A Geometric Revisit to the Trace Quotient Problem

Hao Shen, Klaus Diepold, and Knut Hüper

***Abstract*— This paper studies the problem of trace quotient, or trace ratio maximization, which has enormous applications in computer vision, pattern recognition and machine learning. We provide a geometric revisit to the problem in the framework of optimization on smooth manifolds. The set of critical points of the trace quotient is analyzed. Local quadratic convergence properties of the so-called Iterative Trace Ratio (ITR) scheme, which recently became an attractive solver to the problem, is studied. Based on this result, different from a popular realization of ITR, which requires to solve a symmetric eigenvalue problem at each iteration, we propose a simple, efficient algorithm, which employs only one step of the parallel Rayleigh quotient iteration at each iteration. An numerical experiment demonstrates the local convergence properties of ITR.**

I. INTRODUCTION

In recent years, the problem of *dimensionality reduction* (DR), which aims to uncover certain low-dimensional structure from original high-dimensional data, has attracted enormous research attentions in computer vision, pattern recognition and machine learning. Many popular DR methods, such as linear discriminant analysis [1], locality preserving projection [2], and graph embedding [3], result in the so-called *trace quotient* or *trace ratio* problem, which is usually formulated as the problem of maximizing the following cost function, *cf.* [4], [5],

$$\begin{aligned} f\colon St(k,m) &\to \mathbb{R}, \\ f(X) &:= \frac{\operatorname{tr}(X^\top AX)}{\operatorname{tr}(X^\top BX)}, \end{aligned} \tag{1}$$

where $m > k$. Here, $A \in \mathbb{R}^{m\times m}$ is assumed to be symmetric positive semi-definite, $B \in \mathbb{R}^{m\times m}$ is assumed to be symmetric positive definite and $St(k,m)$ denotes the Stiefel manifold

$$St(k,m) := \{X \in \mathbb{R}^{m\times k} | X^\top X = I_k\}. \tag{2}$$

In many applications, the trace quotient problem as maximizing the cost function f as defined by (1) is often replaced by an alleged simpler problem [4], namely the *quotient trace* problem, which maximizes the following cost function

$$\begin{aligned} g\colon St(k,m) &\to \mathbb{R}, \\ g(X) &:= \operatorname{tr}(X^\top AX(X^\top BX)^{-1}), \end{aligned} \tag{3}$$

which can be used to solve the symmetric generalized eigenvalue problem [6]. It is known, that for specific applications, e.g. dimensionality reduction and classification, solutions given by solving the trace quotient problem often significantly outperform the solutions provided by its counterpart, i.e the quotient trace problem. Therefore, development of efficient algorithms for solving the trace quotient problem is highly attractive in real applications.

Recently, an iterative algorithmic scheme, the so-called Iterative Trace Ratio (ITR) algorithm, became dominant for solving the trace quotient problem [7]. A concrete implementation, *cf.* Algorithm 1 in [7], involves solving a symmetric eigenvalue problem at each iteration. Although such a concrete realization of the ITR scheme has been shown to converge globally and locally quadratically fast [5], it might become prohibitively expensive when the dimensionality of the problem is huge. Moreover, to our best knowledge, local convergence properties of the general ITR scheme have not been published yet.

It is easily seen that the trace quotient f above is invariant with respect to orthonormal basis changes of X, namely

$$X \mapsto X\Theta \quad \text{with} \quad \Theta\Theta^\top = I_k, \tag{4}$$

i.e., it all induces a function on the Graßmann manifold $Gr(k,m)$. In this paper, we revisit the trace quotient problem and the general ITR scheme in the framework of optimization on smooth manifolds, specifically on the Graßmann manifold.

This paper is organized as follows. In Section II, we briefly introduce some basic concepts about the Graßmann manifold, which are needed in our later analysis. Section III characterizes global maximum and critical points of the trace quotient. In Section IV, local quadratic convergence properties of the general ITR scheme are proved. A simple realization of ITR is proposed. Finally in Section V, local convergence properties of ITR algorithms are demonstrated by some numerical experiments.

