# Generalized trace ratio optimization and applications

Mohammed Bellalij, Saïd Hanafi, Rita Macedo and Raca Todosijevic

University of Valenciennes, France

PGMO Days, 2-4 October 2013
ENSTA ParisTech

# Outline



Fondation mathématique Jacques Hadamard

FMJH

UVHC UNIVERSITE DE VALENCIENNES ET DU HAINAUT - CAMBRESIS

# Outline

# Trace Ratio Optimization Problem

$$(TROP) \begin{cases} maximize & \dfrac{Tr(X^T AX)}{Tr(X^T BX)} \\ s.t. & X^T CX = I \\ & X \in \Omega \end{cases}$$

where $A, B, C$ are matrices with appropriate dimensions, and $I$ identity matrix. $Tr(.)$ is the trace of a square matrix.

- Continuous case : $\Omega = \mathbb{R}^{m\times d}$ or $[0,1]^{m\times d}$ or $[-1,1]^{m\times d}$
- Binary case : $\Omega = \{0,1\}^{m\times d}$ or $\{-1,1\}^{m\times d}$
- TROP is a Hard Problem

→ There is no loss of generality in assuming that $C = I$.

# Discrete Trace Ratio Optimization Problem

- Discrete TROP is more difficult than the continuous case.
- Continuous TROP is a relaxation of Discrete TROP.
- Continuous TROP provides an upper bound on the optimal value of the discrete TROP.
- Linear algebra techniques are powerful for solving continuous TROP : much lower computational cost and guaranteed approximation quality.

# Outline

Many real-world application problems can be modeled as trace ratio or more generally as sum of the trace ratio optimization problem. It can arise in :

- The Fisher discriminant analysis in pattern recognition (data mining, machine learning)
- The downlink of a multi-user MIMO system (wireless communication and signal processing)
- The cell formation problem (Cellular manufacturing).

Data Mining is the process of analyzing data in order to extract useful knowledge such as :

- Clustering : Unsupervised learning
- Classification : Supervised learning
- Feature selection : Suppression of irrelevant or redundant features.

# Dimensionality reduction : Major tool of Data Mining

- Most machine learning & data mining techniques may not be effective for high-dimensional data
- Dimensionality reduction plays a fundamental role in data mining :
  - Map the data in high-dimensional space to a low-dimensional space $\Rightarrow$ Reduce noise and redundancy in data before performing a task.
  - Dimensionality reduction usually entails embedding the data in a space of reduced dimension that preserves most of its interesting details.

# Outline

# Some Techniques

Several techniques for dimensionality reduction end with solving a TROP.

- Fisher's Linear Discriminant Analysis
- Support Vector Machines or Kernel Methods
- Graph embedding, and so on.

# Fisher's Linear Discriminant Analysis (LDA)

- LDA attempts to find linear projections of the data that are optimal for discrimination
- The intra-cluster separability : a measure of how well separated two distinct classes are
- The inter-cluster compactness : a measure of how well clustered items of the same class are
- Goal : Search for a transformation matrix that maximizes intra-cluster separability and at the same time minimizes inter-cluster compactness
- The natural model for these dual objectives is to optimize a trace ratio problem

# LDA : Objective deduction

- Let $\mu$ = mean of $X$, and $\mu^{(k)}$ = mean of the $k-$th class (of size $n_k$, $k = 1,\ldots,c$). Define
$A = \sum_{k=1}^{c} n_k(\mu^{(k)}-\mu)(\mu^{(k)}-\mu)^T$, $B = \sum_{k=1}^{c}\sum_{x_i}(x_i-\mu^{(k)})(x_i-\mu^{(k)})^T$.
- Criterion : maximize the ratio of two traces $\frac{Tr(X^TAX)}{Tr(X^TBX)}$
- Constraint : $X^TX = I$ (orthogonal projection)
- …Alternative : Solve the "easier" problem instead :

$$\max_{X^TBX=I} Tr(X^TAX)$$

- Solution : largest generalized eigenvectors $X_i^*$ i.e.

