DOI: 10.14464/gammas.v5i1.540

GAMMAS 2023, Vol. 5, No. 1, p. 1–9

International Association
of Applied Mathematics and Mechanics
– Archive for Students –

Robust Rayleigh quotient minimization and generalized eigenvalue problems

Henrik Schanze[a], [b,⋆]

[a] Technische Universität Braunschweig, Braunschweig, Germany,
Institut für Numerische Mathematik
[b] Current: Helmholtz Centre for Infection Research (HZI), Department of
Epidemiology, Inhoffenstraße 7, Braunschweig, Germany

received 23.10.2022, accepted 08.03.2023, published 24.05.2023

Abstract: *We study the problem of minimizing the non-linear trace quotient* $\operatorname{tr}(V^T G(V)V)/\operatorname{tr}(V^T H(V)V)$ *over the Stiefel manifold* $\operatorname{St}(p,n)$ *of all* $n \times p$ *matrices* V *with orthonormal columns. Hereby we assume* $G(V)$ *and* $H(V)$ *to be symmetric and positive definite for all* $V \in \operatorname{St}(p,n)$*. In this way we generalize the robust Rayleigh quotient optimization. This paper shows a possible way to minimize this quotient by joining and generalizing different known techniques, e.g. the self-consistent-field (SCF) iteration, and examines it by testing a small example.*

Keywords: numeric linear algebra, Rayleigh quotient, SCF Iteration, Newton's method, generalized eigenvalues

1 Introduction

The Rayleigh quotient $\frac{v^T Av}{v^T v}$, for a given matrix $A \in \mathbb{R}^{n\times n}$ and vector $v \in \mathbb{R}^n$, plays an important role in many different branches of mathematics, e.g. in eigenvalue algorithms, machine learning techniques [4], control theory and statistics [2, Sec. 1], [3]. In many cases it is also of interest to find the minimum of the generalized Rayleigh quotient $\frac{v^T Av}{v^T Bv}$ with $B \in \mathbb{R}^{n\times n}$, where A and B have favorable properties such that the minimum exists. For instance, often both matrices are symmetric and B is additionally positive definite. In this case, the solution of

$$\min_{v\in\mathbb{R}^n\setminus\{0\}} \frac{v^T Av}{v^T Bv} \tag{1}$$

is given by a generalized eigenvector v^* of the matrix pencil $\lambda B - A$, i.e. a solution of $Av = \lambda Bv$ for some $\lambda \in \mathbb{R}$ [7]. Given the Cholesky factorization $B = CC^T$, it is easily seen that $Av = \lambda Bv$ is equivalent to the standard symmetric problem $C^{-1}AC^{-T}v = \lambda v$. Therefore, all generalized eigenvalues of $\lambda B - A$ are real and its generalized eigenvectors for different eigenvalues are B-orthogonal. Furthermore, if $Av = \lambda Bv$, it is true that

$$\frac{v^T Av}{v^T Bv} = \frac{\lambda v^T Bv}{v^T Bv} = \lambda,$$

so the minimizer v^* of (1) corresponds to the smallest generalized eigenvalue λ^* of $\lambda B - A$.

In many cases the matrices A and B are generated from real-life data, which can be subject to all kinds of errors. To take this inaccuracy into account, we could use uncertainty-intervals instead of single data-points. In consequence, this leads to matrices that do not have

[⋆] corresponding author: henrik.schanze@web.de
supervisor: Dr. Philip Saltenberger, TU Braunschweig, Germany

constant entries, but entries depending on parameters. To find a solution analogue to the one used for constant matrices, we intend to look for a minimum of the Rayleigh quotient for the worst case of parameters. We will see later that this results in a problem of the form

$$\min_{v\in\mathbb{R}^n\setminus\{0\}} \frac{v^T G(v) v}{v^T H(v) v}, \tag{2}$$

where $G(\cdot)$ and $H(\cdot)$ are matrix-valued functions, though out of simplicity we will call the matrices sometimes, with the same properties as the matrices before (i.e. symmetric positive definiteness) but depend on the vector v. In [2] Bai et al. showed that, under suitable assumptions, a solution v^* of (2) satisfies $G(v^*)v^* = \lambda H(v^*)v^*$, i.e. the minimizer v^* is a nonlinear generalized eigenvector of the matrix pencil $\lambda H(\cdot) - G(\cdot)$. Furthermore, they demonstrated that v^* is additionally an eigenvector corresponding to the smallest positive eigenvalue of a related nonlinear matrix pencil. This fact is motivating the use of the self-consistent-field (SCF) iteration (see [2, Sec.4]) to find the smallest solution for this non-linear eigenvalue problem and by that solve (2).

In some techniques for dimension-reduction (see [6]) we face a similar minimization-problem as in (1):

$$\min_{V\in\mathbb{R}^{n\times p}} \frac{\operatorname{tr}(V^T AV)}{\operatorname{tr}(V^T BV)} \quad \text{s.t.} \quad V^T V = I_p. \tag{3}$$

Here $\operatorname{tr}(\cdot)$ means the trace of a matrix, i.e. the sum of its diagonal entries. As before, A and B are assumed to be symmetric and positive definite. Now we minimize over the Stiefel manifold $\operatorname{St}(p,n)$ of all matrices $V \in \mathbb{R}^{n\times p}$ with orthonormal columns[1].

One way to find an approximate solution to (3) is to replace it by the simpler, but not necessarily equivalent, problem [3, Sec. 2]

$$\min_{V\in\mathbb{R}^{n\times p}} \operatorname{tr}(V^T AV) \quad \text{s.t.} \quad V^T BV = I_p.$$

A minimizer $V^* \in \mathbb{R}^{n\times p}$ of this problem is a solution to the equation $AV = BV\Lambda$, for some $\Lambda \in \mathbb{R}^{p\times p}$. In particular, analogously to the solution of (1), the orthonormal columns of any solution V^* span the same subspace as the generalized eigenvectors of the matrix pencil $\lambda B - A$ corresponding to its p smallest eigenvalues.

