Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A Comparison of Algorithms Related to Trace Minimization to Compute a Small Number of Eigenvalues of a Real Symmetric Matrix

Term Project Thesis

Giuseppe Accaputo
accaputg@ethz.ch

Supervisor: Prof. Dr. Peter Arbenz
Department of Computer Science
ETH Zürich
Zürich, Switzerland

January 27, 2017

Abstract. Eigenvalue problems arise in many computational science and engineering applications. In this thesis, algorithms for computing few of the smallest (or largest) eigenvalues and associated eigenvectors of the large sparse generalized eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{Bx}$ are derived and compared. The trace minimization method by Sameh and Wisniewski [13] is derived and a detailed proof of the trace theorem [13] is presented. A characterization of the trace minimization method as a quasi-Newton method is given by deriving expressions for the Hessian matrix and the gradient. The Jacobi-Davidson method by Sleijpen & van der Vorst [14, 15] is derived and compared to the trace minimization method. The Davidson-type trace minimization method by Sameh and Tong [12] is introduced as a subspace expanding trace minimization method and compared to the block Jacobi-Davidson method.

1. Introduction

We consider the problem of computing a few of the smallest eigenvalues or eigenvectors of the large, sparse, generalized eigenvalue problem

$$\mathbf{Ax} = \lambda\mathbf{Bx}\,, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n,\ \lambda \in \mathbb{R}$ and $\mathbf{A}, \mathbf{B}$ are $n \times n$ symmetric matrices, with $\mathbf{B}$ being positive-definite. The matrix $\mathbf{A} - \lambda\mathbf{B}$ is called a matrix *pencil*, with λ being the *eigenvalue* and $\mathbf{x}$ the *eigenvector* of the pencil $(\mathbf{A}, \mathbf{B})$ in Eq. (1) [10]. In general, only a few of the eigenvalues and the associated eigenvectors are desired. The matrices $\mathbf{A}$ and $\mathbf{B}$ have no general pattern of nonzeros, in which case factorization of either matrix would be impractical.

Throughout this paper we use the notion of Householder [7]. Except for dimensions and indices, or when otherwise indicated, lower case Greek letters represent scalars; lower case Latin letters column vectors; capital letters, Greek or Latin, matrices.

2. The Trace Theorem

The following theorems are instrumental in formulating an extreme eigenspace computation as an optimization problem. Recall that $\mathrm{tr}(\mathbf{A})$, the *trace* of $\mathbf{A}$, denotes the sum of the diagonal elements of $\mathbf{A}$. Further, the trace of a matrix is invariant to similarity transformations.

Theorem 1. [6] *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be symmetric* $n \times n$ *matrices. If* $\mathbf{B}$ *is positive-definite then there is an* $n \times n$ *matrix* $\mathbf{Z}$ *for which*

$$\mathbf{Z}^T\mathbf{BZ} = \mathbf{I}_n \quad \textit{and} \quad \mathbf{Z}^T\mathbf{AZ} = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\,, \tag{2}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of problem (1) *and the columns of* $\mathbf{Z}$ *are their associated eigenvectors. Furthermore, if* $\mathbf{A}$ *is positive-definite, then all of the eigenvalues* λ_i *are positive.*

Theorem 2. (Trace Theorem [13]) *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be given as in Theorem 1 and* $\mathcal{Y}^*$ *be the set of all* $n \times p$ *matrices* $\mathbf{Y}$ *for which* $\mathbf{Y}^T\mathbf{B}\mathbf{Y} = \mathbf{I}_p$*. Then*

$$\min_{\mathbf{Y}\in\mathcal{Y}^*} \operatorname{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) = \sum_{i=1}^{p} \lambda_i \,. \tag{3}$$

In other words,

$$\min_{\mathbf{Y}\in\mathcal{Y}^*} \operatorname{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) = \operatorname{tr}(\mathbf{X}^T\mathbf{A}\mathbf{X}) \tag{4}$$

with

$$\mathbf{X}^T\mathbf{B}\mathbf{X} = \mathbf{I}_p \quad \textit{and} \quad \mathbf{X}^T\mathbf{A}\mathbf{X} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)\,, \tag{5}$$

where $\mathbf{X}$ *corresponds to the first* p *columns of the matrix* $\mathbf{Z}$ *of Theorem 1.*

Proof. For the proof of Theorem 2 we first recall the following theorem:

Theorem 3. (Poincaré Separation Theorem [8, 5]) *Let* $\mathbf{A}$ *be a real symmetric* $n \times n$ *matrix with eigenvalues* $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$*, and let* $\mathbf{G}$ *be a semi-unitary* $n \times k$ *matrix (*$1 \le k \le n$*), so that* $\mathbf{G}^T\mathbf{G} = \mathbf{I}_k$*. Then the eigenvalues* $\mu_1 \le \mu_2 \le \cdots \le \mu_k$ *of* $\mathbf{G}^T\mathbf{A}\mathbf{G}$ *satisfy*

$$\lambda_i \le \mu_i \le \lambda_{n-k+i} \qquad (i = 1, 2, \ldots, k)\,. \tag{6}$$

Since $\mathbf{A}$ and $\mathbf{B}$ are given as in Theorem 1, let $\mathbf{Z}$ be the $n \times n$ matrix for which $\mathbf{Z}^T\mathbf{B}\mathbf{Z} = \mathbf{I}_n$ and $\mathbf{Z}^T\mathbf{A}\mathbf{Z} = \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of the pencil $(\mathbf{A}, \mathbf{B})$. For simplicity we assume that $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$.