$$AX_i^* = \lambda_i BX_i^*$$

- However, its solution may deviate from the original objective

# Outline

# Discriminative hypergraph partitioning

1. Hypergraph partitioning seeks for an optimal partitioning of the vertex set $V$ of a given graph $G = (V, A)$ into $\kappa$ disjoint subsets
2. The discriminative hypergraph partitioning criterion considers both the inter-cluster compactness and the intra-cluster separability, using a similarity matrix $S$, and thus aims to solve

$$\max g(X) = \frac{1}{\kappa} \sum_{j=1}^{\kappa} \frac{X_j^T S X_j}{X_j^T Q X_j}$$

$$\textbf{s.t. } X \in \{0,1\}^{N \times \kappa}, X 1_\kappa = 1_N$$

where $Q = D - S$, with $D$ is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$, and $X_j$ is the $j$-th column of $X$.

$$\max g(X) = \frac{1}{\kappa}\sum_{j=1}^{\kappa} \frac{[X_j(X_j^T X_j)^{-\frac{1}{2}}]^T S[X_j(X_j^T X_j)^{-\frac{1}{2}}]}{[X_j(X_j^T X_j)^{-\frac{1}{2}}]^T Q[X_j(X_j^T X_j)^{-\frac{1}{2}}]}$$

$$\textbf{s.t. } X \in \{0,1\}^{N \times \kappa}, X1_\kappa = 1_N$$

Li et al (2013) show that this problem is equivalent to

$$\max g(X) = \frac{1}{\kappa}\sum_{j=1}^{\kappa} \frac{P_j^T S P_j}{P_j^T Q P_j} \quad \textbf{s.t. } P^T P = I_\kappa \tag{1}$$

where $P = (P_1 P_2 \dots P_k) = X(X^T X)^{-\frac{1}{2}}$.

- The problem (1) consists in maximizing the sum of the trace ratio ( specifically : sum of the generalized Rayleigh quotients of the pencil $(S, Q)$).
- Problem difficult to solve : Existence of multiple local non-global maxima.

So, (1) is approximated as

$$\max f(P) = \frac{1}{\kappa} \frac{\sum_{j=1}^{\kappa} P_j^T S P_j}{\sum_{j=1}^{\kappa} P_j^T Q P_j} = \frac{tr(P^T SP)}{tr(P^T QP)} \quad \textbf{s.t. } P^T P = I_\kappa$$

which is a continuous trace ratio problem.

A new generalized (singular) trace ratio problem :

$$(STROP)\begin{cases} maximize & \dfrac{Tr(X^T AY)}{1+Tr(X^T BY)} \\ s.t. & X \in \Omega \\ & Y \in \Delta \end{cases} \quad (2)$$

We show that the cell formation problem can be formulated as (2).

# 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 |

# Problem formulation : Singular Trace Ratio

$I$ : set of $M$ machines, $J$ : set of $P$ parts, $K$ : set of $C$ cells

$$a_{ij} = \begin{cases} 1, & \text{if machine } i \text{ processes part } j \\ 0, & \text{otherwise} \end{cases}$$

## Measure of autonomy

$$Eff = \frac{a - a_1^{Out}}{a + a_0^{In}}$$

$a = \sum_{i=1}^{M} \sum_{j=1}^{P} a_{ij}$ : total number of entries equal to 1 in matrix $A$

$a_1^{Out}$ :number of exceptional elements

$a_0^{In}$ : number of zero entries in the diagonal blocks

$\Rightarrow$ Objective : Maximize $Eff$

*Consider*

$$x_{ik} = \begin{cases} 1, & \text{if machine } i \text{ belongs to cell } k \\ 0, & \text{otherwise} \end{cases}, \quad \forall i = 1,\dots,M,\ k = 1,\dots,C$$

$$y_{jk} = \begin{cases} 1, & \text{if part } j \text{ belongs to cell } k \\ 0, & \text{otherwise} \end{cases}, \quad \forall j = 1,\dots,P,\ k = 1,\dots,C$$

Some constraints :

- Each machine and each part assigned to exactly one cell :

$$\sum_{k=1}^{C} x_{ik} = 1\,(i = 1,\dots,M) \text{ and } \sum_{k=1}^{C} y_{jk} = 1\,(j = 1,\dots,P)$$