II. MATHEMATICAL PRELIMINARIES

Recall some basic concepts of the Graßmann manifold. We will identify $Gr(k,m)$ as the set of all rank-k symmetric projection operators on $\mathbb{R}^m$ [8], i.e.,

$$Gr(k,m) := \{P \in \mathbb{R}^{m\times m} | P = P^\top, P^2 = P, \operatorname{tr} P = k\}. \tag{5}$$

Let us denote the set of all $m \times m$ skew-symmetric matrices by

$$\mathfrak{so}(m) := \left\{\Omega \in \mathbb{R}^{m\times m} \,\middle|\, \Omega = -\Omega^\top\right\}. \tag{6}$$

The tangent space of $Gr(k,m)$ at $P \in Gr(k,m)$ is given by

$$T_P Gr(k,m) := \{[P,\Omega] \mid \Omega \in \mathfrak{so}(m)\} \tag{7}$$

H. Shen and K. Diepold are with the Institute for Data Processing and CoTeSys at Technische Universität München, Germany. {hao.shen, kldi}@tum.de.

Knut Hüper is with Institut für Mathematik, Universität Würzburg, Germany. hueper@mathematik.uni-wuerzburg.de.

ISBN 978-963-311-370-7

with matrix commutator $[A,B] := AB - BA$. Let $P \in Gr(k,m)$ and let $\Xi \in T_P Gr(k,m)$ be a tangent vector. We consider the Euclidean Riemannian metric on $Gr(k,m)$ induced by the embedding space of symmetric matrices, which is defined by the Frobenius inner product, i.e.

$$\langle \Xi_1, \Xi_2 \rangle := \operatorname{tr}(\Xi_1 \Xi_2) \tag{8}$$

for all $\Xi_1, \Xi_2 \in T_P Gr(k,m)$ The unique geodesic $\gamma_{P,\Xi}$ through P in direction $\Xi \in T_P Gr(k,m)$ is given by

$$\begin{aligned} \gamma_{P,\Xi} &\colon \mathbb{R} \to Gr(k,m), \\ \gamma_{P,\Xi}(t) &:= \mathrm{e}^{t[\Xi,P]}\, P\, \mathrm{e}^{-t[\Xi,P]}. \end{aligned} \tag{9}$$

Let

$$O(m) := \left\{ Q \in \mathbb{R}^{m\times m} \;\middle|\; Q^\top Q = I_m \right\}, \tag{10}$$

be the Lie group of all $m \times m$ orthogonal matrices. As $Gr(k,m)$ is a homogeneous space of $O(m)$, one can represent any point $P \in Gr(k,m)$ by

$$P = Q \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} Q^\top \tag{11}$$

for suitable $Q \in O(m)$, and can accordingly represent

$$[\Xi, P] = Q \begin{bmatrix} 0 & -Z^\top \\ Z & 0 \end{bmatrix} Q^\top, \tag{12}$$

where $Z \in \mathbb{R}^{(m-k)\times k}$, cf. [9].

III. ANALYSIS OF THE TRACE QUOTIENT

With the knowledge about the Graßmann manifold from above, we study the trace quotient, now redefined on $Gr(k,m)$, i.e.,

$$\begin{aligned} f &\colon Gr(k,m) \to \mathbb{R}, \\ f(P) &:= \frac{\operatorname{tr}(PA)}{\operatorname{tr}(PB)}. \end{aligned} \tag{13}$$

In the rest of the paper we refer the trace quotient to the definition as in (13).

