# The trace ratio optimization problem

*Yousef Saad*

**Department of Computer Science and Engineering**

*University of Minnesota*

*Université de Valenciennes*

*May 19, 2016*

## *A personal tribute*

- I was invited by *Mohammed Bellalij* to visit UVHC in May 2009.

**Goal:** To collaborate on linear algebra methods for data mining

- One of my interests at the time: Face recognition..
- Then: Fisher analysis (LDA) $\longrightarrow$ Trace ratio

- Much of what was done in literature was ad-hoc

*Our Aim:* More rigorous techniques + theory

- It was a *hot* month of May, but...
- .. Phenomenal inspiration in our discussions..
- Drafted a paper - Later finalized with Than Ngo [Student]

T. T. Ngo, M. Bellalij, and Y. Saad, *The trace ratio optimization problem for dimensionality reduction.*

- Appeared (quickly!) in 2010 is SIMAX.

- On Nov. 27, 2011, SIAM editor-in-chief sent us an e-mail: Our paper selected as a SIGEST article.
- Appeared in 2012 in SIAM Review ...
- ... Plus there was a formal SIAM award
- SIAM prize at the Siam Annual meeting in July 2013
- Mohammed planned to attend award ceremony (SIAM Annual meeting luncheon)

... but...

Date: Sat, 15 Jun 2013 17:01:03 +0200
From: Mohammed Bellalij <mohammed.bellalij@univ-valenciennes.fr>
To: Mitch Chernoff <Chernoff@siam.org>

....

Dear Mitch,

Because of an important and unplanned meeting in my department on the 11th of July, and my current health, which makes a return trip over 3 days rather complicated, I must unfortunately let you know that I cannot be present for the Awards Lunch on Tuesday, July 9. I am very disappointed to have to miss this ceremony, which was a great honor for me.

Best wishes,

...

...

## ➤ Valenciennes feels empty all of a sudden

*“Un seul être vous manque et tout est dépeuplé.”*
(one person is no longer around and the whole world seems depopulated)
Alphonse de Lamartine.

## *The trace ratio problem*

- Goal of this talk: present this work
- Discuss origin of problem + applications
- Extensions done by Mohammed [MOCASIM'14 talk]

## The trace ratio problem: Origins

- What is data mining?

Set of methods and tools to extract meaningful information or patterns from (big) datasets. Broad area : data analysis, machine learning, pattern recognition, information retrieval, ...

- Blends: linear algebra; Statistics; Graph theory; Approximation theory; Optimization; ...

- A fundamental tool: dimension reduction: Often in the form of an explicit projector that is sought to achieve a certain desirable property, e.g., to separate data well, i.e., to 'discriminate'

## *Major tool of Data Mining: Dimension reduction*

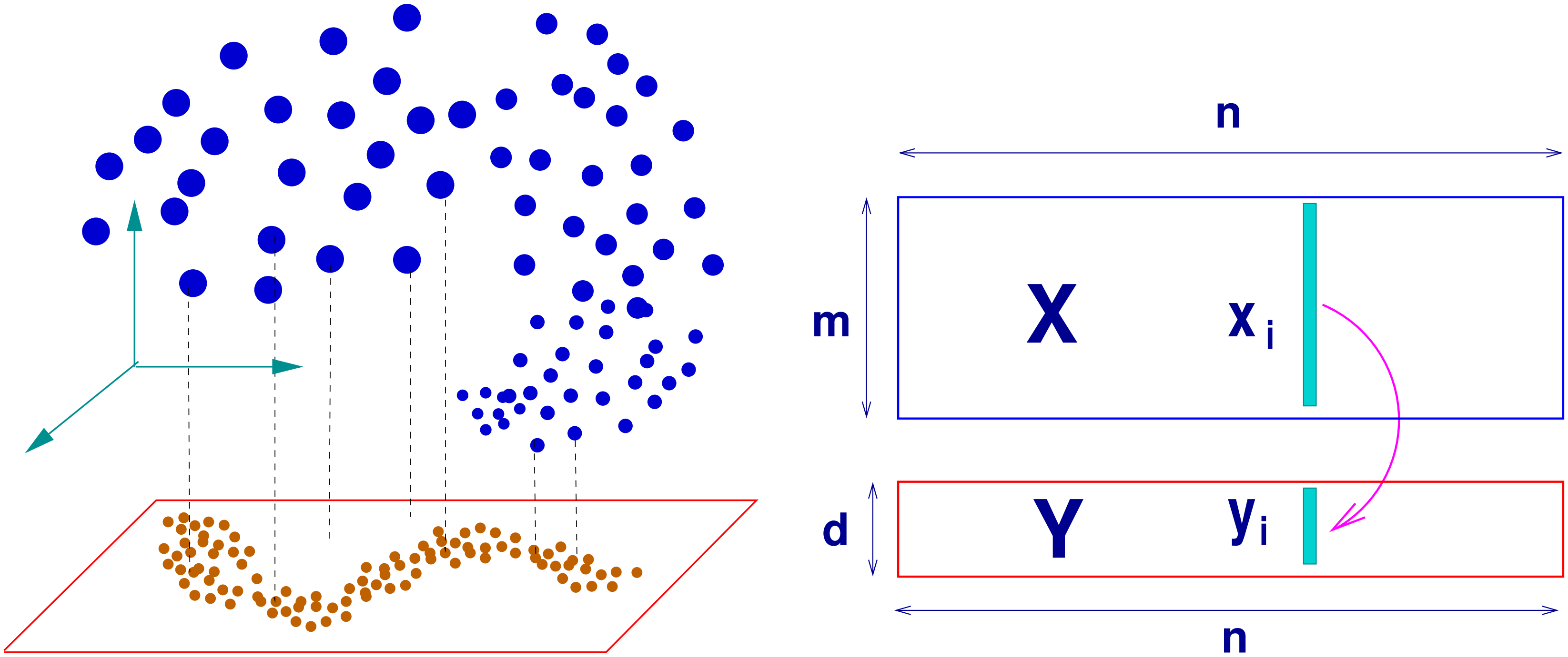
➤ Goal is not as much to reduce size (& cost) but to:

- Reduce noise and redundancy in data before performing a task [e.g., classification as in digit/face recognition]
- Discover important 'features' or 'paramaters'

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

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

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




- $\Phi$ may be linear : $y_i = W^\top x_i$, i.e., $Y = W^\top X$, ..
- ... or nonlinear (implicit).
- Mapping $\Phi$ required to: Preserve proximity? Maximize variance? Preserve a certain graph?