- At least one machine and one part in each cell :

$$\sum_{i=1}^{M} x_{ik} \geq 1\,(k = 1,\dots,C) \text{ and } \sum_{j=1}^{P} y_{jk} \geq 1\,(k = 1,\dots,C)$$

*We verify that*

$$a_1^{Out} = a - \sum_{k=1}^{C}\sum_{i=1}^{M}\sum_{j=1}^{P} a_{ij} x_{ik} y_{jk} \qquad a_0^{In} = \sum_{k=1}^{C}\sum_{i=1}^{M}\sum_{j=1}^{P} (1 - a_{ij}) x_{ik} y_{jk}$$

## Objective function

$$Eff = \frac{\sum_{k=1}^{C}\sum_{i=1}^{M}\sum_{j=1}^{P} a_{ij}x_{ik}y_{jk}}{a+\sum_{k=1}^{C}\sum_{i=1}^{M}\sum_{j=1}^{P}(1-a_{ij})x_{ik}y_{jk}}$$

The cell formation problem may be formulated as generalized (singular) trace ratio problem

$$(STROP)\begin{cases} Maximize & \frac{Tr(X^TA'Y)}{1+Tr(X^TB'Y)} \\ s.t. & X \in \Omega \\ & Y \in \Delta \end{cases}$$

Remarks :

- $A' = (\frac{a_{ij}}{a})$ and $B' = (\frac{1-a_{ij}}{a})$.
- $X \in \Omega$ and $Y \in \Delta \Rightarrow X^TX$ and $Y^TY$ : Regular diagonal matrices.

Another Generalized TROP
Maximization of the Sum of the Trace Ratio on the Stiefel Manifold

L.H. Zhang, R.C. Li 2013

$$(GTROP)\begin{cases} maximize & \frac{Tr(X^T AX)}{Tr(X^T BX)} + Tr(X^T CX) \\ s.t. & X^T X = I \ (I : \text{identity matrix of size } d) \\ & X \in \mathbb{R}^{m\times d} \ (d < m). \end{cases}$$

# Outline

# Closed form solution

Given a symmetric matrix $A$ of dimension $m \times m$ and an arbitrary unitary matrix $X$ of dimension $m \times d$

- Standard eigenvalue problem
  $\max\{Tr(X^T AX) : X^T X = I, X \in \mathbb{R}^{m\times d}\} = \lambda_1 + \cdots + \lambda_d$.
  The eigenvalues $\lambda_1, \cdots, \lambda_d$ are labeled decreasingly
  $X_*$ is an orthogonal basis of the eigenspace of $A$ associated with the $d$ algebraically largest eigenvalues.
- Generalized eigenvalue problem
  Assuming that $B$ positive definite, we have
  $$\max_{\begin{cases} X \in \mathbb{R}^{n\times d} \\ X^T BX = I \end{cases}} Tr(X^T AX) = Tr(X_*^T AX_*) = \lambda_1 + \cdots + \lambda_d$$
  (Eigenvalues labeled decreasingly for the generalized problem $Aw = \lambda Bw$ , $X_*$ $\rightarrow$ eigenvectors associated with the first $d$ eigenvalues, with $X_*^T BX_* = I$)

Fondation mathématique FMJH Jacques Hadamard

UVHC Université de Valenciennes et du Hainaut-Cambrésis

# Existence and Uniqueness of a Solution of TROP

- The Trace Ratio Optimization Problem (TROP) does not have a closed-form solution.
- Trace ratio vs. Ratio trace : Often TROP is simplified into a more accommodating one :

$$\max_{X^T X = I} Tr((X^T B X)^{-1}(X^T A X)).$$

  The obtained solution does not necessarily best maximize the corresponding trace ratio problem.
- TROP admits a finite maximum value $\rho_*$ : It does not have a local non-global solution.
  $\rho_*$ is reached for certain (nonunique) orthogonal matrices $X_*$.
  (T. T. Ngo, M. B. and Y. Saad 2012)

# From trace ratio to trace difference

Solving the trace ratio problem is equivalent to finding the solution of the scalar equation $f(\rho) = 0$, where

$$f(\rho) = \max_{X^T X = I} Tr(X^T (A - \rho B) X)$$

## Proposition

1. $f$ is a strictly decreasing function of $\rho$
2. $f(\rho) = 0$ iff $\rho = \rho_*$
3. $f'(\rho) = -Tr(X(\rho)^T B X(\rho))$

So, finding the optimal solution → a search for the unique root of $f(\rho)$.