The question we are interested in is what we can say about the combination of (2) and (3). In mathematical terms, we want to solve

$$\min_{V\in\mathbb{R}^{n\times p}} \frac{\operatorname{tr}(V^T G(V)V)}{\operatorname{tr}(V^T H(V)V)} \quad \text{s.t.} \quad V^T V = I_p, \tag{4}$$

where $G(V)$ and $H(V)$ are symmetric and positive definite matrices for all $V \in \operatorname{St}(p,n)$. Although this problem seems not to be studied in the literature so far, it can be motivated by finding good dimension reductions for real-life data-points, which might be flawed [3, 6]. Moreover, problem (4) is a direct generalization of the problem (2) studied in [2].

In Section 2, we show how (4) arises from (3) if we assume the matrices A and B not to be constant but parameter-dependent. As in [2] we use this parameter-dependence to model the influence of uncertainty. Thereafter we examine the first properties of (4) in Section 3, before we calculate the derivatives of its numerator and denominator in Section 4. We will use the knowledge from these sections to approach a solution in Section 5. In Section 6 our theoretical insights conclude in an new algorithm, which is finally tested on a small example in Section 7. As we will see, the new algorithm of this paper is capable of finding good solutions for this small example, motivating further investigations.

2 The problem

Let $\Omega \subset \mathbb{R}^m$ and $\Gamma \subset \mathbb{R}^k$ be compact sets of parameters and let the matrix-valued functions

$$A : \mu \in \Omega \mapsto A(\mu) \in \mathbb{R}^{n\times n},$$
$$B : \xi \in \Gamma \mapsto B(\xi) \in \mathbb{R}^{n\times n},$$

be smooth and symmetric positive definite for all $\mu \in \Omega$ and $\xi \in \Gamma$. Subsequently, we will call

$$\frac{\operatorname{tr}(V^T A(\mu)V)}{\operatorname{tr}(V^T B(\xi)V)},$$

the robust trace quotient of $A(\mu)$ and $B(\xi)$. We want to minimize this quotient over $\operatorname{St}(p,n)$ for the "worst choice" of parameters μ and ξ, which leads to the min-max problem

$$\min_{V\in\mathbb{R}^{n\times p}} \max_{\mu\in\Omega,\xi\in\Gamma} \frac{\operatorname{tr}(V^T A(\mu)V)}{\operatorname{tr}(V^T B(\xi)V)} \quad \text{s.t.} \quad V^T V = I_p. \tag{5}$$

Throughout we will assume the parameters μ and ξ to be independent from each other. This enables us to solve the maximization problem by maximizing the numerator and minimizing the denominator separately. Therefore we define for fixed $V \in \mathbb{R}^{n\times p}$

$$\mu^*(V) := \operatorname{argmax}_{\mu\in\Omega} \operatorname{tr}(V^T A(\mu)V),$$
$$\xi^*(V) := \operatorname{argmin}_{\xi\in\Gamma} \operatorname{tr}(V^T B(\xi)V).$$

In the case of non-unique optimizers, μ^* and ξ^* denote any of the optimizers. Recall that, as Ω and Γ are assumed to be compact and $A(\mu)$ and $B(\xi)$ are smooth,

[1] Note that, for $p = 1$ problems (3) and (1) are equivalent.

$\mu^*(V)$ and $\xi^*(V)$ are well-defined for all $V \in \mathbb{R}^{n\times p}$. Via $G(V) := A(\mu^*(V))$ and $H(V) := B(\xi^*(V))$ problem (5) becomes

$$\min_{V\in \mathrm{St}(p,n)} \frac{\max_\mu \mathrm{tr}(V^T A(\mu)V)}{\min_\xi \mathrm{tr}(V^T B(\xi)V)} = \min_{V\in \mathrm{St}(p,n)} \frac{\mathrm{tr}(V^T G(V)V)}{\mathrm{tr}(V^T H(V)V)}. \tag{6}$$

From this point on, we will call $\frac{\mathrm{tr}(V^T G(V)V)}{\mathrm{tr}(V^T H(V)V)}$ the non-linear trace quotient of $(G(\cdot), H(\cdot))$.

3 Basic properties

We start with some elementary properties of the non-linear trace quotient of $(G(\cdot), H(\cdot))$ in (6).

Lemma 1.

1. *For a fixed $V \in \mathrm{St}(p,n)$, $G(V) = G(V)^T$ and $H(V) = H(V)^T$ are both positive definite.*
2. *For a fixed $V \in \mathrm{St}(p,n)$ and $\alpha \in \mathbb{R}\setminus\{0\}$, $G(\alpha V) = G(V)$ and $H(\alpha V) = H(V)$ holds, i.e. $G(\cdot)$ and $H(\cdot)$ are homogeneous in V.*
3. *For a fixed $V \in \mathrm{St}(p,n)$ and orthogonal $Q \in \mathbb{R}^{p\times p}$, $G(VQ) = G(V)$ and $H(VQ) = H(V)$. Furthermore the non-linear trace quotient (6) is invariant under such transformations.*

Proof.

1. Since $A(\mu)$ and $B(\xi)$ are symmetric and positive definite for all $\mu \in \Omega$ and $\xi \in \Gamma$ and $\mu^*(V) \in \Omega$ and $\xi^*(V) \in \Gamma$, $G(V) = A(\mu^*(V))$ and $H(V) = B(\xi^*(V))$ holds and both matrices are symmetric positive definite.
2. For $\alpha \in \mathbb{R}\setminus\{0\}$ we have

$$\begin{aligned}\mu^*(\alpha V) &= \mathrm{argmax}_{\mu\in\Omega} \mathrm{tr}((\alpha V)^T A(\mu)(\alpha V))\\ &= \mathrm{argmax}_{\mu\in\Omega} \alpha^2 \mathrm{tr}(V^T A(\mu)(V))\\ &= \mathrm{argmax}_{\mu\in\Omega} \mathrm{tr}(V^T A(\mu)(V)) = \mu^*(V),\end{aligned}$$

because we are interested in the maximizing parameter and not in the maximum itself. It follows that $G(\alpha V) = A(\mu^*(\alpha V)) = A(\mu^*(V)) = G(V)$. The proof for $H(\cdot)$ is the same.