Let $\mathbf{Y} \in \mathcal{Y}^*$ and set $\mathbf{Y} = \mathbf{Z}\mathbf{G}$ for some $n \times p$ matrix $\mathbf{G}$. From $\mathbf{Y}^T\mathbf{B}\mathbf{Y} = \mathbf{I}_p$ it follows directly that $\mathbf{G}$ is semi-unitary. Hence, we have

$$\mathbf{Y}^T\mathbf{A}\mathbf{Y} = \mathbf{G}^T\mathbf{\Lambda}\,\mathbf{G}\,. \tag{7}$$

Applying Theorem 3 to Eq. (7) with μ_i being the eigenvalues of $\mathbf{G}^T\mathbf{\Lambda}\,\mathbf{G}$ we get $\lambda_i \le \mu_i$ for $i = 1, \ldots, p$ and thus

$$\sum_{i=1}^{p} \lambda_i \le \sum_{i=1}^{p} \mu_i\,. \tag{8}$$

Since $\mathbf{G}^T\mathbf{\Lambda}\mathbf{G}$ is symmetric there exists a spectral decomposition of the form

$$\mathbf{Q}^T(\mathbf{G}^T\mathbf{\Lambda}\mathbf{G})\,\mathbf{Q} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_p)\,, \tag{9}$$

where $\mathbf{Q}$ is a unitary matrix with columns $\mathbf{q}_i$ being the eigenvectors of $\mathbf{G}^T\mathbf{\Lambda}\mathbf{G}$. Further, we have

$$\mathrm{tr}(\mathbf{Q}^T(\mathbf{G}^T\mathbf{\Lambda}\mathbf{G})\,\mathbf{Q}) = \mathrm{tr}(\mathbf{Q}\mathbf{Q}^T(\mathbf{G}^T\mathbf{\Lambda}\mathbf{G})) = \mathrm{tr}(\mathbf{G}^T\mathbf{\Lambda}\mathbf{G}) = \sum_{i=1}^{p}\mu_i\,, \tag{10}$$

thus implying from Eqs. (7) and (8) that

$$\sum_{i=1}^{p}\lambda_i \leq \mathrm{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{Y})\,. \tag{11}$$

By the spectral decomposition theorem equality in Eq. (11) holds if $\mathbf{Y} = \mathbf{Z}_p = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$, where the columns $\mathbf{z}_i$ are the eigenvectors of the pencil $(\mathbf{A}, \mathbf{B})$. The given matrix $\mathbf{Z}_p$ hence diagonalizes the matrix $\mathbf{A}$ from problem (1) and thus leads to

$$\mathbf{Z}_p^T\mathbf{A}\mathbf{Z}_p = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)\,. \tag{12}$$

□

3. The Trace Minimization Method

The Trace Minimization (TRACEMIN) method [13, 12] attempts to compute a few of the largest or smallest eigenvalues and the corresponding eigenvectors of the generalized eigenvalue problem (1), where both $\mathbf{A}$ and $\mathbf{B}$ are positive-definite. In case $\mathbf{A}$ is not positive-definite, problem (1) is replaced by

$$(\mathbf{A} - \nu\mathbf{B})\,\mathbf{x} = (\lambda - \nu)\,\mathbf{B}\mathbf{x}\,, \tag{13}$$

with $\nu < \lambda_1 < 0$, thus resulting in $\mathbf{A} - \nu\mathbf{B}$ being positive-definite.

TRACEMIN is motivated by Theorem 2 and works by treating problem (1) as the quadratic minimization problem

$$\begin{aligned} &\text{minimize} && \mathrm{tr}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) \\ &\text{subject to} && \mathbf{Y}^T\mathbf{B}\mathbf{Y} = \mathbf{I}_p\,. \end{aligned} \tag{14}$$

Given $\mathbf{Y}_k$ as the current approximation to the eigenvectors corresponding to the p smallest eigenvalues where $\mathbf{Y}_k^T\mathbf{B}\mathbf{X} = \mathbf{I}_p$ and $1 \leq p \ll n$, the idea is to compute a correction term $\mathbf{\Delta}_k$ that is chosen as to

$$\begin{aligned} &\text{minimize} && \mathrm{tr}((\mathbf{Y}_k - \mathbf{\Delta}_k)^T\mathbf{A}(\mathbf{Y}_k - \mathbf{\Delta}_k)) \\ &\text{subject to} && \mathbf{Y}_k^T\mathbf{B}\,\mathbf{\Delta}_k = \mathbf{0}\,. \end{aligned} \tag{15}$$

As a result, $\mathbf{Y}_k - \boldsymbol{\Delta}_k$ always satisfies

$$\operatorname{tr}((\mathbf{Y}_k - \boldsymbol{\Delta}_k)^T \mathbf{A}(\mathbf{Y}_k - \boldsymbol{\Delta}_k)) \leq \operatorname{tr}(\mathbf{Y}_k^T \mathbf{A} \mathbf{Y}_k)\,. \tag{16}$$

Further, the next iterate $\mathbf{Y}_{k+1}$ is formed by $\mathbf{B}$-orthonormalizing $\mathbf{Y}_k - \boldsymbol{\Delta}_k$ and thus, by also enforcing $\mathbf{Y}_k^T \mathbf{B} \boldsymbol{\Delta}_k = \mathbf{0}$ in the minimization problem (15) it guarantees that

$$\operatorname{tr}(\mathbf{Y}_{k+1}^T \mathbf{A} \mathbf{Y}_{k+1}) \leq \operatorname{tr}((\mathbf{Y}_k - \boldsymbol{\Delta}_k)^T \mathbf{A}(\mathbf{Y}_k - \boldsymbol{\Delta}_k)) \leq \operatorname{tr}(\mathbf{Y}_k^T \mathbf{A} \mathbf{Y}_k)\,. \tag{17}$$

The solution of the minimization problem (15) can be obtained by introducing Lagrange multipliers to enforce the constraints and by solving the resulting saddle-point problem