First of all, we characterize critical points of the trace quotient f as defined in (13). Taking the first derivative of f at $P \in Gr(k,m)$ in direction $\Xi \in T_P Gr(k,m)$ leads to

$$\begin{aligned} \mathrm{D}\, f(P)\Xi &= \frac{\mathrm{d}}{\mathrm{d}t} f\left(\mathrm{e}^{t[\Xi,P]}\, P\, \mathrm{e}^{-t[\Xi,P]}\right)\Big|_{t=0} \\ &= \frac{1}{\operatorname{tr}(PB)} \operatorname{tr}\Big(\Xi \big(\underbrace{A - \frac{\operatorname{tr}(PA)}{\operatorname{tr}(PB)} B}_{=A - f(P)B =: \Phi(P)} \big)\Big). \end{aligned} \tag{14}$$

Theorem 1: The critical points of f are exactly the solutions of

$$[P, [P, \Phi(P)]] = 0. \tag{15}$$

Proof: The result follows as for any symmetric $m \times m$ matrix S the mapping $S \mapsto [P,[P,S]]$ is the orthogonal projection of S into the tangent space of $Gr(k,m)$ at $P \in Gr(k,m)$, cf. [10]. ■

Recall the representation of $P \in Gr(k,m)$ as given by (11). By invariance we can conclude the following.

Corollary 1: The point P^* being critical is equivalent to $\widetilde{\Phi}(P^*) := Q^\top \Phi(P^*) Q$ being block diagonal. Here the first diagonal block is of dimension $k \times k$, consequently, the second block is of dimension $(m-k) \times (m-k)$.

We certainly expect that characterizing critical points gives us conditions on the eigenvalues of $\widetilde{\Phi}_{11}(P^*)$ compared to those of $\widetilde{\Phi}_{22}(P^*)$, where

$$\widetilde{\Phi}(P^*) = \begin{bmatrix} \widetilde{\Phi}_{11}(P^*) & 0 \\ 0 & \widetilde{\Phi}_{22}(P^*) \end{bmatrix}. \tag{16}$$

Again by invariance, without loss of generality we might choose the orthogonal Q such that $\widetilde{\Phi}(P^*)$ is not only block diagonal, it is even diagonal. To characterize the critical points further we need second order derivative information. For $\Xi \in T_P Gr(k,m)$ arbitrary we get, by using (11), (12), and (15),

$$\begin{aligned} &\frac{\mathrm{d}^2}{\mathrm{d}t^2} f\left(\mathrm{e}^{t[\Xi,P]}\, P\, \mathrm{e}^{-t[\Xi,P]}\right)\Big|_{t=0}\Big|_{P=P^*} \\ &= \frac{1}{\operatorname{tr}(BP^*)} \operatorname{tr}\left([[\Xi,P^*],[[\Xi,P^*],P^*]]\Phi(P^*)\right) \\ &= \frac{1}{\operatorname{tr}(BP^*)} \operatorname{tr}\left(\left[\begin{smallmatrix} -2Z^\top Z & 0 \\ 0 & 2ZZ^\top \end{smallmatrix}\right] \widetilde{\Phi}(P^*)\right) \\ &= \underbrace{\frac{2}{\operatorname{tr}(BP^*)}}_{>0 \text{ as } B \succ 0} \sum_{i=1}^{k} \sum_{j=1}^{m-k} z_{ij}^2 (\sigma_j - \lambda_i), \end{aligned} \tag{17}$$

where $\{\lambda_i\}_{i=1}^k$ denotes the spectrum of $\widetilde{\Phi}_{11}(P^*)$, whereas $\{\sigma_j\}_{j=1}^{m-k}$ denotes that of $\widetilde{\Phi}_{22}(P^*)$.

Corollary 2: Local maxima of f are characterized by the condition that $\{\lambda_i\}_{i=1}^k \succeq \{\sigma_j\}_{j=1}^{m-k}$.

Now let us characterize the global maximum of the trace quotient f. Obviously, for any $P \in Gr(k,m)$,

$$\begin{aligned} \operatorname{tr}(P\,\Phi(P)) &= \operatorname{tr}(PA) - f(P)\operatorname{tr}(PB) \\ &= 0. \end{aligned} \tag{18}$$

Theorem 2: Let P^* be a local maximum of f. Then P^* is even a global maximum.

