

THE TRACE RATIO OPTIMIZATION PROBLEM

Mohammed Bellalij

LAMAV - Valenciennes - FRANCE

Mohammed.Bellalij@univ-valenciennes.fr

(joint work with Yousef Saad and Tanh Ngo - University of Minnesota, USA)

Seminar PGMO, January 2013

Ecole Polytechnique



- 1 Introduction
 - Preliminaries
 - Some examples
- 2 Mathematical analysis
- 3 Direct and efficient iterative procedure
- 4 Convergence to global optimum - Newton's method
- 5 Conclusion
- 6 References

Drowning in data :

- Data accumulates in an unprecedented speed :
 - 90 % of data in world today was created in last two years.
 - Every day, 2.3 Million terabytes (2.3×10^{18} bytes) created.
- Data preprocessing is an important part for effective machine learning & data mining.
- Trend \Rightarrow re-shaping & energizing many research areas (including : numerical linear algebra).



Data Mining

- 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.
- Data Mining \Rightarrow A broad discipline which includes such different areas as machine learning, data analysis, pattern recognition, etc
- Tools used : Optimization ; statistics ; linear algebra ; graph theory ; approximation theory ; ...

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.
 - Area of data mining where numerical linear algebra techniques play a crucial role.

Applications

- Face recognition
- Handwritten digit recognition
- Protein classification
- Customer relationship management
- Intrusion detection
- and so on

Linear Dimensionality Reduction

- Given : a data set of n points in m -dimensional space, i.e.

$$X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{m \times n}$$

- Problem : Map the set X to reduce space of dimension $d \ll m$, i.e.

$$Y = [y_1, y_2, \dots, y_n] \in \mathbb{R}^{d \times n}$$

- Using linear transformation(projector) :

$$Y = W^T X.$$

- Practical Constraint : The Y or W must satisfy certain properties

Some Techniques

Several techniques for dimensionality reduction end with solving a Trace Ratio Optimization Problem in the form

$$W_* = \arg \max \left\{ \frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)}; W^T C W = I \right\}$$

where A, B, C are matrices with appropriate dimensions, and I is the identity matrix.

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

Why these projectors are sought in practice ?

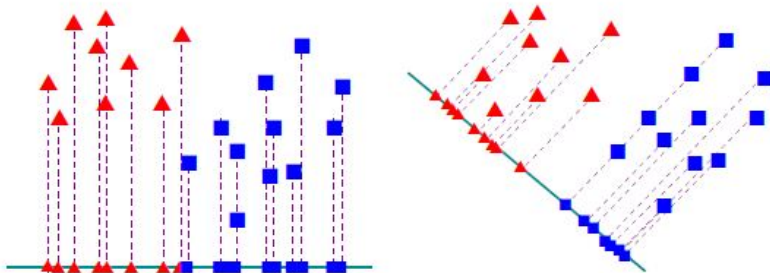


FIG. 1.1. Points on a 2-D plane. The projection on the x -axis does not separate the classes well. The one on the right figure does a much better job at separating between squares and triangles.



A typical example : identifying spam e-mail.

- We have a set X of messages each of which has been classified (e.g., by a human) as spam or non-spam (2 classes).
- Next comes a new message and we would like to classify it (automatically) as spam or non-spam by exploiting the available training set X .
- We could just use the Euclidean distance and find the closest set to x in some sense and this will then help determine a label.
- This does not work well for various reasons. Instead, it is better to perform the comparison on the low-dimensional space obtained by using the optimal projector found by LDA.

Fisher's Linear Discriminant Analysis (LDA)

- LDA attempts to find linear projections of the data that are optimal for discrimination 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.
- Goal : Search for a projector that maximizes "between scatter" and at the same time minimizes "within scatter".
- The natural model for these dual objectives is to optimize a trace ratio problem.

LDA : One-dimensional projector

- The between scatter measure of the projected points :

$$\Phi_b = \sum_{k=1}^c n_k |v^T(\mu^{(k)} - \mu)|^2 = v^T \left[\sum_{k=1}^c n_k (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T \right] v.$$

- The within scatter measure of the projected points :

$$\Phi_w = \sum_{k=1}^c \sum_{i \in C^{(k)}} (y_i - \tilde{\mu}^{(k)})^2 = v^T \left[\sum_{k=1}^c \sum_{i \in C^{(k)}} (x_i - \mu^{(k)})(x_i - \mu^{(k)})^T \right] v.$$

- Matrix formulation :

$$S_b = \sum_{k=1}^c n_k (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T; S_w = \sum_{k=1}^c \sum_{i \in C^{(k)}} (x_i - \mu^{(k)})(x_i - \mu^{(k)})^T ;.$$

- So, the best one-dimensional projector v can be found by maximizing the

$$\text{ratio } \frac{\Phi_b}{\Phi_w} : \max_v \frac{v^T S_b v}{v^T S_w v}.$$

- Optimal v : eigenvector associated with the largest (generalized) eigenvalue of $S_b v_* = \lambda S_w v_*$.