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}\mathbf{Y}_k \\ \mathbf{Y}_k^T \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta}_k \\ \mathbf{L}_k \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{Y}_k \\ \mathbf{0} \end{pmatrix}, \tag{18}$$

where the matrix $\mathbf{L}_k$ represents the Lagrange multipliers. This system can be rewritten as the positive-semidefinite system

$$(\mathbf{PAP})\,\boldsymbol{\Delta}_k = \mathbf{PA}\mathbf{Y}_k, \;\; \text{subject to } \mathbf{Y}_k^T \mathbf{B} \boldsymbol{\Delta}_k = \mathbf{0}\,, \tag{19}$$

where $\mathbf{P}$ is the orthogonal projector onto the space $\mathbf{B}$-orthogonal to $\mathbf{Y}_k$ and is defined as $\mathbf{P} = \mathbf{I} - \mathbf{B}\mathbf{Y}_k(\mathbf{Y}_k^T \mathbf{B}^2 \mathbf{Y}_k)^{-1}\mathbf{Y}_k^T \mathbf{B}$. Eq. (19) is solved by the conjugate gradient (CG) method. By choosing zero as the initial iterate, the linear constraint $\mathbf{Y}_k^T \mathbf{B} \boldsymbol{\Delta}_k^{(l)} = \mathbf{0}$ is automatically satisfied for any intermediate $\boldsymbol{\Delta}_k^{(l)}$. In [12], the authors show that the update $\boldsymbol{\Delta}_k$ is determined by the exact solution of Eq. (19), where the solution is given by

$$\boldsymbol{\Delta}_k = \mathbf{Y}_k - \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}_k(\mathbf{Y}_k^T \mathbf{B}\mathbf{A}^{-1}\mathbf{B}\mathbf{Y}_k)^{-1}\,. \tag{20}$$

This means that the subspace spanned by $\mathbf{Y}_k - \boldsymbol{\Delta}_k$ is the same subspace spanned by $\mathbf{A}^{-1}\mathbf{B}\mathbf{Y}_k$ and thus, if Eq. (19) is solved exactly at each iteration step (which happens when an exact factorization of $\mathbf{A}$ is used) the basic TRACEMIN algorithm in Algorithm 1 is mathematically equivalent to (block) inverse iteration [1, 4]. Therefore, the basic TRACEMIN algorithm can be thought of as an inexact inverse iteration, while still preserving global convergence [12].

Due to the relation of the basic TRACEMIN algorithm with inverse iteration, alongside of its robust global convergence property, basic TRACEMIN also inherits the linear convergence rate from the inverse iteration method [4]. Additionally, for problems in which the desired eigenvalues are poorly separated from the remaining part of the spectrum, basic TRACEMIN converges too slowly. To counter this, TRACEMIN tries to improve the rate of

convergence by shifting the system shown in Eq. (19). In [13], TRACEMIN is accelerated using *multiple dynamic shifts*, where Eq. (19) becomes

$$(\mathbf{P}(\mathbf{A}-\sigma_{k,i}\mathbf{B})\mathbf{P})\,\mathbf{d}_{k,i} = \mathbf{P}\mathbf{A}\mathbf{y}_{k,i}, \text{ subject to } \mathbf{Y}_k^T\mathbf{B}\,\mathbf{d}_{k,i} = \mathbf{0}\,, \quad 1 \le i \le s\,, \tag{21}$$

with $\mathbf{d}_{k,i}$ and $\mathbf{y}_{k,i}$ being the i-th columns of $\mathbf{\Delta}_k$ and $\mathbf{Y}_k$, respectively, and $\sigma_{k,i}$ being the associated shift at step k. At the beginning of the algorithm, a single shift is used for all the columns of $\mathbf{Y}_k$. As the algorithm moves closer to convergence, multiple shifts are introduced dynamically and the CG process is modified to handle possible breakdown. For one, the CG process is terminated when the error $(\mathbf{x}_{k,i}-\mathbf{d}_{k,i}^{(l)})^T\mathbf{A}(\mathbf{x}_{k,i}-\mathbf{d}_{k,i}^{(l)})$ increases by a small factor [13, § 2.4], which helps maintain global convergence in the presence of shifting.

Algorithm 1. The basic TRACEMIN algorithm.

Choose a block size $s \ge p$ and an $n \times s$ matrix $\mathbf{V}_1$ of full rank such that $\mathbf{V}_1^T\mathbf{B}\mathbf{V}_1 = \mathbf{I}_s$.
For $k = 1, 2, \ldots$ until convergence, do

1. Compute $\mathbf{W}_k = \mathbf{A}\mathbf{V}_k$ and the interaction matrix $\mathbf{H}_k = \mathbf{V}_k^T\mathbf{W}_k$.
2. Compute the eigenpairs $(\mathbf{X}_k, \mathbf{\Theta}_k)$ of $\mathbf{H}_k$. The eigenvalues are arranged in ascending order and the eigenvectors are chosen to be orthogonal.
3. Compute the corresponding Ritz vectors $\mathbf{Y}_k = \mathbf{V}_k\mathbf{X}_k$.
4. Compute th residuals $\mathbf{R}_k = \mathbf{A}\mathbf{Y}_k - \mathbf{B}\mathbf{Y}_k\mathbf{\Theta}_k = \mathbf{W}_k\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\mathbf{\Theta}_k$.
5. Test for convergence.
6. Solve the positive-semidefinite linear system (19) approximately via CG scheme.
7. $\mathbf{B}$-orthonormalize $\mathbf{Y}_k - \mathbf{\Delta}_k$ into $\mathbf{V}_{k+1}$ by the Gram-Schmidt process with reorthogonalization.

