - \sum_j w_{ij} y_j \right\|^2 \quad \text{subject to:} \quad \boldsymbol{Y}\,\mathbb{1} = 0, \quad \boldsymbol{Y}\boldsymbol{Y}^\top = \boldsymbol{I}$$

Solution:

$$\boxed{(\boldsymbol{I} - \boldsymbol{W}^\top)(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{u}_i = \boldsymbol{\lambda}_i \boldsymbol{u}_i; \qquad \boldsymbol{y}_i = \boldsymbol{u}_i^\top \ .}$$

➤ $(\boldsymbol{I} - \boldsymbol{W}^\top)(\boldsymbol{I} - \boldsymbol{W})$ replaces the graph Laplacean of eigenmaps

Implicit vs explicit mappings

➤ In Eigenmaps and LLE we only determine a set of $y_i's$ in $\mathbb{R}^d$ from the data points $\{x_i\}$.

➤ The mapping $y_i = \phi(x_i), i = 1, \cdots, n$ is implicit

➤ Difficult to compute a y for an x that is not one of the x_i's

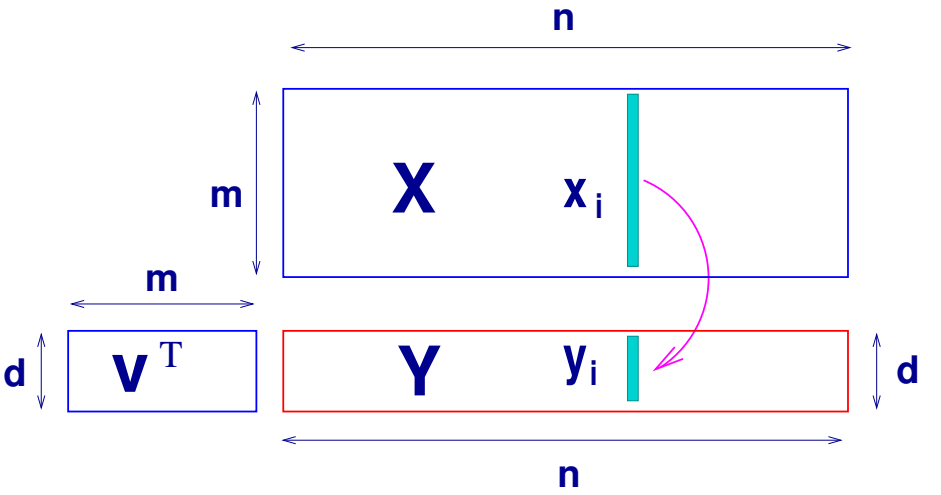
➤ Inconvenient for classification. Thus is known as the "The out-of-sample extension" problem

➤ In Explicit (also known as linear) methods: mapping ϕ is known explicitly (and it is linear.)

Locally Preserving Projections (He-Niyogi-03)

➤ LPP is a linear dimensionality reduction technique

➤ Recall the setting:
Want $\boldsymbol{V} \in \mathbb{R}^{m \times d}$; $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$



➤ Starts with the same neighborhood graph as Eigenmaps: $\boldsymbol{L} \equiv \boldsymbol{D} - \boldsymbol{W}$ = graph 'Laplacean'; with $\boldsymbol{D} \equiv \boldsymbol{diag}(\{\Sigma_i \boldsymbol{w_{ij}}\})$.

- Optimization problem is to solve

$$\min_{Y \in \mathbb{R}^{d \times n},\ YDY^\top = I} \quad \Sigma_{i,j} w_{ij} \|y_i - y_j\|^2, \quad Y = V^\top X.$$

- Difference with eigenmaps: Y is an explicit projection of X
- Solution (sort eigenvalues increasingly)

$$XLX^\top v_i = \lambda_i XDX^\top v_i \quad y_{i,:} = v_i^\top X$$

- Note: essentially same method in [Koren-Carmel'04] called 'weighted PCA' [viewed from the angle of improving PCA]

ONPP (Kokiopoulou and YS '05)