## *Example: Principal Component Analysis (PCA)*

In *Principal Component Analysis* $W$ is computed to maximize variance of projected data:

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

➤ Leads to maximizing

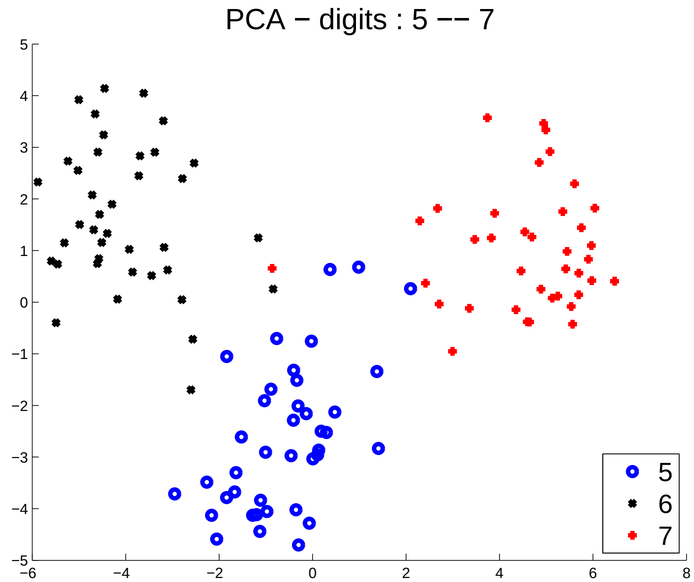
$$\text{Tr}\left[W^\top (X - \mu e^\top)(X - \mu e^\top)^\top W\right], \quad \mu = \tfrac{1}{n}\Sigma_{i=1}^{n} x_i$$

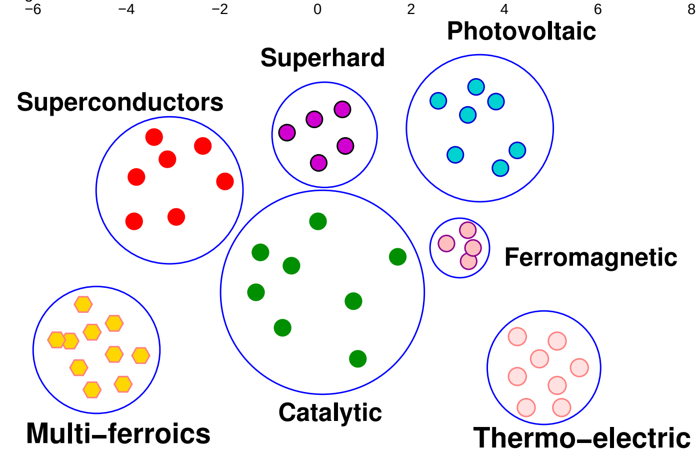
➤ Solution $W = \{$ dominant eigenvectors $\}$ of the covariance matrix $\equiv$ Set of left singular vectors of $\bar{X} = X - \mu e^\top$

# *Unsupervised learning*

*"Unsupervised learning"* : methods that do not exploit known labels

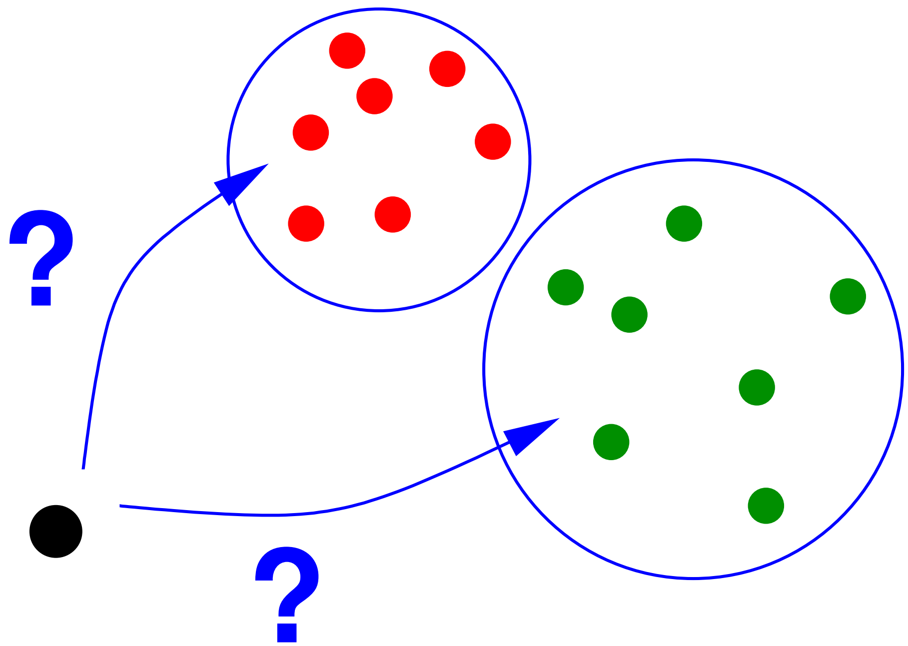
- Example of digits: perform a 2-D projection
- Images of same digit tend to cluster (more or less)
- Such 2-D representations are popular for visualization
- Can also try to find natural clusters in data, e.g., in materials
- Basic clusterning technique: K-means



## *Supervised learning: classification*

*Problem:* Given labels (say “A” and “B”) for each item of a given set, find a mechanism to classify an unlabelled item into either the “A” or the “B" class.