Proof: For any $P \in Gr(k,m)$, it holds

$$\begin{aligned} \operatorname{tr}(P^*\,\Phi(P^*)) &= 0 \\ &\geq \operatorname{tr}(P\,\Phi(P^*)) \\ &= \operatorname{tr}(PA) - f(P^*)\operatorname{tr}(PB) \end{aligned} \tag{19}$$

i.e.

$$f(P^*) \geq f(P), \tag{20}$$

which implies that P^* is a global maximum of f. ■

Lemma 1: Let P^* be a global maximum of f. Assume $\{\lambda_i\}_{i=1}^k \succ \{\sigma_j\}_{j=1}^{m-k}$, meaning that there is a gap between the k-th largest eigenvalue of $\Phi(P^*)$ and the $k+1$-st largest eigenvalue of $\Phi(P^*)$. Then the global maximum is unique with nondegenerated Hessian.

Proof: [Sketch] Uniqueness follows from the standard discussion of the generalized Rayleigh quotient function defined on the Grassmann manifold, cf. [8], [11]. The second statement is easily seen from (17). ■

IV. GENERAL ITR AND TWO REALIZATIONS

In this section, we study local convergence properties of the general iterative trace ratio scheme. The techniques used here are similar to the work in [12].

The general ITR algorithm, which was originally proposed in [7], can be summarized as follows.

Algorithm 1: General ITR Algorithm

Step 1: Given an initial guess $P_0 \in Gr(k,m)$, and set $i = 0$.

Step 2: Update

$$P_{i+1} \leftarrow \underset{\widetilde{P} \in Gr(k,m)}{\operatorname{argmax}} \operatorname{tr}\left(\widetilde{P}\left(A - \frac{\operatorname{tr}(P_i A)}{\operatorname{tr}(P_i B)} B\right)\right).$$

Step 3: If $\|P_{i+1} - P_i\|_F$ is small enough, stop. Otherwise, set $i = i + 1$ and go to Step 2.

Here, $\|\cdot\|_F$ is the Frobenius norm of matrices.

Obviously, the key point of Algorithm 1 is how to solve the maximization problem in Step 2 efficiently. Similar to theorem 4.1 in [12], we have the following result.

Theorem 3: For a given $\eta \in \mathbb{R}$, let

$$s_\eta \colon Gr(k,m) \to Gr(k,m), \tag{21}$$

with

$$P_{i+1} = s_\eta(P_i), \qquad i \in \mathbb{N}_0 \tag{22}$$

be a quadratically fast and locally differentiable algorithm, or alternatively, one step of such an algorithm to compute

$$\max_{\widetilde{P} \in Gr(k,m)} \operatorname{tr}\left(\widetilde{P}\,(A - \eta B)\right). \tag{23}$$

Consider the sequence $\{(P_i, \eta_i)\}$ with $i = 0, 1, 2, \dots$, generated by the recursions

$$\begin{cases} P_{i+1} = s_{\eta_i}(P_i), \\ \eta_{i+1} = f(P_i), \end{cases} \tag{24}$$

with f being the trace quotient defined in (13), $X_0 \in Gr(k,m)$ arbitrary, and $\eta_0 = f(X_0)$. If the sequence $\{(P_i, \eta_i)\}$ converges to $(P^*, f(P^*))$, where P^* is the global maximum of f, then it converges locally quadratically fast.

Proof: Let us define

$$s \colon Gr(k,m) \to Gr(k,m), \qquad P \mapsto s_{\eta(P)}(P). \tag{25}$$

We need to show that the first derivative of this algorithmic map s, i.e.

$$\mathrm{D}\, s(P) \colon T_P Gr(k,m) \to T_{s(P)} Gr(k,m) \tag{26}$$

vanishes at P^*.