- Orthogonal Neighborhood Preserving Projections
- A linear (orthogonoal) version of LLE obtained by writing $\boldsymbol{Y}$ in the form $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$
- Same graph as LLE. Objective: preserve the affinity graph (as in LLE) *but* with the constraint $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$
- Problem solved to obtain mapping:

$$\min_{\boldsymbol{V}} \text{Tr} \left[\boldsymbol{V}^\top \boldsymbol{X} (\boldsymbol{I} - \boldsymbol{W}^\top)(\boldsymbol{I} - \boldsymbol{W}) \boldsymbol{X}^\top \boldsymbol{V} \right]$$

s.t. $\boldsymbol{V}^T \boldsymbol{V} = \boldsymbol{I}$

- In LLE replace $\boldsymbol{V}^\top \boldsymbol{X}$ by $\boldsymbol{Y}$

More recent methods

- Quite a bit of recent work - e.g., methods: node2vec, DeepWalk, GraRep, ... See the following papers ... among many others :

[1] *William L. Hamilton, Rex Ying, and Jure Leskovec* *Representation Learning on Graphs: Methods and Applications* arXiv:1709.05584v3

[2] *Shaosheng Cao, Wei Lu, and Qiongkai Xu* *GraRep: Learning Graph Representations with Global Structural Information*, CIKM, ACM Conference on Information and Knowledge Management, 24

[3] *Amr Ahmed, Nino Shervashidze, and Shravan Narayanamurthy*, *Distributed Large-scale Natural Graph Factorization* [Proc. WWW 2013, May 1317, 2013, Rio de Janeiro, Brazil]

Example: Graph factorization

- Line of work in Papers [1] and [3] above + others
- Instead of minimizing $\sum w_{ij} \|y_i - y_j\|_2^2$ as before

... try to minimize

$$\sum_{ij} |w_{ij} - y_i^T y_j|^2$$

- In other words solve: $\min_Y \|W - Y^T Y\|_F^2$
- Referred to as *Graph factorization*
- Common in knowledge graphs

Major tool of Data Mining: Dimension reduction

- Eigenmaps and LLE are a form of dimension reduction:

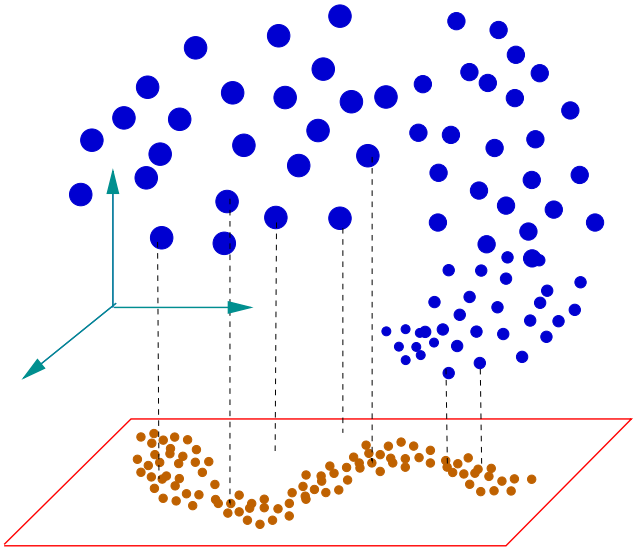
$$\text{Data in } \mathbb{R}^m \rightarrow \text{graph} \rightarrow \text{Data in } \mathbb{R}^d$$

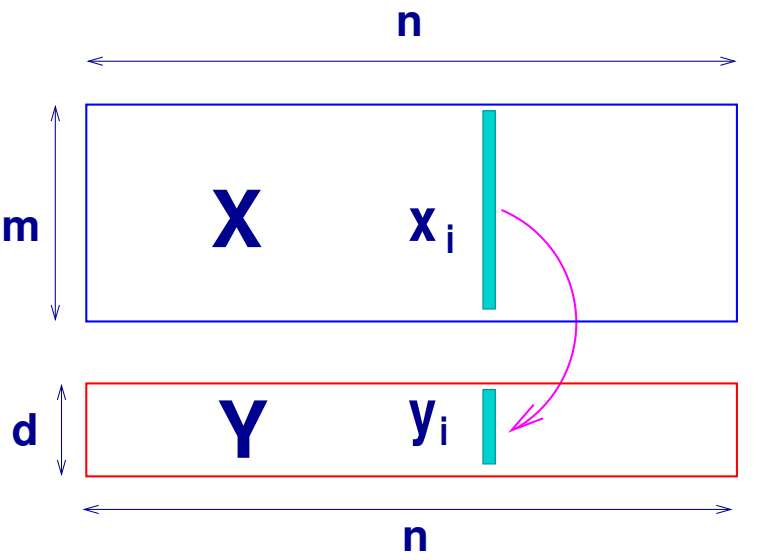
- So are the explicit (linear) methods (LPP, ONPP), ...,