3. Let $Q \in \mathbb{R}^{p\times p}$ be orthogonal, then

$$\begin{aligned}\mu^*(VQ) &= \mathrm{argmax}_\mu \mathrm{tr}((VQ)^T A(\mu)(VQ))\\ &= \mathrm{argmax}_\mu \mathrm{tr}(Q^T V^T A(\mu)VQ)\\ &= \mathrm{argmax}_\mu \mathrm{tr}(V^T A(\mu)VQQ^T)\\ &= \mathrm{argmax}_\mu \mathrm{tr}(V^T A(\mu)V) = \mu^*(V)\end{aligned}$$

due to the cyclic invariance of the trace function. As before, this implies

$$G(VQ) = A(\mu^*(VQ)) = A(\mu^*(V)) = G(V).$$

The proof for $H(\cdot)$ follows similarly. We now obtain

$$\mathrm{tr}((VQ)^T G(VQ)(VQ)) = \mathrm{tr}(V^T G(V)V)$$

and

$$\mathrm{tr}((VQ)^T H(VQ)(VQ)) = \mathrm{tr}(V^T H(V)V)$$

and therefore the non-linear trace quotient is invariant under such transformations.

□

For a rigorous mathematical treatment of (6) and inspired by [2, Def. 2.2] we restrict our choices for $V \in \mathrm{St}(p,n)$ even more by the following definition:

Definition 2. We call $V \in \mathrm{St}(p,n)$ *regular* if $\mu^*(V)$ and $\xi^*(V)$ are twice continuously differentiable at V. Moreover, we call problem (6) regular, if $\mu^*(V)$ and $\xi^*(V)$ are twice continuously differentiable for all $V \in \mathbb{R}^{n\times p}$.

Regularity can a priori not be guaranteed from the formulation of problem (5). However, if the parameters $\mu^*(V)$ and $\xi^*(V)$ have analytic expressions, it will not be a severe restriction (see [2, Sec. 2]). However, this does not mean that explicit expressions for $\mu^*(V)$ and $\xi^*(V)$ can be easily evaluated or are even available. From now on, we will assume problem (6) to be regular. In particular, this implies that $\mathrm{tr}(V^T G(V)V)$ and $\mathrm{tr}(V^T H(V)V)$ are differentiable for V. We will calculate these differentials in Section 4 as they will be important for the derivation of our solution strategy for problem (6) in the subsequent sections.

4 Derivatives of numerator and denominator

We now want to calculate the derivative of the numerator $g(V) = \mathrm{tr}(V^T G(V)V)$ and the denominator $h(V) = \mathrm{tr}(V^T H(V)V)$ of our non-linear trace quotient. As both terms only differ in the label of their central matrix, we focus on the numerator and apply the same results to the denominator.

We start with the derivative of a single entry of V. For $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$ we get

$$\begin{aligned}\frac{\partial g(V)}{\partial v_{ij}} &= \mathrm{tr}\left(\frac{\partial}{\partial v_{ij}}(V^T G(V)V)\right)\\ &= \mathrm{tr}\left(\frac{\partial V^T}{\partial v_{ij}} G(V)V + V^T \frac{\partial G(V)}{\partial v_{ij}} V + V^T G(V)\frac{\partial V}{\partial v_{ij}}\right)\\ &= \mathrm{tr}\left(e_j e_i^T G(V)V + V^T G(V) e_i e_j^T + V^T \frac{\partial G(V)}{\partial v_{ij}} V\right)\\ &= 2e_i^T G(V)Ve_j + \mathrm{tr}\left(V^T \frac{\partial G(V)}{\partial v_{ij}} V\right),\end{aligned} \tag{7}$$

where $e_i \in \mathbb{R}^n$ and $e_j \in \mathbb{R}^p$ are the i-th and j-th unit-vectors of the respective spaces[2]. Before we continue, we present a elegant property that results from the assumption of our problem being regular and which will be helpful in simplifying the derivative in (7).

Lemma 3. *Let $V \in \mathrm{St}(p,n)$ be regular, then*

$$\mathrm{tr}\left(V^T \frac{\partial G(V)}{\partial v_{ij}} V\right) \equiv 0 \quad \textit{and} \quad \mathrm{tr}\left(V^T \frac{\partial H(V)}{\partial v_{ij}} V\right) \equiv 0$$

holds for $1 \le i \le n$ and $1 \le j \le p$.

Proof. We will only show the first equation since the second follows in the same way.

We define $f(t) = \mathrm{tr}(V^T A(\mu^*(V + te_i e_j^T))V)$ and observe that

$$\begin{aligned} f(0) &= \mathrm{tr}(V^T A(\mu^*(V))V) = \max_{\mu} \mathrm{tr}(V^T A(\mu)V) \\ &\ge \mathrm{tr}(V^T A(\mu^*(V + te_i e_j^T))V) = f(t) \end{aligned}$$

holds. Moreover, according to our assumptions, $f(t)$ is continuously differentiable at $t = 0$. Thus $f(t)$ has a local maximum at $t = 0$ and therefore we obtain

$$\begin{aligned} 0 = f'(0) &= \frac{d\,\mathrm{tr}(V^T A(\mu^*(V + te_i e_j^T))V)}{dt} \\ &= \mathrm{tr}\left(V^T \frac{dA(\mu^*(V + te_i e_j^T))}{dt}\bigg|_{t=0} V\right) = \mathrm{tr}\left(V^T \frac{\partial G(V)}{\partial v_{ij}} V\right). \end{aligned}$$

□

We can use the result from Lemma 3 to simplify the derivative of our numerator in (7) to

$$\frac{\partial g(V)}{\partial v_{ij}} = e_i^T 2G(V)Ve_j,$$

for regular V, which in matrix notation directly leads to $\nabla_V g(V) = 2G(V)V$. Similarly we have $\nabla_V H(V) = 2H(V)V$. Notice that these compact expression resemble the ones for the derivative in case of constant matrices, i.e.

$$\nabla_V \mathrm{tr}(V^T AV) = 2AV$$

whenever $A \in \mathbb{R}^{n\times n}$. However, do not overlook that this is due to the assumed regularity of the problem.