LDA : Objective function deduction

- Let μ = mean of X , and $\mu^{(k)}$ = mean of the k -th class (of size n_k , $k = 1, \dots, c$). Define

$$A = \sum_{k=1}^c n_k (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T, \quad B = \sum_{k=1}^c \sum_{x_i} (x_i - \mu^{(k)})(x_i - \mu^{(k)})^T.$$

- Criterion : maximize the ratio of two traces $\frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)}$
- Constraint : $W^T W = I$ (orthogonal projection).
- ... Alternative : Solve instead the "easier" problem : $\max_{W^T B W = I} \text{Tr}(W^T A W)$.
- Solution : largest eigenvectors of $A w_j = \lambda_j B w_j$.
- However, its solution may deviate from the original objective.

Closed form solution : standard eigenvalue problem

- Given a symmetric matrix A of dimension $m \times m$ and an arbitrary unitary matrix W of dimension $m \times d$, it is known that the trace of $W^T A W$ reaches its maximum when W is an orthogonal basis of the eigenspace of A associated with the d algebraically largest eigenvalues.
- $\max\{Tr(W^T A W) : W^T W = I, W \in \mathbb{R}^{m \times d}\} = \lambda_1 + \dots + \lambda_d$
 where the eigenvalues $\lambda_1, \dots, \lambda_d$ are labeled decreasingly

Closed form solution : generalized eigenvalue problem

- Assuming that B positive definite, we have

$$\begin{cases} \max & \text{Tr}(W^T A W) = \text{Tr}(W_*^T A W_*) = \lambda_1 + \dots + \lambda_d \\ W \in \mathbb{R}^{m \times d} \\ W^T B W = I \end{cases}$$

(Eigenvalues labeled decreasingly for the generalized problem

$A w = \lambda B w$, W_* \rightarrow eigenvectors associated with the first d eigenvalues, with $W_*^T B W_* = I$)

- This problem is a simplification of Trace Ratio Optimization Problem.

Existence and uniqueness of a solution

- The trace ratio problem may not have a solution when B is not positive definite : $\rightarrow \text{Tr}(W^T B W) = 0!$
- A simple example : $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $W = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

Lemma

Assume that B is positive semi-definite and let d be the number of columns of W . If B has at most $d - 1$ zero eigenvalues then $\text{Tr}(W^T B W)$ is nonzero for any unitary W .

- The problem is well-posed under the condition that the null space of B is of dimension less than d , i.e., that its rank be at least $n - d + 1$. In this case the maximum is finite.

Existence and uniqueness of a solution

Proposition

Let A, B be two symmetric matrices and assume that B is semi-positive definite with rank greater than $n - d$. Then the ratio $\frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)}$ admits a finite maximum value. The maximum is reached for a certain W that is unique up to unitary transforms of the columns.

- Mild condition as it is the case for all real datasets in experiments on dimensionality reduction
- In the remainder of the talk we will assume that B satisfies the conditions of the proposition
- There is no loss of generality in assuming that $C = I$.

Trace ratio vs. Ratio trace

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

$$\max_{W^T W=I} \text{Tr}((W^T B W)^{-1} (W^T A W)).$$
- This optimization problem is easily solved as generalized eigenvalue problem.
- The obtained solution does not necessarily best maximize the corresponding trace ratio problem. This can actually be viewed as a greedy algorithm that essentially maximizes $\sum_{i=1}^d \frac{w_i^T A w_i}{w_i^T B w_i}$ but not

$$\frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)} = \frac{\sum_{i=1}^d w_i^T A w_i}{\sum_{i=1}^d w_i^T B w_i}.$$

From trace ratio to trace difference

- Suppose we have the global maximum solution W_* = $\arg \max \left\{ \frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)}; W^T W = I \right\}$, then
- $\frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)} \leq \rho_* = \frac{\text{Tr}(W_*^T A W_*)}{\text{Tr}(W_*^T B W_*)}$, $\forall W$ s.t. $W^T W = I$.
- It follows that $\text{Tr}(W^T (A - \rho_* B) W) \leq 0$.
- So we have $\max_{W^T W = I} \text{Tr}(W^T (A - \rho_* B) W) = 0$.
- Define the trace difference function

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

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

First useful expression for $f(\rho)$

We have

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

Denote :

- $G(\rho) = A - \rho B$.
- $W(\rho)$ a set of the d eigenvectors which reach the above maximum.
- The n eigenvalues of $G(\rho)$ labeled decreasingly :
 $\lambda_1(\rho) \geq \lambda_2(\rho) \geq \dots \geq \lambda_n(\rho)$.