Dimenson reduction: Given: $\boldsymbol{X} = [x_1, \cdots, x_n] \in \mathbb{R}^{m \times n}$, find a low-dimens. representation $\boldsymbol{Y} = [y_1, \cdots, y_n] \in \mathbb{R}^{d \times n}$ of $\boldsymbol{X}$

- Achieved by a mapping $\Phi : x \in \mathbb{R}^m \longrightarrow y \in \mathbb{R}^d$ so:

$$\phi(x_i) = y_i, \quad i = 1, \cdots, n$$





- Φ may be linear : $y_j = W^\top x_j,\ \forall j$, *or,* $Y = W^\top X$
- ... or nonlinear (implicit).
- Mapping Φ required to: Preserve proximity? Maximize variance? Preserve a certain graph?

Basics: Principal Component Analysis (PCA)

In *Principal Component Analysis* $\boldsymbol{W}$ is computed to:

Maximize variance of projected data:

$$\max_{\boldsymbol{W}\in\mathbb{R}^{m\times d};\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}} \sum_{i=1}^{n}\left\|\boldsymbol{y}_i-\frac{1}{\boldsymbol{n}}\sum_{j=1}^{n}\boldsymbol{y}_j\right\|_2^2, \quad \boldsymbol{y}_i=\boldsymbol{W}^\top\boldsymbol{x}_i.$$

➤ Leads to maximizing

$$\mathrm{Tr}\left[\boldsymbol{W}^\top(\boldsymbol{X}-\boldsymbol{\mu}\boldsymbol{e}^\top)(\boldsymbol{X}-\boldsymbol{\mu}\boldsymbol{e}^\top)^\top\boldsymbol{W}\right], \quad \boldsymbol{\mu}=\tfrac{1}{n}\Sigma_{i=1}^{n}\boldsymbol{x}_i$$

➤ Solution $\boldsymbol{W}=\{$ dominant eigenvectors $\}$ of the covariance matrix $\equiv$ Set of left singular vectors of $\bar{\boldsymbol{X}}=\boldsymbol{X}-\boldsymbol{\mu}\boldsymbol{e}^\top$

SVD:

$$\bar{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}, \quad \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}, \quad \boldsymbol{\Sigma} = \text{Diag}$$

- Optimal $\boldsymbol{W} = \boldsymbol{U}_d \equiv$ matrix of first d columns of $\boldsymbol{U}$
- Solution $\boldsymbol{W}$ also minimizes 'reconstruction error' ..

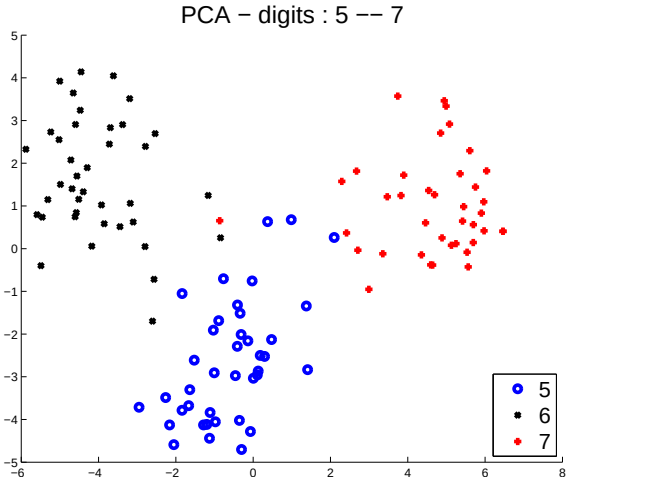
$$\sum_i \|\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{x}_i\|^2 = \sum_i \|\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i\|^2$$