In the next section we reformulate our minimization problem of the non-linear trace quotient to obtain a non-linear eigenvalue problem. Solving this eigenvalue problem will be an intermediate step in our solution strategy to problem (6) presented in Section 6.

5 Approaching a solution

We now want to follow an approach similar to the one Saad et al. took in [3] for the case of constant matrices G and H. Over all $V \in \mathrm{St}(p,n)$ we want to minimize

$$\frac{\mathrm{tr}(V^T G(V)V)}{\mathrm{tr}(V^T H(V)V)}$$

and assume the problem is regular. Since $H(V)$ is positive definite the denominator is never zero. Moreover, as $G(V)$ and $H(V)$ depend continuously on V, the non-linear trace quotient is continuous in V. As $\mathrm{St}(p,n)$ is compact [1], we know that there exists some $V^* \in \mathrm{St}(p,n)$ such that for all $V \in \mathrm{St}(p,n)$

$$\rho^* := \frac{\mathrm{tr}(V^{*T} G(V^*)V^*)}{\mathrm{tr}(V^{*T} H(V^*)V^*)} \le \frac{\mathrm{tr}(V^T G(V)V)}{\mathrm{tr}(V^T H(V)V)},$$

i.e. $\rho^* = \min_{V\in\mathrm{St}(p,n)} \frac{\mathrm{tr}(V^T G(V)V)}{\mathrm{tr}(V^T H(V)V)}$ and V^* is a corresponding minimizer. Using the linearity of the trace, this leads to

$$0 \le \mathrm{tr}(V^T (G(V) - \rho^* H(V))V) \tag{8}$$

for all $V \in \mathrm{St}(p,n)$ where zero is attained for $V = V^*$. From (8) we define the scalar-valued function

$$f(\rho) = \min_{V\in\mathrm{St}(p,n)} \mathrm{tr}(V^T (G(V) - \rho H(V))V) \tag{9}$$

and notice that (8) implies that (the value of) our minimal trace quotient ρ^* is a root of f. We will show in corollary 5 that in fact ρ^* is the only root of this function.

Our search for a matrix $V \in \mathrm{St}(p,n)$, such that the non-linear trace quotient corresponding to $G(\cdot)$ and $H(\cdot)$ is minimal, has become the search for a root of a scalar function with, as we will see below, helpful properties. Before we continue on this path, we will have to bring the function f in (9) into a more elegant shape, as for now it consists of an implicit optimization problem.

5.1 The evaluation of $f(\rho)$

As a first step, we consider the problem of finding $f(\rho)$ for a given $\rho \in \mathbb{R}$. That is, we intend to solve

$$\min_{V\in\mathbb{R}^{n\times p}} \mathrm{tr}(V^T (G(V) - \rho H(V))V), \quad \text{s.t. } V^T V = I_p, \tag{10}$$

where $G(V) \in \mathbb{R}^{n\times n}$ and $H(V) \in \mathbb{R}^{n\times n}$ are symmetric positive definite and $\rho \in \mathbb{R}$. Again, as this is a continuous function on $\mathrm{St}(p,n)$, there has to exist a minimal value and a corresponding minimizer V^*. The Lagrangian function for this problem is

$$L(V,\Lambda) = \mathrm{tr}(V^T (G(V) - \rho H(V))V) - \mathrm{tr}(\Lambda(V^T V - I_p)),$$

[2] Although from different spaces, we want to denote both this way out of simplicity, ignoring the danger of confusion.

where Λ denotes the matrix of Lagrange multipliers [8]. For the derivative of L we need the derivatives of both trace terms. Due to the linearity of the trace, we know the former from Section 4 and conclude

$$\nabla_V \operatorname{tr}(V^T(G(V)-\rho H(V))V) = 2(G(V)-\rho H(V))V.$$

In a similar way we calculate the derivative of the latter term and since Λ has to be symmetric due to the symmetric constraint, we obtain

$$\nabla_V \operatorname{tr}(\Lambda(V^T V - I_p)) = 2V\Lambda.$$

This leads to the condition that a minimizer V^* of (10) satisfies

$$(G(V^*)-\rho H(V^*))V^* = V^*\hat{\Lambda} \quad \text{and} \quad (V^*)^T V^* = I_p, \tag{11}$$

for a matrix $\hat{\Lambda} \in \mathbb{R}^{p\times p}$ (notice that $\hat{\Lambda}$ is necessarily symmetric). In other words, the columns of the minimizer span an invariant eigenspace of the orthonormal eigenvectors of the symmetric matrix $G(V^*)-\rho H(V^*)$. Since we are looking for a minimum, the corresponding eigenvalues (i.e. the eigenvalues of $\hat{\Lambda}$) should be as small as possible, since it follows from (11) that

$$(V^*)^T(G(V^*)-\rho H(V^*))V^* = \hat{\Lambda} \tag{12}$$

and, according to (8),

$$\operatorname{tr}((V^*)^T(G(V^*)-\rho H(V^*))V^*) = \operatorname{tr}(\hat{\Lambda}) = \lambda_1+\cdots+\lambda_p$$

where $\lambda_1,\ldots,\lambda_p \in \mathbb{R}$ are the eigenvalues of $\hat{\Lambda}$. In other words, this condition could be interpreted as a nonlinear eigenproblem of the matrix $G(\cdot)-\rho H(\cdot)$. One option to find V^* is the use of the SCF iteration. As our solution strategy to problem (6), which will be outlined in Section 6, is based on Newton's method for finding a root of $f(\rho)$, we will discuss the derivative of $f(\rho)$ in the next section and postpone the discussion of the SCF iteration for the solution of (11) to Section 6.1.