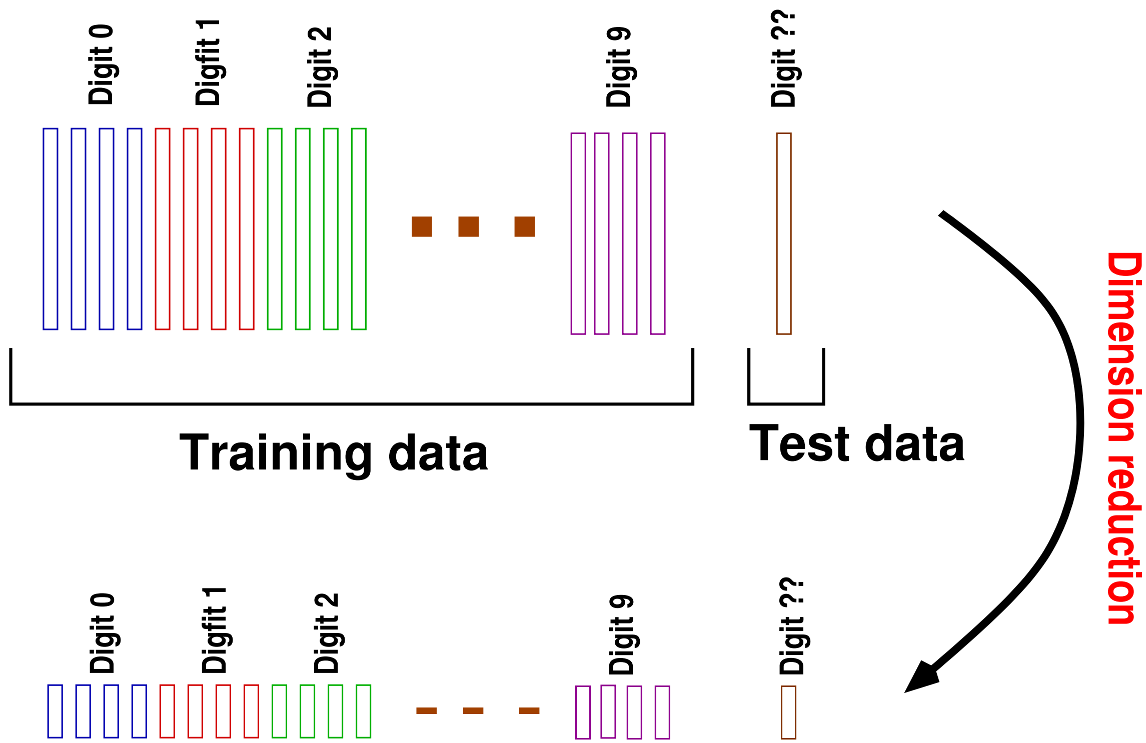



- Many applications.
- Example: distinguish SPAM and non-SPAM messages
- Can be extended to more than 2 classes.

## Supervised learning: classification

➤ Best illustration: written digits recognition example

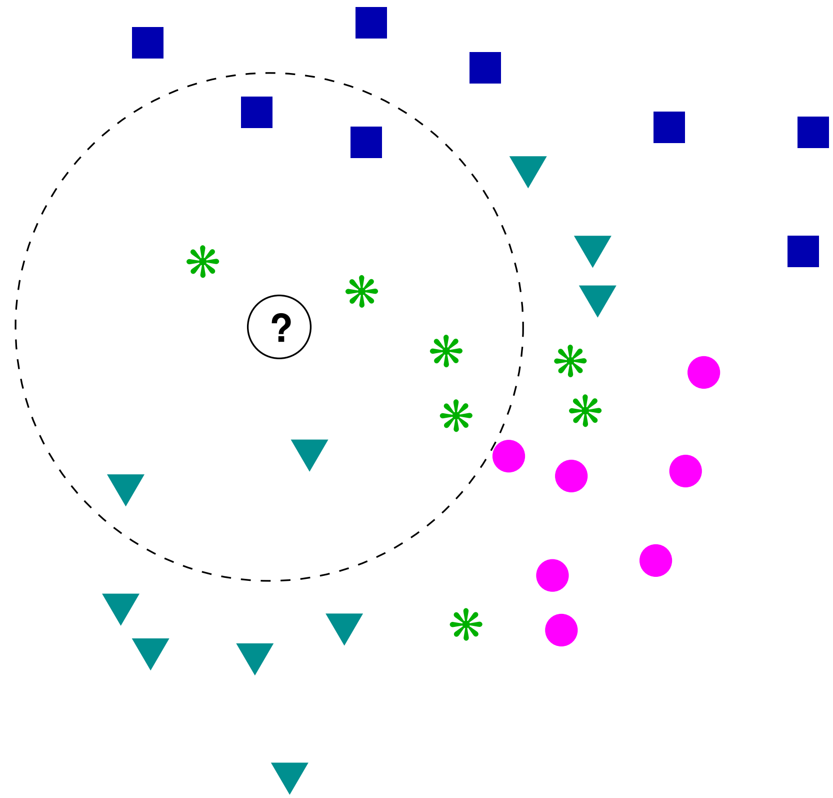
*Given:* a set of labeled samples (training set), and an (unlabeled) test image.
*Problem:* find label of test image




➤ Roughly speaking: we seek dimension reduction so that recognition is 'more effective' in low-dim. space

## *Basic method: K-nearest neighbors (KNN) classification*

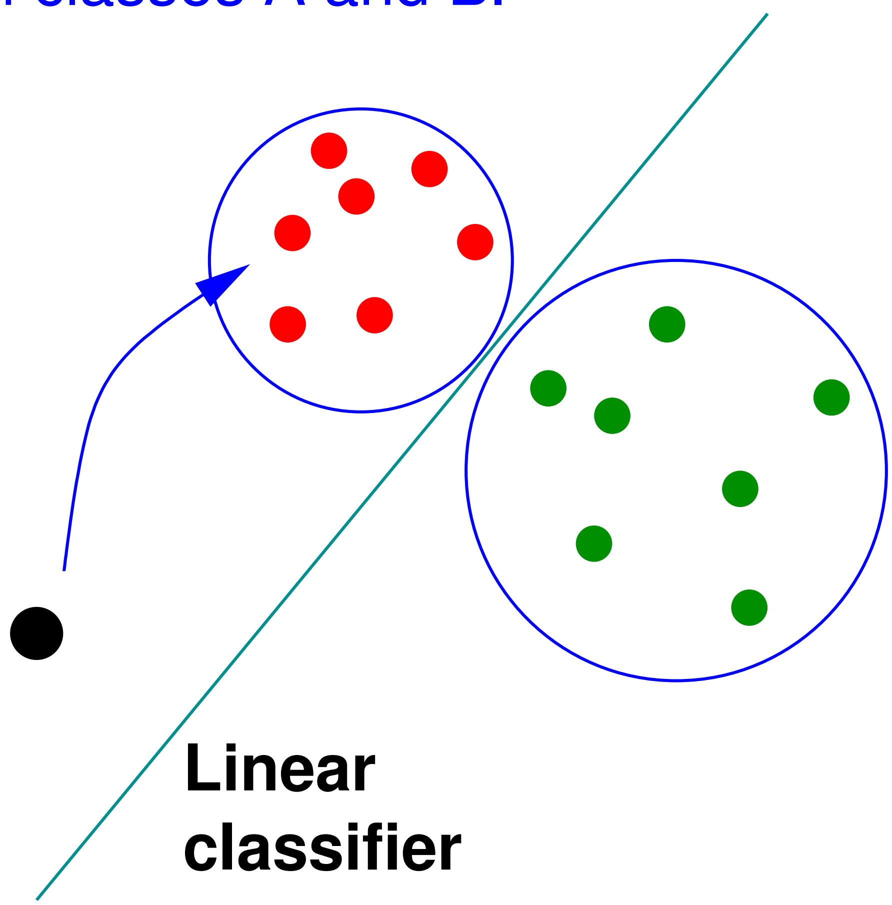
- Idea of a voting system: get distances between test sample and training samples
- Get the $k$ nearest neighbors (here $k = 8$)
- Predominant class among these $k$ items is assigned to the test sample ("$*$" here)



# MATLAB DEMO

## *Linear classifiers and Fisher's LDA*

➤ Idea for two classes: Find a hyperplane which best separates the data in classes A and B.