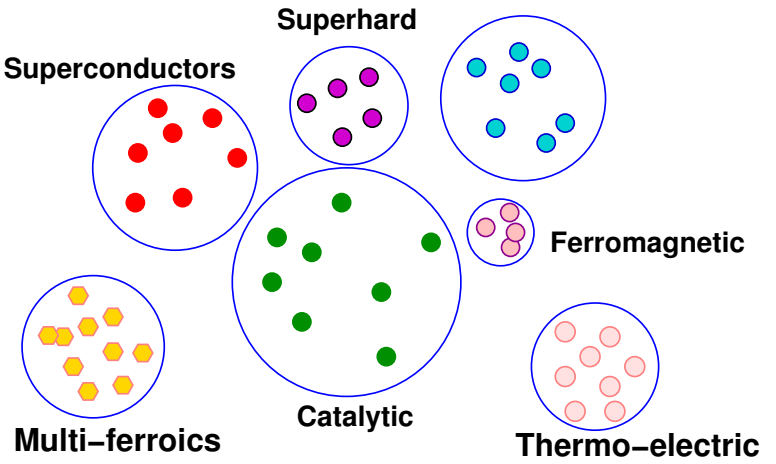
- In some methods recentering to zero is not done, i.e., $\bar{\boldsymbol{X}}$ replaced by $\boldsymbol{X}$.

Recall: Unsupervised learning

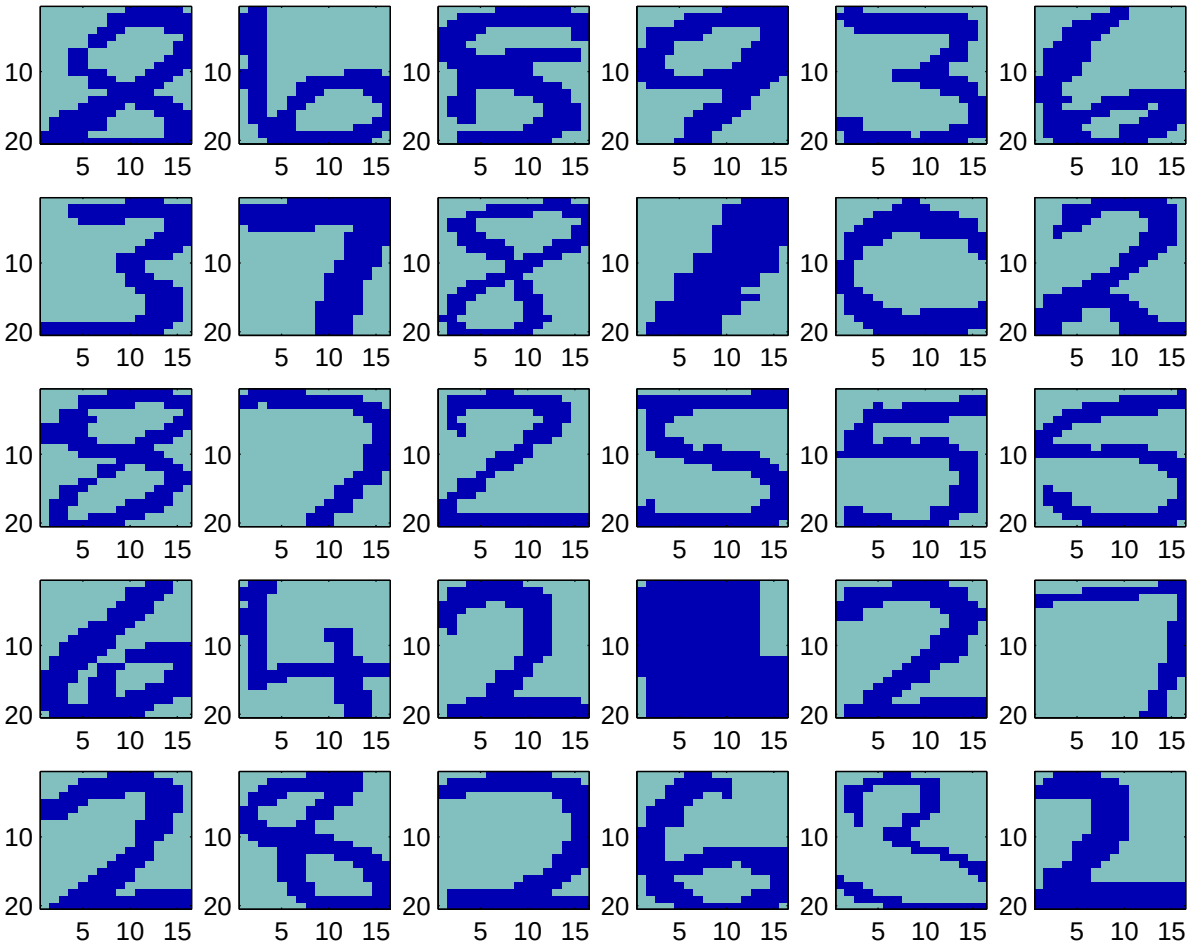
"Unsupervised learning" : methods do not exploit labeled data