If $P^* \in Gr(k,m)$ is the unique global maximum of f, then it can be shown that P^* is a fixed point of s, i.e. $s(P^*) = P^*$, following the result from Lemma 1. Since $s_\eta(P)$ solves the maximization problem defined in (23), the first derivative of s_η at P^* in direction $\Xi \in T_{P^*}Gr(k,m)$ must vanish, i.e.

$$\mathrm{D}\, s_\eta(P)\Xi|_{P=P^*} = 0. \tag{27}$$

Then, we compute the first derivative of s at P^* in direction $\Xi \in T_{P^*}Gr(k,m)$ as

$$\begin{aligned} \mathrm{D}\, s(P)\Xi|_{P=P^*} &= \left.\frac{\partial s_\eta}{\partial \eta}\right|_{\eta = f(P^*)} \cdot \mathrm{D}\,\eta(P)\Xi|_{P=P^*} \\ &\quad + \mathrm{D}\, s_\eta(P)\Xi|_{P=P^*} \\ &= \left.\frac{\partial s_\eta}{\partial \eta}\right|_{\eta = f(P^*)} \cdot \mathrm{D}\,\eta(P)\Xi|_{P=P^*}. \end{aligned} \tag{28}$$

Since η is chosen to be the trace quotient f as in (24), by the critical point condition of f, i.e., $\mathrm{D}\,\eta(P)\Xi|_{P=P^*} = 0$, we just show

$$\mathrm{D}\, s(P)\Xi|_{P=P^*} = 0. \tag{29}$$

Thus, the result follows. ■

It can be shown that, e.g. at the i-th iteration, a global maximizer to this sub-problem is the orthogonal projector corresponding to the k largest eigenvalues of $A - f(P_i)B$. This strategy leads to the following algorithm, referred to here as EIG-ITR algorithm, which is the most dominating implementation of ITR in applications [5], [7].

Algorithm 2: EIG-ITR Algorithm

Step 1: Given an initial guess $P_0 = X_0 X_0^\top \in Gr(k,m)$, where $X_0 \in St(k,m)$, and set $i = 0$.

Step 2: Compute an orthonormal eigenbasis $X_{i+1} \in St(k,m)$ corresponding to the k largest eigenvalues of

$$\Phi(P_i) = A - f(P_i)B.$$

Step 3: Update $P_{i+1} \leftarrow X_{i+1} X_{i+1}^\top$.

Step 4: If $\|P_{i+1} - P_i\|_F$ is small enough, stop. Otherwise, set $i = i + 1$ and go to Step 2.

Local convergence properties of EIG-ITR follows directly from Theorem 3.

Lemma 2: If $P^* \in Gr(k,m)$ is the unique global maximizer of the trace quotient f as defined in (13), then the EIG-ITR algorithm converges locally quadratically fast to P^*.

It is clear, that if the dimension m of the problem increases, computing all eigenvalues of an $m \times m$ symmetric matrix at each iteration becomes prohibitively expensive. In the rest of this section, we propose a simple, alternative approach to solve the sub-problem in Step 2 of Algorithm 1. At the i-th iteration, we apply only one step of the parallel Rayleigh Quotient Iteration (RQI), *cf.* Algorithm 4.1 in [13], which is locally cubically convergent to k eigenvectors of $\Phi(P_i)$. We refer to our proposed algorithm as RQI-ITR.

Algorithm 3: RQI-ITR Algorithm

Step 1: Given an initial guess $P_0 = X_0 X_0^\top \in Gr(k,m)$, where $X_0 \in St(k,m)$, and set $i = 0$.
Step 2: Compute $\Phi(P_i) = A - f(P_i)B$.
Step 3: Solve for $Z \in \mathbb{R}^{m\times k}$
$$\Phi(P_i)Z - Z\,\mathrm{diag}(X^\top \Phi(P_i)X) = X.$$
Step 4: Update $X_{i+1} \leftarrow (Z)_Q$, and $P_{i+1} \leftarrow X_{i+1}X_{i+1}^\top$.
Step 5: If $\|P_{i+1} - P_i\|_F$ is small enough, stop. Otherwise, set $i = i+1$ and go to Step 2.

Here, $(Z)_Q$ gives the orthogonal Q-factor from Z after performing Gram-Schmidt orthonormalization.

Local convergence properties of RQI-ITR can be summarized as follows.