5.2 The derivative of f

Since it is obvious that the solution V^* of (11) depends on the chosen parameter ρ, we may write $V(\rho)$ instead of V^* and express f in the form

$$f(\rho) = \operatorname{tr}(V(\rho)^T(G(V(\rho))-\rho H(V(\rho)))V(\rho)),$$

where $V(\rho) \in \operatorname{St}(p,n)$ denotes a matrix, which satisfies the eigenproblem condition (12) from above for the p smallest eigenvalues. To continue our analysis, we want $f(\rho)$ to be a differentiable function. In particular, this means that the solution matrix $V(\rho)$ of (12) has to vary in a differentiable way with ρ. Certainly, whenever $V(\rho)$ is a solution to (10), then so is $V(\rho)Q$ for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$. However, Kato showed for a more general case in [5, Ch.II,Sec.6.2] that in fact a basis of orthogonal eigenvectors, varying in a differentiable way with ρ, exists. Therefore we can calculate the derivative of f. Before we get to do this, we want to note a result, which is a small variation of Theorem 3 and will be quite useful in a moment.

Corollary 4. *Let $V(\rho) \in \operatorname{St}(p,n)$ be regular and differentiable in ρ, then*

$$\operatorname{tr}\left(V(\rho)^T \frac{dG(V(\rho))}{d\rho} V(\rho)\right) \equiv 0 \quad \textit{and}$$
$$\operatorname{tr}\left(V(\rho)^T \frac{dH(V(\rho))}{d\rho} V(\rho)\right) \equiv 0$$

holds.

Proof. The proof is very similar to the one for Theorem 3. Since $V(\rho)$ is continuous in ρ, we examine the function $h(t) = \operatorname{tr}(V(\rho)^T A(\mu^*(V(\rho+t)))V(\rho))$ and follow the same steps as before. □

With this in mind we can start differentiating f. We start with a useful observation. Since $V(\rho)$ consists of orthonormal columns $V(\rho)^T V(\rho) = I_p$ holds and we get

$$0 = \frac{dI_p}{d\rho} = \frac{dV(\rho)^T V(\rho)}{d\rho} = \frac{dV(\rho)^T}{d\rho}V(\rho) + V(\rho)^T\frac{dV(\rho)}{d\rho}, \tag{13}$$

which implies that $\frac{dV(\rho)^T}{d\rho}V(\rho)$ is skew-symmetric and thus its diagonal elements must all be zero. Since the columns of $V(\rho)$ span an invariant eigenspace, $V(\rho)$ fulfils

$$(G(V(\rho))-\rho H(V(\rho)))V(\rho) = V(\rho)\Lambda(\rho),$$

where $\Lambda(\rho)$ is a diagonal matrix with the corresponding eigenvalues. We now conclude

$$
\begin{aligned}
\frac{df(\rho)}{d\rho} &= \frac{d}{d\rho}\operatorname{tr}(V(\rho)^T(G(V(\rho))-\rho H(V(\rho)))V(\rho)) \\
&= \operatorname{tr}\left(\frac{dV(\rho)^T}{d\rho}(G(V(\rho))-\rho H(V(\rho)))V(\rho)\right) \\
&\quad + \operatorname{tr}\left(V(\rho)^T\frac{d(G(V(\rho))-\rho H(V(\rho)))}{d\rho}V(\rho)\right) \\
&\quad + \operatorname{tr}\left(V(\rho)^T(G(V(\rho))-\rho H(V(\rho)))\frac{dV(\rho)}{d\rho}\right) \\
&= \operatorname{tr}\left(\frac{dV(\rho)^T}{d\rho}V(\rho)\Lambda(\rho)\right) + \operatorname{tr}\left(\Lambda(\rho)V(\rho)^T\frac{dV(\rho)}{d\rho}\right) \\
&\quad + \operatorname{tr}\left(V(\rho)^T\frac{dG(V(\rho))}{d\rho}V(\rho)\right) \\
&\quad - \rho\operatorname{tr}\left(V(\rho)^T\frac{dH(V(\rho))}{d\rho}V(\rho)\right) \\
&\quad - \operatorname{tr}(V(\rho)^T H(V(\rho))V(\rho)) \\
&= -\operatorname{tr}(V(\rho)^T H(V(\rho))V(\rho)). \qquad (14)
\end{aligned}
$$

The last equality follows from corollary 4 and our previous observation from (13)[3]. This leads to the following corollary:

Corollary 5. *The function* $f(\rho) = \operatorname{tr}(V(\rho)^T(G(V(\rho)) - \rho H(V(\rho)))V(\rho))$ *has exactly one root.*

Proof. Since $H(V(\rho))$ is positive definite for every ρ by assumption, we easily see that the derivative $f'(\rho)$ is always less than zero and thus f is monotonically decreasing. So f can have at most one root, which (as we saw before) it has. □

6 The Newton-SCF algorithm

The next natural question is: How do we find the root of $f(\rho)$ and thus the minimum of our non-linear trace quotient? Since we have just seen that f is differentiable in ρ, we want to utilize one of the most classic methods for such a task, Newton's method. As an iteration we get directly from (14)

$$\rho_{k+1} = \rho_k - \frac{f(\rho_k)}{f'(\rho_k)} = \frac{\operatorname{tr}(V(\rho_k)^T G(V(\rho_k))V(\rho_k))}{\operatorname{tr}(V(\rho_k)^T H(V(\rho_k))V(\rho_k))},$$

which is surprisingly the non-linear trace quotient at $V(\rho_k)$. At this point it is important to notice that ρ_{k+1} is a root of the function

$$\tilde{f}_k(\rho) = \operatorname{tr}(V(\rho_k)^T(G(V(\rho_k)) - \rho H(V(\rho_k)))V(\rho_k)),$$

as one can see by simply plugging ρ_{k+1} into $\tilde{f}_k$. Therefore one needs to be cautious when checking the size of $f(\rho_{k+1})$ (for example, to examine the quality of our solution), to not accidentally calculate $\tilde{f}_k(\rho_{k+1})$, which will in finite arithmetic always give a value near zero, whether ρ_{k+1} is a good approximation to the root of f or not.