- Example of digits: perform a 2-D projection
- Images of same digit tend to cluster (more or less)
- Such 2-D representations are popular for visualization
- Can also try to find natural clusters in data, e.g., in materials
- Basic clusterning technique: K-means

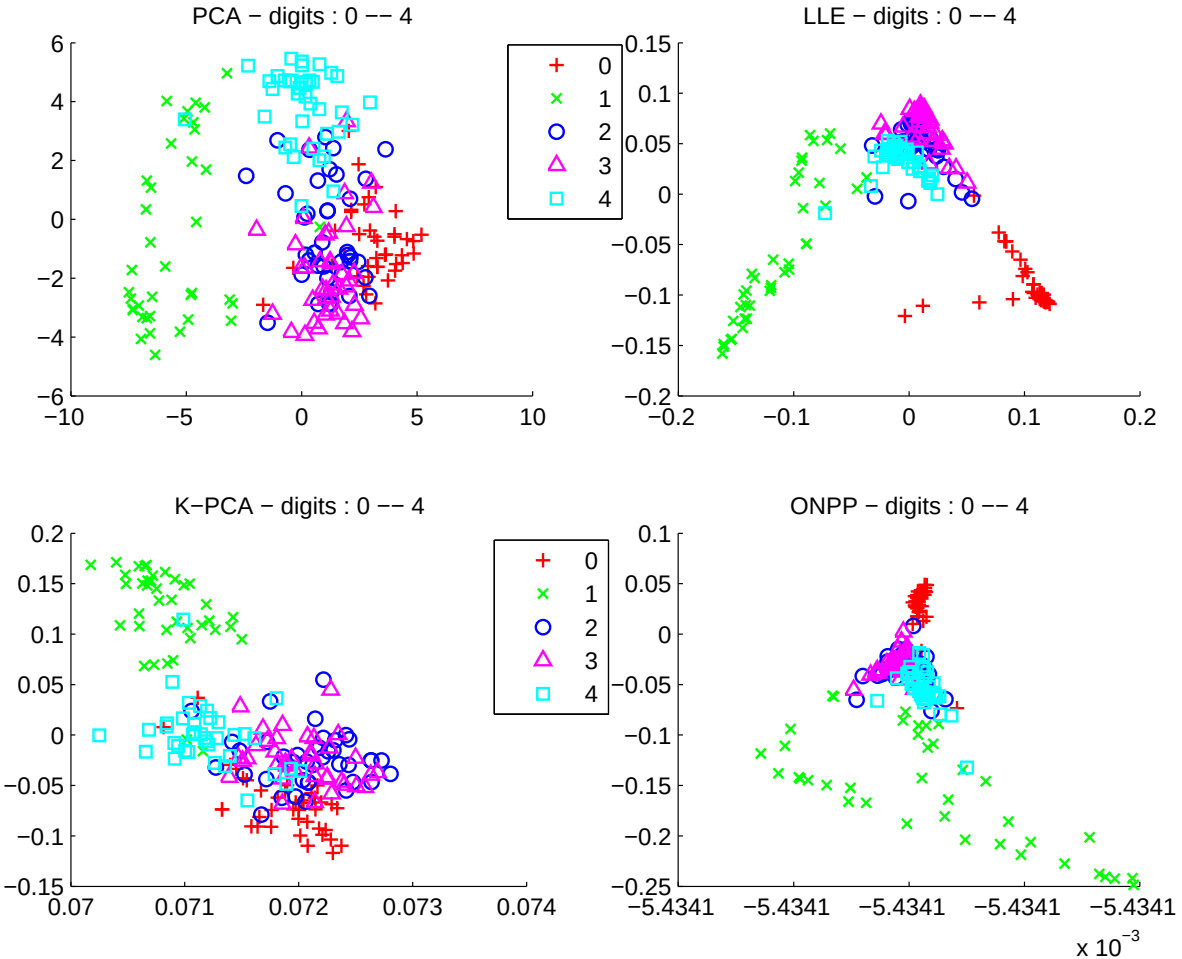




Example: Digit images (a random sample of 30)



2-D 'reductions':



SPECTRAL DENSITIES AND RANK ESTIMATION

What dimension to use in dimension reduction?

- Important question – but a hard one.
- Often, dimension k is selected in an ad-hoc way.
- k = intrinsic rank of data.
- Can we estimate it?

Two scenarios:

1. We know the magnitude of the noise, say τ. Then, ignore any singular value below τ and count the others.

2. We have no idea on the magnitude of noise. Determine a good threshold τ to use and count singular values $> \tau$.

Determining rank by eigenvalue counts

- Idea: count eigenvalues of $A^T A$ (or AA^T) that are $> \tau$.
- Use technique in [E. Di Napoli, E. Polizzi, and Y.S., 2013] based on trace estimators.
- Summarized next for general situation of a symmetric real (or Hermitian complex) matrix A

Eigenvalue counts [E. Di Napoli, E. Polizzi, YS]

The problem:

Estimate number of eigenvalues of A in given interval $[a, \quad b]$

Motivation:

➤ Eigensolvers based on splitting the spectrum intervals and extracting eigenpairs from each interval independently.