End for

---

4. Characterization of the Trace Minimization Method as a Quasi-Newton Method

In this section we approach the TRACEMIN method as a quasi-Newton method by deriving the appropriate correction equation.

Newton's method is a root-finding algorithm that uses the first few terms of the Taylor series of a function F close to a suspected root. Given a function

$F : \mathbb{R}^n \to \mathbb{R}$, the method uses the the first-order Taylor expansion of F around $\mathbf{x}_k$ [9, Theorem 2.1]

$$F(\mathbf{x}_k + \mathbf{p}_k) \approx F(\mathbf{x}_k) + \mathbf{p}_k^T \nabla F(\mathbf{x}_k) \tag{22}$$

and chooses a correction $\mathbf{p}_k$ such that $F(\mathbf{x}_k + \mathbf{p}_k) = 0$.

As required by the first-order necessary optimality condition [9, Theorem 2.2], a local minimizer $\mathbf{x}_*$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies $\nabla f(\mathbf{x}_*) = 0$, meaning that $\mathbf{x}_*$ is a root of the function $F(\mathbf{x}) = \nabla f(\mathbf{x})$. Hence, it is possible to apply Newton's method to $F(\mathbf{x})$ in order to find a critical point of the objective function. This is accomplished by inserting $F(\mathbf{x}_k) = \nabla f(\mathbf{x}_k)$ into Eq. (22) and solving for $\mathbf{p}_k$, thus resulting in the Newton step

$$\mathbf{p}_k = -\mathbf{H}_f(\mathbf{x}_k)^{-1} \ \nabla f(\mathbf{x}_k)\,, \tag{23}$$

where $\nabla f(\mathbf{x}_k)$ is the gradient of f at $\mathbf{x}_k$ and $\mathbf{H}_f$ is the Hessian matrix of f at $\mathbf{x}_k$ defined as $\mathbf{H}_f(\mathbf{x}_k) \equiv \nabla^2 f(\mathbf{x}_k)$. If $\mathbf{x}$ is *close enough* to a local minimizer $\mathbf{x}_*$ and the Hessian $\mathbf{H}_f(\mathbf{x})$ is positive-definite, Newton's method converges to the local minimizer with a quadratic rate of convergence [4, § 1.2.2].

Quasi-Newton methods instead calculate the search direction $\mathbf{p}_k$ by replacing the true Hessian $\mathbf{H}_f(\mathbf{x}_k)$ with an approximation $\mathbf{B}_k$. The quasi-Newton update is thus defined as

$$\mathbf{p}_k = -\mathbf{B}_k^{-1} \ \nabla f(\mathbf{x}_k)\,. \tag{24}$$

Since we present here an approach that does not require a detailed knowledge of quasi-Newton schemes, we refer to [9] for more information on the schemes.

For the TRACEMIN method, the objective function is given by

$$f : \mathbb{R}_*^{n\times p} \to \mathbb{R} : \mathbf{Y} \mapsto \mathrm{tr}((\mathbf{Y}^T\mathbf{B}\mathbf{Y})^{-1}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}))\,, \tag{25}$$

where $\mathbb{R}_*^{n\times p}$ denotes the set of full-rank $n \times p$ matrices.

In each iteration, the TRACEMIN method tries to find a correction term $\boldsymbol{\Delta}_k$ such that

$$f(\mathbf{Y}_k - \boldsymbol{\Delta}_k) \leq f(\mathbf{Y}_k)\,, \tag{26}$$

which is accomplished by additionally requiring that the correction term is $\mathbf{B}$-orthogonal to $\mathbf{Y}_k$. Thus, we require $\boldsymbol{\Delta}_k$ to satisfy $\mathbf{Y}_k\mathbf{B}\boldsymbol{\Delta}_k = \mathbf{0}$, i.e. $\boldsymbol{\Delta}_k \in \mathcal{H}_{Y_k}$, where

$$\mathcal{H}_{Y_k} := \{\mathbf{Z} \in \mathbb{R}^{n\times p} : \mathbf{Y}_k^T\mathbf{B}\mathbf{Z} = \mathbf{0}\}\,. \tag{27}$$

A second-order expansion of f around $\boldsymbol{\Delta}_k = \mathbf{0}$ gives:

$$\begin{aligned} f(\mathbf{Y}_k + \boldsymbol{\Delta}_k) &= \operatorname{tr}(((\mathbf{Y}_k + \boldsymbol{\Delta}_k)^T\mathbf{B}(\mathbf{Y}_k + \boldsymbol{\Delta}_k))^{-1}(\mathbf{Y}_k + \boldsymbol{\Delta}_k)^T\mathbf{A}(\mathbf{Y}_k + \boldsymbol{\Delta}_k)) \\ &= \operatorname{tr}((\mathbf{I} + (\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}(\boldsymbol{\Delta}_k^T\mathbf{B}\boldsymbol{\Delta}))^{-1}(\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}(\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k + 2\boldsymbol{\Delta}_k^T\mathbf{A}\mathbf{Y}_k + \boldsymbol{\Delta}_k^T\mathbf{A}\boldsymbol{\Delta}_k)) \\ &\overset{(*)}{=} \operatorname{tr}((\mathbf{I} - (\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}(\boldsymbol{\Delta}_k^T\mathbf{B}\boldsymbol{\Delta}))(\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}(\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k + 2\boldsymbol{\Delta}_k^T\mathbf{A}\mathbf{Y}_k + \boldsymbol{\Delta}_k^T\mathbf{A}\boldsymbol{\Delta}_k)) \\ &\qquad + H.O.T. \\ &= \operatorname{tr}((\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}(\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k)) + \operatorname{tr}((\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\boldsymbol{\Delta}_k^T 2\mathbf{A}\mathbf{Y}_k) \\ &\quad + \frac{1}{2}\operatorname{tr}((\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\boldsymbol{\Delta}_k^T 2(\mathbf{A}\boldsymbol{\Delta}_k - \mathbf{B}\boldsymbol{\Delta}_k(\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k)) + H.O.T. \end{aligned} \tag{28}$$