The question remains how to get $V(\rho)$ for a certain ρ. This is exactly the problem (10) we derived in Section 5.1. We want $V(\rho)$ to satisfy the eigenproblem condition (11) for $G(\cdot) - \rho H(\cdot)$. To find such a matrix, we want to construct a generalization of the SCF iteration.

6.1 The Generalized SCF Iteration

We are now looking at a broader problem which includes our problem as a special case. We want to find a solution $V \in \mathbb{R}^{n\times p}$ for

$$A(V)V = B(V)V\Lambda \quad \text{with} \quad V^T B(V)V = I_p, \qquad (15)$$

where $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix and $A(V), B(V) \in \mathbb{R}^{n\times n}$ are smooth functions of $V \in \mathbb{R}^{n\times p}$. In other words, we want to solve the generalized non-linear eigenvalue problem for the matrix pencil $\lambda B(\cdot) - A(\cdot)$. Notice that this is exactly our problem from (11) for $A(V) = G(V) - \rho H(V)$ and $B(V) = I_n$.

To do this, we want to generalize the SCF Iteration, which is used to solve this problem for the case $p = 1$, see [2, Sec.4]. The idea is rather simple. We start with a matrix $V_0 \in \mathbb{R}^{n\times p}$, which could be arbitrary or an initial guess, with $B(V_0)$-orthonormal columns and plug it into our matrices $A(\cdot)$ and $B(\cdot)$. Then we calculate p generalized eigenvalues and corresponding eigenvectors of this standard matrix pencil $\lambda B(V_0) - A(V_0)$ and use them to build V_1. In general, we follow the rule

$$V_{k+1} \leftarrow B(V_k)\text{-orth. eigenvec. of } \lambda B(V_k) - A(V_k). \qquad (16)$$

Notice that generalized eigenvectors $v_i, v_j \in \mathbb{R}^n$ for eigenvalues $\lambda_i \neq \lambda_j$ of $\lambda H(V_k) - G(V_k)$ are indeed $B(V_k)$-orthogonal, since

$$
\begin{aligned}
\lambda_j(v_i^T B(V_k)v_j) = v_i^T A(V_k)v_j = v_j^T A(V_k)v_i &= \lambda_i(v_j^T B(V_k)v_i) \\
&= \lambda_i(v_i^T B(V_k)v_j),
\end{aligned}
$$

so if $\lambda_i \neq \lambda_j$ we necessarily have $v_i^T B(V_k)v_j = 0$. Furthermore, recall that all eigenvalues of $\lambda B(V_k) - A(V_k)$ are real (see Section 1). Since the non-linear eigenvalues we are looking for should be as small as possible, it seems reasonable to choose the eigenvectors corresponding to the smallest p eigenvalues of $\lambda B(V_k) - A(V_k)$ in every iteration.

[3] Since their diagonals are zero and $\Lambda(\rho)$ is diagonal, the same holds for the diagonals of the products and thus the trace vanishes.

A necessary condition for the iteration (16) to converge is that there exists a matrix $\tilde{V} \in \mathbb{R}^{n\times p}$ consisting of eigenvectors corresponding to the p smallest generalized eigenvalues of $\lambda B(\tilde{V}) - A(\tilde{V})$. It could be the case that the columns of V are indeed generalized eigenvectors of this matrix pencil, but the corresponding eigenvalues are not the p smallest, in which case our iteration could get stuck in a loop. For a detailed analysis of this phenomenon in the case $p = 1$ see [2, Sec. 4]. To prevent this, we propose the following transformation of (15), generalizing [2, Sec.4.1]:

$$A_\sigma(V) = B(V)V\Lambda_\sigma \quad \text{with} \quad V^T B(V)V = I_p,$$

with

$$A_\sigma(V) := A(V) - \sigma(V)\sum_{i=1}^{p} \frac{B(V)v_i v_i^T B(V)}{v_i^T B(V)v_i},$$

where v_i denotes the i-th column of V and $\sigma(V)$ is a scalar function. Let us see what happens if we multiply $A_\sigma(V)$ with v_j if V is a solution of (15):

$$\begin{aligned} A_\sigma(V)v_j =& A(V)v_j - \sigma(V)\sum_{i=1}^{p} \frac{B(V)v_i v_i^T B(V)v_j}{v_i^T B(V)v_i} \\ =& B(V)v_j\lambda_j - \sigma(V)\sum_{i=1}^{p} B(V)v_i\delta_{ij} \\ =& B(V)v_j\lambda_j - \sigma(V)B(V)v_j \\ =& B(V)v_j(\lambda_j - \sigma(V)) \end{aligned}$$

We see that the columns of V are also eigenvectors of $\lambda B(V) - A_\sigma(V)$, but with their eigenvalue shifted by $-\sigma(V)$. We can also conclude that generalized eigenvalues of $\lambda B(V) - A(V)$, of which eigenvectors are not columns[4] of V, are not effected by this shift, because the latter term in the first row of the equation chain above is equal to zero. The following theorem tells us how we can use this to our advantage.

Theorem 6. *Let the columns of $V \in \mathbb{R}^{n\times p}$ be non-linear generalized eigenvectors of $\lambda B(\cdot) - A(\cdot)$ with corresponding eigenvalues $\lambda_1, \dots, \lambda_p$. Let*

$$\sigma(V) = \beta\lambda_{max} - \lambda_{min},$$

where λ_{max} and λ_{min} denote the largest and smallest generalized eigenvalues of $\lambda B(V) - A(V)$ respectively and $\beta > 1$. Then the columns of V are generalized eigenvectors corresponding to the p smallest eigenvalues of $\lambda B(V) - A_\sigma(V)$.