- Contour integration-type methods, e,g., FEAST [Polizzi 2011], Sakurai-Sugiura - method [2003, 2007, ..]
- Polynomial filtering, e.g.,: Schofield, Chelikowsky, YS'2011.

➤ Problems related to rank estimation in many applications

Eigenvalue counts: Standard approach and an alternative

Problem: A Hermitian with eigenpairs $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. *Want:* number $\mu_{[a,b]}$ of λ_i's $\in [a,\ b]$ - where $\lambda_1 \leq a \leq b \leq \lambda_n$.

- Standard method: Use Sylvester inertia theorem. $\rightarrow$ Expensive

- Alternative: Exploit trace of the eigen-projector:

$$P = \sum_{\lambda_i \in [a\ b]} u_i u_i^T.$$

- We know that :

$$\mathrm{Tr}\,(P) = \mu_{[a,b]}$$

- Goal: calculate an approximation to : $\mathrm{Tr}\,(P)$

➤ P is not available ... but can be approximated by: • (1) a polynomial in A, or • (2) a rational function in A.

Approximation theory viewpoint:

➤ Interpret P as a step function of A, namely:

$$P = h(A) \quad \text{where} \quad h(t) = \begin{cases} 1 & \text{if } t \in [a\ b] \\ 0 & \text{otherwise} \end{cases}$$

➤ Approximate $h(t)$ by a polynomial ψ. Then use statistical estimator for approximating Tr $(\psi(A))$ – to be discussed next

➤ Hutchinson's unbiased estimator uses only matrix-vector products to approximate the trace of a generic matrix A.

How to estimate the trace of a matrix

Trace estimation: A few examples of applications

Problem 1: Compute Tr[inv[A]] the trace of the inverse.