where in $(*)$ we used the approximation $(\mathbf{I} + \mathbf{A})^{-1} = \sum_{n=0}^{\infty}(-1)^n\mathbf{A}^n$ [11, Eq. (187)]. By further introducing $\mathbf{P} = \mathbf{I} - \mathbf{B}\mathbf{Y}_k(\mathbf{Y}_k^T\mathbf{B}^2\mathbf{Y}_k)^{-1}\mathbf{Y}_k^T\mathbf{B}$ as the orthogonal projector onto the space $\mathbf{B}$-orthogonal to $\mathbf{Y}_k$, and using the inner product [1, 4]

$$\langle \mathbf{Z}_1, \mathbf{Z}_2 \rangle := \operatorname{tr}((\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\mathbf{Z}_1^T\mathbf{Z}_2), \quad \mathbf{Z}_1, \mathbf{Z}_2 \in \mathcal{H}_{Y_k} \tag{29}$$

we can rewrite Eq. (28) as follows:

$$f(\boldsymbol{Y}_k + \boldsymbol{\Delta}_k) = f(\mathbf{Y}_k) + \langle \boldsymbol{\Delta}_k, 2\mathbf{P}\mathbf{A}\mathbf{Y}\rangle + \frac{1}{2}\langle \boldsymbol{\Delta}_k, 2\mathbf{P}(\mathbf{A}\boldsymbol{\Delta}_k - \mathbf{B}\boldsymbol{\Delta}_k(\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k)\rangle + H.O.T. \tag{30}$$

From Eq. (30) we now identify $2\mathbf{P}\mathbf{A}\mathbf{Y}$ to be the gradient of f at $\boldsymbol{\Delta}_k = \mathbf{0}$ and the operator

$$\mathbf{H}_f : \mathcal{H}_{Y_k} \to \mathcal{H}_{Y_k} : \boldsymbol{\Delta}_k \mapsto 2\mathbf{P}(\mathbf{A}\boldsymbol{\Delta}_k - \mathbf{B}\boldsymbol{\Delta}_k(\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k) \tag{31}$$

to be the Hessian of f at $\boldsymbol{\Delta}_k = \mathbf{0}$ [1, 4]. The Newton correction equation in Eq. (23) thus yields the equation

$$\mathbf{P}(\mathbf{A}\boldsymbol{\Delta}_k - \mathbf{B}\boldsymbol{\Delta}_k(\mathbf{Y}_k^T\mathbf{B}\mathbf{Y}_k)^{-1}\mathbf{Y}_k^T\mathbf{A}\mathbf{Y}_k) = -\mathbf{P}\mathbf{A}\mathbf{Y}\,. \tag{32}$$

By substituting the Hessian of f with the approximate Hessian $2\mathbf{P}\mathbf{A}\mathbf{P}$, the correction equation becomes

$$(\mathbf{P}\mathbf{A}\mathbf{P})\boldsymbol{\Delta}_k = -\mathbf{P}\mathbf{A}\mathbf{Y}_\mathbf{k}, \quad \boldsymbol{\Delta}_k \in \mathcal{H}_{Y_k}\,, \tag{33}$$

which is the same as Eq. (19) solved in the TRACEMIN method [4, § 4.3.2]. Further, since TRACEMIN is only described for positive-definite $\mathbf{A}$, this linear system is positive-definite for all vectors in $\mathcal{H}_{Y_k}$, which implies that the Newton step is well defined.

However, it is important to note that the characterization of TRACEMIN as a quasi-Newton method does not capture the global convergence theory which the authors of [13] established for TRACEMIN; we refer to [4] for further details.

5. The Jacobi-Davidson Method

The Jacobi-Davidson (JD) method [15, 14] calculates the eigenvectors and eigenvalues of the pencil $(\mathbf{A}, \mathbf{B})$ by constructing a correction, for a given eigenvector approximation, in a subspace orthogonal to the given approximation. The correction is chosen orthogonal since we want to expand the current search space in a profitable and unexplored direction.

The name of the JD method follows from a combination of two principles [3]. The first principle, i.e., the computation of the correction in a given subspace is done in a Davidson manner, since Davidson suggested the usage of other subspaces than Krylov subspaces for the construction of orthonormal basis vectors. The second principle is based on an approach suggested by Jacobi, where the idea is to compute orthogonal corrections.

Let $\mathbf{u}$ be a non-zero Ritz approximation of an eigenvector $\mathbf{x}$ with Ritz value θ corresponding to the eigenvalue λ associated to $\mathbf{x}$. Then from the Ritz-Galerkin condition it follows that

$$\mathbf{r} \equiv \mathbf{A}\mathbf{u} - \theta\mathbf{B}\mathbf{u} \perp \mathbf{u}\,. \tag{34}$$

The goal is now to find a correction vector $\mathbf{t}$ for $\mathbf{u}$ in the space $\mathbf{B}$-orthogonal to $\mathbf{u}$, such that

$$\mathbf{A}(\mathbf{u}+\mathbf{t}) = \lambda\mathbf{B}(\mathbf{u}+\mathbf{t}), \quad \mathbf{u}^T\mathbf{B}\mathbf{t} = 0\,, \tag{35}$$

and $\mathbf{x}$ is a scalar multiple of $\mathbf{u}+\mathbf{t}$. Since the correction $\mathbf{t}$ is required to be $\mathbf{B}$-orthogonal to $\mathbf{u}$, it follows that

$$\left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T\mathbf{B}}{\mathbf{u}^T\mathbf{B}\mathbf{u}}\right)\mathbf{t} = \mathbf{t}\,. \tag{36}$$

and the correction equation becomes [14]

$$\left(\mathbf{I} - \frac{\mathbf{B}\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{B}\mathbf{u}}\right)(\mathbf{A}-\theta\mathbf{B})\left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T\mathbf{B}}{\mathbf{u}^T\mathbf{B}\mathbf{u}}\right)\mathbf{t} = -\mathbf{r}\,. \tag{37}$$