# Newton-Lanczos algorithm for TROP

- Select initial $m \times d$ unitary matrix $X$
- compute $\rho = \dfrac{Tr(X^T AX)}{Tr(X^T BX)}$
- While Not convergence Do
  - Call the Lanczos algorithm to compute the $d$ largest eigenvalues of $A - \rho B$ and associated eigenvectors $[w_1, w_2, \cdots, w_d] \equiv X$
  - Set $\rho := \dfrac{Tr(X^T AX)}{Tr(X^T BX)}$
- End While

# Singular trace ratio optimization problem

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 $(X, Y) \in \mathcal{O}_{m,n,k}$.

# Singular value decomposition (SVD) for trace optimization

Let $C$ be a $m \times n$ real matrix with rank $(C) = r \leq n < m$.

1. There are positive real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and square orthogonal matrices $U \in \mathcal{O}_{m,m}$ and $V \in \mathcal{O}_{n,n}$, such that
$$U^T CV = \Sigma = \begin{pmatrix} \Sigma_r & O_{r,n-r} \\ O_{m-r,r} & O_{m-r,n-r} \end{pmatrix} \in \mathbb{R}^{m \times n},$$
$\rightarrow \Sigma_r = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$
$\rightarrow \sigma_1, \sigma_2, \cdots, \sigma_r$ : the singular values of $C$.
2. The trace of $X^T CY$ reaches its maximum (resp., minimum) when $X$ and $Y$ are an orthogonal basis of the left and right singular eigenspaces of $C$ associated with the $k$ algebraically largest (resp., smallest) singular values.

# Closed form solution : singular value problem

## Theorem

Let $C$ be a $m \times n$ real matrix with $m > n \geq r$. Suppose $U^T CV = \Sigma$ is the SVD of $C$. Let the columns of the matrix $U$ be partitioned in two blocks, $U = [U_1\ ,\ U_2]$ with $U_1 \in \mathbb{R}^{m \times n}$ and $U_2 \in \mathbb{R}^{m \times (m-n)}$.

Then, $\max_{(X,Y) \in \mathcal{O}_{m,n,k}} Tr\left[X^T CY\right] = \sum_{i=1}^{k} \sigma_i = Tr\left[U_{1,k}^T\ C\ V_k\right]$.

# 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, we just need 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 $f'(\rho) = -Tr\left[X(\rho)^T B\ Y(\rho)\right] - 1$ 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 $\rho = \dfrac{Tr[X^TAY]}{1+Tr[X^TBY]}$ ;
  2. Run the SVD algorithm to compute the $k$ largest singular values of $A-\rho B$ as well as their associated singular eigenvectors $X_i$ and $Y_i$ ;
  3. Repeat the above two steps until convergence.

# Outline

## SUMMARY

- The formulation of trace ratio is shared by many real-world application problems.
- A new formulation for the cell formation problem as the trace ratio : the Singular Trace Ratio Optimization Problem (STROP).
- Continuous TROP and STROP do not admit local non-global maxima.
- A direct solution to TROP or STROP is presented : The maximum is the unique zero of a scalar function.
- Both can be solved numerically by a judicious use of the Lanczos procedure, a good initialization, and inexact eigenvector calculations in the early stages of the Newton procedure.

## REFERENCES


- T. T. Ngo, **M. B.** and Y. Saad. *The trace ratio optimization problem,* SIAM Review, 54(3) (2012), pp.545-569.
- X. Li, W. Hu, C. Shen, A. Dick and Z. Zhang. *Context-Aware Hypergraph Construction for Robust Spectral Clustering,* Knowledge and Data Engineering, IEEE Transactions on (Volume :PP , Issue : 99, 2013 )
- L.-H. Zhang, and R.-C. Li. *Maximization of the Sum of the Trace Ratio on the Stiefel,* Technical Report 2013-04.


# Thanks !

Fondation mathématique FMJH Jacques Hadamard

UVHC UNIVERSITÉ DE VALENCIENNES ET DU HAINAUT - CAMBRESIS