## *Linear classifiers*

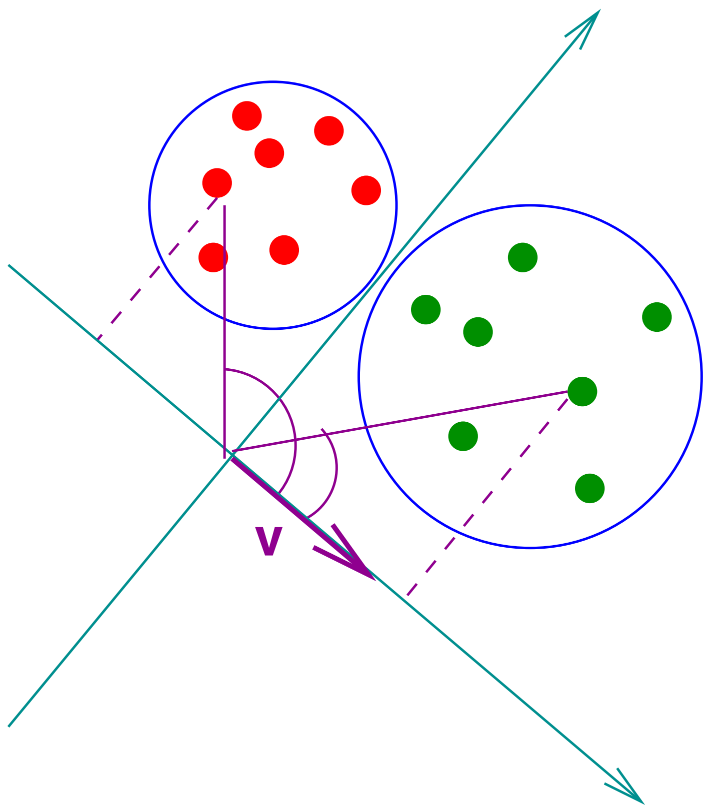
Given:
- $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ – the data matrix.
- $\boldsymbol{L} = [\boldsymbol{l}_1, \cdots, \boldsymbol{l}_n]$ – the data labels: +1 or -1.

➤ 1st Solution: Find a vector $\boldsymbol{v}$ such that $\boldsymbol{v}^T \boldsymbol{x}_i$ close to $\boldsymbol{l}_i$ $\forall \boldsymbol{i}$

➤ Common solution: (1) SVD to reduce dimension of data [e.g. 2-D]; (2) Do comparison in this space, e.g.:

*A:* $\boldsymbol{v}^T \boldsymbol{x}_i \geq 0$ , *B:* $\boldsymbol{v}^T \boldsymbol{x}_i < 0$.

[Note: $\boldsymbol{v}$ principal axis drawn below where it should be]



## *Fisher's Linear Discriminant Analysis (LDA)*

*Goal:* Use label information to define a good projector, i.e., one that can 'discriminate' well between given classes

- Define "between scatter": a measure of how well separated two distinct classes are.
- Define "within scatter": a measure of how well clustered items of the same class are.
- Objective: make "between scatter" measure large and "within scatter" small.

Idea: Find projector that maximizes the ratio of the "between scatter" measure over "within scatter" measure

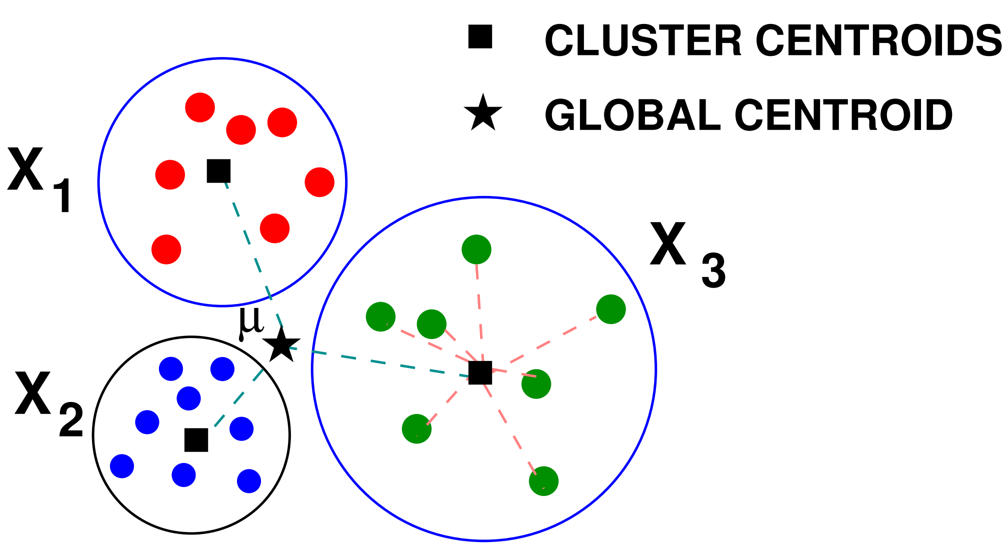
Define:

$$S_B = \sum_{k=1}^{c} n_k(\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T,$$

$$S_W = \sum_{k=1}^{c} \sum_{x_i \in X_k} (x_i - \mu^{(k)})(x_i - \mu^{(k)})^T$$

Where:

- $\mu$ = mean $(X)$
- $\mu^{(k)}$ = mean $(X_k)$
- $X_k$ = $k$-th class
- $n_k = |X_k|$



- Consider 2nd moments for a vector $a$: Project on one-dimensional space span$\{a\}$:

$$a^T S_B a = \sum_{i=1}^{c} n_k \, |a^T(\mu^{(k)} - \mu)|^2,$$

$$a^T S_W a = \sum_{k=1}^{c} \sum_{x_i \in X_k} |a^T(x_i - \mu^{(k)})|^2$$

- $a^T S_B a \equiv$ weighted variance of projected $\mu_j$'s
- $a^T S_W a \equiv$ w. sum of variances of projected classes $X_j$'s
- LDA projects the data so as to maximize the ratio of these two numbers:

$$\boxed{\max_a \frac{a^T S_B a}{a^T S_W a}}$$

- Optimal $a$ = eigenvector associated with the largest eigenvalue of: $S_B u_i = \lambda_i S_W u_i$ .

## *LDA – Extension to arbitrary dimensions*

- Criterion: maximize the ratio of two traces: $\frac{\text{Tr}\,[U^T S_B U]}{\text{Tr}\,[U^T S_W U]}$

- Constraint: $U^T U = I$ (orthogonal projector).

- Reduced dimension data: $Y = U^T X$.

*Common viewpoint:* hard to maximize, therefore ...

- ... alternative: Solve instead the ('easier') problem: $\max_{U^T S_W U = I} \text{Tr}\,[U^T S_B U]$

- Solution: largest eigenvectors of $S_B u_i = \lambda_i S_W u_i$ .