Further, Eq. (37) is the same as the augmented correction equation given by [14, Theorem 3.5]

$$\begin{pmatrix} \mathbf{A}-\theta\mathbf{B} & \mathbf{B}\mathbf{u} \\ \mathbf{u}^T\mathbf{B} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{t} \\ \epsilon \end{pmatrix} = \begin{pmatrix} -\mathbf{r} \\ 0 \end{pmatrix}, \tag{38}$$

where ϵ is a Lagrange multiplier enforcing the $\mathbf{B}$-orthogonality of $\mathbf{t}$ against $\mathbf{u}$.

One can also try to derive the correction equation in Eq. (37) by exploiting the fact that the eigenvectors and eigenvalues of the pencil $(\mathbf{A}, \mathbf{B})$ can be identified as the stationary points of the generalized Rayleigh quotient [2]

$$\rho(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{B}\mathbf{x}}\,, \quad \forall\mathbf{x} \in \mathbb{R}^n\backslash\{\mathbf{0}\}\,. \tag{39}$$

For this, we are going to compute the zeros of the function

$$F(\mathbf{x}) = \nabla \rho(\mathbf{x}) \tag{40}$$

with the help of Newton's method. Hence, the Newton step for finding a solution to $F(\mathbf{x}) = \mathbf{0}$ is given by

$$\mathbf{t}_k = -\,\mathbf{H}_\rho(\mathbf{x}_k)^{-1}\ \nabla \rho(\mathbf{x}_k)\,, \tag{41}$$

where $\mathbf{H}_\rho$ is the Hessian of ρ defined by [16]

$$H_\rho(\mathbf{x}) = \frac{2}{\mathbf{x}^T\mathbf{B}\mathbf{x}}\left[\left(\mathbf{I} - \frac{2}{\mathbf{x}^T\mathbf{B}\mathbf{x}}\mathbf{B}\mathbf{x}\mathbf{x}^T\right)(\mathbf{A} - \rho(\mathbf{x})\mathbf{B})\left(\mathbf{I} - \frac{2}{\mathbf{x}^T\mathbf{B}\mathbf{x}}\mathbf{x}\mathbf{x}^T\mathbf{B}\right)\right] \tag{42}$$

and the gradient of ρ is [16]

$$\nabla \rho(\mathbf{x}) = 2\frac{\mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{x}\rho(\mathbf{x})}{\mathbf{x}^T\mathbf{B}\mathbf{x}}\,. \tag{43}$$

Thus, the Newton equation $\mathbf{H}_\rho(\mathbf{x}_k)\mathbf{t}_k = -\nabla \rho(\mathbf{x}_k)$ becomes

$$\left(\mathbf{I} - \frac{2}{\mathbf{x}_k^T\mathbf{B}\mathbf{x}_k}\mathbf{B}\mathbf{x}_k\mathbf{x}_k^T\right)(\mathbf{A} - \rho(\mathbf{x}_k)\mathbf{B})\left(\mathbf{I} - \frac{2}{\mathbf{x}_k^T\mathbf{B}\mathbf{x}_k}\mathbf{x}_k\mathbf{x}_k^T\mathbf{B}\right) = -(\mathbf{A}\mathbf{x}_k - \mathbf{B}\mathbf{x}_k\rho(\mathbf{x}_k)) =: -\mathbf{r}(\mathbf{x}_k)\,. \tag{44}$$

It is important to note that the Hessian is singular if $\mathbf{x}$ is an eigenvector, since $\mathbf{H}_\rho(\mathbf{x})\mathbf{x} = -F(\mathbf{x}) = \mathbf{0}$ [2]. If we instead apply the Newton method to [17]

$$F(\mathbf{x}, \lambda) := \begin{pmatrix} (\mathbf{A} - \lambda\mathbf{B})\mathbf{x} \\ \mathbf{x}^T\mathbf{B}\mathbf{x} - 1 \end{pmatrix} \tag{45}$$

we get the Newton step

$$\begin{pmatrix} \mathbf{A} - \lambda_k\mathbf{B} & \mathbf{B}\mathbf{x}_k \\ \mathbf{x}_k^T\mathbf{B} & 0 \end{pmatrix}\begin{pmatrix} \mathbf{t}_k \\ \epsilon_k \end{pmatrix} = \begin{pmatrix} -\mathbf{r}_k \\ 0 \end{pmatrix}, \tag{46}$$

which is nonsingular unless λ_k is a multiple eigenvalue of the pencil $(\mathbf{A}, \mathbf{B})$ [2].

In case of the JD method, the targeted eigenvalue λ_k in Eq. (46), which is not available during the iteration, is replaced by a shift θ_k [14], thus finally resembling the augmented correction equation in Eq. (38).

After having defined a way to calculate an orthogonal correction vector $\mathbf{t}_k$ by following Jacobi's idea, we now apply Davidson's approach. The consecutive corrections $\mathbf{t}_k$ are now used to build the search space. The solution $\mathbf{t}_k$ of the correction equation (37) is appended to $\mathbf{V}_k$, resulting in $\mathbf{V}_{k+1} = [\mathbf{V}_k, \mathbf{t}_k]$ and thus accelerating the convergence by increasing the dimension of the trial space by one [2].