➤ Arises in cross validation methods [Stats]. Motivation for the work [Golub & Meurant, "Matrices, Moments, and Quadrature", 1993, Book with same title in 2009]

Problem 2: Compute Tr [f (A)], f a certain function

➤ Arises in many applications in Physics, e.g., Stochastic estimations of Tr (f(A)) extensively used by quantum chemists to estimate Density of States, see

[H. Röder, R. N. Silver, D. A. Drabold, J. J. Dong, Phys. Rev. B. 55, 15382 (1997)].

Problem 3: Compute diag[inv(A)] the diagonal of the inverse

➤ Dynamic Mean Field Theory [DMFT, motivation for our work on this topic]. Related approach: Non Equilibrium Green's Function (NEGF) approach used to model nanoscale transistors.

➤ Uncertainty quantification: diagonal of the inverse of a covariance matrix needed [Bekas, Curioni, Fedulova '09]

Problem 4: Compute diag[f (A)] ; f = a certain function.

➤ Arises in density matrix approaches in quantum modeling

$$f(\epsilon) = \frac{1}{1 + \exp(\frac{\epsilon - \mu}{k_B T})}$$

Here, f = Fermi-Dirac operator
Note: when $T \rightarrow 0$ then $f \rightarrow$ a step function.

➤ Linear-Scaling methods

Problem 5: Estimate the numerical rank.

➤ Amounts to counting the number of singular values above a certain threshold τ == Trace $(\phi_\tau(A^T A))$..

$\phi_\tau(t)$ is a certain step function.

Problem 6: Estimate the log-determinant (common in statistics)

$$\log\det(A) = \text{Trace}(\log(A)) = \sum_{i=1}^{n} \log(\lambda_i).$$

.... many others

Important tool: Stochastic Trace Estimator

➤ To estimate diagonal of $B = f(A)$ (e.g., $B = A^{-1}$), let:

Notation:

- $d(B) = \text{diag}(B)$ [matlab notation]
- $\odot$ and $\oslash$: Elementwise multiplication and division of vectors
- $\{v_j\}$: Sequence of s random vectors

Result:

$$d(B) \approx \left[\sum_{j=1}^{s} v_j \odot Bv_j\right] \oslash \left[\sum_{j=1}^{s} v_j \odot v_j\right]$$

C. Bekas , E. Kokiopoulou & YS ('05); C. Bekas, A. Curioni, I. Fedulova '09; ...

Trace of a matrix

- For the trace - take vectors of unit norm and

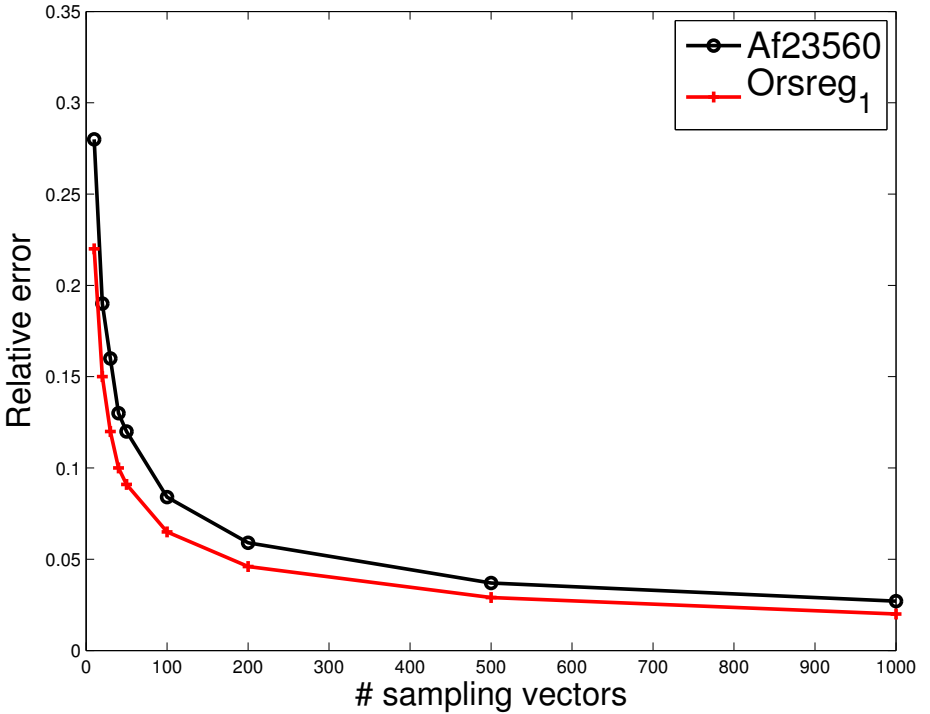
$$\text{Trace}(B) \approx \frac{1}{s} \sum_{j=1}^{s} v_j^T B v_j$$