## *LDA – Extension to arbitrary dimensions (cont.)*

➤ Consider the original problem:

$$\max_{U \in \mathbb{R}^{n\times p},\ U^TU=I} \frac{\text{Tr}\,[U^TAU]}{\text{Tr}\,[U^TBU]}$$

Let $A, B$ be symmetric & assume that $B$ is semi-positive definite with rank$(B) > n-p$. Then Tr $[U^TAU]$/Tr $[U^TBU]$ has a finite maximum value $\rho_*$. The maximum is reached for a certain $U_*$ that is unique up to unitary transforms of columns.

➤ Consider the function:

$$\boxed{f(\rho) = \max_{V^TV=I} \text{Tr}\,[V^T(A-\rho B)V]}$$

➤ Call $V(\rho)$ the maximizer for an arbitrary given $\rho$.

➤ Note: $V(\rho)$ = Set of eigenvectors - not unique

- Define $\boxed{G(\rho) \equiv A - \rho B}$ and its $n$ eigenvalues:

$$\mu_1(\rho) \geq \mu_2(\rho) \geq \cdots \geq \mu_n(\rho) \; .$$

- Clearly:

$$\boxed{f(\rho) = \mu_1(\rho) + \mu_2(\rho) + \cdots + \mu_p(\rho) \; .}$$

- Can express this differently. Define eigenprojector:

$$\boxed{P(\rho) = V(\rho) V(\rho)^T}$$

- Then:

$$\begin{aligned} f(\rho) &= \mathrm{Tr}\,[V(\rho)^T G(\rho) V(\rho)] \\ &= \mathrm{Tr}\,[G(\rho) V(\rho) V(\rho)^T] \\ &= \mathrm{Tr}\,[G(\rho) P(\rho)]. \end{aligned}$$

- Recall [e.g. Kato '65] that:

$$\boldsymbol{P(\rho) = \frac{-1}{2\pi i}\int_\Gamma (G(\rho) - zI)^{-1}\, dz}$$

$\boldsymbol{\Gamma}$ is a smooth curve containing the $\boldsymbol{p}$ eivenvalues of interest and $\boldsymbol{R_\rho(z)}$ is the resolvant

$$\boldsymbol{R_\rho(z) = (G(\rho) - zI)^{-1} = (A - \rho B - zI)^{-1}.}$$

- Hence: $\boldsymbol{f(\rho) = \frac{-1}{2\pi i}\text{Tr}\int_\Gamma G(\rho)(G(\rho) - zI)^{-1}\, dz = ...}$

$$\boldsymbol{= \frac{-1}{2\pi i}\text{Tr}\int_\Gamma z(G(\rho) - zI)^{-1}\, dz}$$

- With this, can prove :

1. $f$ is a non-increasing function of $\rho$;
2. $f(\rho) = 0$ iff $\rho = \rho_*$;
3. $f'(\rho) = -\text{Tr}\left[V(\rho)^T B V(\rho)\right]$

➤ Careful when defining $V(\rho)$: define the eigenvectors so the mapping $V(\rho)$ is differentiable. But $\exists$ Differentiable branch of eigenvectors

## *Can now use Newton's method.*

$$\rho_{new} = \rho - \frac{\text{Tr}\,[V(\rho)^T(A - \rho B)V(\rho)]}{-\text{Tr}\,[V(\rho)^T BV(\rho)]} = \frac{\text{Tr}\,[V(\rho)^T AV(\rho)]}{\text{Tr}\,[V(\rho)^T BV(\rho)]}$$

- Newton's method to find the zero of $f$ $\equiv$ a fixed point iteration with:

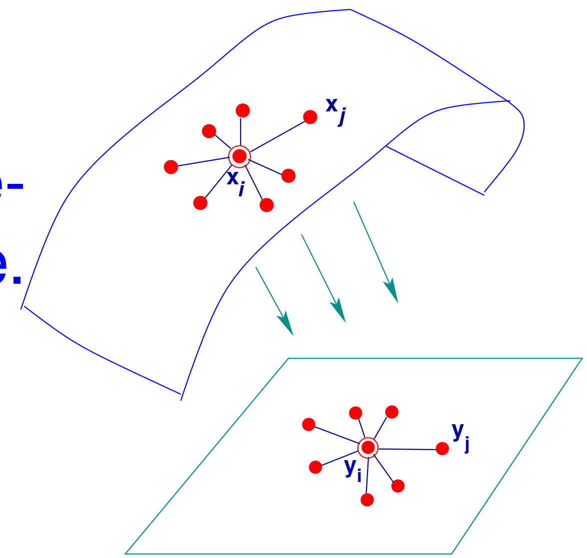
$$g(\rho) = \frac{\text{Tr}\,[V^T(\rho)AV(\rho)]}{\text{Tr}\,[V^T(\rho)BV(\rho)]}.$$

- Idea: Compute $V(\rho)$ by a Lanczos-type procedure
- Note: Standard problem - [not generalized] $\longrightarrow$ inexpensive
- See T. Ngo, M. Bellalij, and Y.S. 2010 for details

# GRAPH-BASED TECHNIQUES

## Graph-based methods

- Start with a graph of data. e.g.: graph of $k$ nearest neighbors (k-NN graph)




**Want:** Perform a projection which preserves the graph in some sense.

- Define a *graph Laplacean:*

$$L = D - W$$

e.g.,: $w_{ij} = \begin{cases} 1 & \text{if } j \in Adj(i) \\ 0 & \text{else} \end{cases} \qquad D = \operatorname{diag}\left[d_{ii} = \sum_{j \neq i} w_{ij}\right]$

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

## *Example: The Laplacean eigenmaps approach*

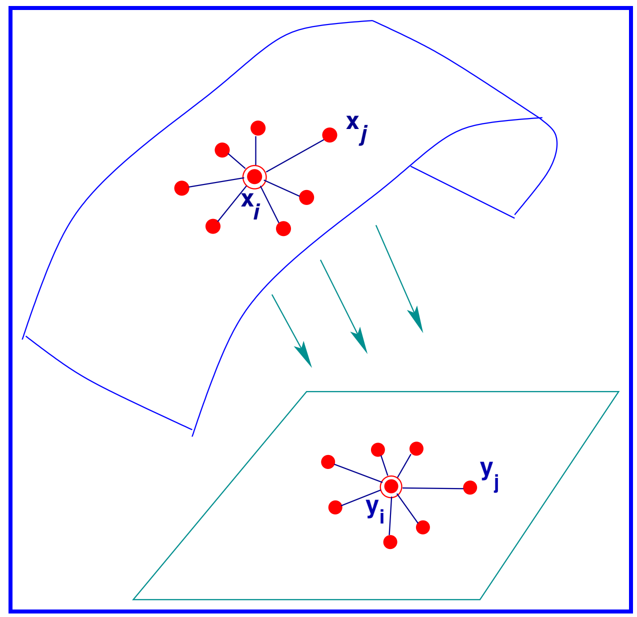
Laplacean Eigenmaps [Belkin-Niyogi '01] *minimizes*