Proof. We have seen that the j-th column of V is an eigenvector of $\lambda B(V) - A_\sigma(V)$, corresponding to the eigenvalue $\lambda_j - \sigma(V)$ and that other eigenvectors w of this matrix pencil keep their eigenvalues μ. We now get

$$\lambda_j - \sigma(V) = \lambda_j - \beta\lambda_{max} + \lambda_{min} < \underbrace{\lambda_j - \lambda_{max}}_{\leq 0} + \mu \leq \mu,$$

which shows that the eigenvalues corresponding to the columns of V are smaller than the remaining eigenvalues of $\lambda B(V) - A_\sigma(V)$. □

As we have seen $\lambda B(V) - A(V)$ and $\lambda B(V) - A_\sigma(V)$ share the same eigenvectors but with shifted eigenvalues corresponding to the eigenvectors in V. By this we can ensure, that the eigenvectors in V belong to the p smallest eigenvalues of the eigenproblem and so preventing the iteration (16) to get stuck in a loop.

Summarizing our results, we get the generalized SCF Iteration (see Algorithm 1).

Algorithm 1 Generalized SCF Iteration

```
Input: $A(\cdot), B(\cdot), p, V_0, \beta$
Output: $V_k = SCF(A, B, p, V_0)$
**while** no convergence **do**
    $\lambda_{max} = \lambda_{max}(\lambda B(V_{k-1}) - A(V_{k-1}))$
    $\lambda_{min} = \lambda_{min}(\lambda B(V_{k-1}) - A(V_{k-1}))$
    $\sigma = \beta \cdot \lambda_{max} - \lambda_{min}$
    $A_\sigma(V_{k-1}) = A(V_{k-1}) - \sigma\sum_{i=1}^{p} \frac{B(V_{k-1})v_i v_i^T B(V_{k-1})}{v_i^T B(V_{k-1})v_i}$
    $V_k \leftarrow B(V_{k-1})$-orth. eigenvectors to the $p$ smallest eigenvalues of $(A_\sigma(V_{k-1}), B(V_{k-1}))$
**end while**
**return** $V_k$
```

In practice we use $\beta = 1.01$, but other choices might work as good or even better.

6.2 Putting the pieces together

As we mentioned at the beginning of this section, we want to use Newton's method to find the root of f. To calculate $V(\rho)$, which corresponds to the minimizer we are looking for, we want to use the generalized SCF Iteration. Basically our algorithm only consists of these two parts, therefore we want to call it the 'Newton-SCF Algorithm' (see Algorithm 2).

[4] To be more precise, we mean the linear span of the columns.

Algorithm 2 Newton-SCF

```
Input: $G(\cdot), H(\cdot), p, V_0, \beta$
Output: $V_k = \text{Newton-SCF}(G, H, p, V_0, \beta)$
$\rho \leftarrow \frac{\mathrm{tr}(V_0^T G(V_0) V_0)}{\mathrm{tr}(V_0^T H(V_0) V_0)}$
while no convergence do
    $V_k \leftarrow SCF(G(\cdot) - \rho H(\cdot), I_n, p, V_{k-1})$
    $\rho \leftarrow \frac{\mathrm{tr}(V_k^T G(V_k) V_k)}{\mathrm{tr}(V_k^T H(V_k) V_k)}$
end while
return $V_k$
```

Notice that, since we want to have non-linear eigenvectors of a single matrix instead of generalized eigenvectors of a matrix pencil, we initialize our SCF iteration with $G(\cdot) - \rho H(\cdot)$ and the constant identity matrix I_n.

6.3 Effort management

Since we call an SCF iteration in every iteration of the Newton-SCF-algorithm, it is reasonable to ask how much we should invest in these iterations and whether it makes sense to adjust this investment over time.
The first thought could be that in the beginning we invest less in the SCF iteration, since we do not need high precision when our main iteration is far from converging. However, small and easy examples suggest that the opposite might be the case and this could be explained as follows:

If ρ_{k-1} and ρ_k are not far apart, the matrices $V(\rho_{k-1})$ and $V(\rho_k)$ will also be "close" to each other, since $V(\rho)$ depends continuously on ρ. Also the non-linear eigenvalue problems $G(\cdot) - \rho_{k-1} H(\cdot)$ and $G(\cdot) - \rho_k H(\cdot)$ are pretty similar. But then $V(\rho_{k-1})$ should be a good approximation for the eigenvectors of $G(\cdot) - \rho_k H(\cdot)$ and since we start our SCF iteration with $V(\rho_{k-1})$, we should only need a few steps to get $V(\rho_k)$. So, approaching our solution ρ^*, we should need fewer and fewer steps for the SCF iterations.

7 A small experiment

In this section, we want to give a proof of concept how the Newton-SCF-Algorithm works in a small experiment. We also want to show that the classic SCF Iteration with the generalized SCF-Algorithm is not capable to minimize the non-linear trace quotient in general.

To set up our experiment, we construct the parameter dependent matrices as follows: We choose $A_0, A_1, A_2 \in \mathbb{R}^{n \times n}$ with random entries uniformly distributed between 0 and 1 and define

$$\tilde{A}(\mu) := A_0 + g_1(\mu) A_1 + g_2(\mu) A_2 \quad \text{and} \quad A(\mu) := \tilde{A}^T(\mu) \tilde{A}(\mu).$$

We proceed with B in the same way with functions denoted $h_1(\xi)$ and $h_2(\xi)$. To continue, we have to solve two optimization problems to obtain $\mu^*(V)$ and $\xi^*(V)$. We determine $\mu^*(V)$ and $\xi^*(V)$ numerically, i.e. we use a discretization of the parameter spaces, calculate the value for each of the candidates and determine the parameter with the smallest or largest value.