A block JD, as described in [14, 12], is given in Algorithm 2. The block JD tries to obtain approximations for s eigenvalues simultaneously. Further, at every outer iteration the dimension of the subspace $\mathbf{V}_k$ is increased by s, with the maximum dimension of the subspace being m [14, § 9.5].

If Eq. (47) is solved to high-order accuracy, it is reduced to the Rayleigh quotient iteration with expanding subspaces, and thus converges cubically in that case [12, § 5.1]. In general, the algorithm's convergence rate is between superlinear and cubic [14, § 3].

Algorithm 2. The block Jacobi-Davidson algorithm.

Choose a block size $s \geq p$ and an $n \times s$ matrix $\mathbf{V}_1$ of full rank such that $\mathbf{V}_1^T \mathbf{B} \mathbf{V}_1 = \mathbf{I}_s$.
For $k = 1, 2, \ldots$ until convergence, do

1. Compute $\mathbf{W}_k = \mathbf{A}\mathbf{V}_k$ and the interaction matrix $\mathbf{H}_k = \mathbf{V}_k^T \mathbf{W}_k$.
2. Compute the eigenpairs $(\mathbf{X}_k, \mathbf{\Theta}_k)$ of $\mathbf{H}_k$. The eigenvalues are arranged in ascending order and the eigenvectors are chosen to be orthogonal.
3. Compute the corresponding Ritz vectors $\mathbf{Y}_k = \mathbf{V}_k \mathbf{X}_k$.
4. Compute th residuals $\mathbf{R}_k = \mathbf{A}\mathbf{Y}_k - \mathbf{B}\mathbf{Y}_k\mathbf{\Theta}_k = \mathbf{W}_k\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\mathbf{\Theta}_k$.
5. Test for convergence.
6. For $1 \leq i \leq s$, solve the indefinite system

$$\begin{pmatrix} \mathbf{A} - \theta_{k,i}\mathbf{B} & \mathbf{B}\mathbf{x}_{k,i} \\ \mathbf{x}_{k,i}^T \mathbf{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{t}_{k,i} \\ \epsilon_{k,i} \end{pmatrix} = \begin{pmatrix} -\mathbf{r}_{k,i} \\ 0 \end{pmatrix}, \tag{47}$$

   where $\mathbf{r}_{k,i} = \mathbf{A}\mathbf{x}_{k,i} - \theta_{k,i}\mathbf{B}\mathbf{x}_{k,i}$ is the residual corresponding to the Ritz pair $(\mathbf{x}_{k,i}, \theta_{k,i})$.
7. If $\dim(\mathbf{V}_k) \leq m - s$, then

$$\mathbf{V}_{k+1} = \mathrm{Mod}GS_B(\mathbf{V}_k, \mathbf{\Delta}_k)\,, \tag{48}$$

   else

$$\mathbf{V}_{k+1} = \mathrm{Mod}GS_B(\mathbf{X}_k, \mathbf{\Delta}_k)\,. \tag{49}$$

   Here, $\mathrm{Mod}GS_B$ stands for the Gram-Schmidt process with reorthogonalization with respect to the B-inner products, i.e. $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{B} \mathbf{y}$.

End for

---

6. The Davidson-Type Trace Minimization Algorithm

Block JD's performance depends on how good the initial guess is and how efficiently and accurately the inner system (47) is solved. Further, block JD suffers from the following problems [12, § 5.1]:

1. The Ritz shifting strategy forces the algorithm to converge to eigenvalues closest to the Ritz values that are often far away from the desired eigenvalues at the beginning of the iteration;
2. Due to the subspace expanding, the Ritz values are decreasing and the algorithm is forced to converge to the smallest eigenpairs;
3. If a Ritz value approaches a multiple eigenvalue or a cluster of eigenvalues, the inner system (47) becomes poorly conditioned.

In [12], Sameh and Tong present the *Davidson-type trace minimization algorithm* that partially solves the above mentioned problems by employing the techniques developed in the TRACEMIN method, i.e., the multiple dynamic shifting strategy [12, § 4.2], the implicit deflation technique, where $\mathbf{d}_{k,i}$ is required to be $\mathbf{B}$-orthogonal to all the Ritz vectors obtained in the previous iteration step (which is essential in the original TRACEMIN algorithm for maintaining the trace reduction property (17)), and the dynamic stopping strategy [12, § 4.3].

Let $s \geq p$ be the block size and let $m \geq s$ be a given integer that limits the dimension of the subspaces. The Davidson-type TRACEMIN algorithm is given as follows:

Algorithm 3. The Davidson-type trace minimization algorithm.

Choose a block size $s \geq p$ and an $n \times s$ matrix $\mathbf{V}_1$ of full rank such that $\mathbf{V}_1^T \mathbf{B} \mathbf{V}_1 = \mathbf{I}_s$.
For $k = 1, 2, \ldots$ until convergence, do

1. Compute $\mathbf{W}_k = \mathbf{A}\mathbf{V}_k$ and the interaction matrix $\mathbf{H}_k = \mathbf{V}_k^T \mathbf{W}_k$.
2. Compute the eigenpairs $(\mathbf{X}_k, \mathbf{\Theta}_k)$ of $\mathbf{H}_k$. The eigenvalues are arranged in ascending order and the eigenvectors are chosen to be orthogonal.
3. Compute the corresponding Ritz vectors $\mathbf{Y}_k = \mathbf{V}_k \mathbf{X}_k$.
4. Compute th residuals $\mathbf{R}_k = \mathbf{A}\mathbf{Y}_k - \mathbf{B}\mathbf{Y}_k\mathbf{\Theta}_k = \mathbf{W}_k\mathbf{X}_k - \mathbf{B}\mathbf{Y}_k\mathbf{\Theta}_k$.
5. Test for convergence.
6. For $1 \leq i \leq s$, solve the indefinite system

$$(\mathbf{P}(\mathbf{A} - \sigma_{k,i}\mathbf{B})\mathbf{P})\,\mathbf{d}_{k,i} = \mathbf{P}\mathbf{r}_{k,i}, \quad \mathbf{Y}_k^T \mathbf{B}\,\mathbf{d}_{k,i} = 0 \tag{50}$$

   to a certain accuracy determined by the stopping criterion described in [12, § 4.3]. The shift parameters $\sigma_{k,i},\ 1 \leq i \leq s$, are determined according to the dynamic shifting strategy described in [12, § 4.2].