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

*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

➤ Leads to $n \times n$ sparse eigenvalue problem [In 'sample' space]



- Problem translates to:

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

- Solution (sort eigenvalues increasingly):

$$(D-W)u_i = \lambda_i D u_i \; ; \quad y_i = u_i^\top ; \quad i = 1, \cdots, d$$

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

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

$$\boxed{(I-W)u_i = \lambda_i u_i \; ; \quad y_i = u_i^\top ; \quad i = 1, \cdots, d}$$

## *Implicit vs explicit mappings*

- In PCA the mapping $\Phi$ from high-dimensional space ($\mathbb{R}^m$) to low-dimensional space ($\mathbb{R}^d$) is explicitly known:

$$\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{x}) \equiv \boldsymbol{V}^T \boldsymbol{x}$$

- In Eigenmaps and LLE we only know

$$\boldsymbol{y_i} = \boldsymbol{\phi}(\boldsymbol{x_i}), \boldsymbol{i} = \boldsymbol{1}, \cdots, \boldsymbol{n}$$

- Mapping $\phi$ is complex, i.e.,
- Difficult to get $\phi(x)$ for an arbitrary $x$ not in the sample.
- Inconvenient for classification
- "The out-of-sample extension" problem

## *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 LEE) *but* with the constraint $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$
- Problem solved to obtain mapping:

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

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

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

## A unified view: two types of problems encountered

*First :* $Y$ obtained from computing eigenvectors

➤ LLE, Eigenmaps, ...

$$\min_{\begin{cases} Y \in \mathbb{R}^{d \times n} \\ YY^\top = I \end{cases}} \mathrm{Tr}\left[ Y M Y^\top \right]$$

*Second:* Low-dim. data: $Y = V^\top X$

➤ $G$ == identity, or $XDX^\top$, or $XX^\top$

$$\min_{\begin{cases} V \in \mathbb{R}^{m \times d} \\ V^\top G V = I \end{cases}} \mathrm{Tr}\left[ V^\top X M X^\top V \right]$$

*Observation:* 2nd is just a projected version of the 1st.

* Joint work with Efi Kokiopoulou and J. Chen

## A unified view: two types of problems

➤ In essence we select two matrices

- $A$ : represents a similarity, distance
- $B$ : represents dissimilarity, separate groups, ...

**Want:** Projected data $Y$ such that $\mathrm{Tr}\, Y^T A Y$ is small while $\mathrm{Tr}\, Y^T B Y$ is kept large (or normalized).

➤ Encapsulates: graph partitioning, LDA, PCA, ... (almost everything!)

➤ Can select $A$, $B$ from 'local' information: kNN graphs

➤ Can select $A$, $B$ from 'global' information: use all of data $X$: LLE, ONPP, PCA, ...

| Method | Object. (min) | Constraint |
|---|---|---|
| LLE | Tr $[Y(I - W^T)(I - W)Y^T]$ | $YY^T = I$ |
| Eigenmaps | Tr $[Y(D - W)Y^T]$ | $YDY^T = I$ |
| PCA/MDS | Tr $[-V^T X(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X^T V]$ | $V^T V = I$ |
| LPP | Tr $[V^T X(D - W)X^T V]$ | $V^T X D X^T V = I$ |
| OLPP | Tr $[V^T X(D - W)X^T V]$ | $V^T V = I$ |
| NPP | Tr $[V^T X(I - W^T)(I - W)X^T V]$ | $V^T X X^T V = I$ |
| ONPP | Tr $[V^T X(I - W^T)(I - W)X^T V]$ | $V^T V = I$ |
| LDA | Tr $[V^T X(I - H)X^T V]$ | $V^T X X^T V = I$ |
| Spect. Clust. (ratio cut) | Tr $[Z^T(D - W)Z]$ | $Z^T Z = I$ |
| Spect. Clust. (normalized cut) | Tr $[Z^T(D - W)Z]$ | $Z^T D Z = I$ |

- See: (survey paper)

E. Kokiopoulou, J. Chen, Y. S., " *Trace optimization and eigenproblems in dimension reduction methods*," Numerical Linear Algebra with Applications; vol. 18, pages 565-602 (2011).

## *Tests*

Notation for the various methods tested:

- LDA and LDE == methods that rely on the eigenvectors of $B^{-1}A$. LDA : non-local matrices, LDE : local matrices.
- LDA-ITER and LDE-ITER == methods that optimize the trace ratio [Newton scheme]. Matrices $A$ and $B$ are ‘non-local’ for LDA-ITER and ‘local’ for LDE-ITER.

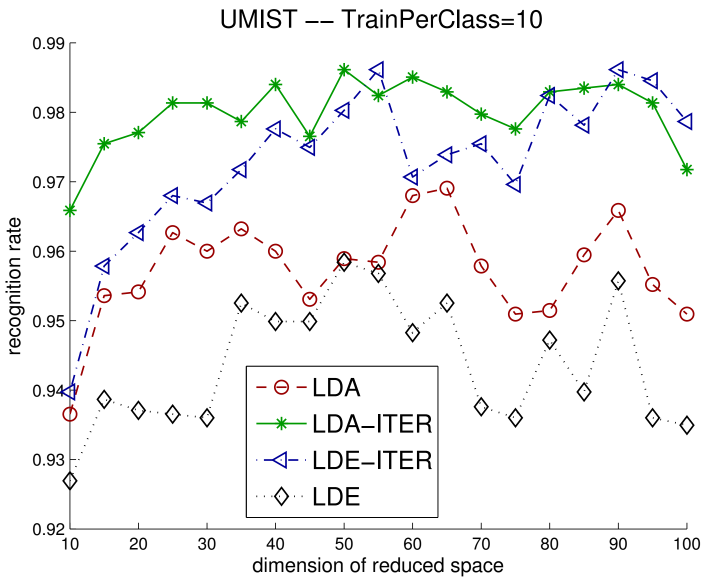
- First: compare trace ratios

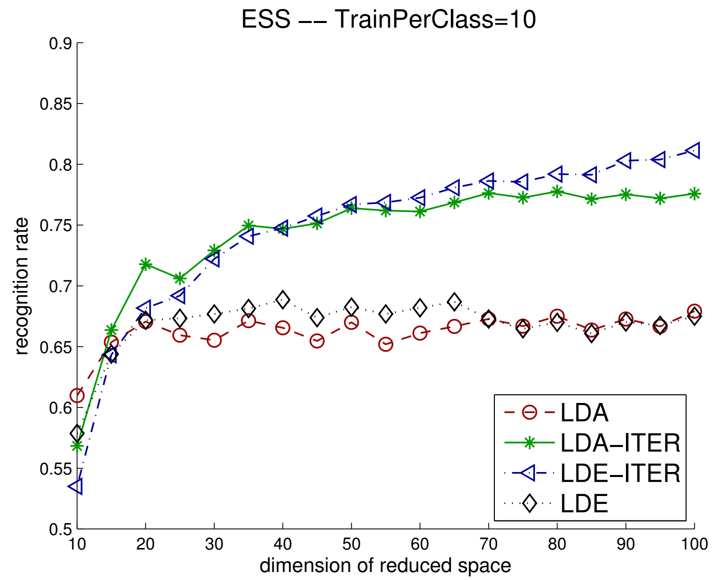
*Values of* $Tr[V^TAV]/Tr[V^TBV]$.