- Hutchinson's estimator : take random vectors with components of the form $\pm 1/\sqrt{n}$ [Rademacher vectors]

- Extensively studied in literature. See e.g.: *Hutchinson '89; H. Avron and S. Toledo '11; G.H. Golub & U. Von Matt '97; Roosta-Khorasani & U. Ascher '15; ...*

Typical convergence curve for stochastic estimator

- Estimating the diagonal of inverse of two sample matrices



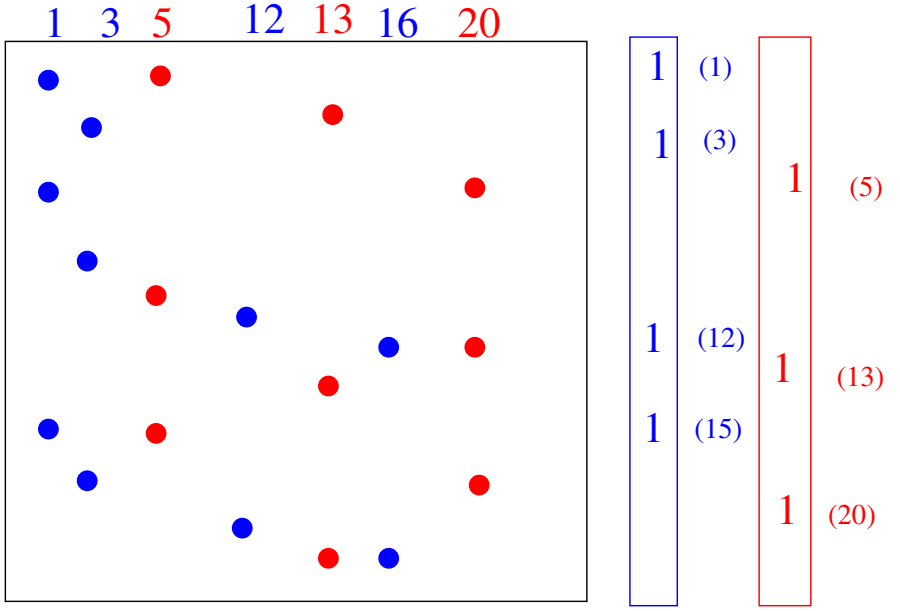
Alternative: standard probing

Basis of the method: Color columns of matrix so that no two columns of the same color overlap.

Entries of same color can be computed with 1 matvec

➤ Corresponds to coloring graph of $A^T A$.

➤ For problem of diag(A) need only color graph of A



In summary:

- Probing much more powerful when $f(A)$ is known to be nearly sparse (e.g. banded)..
- Approximate pattern (graph) can be obtained inexpensively
- Generally just a handful of probing vectors needed – Can be obtained by coloring graph
- However:
- Not as general: need $f(A)$ to be ' ϵ – sparse '

References:

- *J. M. Tang and YS*, *A probing method for computing the diagonal of a matrix inverse*, Numer. Lin. Alg. Appl., 19 (2012), pp. 485–501.

See also (improvements)

- *Andreas Stathopoulos, Jesse Laeuchli, and Kostas Orginos Hierarchical Probing for Estimating the Trace of the Matrix Inverse on Toroidal Lattices* SISC, 2012. [somewhat specific to Lattice QCD]

- *E. Aune, D. P. Simpson, J. Eidsvik* [Statistics and Computing 2012] combine probing with stochastic estimation. Good improvements reported.