7. If $\dim(\mathbf{V}_k) \leq m - s$, then

$$\mathbf{V}_{k+1} = \mathrm{Mod}GS_B(\mathbf{V}_k, \mathbf{\Delta}_k)\,, \tag{51}$$

else

$$\mathbf{V}_{k+1} = \mathrm{Mod}GS_B(\mathbf{X}_k, \mathbf{\Delta}_k)\,. \tag{52}$$

End for

In [12, § 5.3], both the block JD and the Davidson-type TRACEMIN algorithm are compared by doing numerical experiments on a variety of problems. From the results it is observed that the difference between both algorithms becomes clear when the number of inner iteration steps is increased, in which case the Davidson-type TRACEMIN algorithm needs fewer outer iteration steps for most of the solved problems. This behavior comes from the dynamic shifting strategy deployed by the Davidson-type TRACEMIN algorithm, which accelerates the algorithm significantly. It is also observed that block JD converges to wrong eigenpairs when the inner systems are solved to high accuracy. If the inner systems are solved crudely, both algorithms actually perform the same. Further, it is observed that the success of the block JD method depends on good starting spaces.

7. Conclusion

The goal of this thesis was to derive and compare a number of algorithms for computing a few eigenvalues and associated eigenvectors of the generalized eigenvalue problem. Since the trace theorem is an important ingredient for the trace minimization method, we provided a detailed proof of the theorem. Further, the trace minimization method has been derived and in a next step characterized as a quasi-Newton method without involving any differential geometry in our description. Relations between some of the algorithms have been elaborated, showing that the trace minimization method and Jacobi-Davidson method, albeit starting with different derivations are algorithmically still quite similar.

The Jacobi-Davidson method for the generalized eigenvalue problem has been derived using two approaches: a) by starting from the Ritz-Galerkin condition and; b) by deriving a Newton step from a suitable function $F(\mathbf{x}, \lambda)$. Furthermore, a block-version of the Jacobi-Davidson algorithm has been provided.

Finally, after having analyzed the trace minimization method and Jacobi-Davidson method, we have supplied a description of the Davidson-type trace minimization method, which combines techniques developed in the trace

minimization method and Davidson's approach of expanding subspaces, resulting in a rather robust method in regard to some problems compared to the block Jacobi-Davidson method.

Acknowledgments

I want to thank Professor Arbenz for the possibility to work on this project, for providing me with an office place, for his knowledge, patience and for his time.

References

[1] P.-A. ABSIL, C. G. BAKER, K. A. GALLIVAN, AND A. SAMEH, *Adaptive model trust region methods for generalized eigenvalue problems*, in International Conference on Computational Science, Springer, 2005, pp. 33–41.

[2] P. ARBENZ AND R. B. LEHOUCQ, *A comparison of algorithms for modal analysis in the absence of a sparse direct method*, Technical Report SAND2003-1028J, Sandia National Laboratories, 2003.

[3] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. VAN DER VORST, *Templates for the solution of algebraic eigenvalue problems: a practical guide*, Society for Industrial and Applied Mathematics, 2000.

[4] C. G. BAKER, *Riemannian manifold trust-region methods with applications to eigenproblems*, ProQuest, 2008.

[5] R. BELLMAN, *Introduction to matrix analysis*, Society for Industrial and Applied Mathematics, 1970.

[6] J. N. FRANKLIN, *Matrix theory*, Courier Corporation, 2012.

[7] A. HOUSEHOLDER, *The theory of matrices in numerical analysis*, Dover Publications, 1975.

[8] J. R. MAGNUS, H. NEUDECKER, ET AL., *Matrix differential calculus with applications in statistics and econometrics*, John Wiley, 1995.

[9] J. NOCEDAL AND S. WRIGHT, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer, 2006.

[10] B. N. PARLETT, *The symmetric eigenvalue problem*, Society for Industrial and Applied Mathematics, 1998.

[11] K. B. PETERSEN, M. S. PEDERSEN, ET AL., *The matrix cookbook*, Technical University of Denmark, (2012).

[12] A. SAMEH AND Z. TONG, *The trace minimization method for the symmetric generalized eigenvalue problem*, Journal of Computational and Applied Mathematics, 123 (2000), pp. 155–175.

[13] A. H. SAMEH AND J. A. WISNIEWSKI, *A trace minimization algorithm for the generalized eigenvalue problem*, SIAM Journal on Numerical Analysis, 19 (1982), pp. 1243–1259.

[14] G. L. SLEIJPEN, A. G. BOOTEN, D. R. FOKKEMA, AND H. A. VAN DER VORST, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT Numerical Mathematics, 36 (1996), pp. 595–633.

[15] G. L. SLEIJPEN AND H. A. VAN DER VORST, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.

[16] C. VÖMEL, *Harmonic Ritz values and their reciprocals*, Technical Report 610, ETH Zurich, 2008.

[17] Y. ZHOU, *Studies on Jacobi–Davidson, rayleigh quotient iteration, inverse iteration generalized Davidson and newton updates*, Numerical Linear Algebra with Applications, 13 (2006), pp. 621–642.