| Dims | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| LDE-ITER | 32.46 | 19.37 | 13.67 | 11.71 | 28.29 | 16.96 |
| LDE | 23.54 | 13.55 | 9.46 | 8.00 | 20.08 | 12.74 |

- Significantly bigger ratios with trace optimization

# ➤ Face recognition: a couple of comparisons [from paper]



## References:

[1] T. T. Ngo and M. Bellalij and Y. Saad, *“The Trace Ratio Optimization Problem”*, SIAM review, vol.54, (3), pp 545–569, (2012)

[2] T. T. Ngo and M. Bellalij and Y. Saad, *“The Trace Ratio Optimization Problem for dimensionality reduction"*, SIMAX, vol. 31,pp. 2950-2971.

# A FEW EXTENSIONS

## *Extension 1: Application to Hypergraph clustering*

- See: *Context-Aware Hypergraph Construction for Robust Spectral Clustering* Xi Li, Weiming Hu, Chunhua Shen, Anthony Dick, and Zhongfei Zhang, IEEE TKDE,
- Issue: construct equivalent of kNN graph + do clustering on hypergraph.
- Step 1: construct a similarity matrix $S$ that captures similarity between groups of vertices.
- $D \equiv diag(Se), Q = D - S$.
- *Step 2:* Trace ratio maximization

$$\boxed{\max_{P s.t. P^T P = I_k} \frac{\text{Tr}\,[P^T S P]}{\text{Tr}\,[P^T Q P]}}$$

## *Extension 2: Application to cell Formation Problem*

Date: Sat, 05 Jan 2013 02:44:59 +0100
From: Mohammed Bellalij <mohammed.bellalij@univ-valenciennes.fr>
To: Yousef Saad <saad@cs.umn.edu>
Subject: Meilleurs voeux pour 2013 et nouveau probleme de trace

Salut Yousef,

(....)

J'ai recemment trouve un problème d'optimisation discrète (the cell formation problem - cellular manufacturing) dans le domaine de la conception des cellules de production qui peut s'ècrire sous forme de rapport de traces (singulières). Sa forme relaxée est de maximiser Tr[XTAY ]/($\mu$ +Tr[XTBY ]) sous les contraintes X et Y matrices orthogonales avec p colonnes . J'y travaille en ce moment et dès que je rédigerai une note, je te l'enverrai.

## THE SINGULAR TRACE RATIO OPTIMIZATION PROBLEM

M. BELLALIJ* ‡, S. HANAFI† ‡, AND Y. SAAD§

**Abstract.** This paper considers the problem of optimizing the ratio $\frac{Tr[X^T AY]}{\mu + Tr[X^T BY]}$ over all orthogonal ma
and $Y$ with $k$ columns, where $A, B$ are two rectangular matrices................

**Key words.**

**1. Introduction.** Throughout this paper, we use the following mathematical notations.
vill assume that $n$ and $m$ are naturel numbers and $m < n$. Let $\mathbb{R}$ denote the set of all real num
$_p \in \mathbb{R}^{p\times p}$ denotes the identity matrix. Let us denote by $O_{p,q}$ the null matrix of $\mathbb{R}^{p\times q}$. The supers
stands for the transposition. The trace of a square matrix $M$, i.e., the sum of f the diagonal en
f $M$, is denoted by $Tr(M)$. If $u \in \mathbb{R}^p$(resp. ?), then diag$(u)$ is an $p \times p$ diagonal matrix
on the main diagonal. We will us the symbol $\mathcal{O}_{p,q}$ to denote the set of orthogonal matric
$\mathbb{R}^{p\times q}$. $\mathcal{O}_{p,q} = \{M \in \mathbb{R}^{p\times q} \ : \ M^T M = I_q\}$ is often referred to as the Stiefel manifold.
The rest of this paper is organized as follows.

**2. Eigenvalues and trace maximization revisited.** Given a symmetric matrix $A$ c
nension $n \times n$ with spectrum $\sigma(A) = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$, there is an orthogonal matrix $U$
$u_1, u_2, \cdots, u_n] \in \mathbb{R}^{n\times n}$ $(U^T U = UU^T = I_n)$ such that $A = UDU^T$, where $D =$diag$(\lambda_1, \lambda_2, \cdots,$
ve assume that eigenvalues are labeled decreasingly, the notation $U_k = [u_1, u_2, \cdots, u_k]$ $(k \leq$
s the orthogonal matrix consisting of the eigenvectors corresponding to the first $k$ eigenva
$\lambda_1, \lambda_2, \cdots, \lambda_k$. Let $V$ denote an arbitrary orthogonal matrix of dimension $n \times p$, it is known
he trace of $V^T AV$ reaches its maximum (resp., minimum) when $V$ is an orthogonal basis o
eigenspace of $A$ associated with the $p$ algebraically largest (resp., smallest) eigenvalues. This r
s seldom stated in standard textbooks, but it is an immediate consequence of the courant-F
haracterization; see, e.g. [2, 3].
Here is another way to proving this result stated in the following theorem.