As parameter spaces we use $\Omega = \Gamma = [-1, 1]^2$ and $\mu = \xi = (x, y) \in [-1, 1]^2$. The discretization of the parameter spaces for the determination of $\mu^*(V)$ and $\xi^*(V)$ is done with a $\Delta = 0.02$ step-size. For A we use $g_1(\mu) = e^{x+y}$ and $g_2(\mu) = -e^{x+y}$ as "parameter functions" and for B we use $h_1(\xi) = \frac{x+10}{y+10}$ and $h_2(\xi) = \frac{y+10}{x+10}$. To give our solutions found by the Newton-SCF-Algorithm something to compare with, we also try to calculate a solution by simply applying the (generalized) SCF iteration to our problem (6), in the same way [2] did for (2). In [2] it is shown that this leads to good results for $p = 1$. Both procedures need (approximately) the same number of iterations to converge. The SCF iteration does 500 iterations and the Newton-SCF-Algorithm does 25 newton steps and within each step up to 20 SCF-Iterations. We also show the value of our non-linear trace quotient for 100 random matrices $W \in \mathbb{R}^{n \times p}$ with orthonormal columns, to check whether the Newton-SCF-Algorithm is better or worse than simply guessing a solution.

As we can see, the Newton-SCF-Algorithm is superior to simply applying the SCF iteration or guessing random matrices (see Figure 1 left plot). Inspecting the development of the trace quotient values for each iteration for the SCF iteration, it seems that they indeed converge, but very slowly and maybe not to a value which could be considered a minimum for the trace-quotient (see Figure 1 central plot). However the Newton-SCF-Algorithm has a very decent convergence speed and the values seem to be (in comparison to the random values) pretty small (see Figure 1 right plot).

Since this example is rather small, further investigations and experiments in this regard should be made to validate this approach.

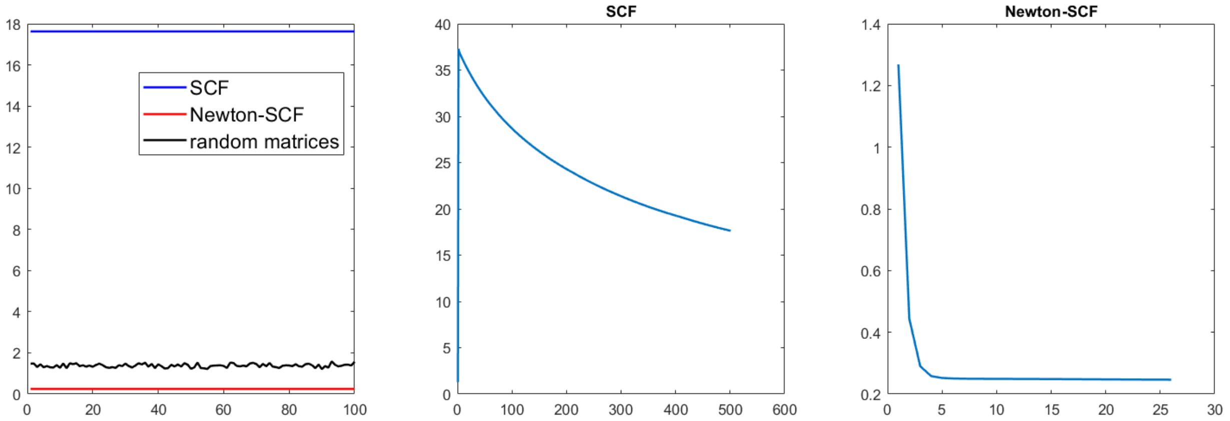

Figure 1 – *Left:* Comparison of found solutions of the SCF Iteration (blue), our Newton-SCF-Alg. (red) and in comparison the value of the non-linear trace quotient for random matrices (black). *Middle:* Values of the trace quotient in each iteration of the SCF Iteration. *Right:* Values of the trace quotient in each iteration of the Newton-SCF-Alg..

8 Summary

In this paper we investigated the optimization of the non-linear trace quotient (6) and derived a new algorithm to solve this problem. We reformulated the problem as a non-linear eigenvalue problem and used an approach based on the ideas of [2] and [3] for its solution. We gave proof to some results regarding this non-linear eigenvalue problem in a broader context and showed how the SCF Iteration and an involved shift technique from [2] can be adapted to the multidimensional case. Finally, we demonstrated the performance of the algorithm in a small experiment.

Code Availability: The MATLAB source code of the implementation used to compute the presented results is available as supplementary material and can be obtained under

DOI:10.14464/gammas.v5i1.540.

References

[1] P.-A. Absil, R. Mahony, and H. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Pres, 2008. ISBN 9780691132983.

[2] Zhaojun Bai, Ding Lu, and Bart Vandereycken. Robust rayleigh quotient minimization and nonlinear eigenvalue problems. *SIAM J. SCI. COMPUT.*, 40(5), 2018.

[3] Mohammed Bellalij and Yousef Saad. The trace ratio optimization problem for dimensionaliy reduction. *SIAM J. MATRIX ANAL. APPL.*, 31(5), 2010.

[4] M. R. Guarracino, C. Cifarelli, O. Seref, and P. M. Pardalos. A classification method based on generalized eigenvalue problems. *Optimization Methods and Software*, 22(1):73–81, 2005.

[5] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 1976. ISBN 3-540-07558-5.

[6] Effrosyni Kokiopoulou, Jie Chen, and Yousef Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18, 2011.

[7] Ahmed Sameh and Zhanye Tong. The trace minimization method for the symmetric generalized eigenvalue problem. *Journal of Computational and Applied Mathematics*, 123:155–175.

[8] Willi-Hans Steeb and Yorick Hardy. *Matrix Calculus and Kronecker Product*. WORLD SCIENTIFIC, 2nd edition, 2011.