It is clear that $f(\rho) = \lambda_1(\rho) + \lambda_2(\rho) + \dots + \lambda_d(\rho)$.

Second useful expression for $f(\rho)$

- Based on the eigenprojector : $P(\rho) = W(\rho)W(\rho)^T$.
- Clearly

$$f(\rho) = \text{Tr}(W(\rho)^T G(\rho) W(\rho)) = \text{Tr}(G(\rho) W(\rho) W(\rho)^T) = \text{Tr}(G(\rho) P(\rho)).$$

- By exploiting the Dunford integral,

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

(where Γ is a Jordan curve containing the d eigenvalues of interest).

- We obtain

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

Properties of the trace difference function : f

Proposition

- 1 f is a strictly decreasing function of ρ ;
- 2 $f(\rho) = 0$ iff $\rho = \rho_*$.
- 3 $f'(\rho) = -\text{Tr}(W(\rho)^T B W(\rho))$.

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

Localization of the optimum

Proposition

- The root ρ_* of $f(\rho)$ is located in the interval $[\lambda_d, \lambda_1]$, where λ_i is the i -th largest eigenvalue of the pair (A, B) .
- Assume that B is positive definite. Then the root ρ_* of $f(\rho)$ is such that
$$\frac{\sum_{i=1}^d \lambda_i(A)}{\sum_{i=1}^d \lambda_i(B)} \leq \rho_* \leq \frac{\sum_{i=1}^d \lambda_i(A)}{\sum_{i=1}^d \lambda_{n-i+1}(B)}$$
, where $\lambda_i(A)$, and $\lambda_i(B)$ are the i -th largest eigenvalues of the matrices A and B respectively.

Practical implementation via Newton's method

From the expression of the differential of f , Newton's method takes the form

$$\rho_{new} = \rho - \frac{\text{Tr}(W(\rho)^T (A - \rho B) W(\rho))}{-\text{Tr}(W(\rho)^T B W(\rho))} = \frac{\text{Tr}(W(\rho)^T A W(\rho))}{\text{Tr}(W(\rho)^T B W(\rho))}$$

→ Newton's method for finding the zero of f amounts to a form of fixed point

iteration : $g(\rho) = \frac{\text{Tr}(W(\rho)^T A W(\rho))}{\text{Tr}(W(\rho)^T B W(\rho))}$.

→ Exploit the Lanczos algorithm to provide a highly effective procedure.

Newton-Lanczos algorithm for TROP

- Input : A, B and a dimension m .
- Select initial $m \times d$ unitary matrix W ;
- compute $\rho = \frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)}$.
- While Not convergence Do :
- Call the Lanczos algorithm to compute the d largest eigenvalues of $A - \rho B$ and associated eigenvectors $[w_1, w_2, \dots, w_d] \equiv W$
- Set $\rho := \frac{\text{Tr}(W^T A W)}{\text{Tr}(W^T B W)}$.
- End While.

CONCLUSION

- TROP is more expressive model than common simplified approaches.
- We show that TROP can be solved by a judicious use of the Lanczos procedure, a good initialization, and inexact eigenvector calculations in the early stages of the Newton procedure,
- Our procedure is much less expensive than common approach which relies on solving the generalized eigenvalue problem.
- Numerical Linear Algebra and Optimization techniques are powerful for solving TROP.

References

Recent papers advocated similar or related techniques



C. Shen, H. Li, and M. J. Brooks,

A convex programming approach to the trace quotient problem.

In ACCV (2) 2007.



H. Wang, S.C. Yan, D.Xu, X.O. Tang, and T. Huang.

Trace ratio vs. ratio trace for dimensionality reduction.

In IEEE Conference on Computer Vision and Pattern Recognition, 2007






S. Yan and X. O. Tang,

Trace ratio revisited

Proceedings of the European Conference on Computer Vision, 2006.

References

-  K. Fukunaga, Introduction to Statistical Pattern Recognition. 2nd ed. *Academic Press* , San Diego, CA, 1991.
-  T. T. Ngo, **M. B.** and Y. Saad.
The trace ratio optimization problem.
SIAM Review, 54, issue number 3, 2012.
-  T. T. Ngo, **M. B.** and Y. Saad.
The trace ratio optimization problem for dimensionality reduction.
SIAM J. Matrix Anal. and Appl., 31, pp. 2950-2971, 2010.

Thank you for your attention