THEOREM 2.1.

$$\max_{\begin{cases} V^T V = I_k \\ V \in \mathbb{R}^{n\times k} \end{cases}} Tr[V^T AV] = \sum_{i=1}^{k} \lambda_i \ = Tr[U_k^T \ A \ U_k] \ .$$

# Generalized trace ratio optimization and applications

Mohammed Bellalij

University of Valenciennes, France

MOCASIM, 19-22 November 2014
Marrakech

# Cell formation problem

- Application : Group technology or cellular manufacturing
- System : machines and parts interacting
- Partition the system into subsystems to maximize efficiency :
  - Interactions between the machines and the parts within a subsystem are maximized
  - Interactions between the parts of other systems are reduced as much as possible

| Machines \ Parts | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ |
|---|---|---|---|---|---|
| $M_1$ | 1 | 0 | 0 | 1 | 0 |
| $M_2$ | 0 | 1 | 1 | 0 | 1 |
| $M_3$ | 1 | 0 | 0 | 1 | 0 |
| $M_4$ | 0 | 1 | 1 | 0 | 1 |
| $M_5$ | 1 | 0 | 0 | 1 | 0 |

| Machines \ Parts | $P_2$ | $P_3$ | $P_5$ | $P_1$ | $P_4$ |
|---|---|---|---|---|---|
| $M_1$ | 1 | 1 | 1 | 0 | 0 |
| $M_2$ | 1 | 1 | 1 | 0 | 0 |
| $M_3$ | 0 | 0 | 0 | 1 | 1 |
| $M_4$ | 0 | 0 | 0 | 1 | 1 |
| $M_5$ | 0 | 0 | 0 | 1 | 1 |

## *The Cell Manufacturing Formation Problem (MCFP)*

*Goal:* Identify families of parts and group of machines on which these parts are to be processed.

*"If the number, types, and capacities of production machines, the number and types of parts to be manufactured, and the routing plans and machine standards for each part are known, which machines and their associated parts should be grouped together to form manufacturing cells?" (Wu and Salvendy, 1993).*

➤ Very rich literature.. Rich variety of methods [metaheuristics, PCA, Simulated annealing, graph partitioning, ....]

## *Leads to: Singular-value TRace OPtimization (STROP)*

- Mohammed formulated the problem ...
- .. and an algorithm for solving it.
- Note: (1) Modified ratio;
  (2) SVD is now needed instead of eigen-decomposition
- SVD analogue of the trace ratio problem..
- Wrote a short note – and presented the work in MOCASIM-14

The problem consists of finding the matrices $X$ and $Y$ which solve the following discrete generalized trace ratio problem :

## Discrete STROP

$$\begin{cases} maximize & \frac{Tr(X^T AY)}{1+Tr(X^T BY)} \\ s.t. & X = (x_{ik}) \in \{0,1\}^{M\times C}, Y = (y_{jk}) \in \{0,1\}^{P\times C} \\ & \sum_{k=1}^{C} x_{ik} = 1; i = 1,\dots,M \text{ and } \sum_{i=1}^{M} x_{ik} \geq 1; k = 1,\dots,C \\ & \sum_{k=1}^{C} y_{jk} = 1; j = 1,\dots,P \text{ and } \sum_{j=1}^{P} y_{jk} \geq 1; k = 1,\dots,C. \end{cases}$$

To obtain an optimal solution we would need first to maximize the relaxed problem

$$\max_{(X,Y)\in\mathcal{O}} \frac{Tr[X^T AY]}{1 + Tr[X^T BY]}.$$

# STROP

Given real matrices $A$ and $B$ of dimension $m \times n$.
Let $\mathcal{O}_{p,k} = \{Z \in R^{p\times k} \ : \ Z^T Z = I_k\}$ and $\mathcal{O}_{m,n,k} = \mathcal{O}_{m,k} \times \mathcal{O}_{n,k}$.
Goal : Find a pair of orthogonal matrices $X_* \in \mathcal{O}_{m,k}$ and $Y_* \in \mathcal{O}_{n,k}$ optimal solution of the problem :

$$\max_{(X,Y)\in\mathcal{O}_{m,n,k}} \frac{Tr[X^T AY]}{1 + Tr[X^T BY]}.$$

We will assume that the matrix $B$ verifies $1 + Tr[X^T BY] > 0$ for any $(X, Y) \in \mathcal{O}_{m,n,k}$.

# Existence and Uniqueness of a Solution of STROP

- The problem STROP admits a finite maximum value $\rho_*$. It is reached for certain (nonunique) orthogonal matrices : $X_*$ and $Y_*$ .
  Thanks to the cyclic property of the trace, any simultaneous orthogonal transformation of the columns of $X_*$ and $Y_*$ will not change the objective function ( $U = X_* R, V = Y_* R$ for any regular matrix $R \in \mathbb{R}^{k\times k}$ such that $R^{-1} = R^T$ ).
- We have $Tr[X^T(A - \rho_* B)Y] \leq \rho_*$ because $1 + Tr[X^T BY] > 0$. Therefore, we have the following necessary condition for the triplet $\rho_*, X_*$ and $Y_*$ to be optimal :

$$\max_{(X,Y)\in\mathcal{O}_{m,n,k}} Tr[X^T(A - \rho_* B)Y] = Tr[X_*^T(A - \rho_* B)Y_*] = \rho_*.$$

- Let $g(\rho) = \max_{(X,Y)\in\mathcal{O}_{m,n,k}} Tr[X^T(A - \rho\ B)Y]$.
  Then, it is equivalent to solve the scalar equation $g(\rho) = \rho$.

# Properties of $f(\rho) = g(\rho) - \rho$.

- Evaluating $g(\rho)$ consists in computing the left and right singular vectors $X(\rho)$ and $Y(\rho)$ associated with the $k$ largest singular values of $A - \rho B$. So, $g(\rho) = Tr[X^T(\rho)(A - \rho\ B)Y(\rho)]$.
- Under the assumption that $B$ verifies $1 + Tr[X^T BY] > 0$, we have
  1. $f$ is differentiable at $\rho$ with $\frac{df(\rho)}{d\rho} = -Tr\,[X(\rho)^T B\ Y(\rho)] - 1$ and $f$ is a strictly decreasing function.
  2. $f$ is convex.
  3. $f(\rho) = 0$ iff $\rho = \rho_*$.

# Fractional iteration of the Newton-approximation-formula

- Newton's method to approximate the unique fixed point of $g$ :

$$\rho_{new} = \rho - \frac{Tr[X^T(\rho)(A-\rho\ B)Y(\rho)] - \rho}{-Tr[X^T(\rho)\ B\ Y(\rho)] - 1} = \frac{Tr[X^T(\rho)\ A\ Y(\rho)]}{1 + Tr[X^T(\rho)\ B\ Y(\rho)]}.$$

- The Newton-SVD algorithm includes the following three iterative steps :
  1. Compute the trace ratio $\rho = \dfrac{Tr[X^T AY]}{1 + Tr[X^T BY]}$ ;
  2. Run the SVD algorithm to compute the $k$ largest singular values of $A - \rho B$ as well as their associated singular eigenvectors $X$ and $Y$ ;
  3. Repeat the above two steps until convergence.

## *Conclusion*

- Many interesting applications of trace ratio optimization problems.

- Some recent following on our work - for some applications - More to come?

- Mohammed started to work on some nice extensions ..

  .. But his work was left unfinished

- Mohammed was a very amiable, human being
- The example to retain from his character is to

*BE KIND to OTHERS*

*BE COOL and RELAX*

and

*BE HAPPY and OPTIMIST*