Lemma 3: If $P^* \in Gr(k,m)$ is the unique global maximizer of the trace quotient f as defined in (13), then the RQI-ITR algorithm converges locally quadratically fast to P^*.

V. Numerical Experiments

In this section, we demonstrate the local quadratic convergence properties of both EIG-ITR and RQI-ITR. In our experiment, we set $m = 10$ and $k = 7$. The convergence is measured by the distance of the accumulation point $P^* \in Gr(k,m)$ to the i-th iterate $P_i \in Gr(k,m)$, i.e. by the Frobenius norm $\|P^* - P_i\|_F$. It can be seen from Fig. 1 that both EIG-ITR and RQI-ITR are locally quadratically convergent to the unique global maximizer P^* of the trace quotient f as defined in (13).

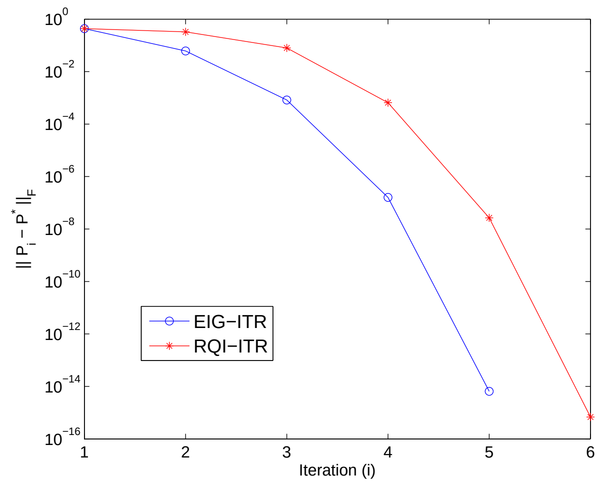

Fig. 1. Convergence properties: EIG-ITR vs. RQI-ITR.

Acknowledgments

This work has been supported in parts by the German DFG funded Cluster of Excellence CoTeSys - Cognition for Technical Systems.

References

[1] K. Fukunaga, *Introduction to Statistical Pattern Classification*, Academic Press, 1990.
[2] X. He, D. Cai, and W. Min, "Statistical and computational analysis of locality preserving projection," in *Proceedings of the* 22^{nd} *International Conference on Machine Learning (ICML 2005)*, 2005, pp. 281–288.
[3] S. Yan, D. Xu, B. Zhang, and H. Zhang, "Graph embedding: A general framework for dimensionality reduction," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 2005, pp. 830–837.
[4] T. Ngo, M. Bellalij, and Y. Saad, "The trace ratio optimization problem," University of Minnesota Technical Report (umsi-2009-tmp4), 2009.
[5] L. Zhang, L. Liao, and M. Ng, "Fast algorithms for the generalized Foley-Sammon discriminant analysis," *SIAM Journal of Matrix Analysis and Application*, vol. 31, no. 4, pp. 1584–1605, 2010.
[6] A. Sameh and Z. Tong, "The trace minimization method for the symmetric generalized eigenvalue problem," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1-2, pp. 155–175, 2000.
[7] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, 2007, pp. 136–139.
[8] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[9] K. Hüper and F. Silva Leite, "On the geometry of rolling and interpolation curves on S^n, SO_n, and Grassmann manifolds.," *J. Dyn. Control Syst.*, vol. 13, no. 4, pp. 467–502, 2007.
[10] U. Helmke, K. Hüper, and J. Trumpf, "Newton's method on Graßmann manifolds," arXiv:0709.2205v2, 2007.
[11] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds.*, Princeton, NJ: Princeton University Press. xv, 224 p., 2008.
[12] K. Hüper and U. Helmke, "A new algorithm for the generalized eigenvalue problem," in *Proceedings of the* 21^{st} *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, 1997, pp. 35–38.
[13] K. Hüper, *A Calculus Approach to Matrix Eigenvalue Algorithms*, Habilitation Dissertation, Department of Mathematics, University of Würzburg, Germany, July 2002.