Block Diagonalization of Nearly Diagonal Matrices

E. KOVAČ STRIKO[*]and K. VESELIĆ [†]

February 19, 2001

Abstract

In this paper we study the effect of block diagonalization of a nearly diagonal matrix by iterating the related Riccati equations. We show that the iteration is fast, if a matrix is diagonally dominant or scaled diagonally dominant and the block partition follows an appropriately defined spectral gap. We also show that both kinds of diagonal dominance are not destroyed after the block diagonalization.

1 Introduction

In this note we consider block analoga of plane transformations in order to block diagonalize a given matrix. Iterations of plane rotations were introduced by Jacobi [4], however it does not seem to be broadly known (see Sleipen, van der Vorst [6]) that Jacobi himself was not satisfied with the convergence of his method and tried to accelerate it by using generalized rotations which annihilate a whole off-diagonal block. Of course, this complicates the computation of rotation parameters which are easily computable

[*]Faculty of Transport and Traffic Engineering, Vukelićeva 4 HR-10000 Zagreb, Croatia, email: kovacm@mafpz.fpz.hr (this work was done during the first author's stay at the Fernuniversität Hagen).

[†]Lehrgebiet Mathematische Physik, Fernuniversität Hagen, Postf. 940, 58084 Hagen, Germany, email: kresimir.veselic@fernuni-hagen.de

only if the matrix is already diagonal enough in which case the corresponding Riccati equations are solvable by a couple of iterations.[1]

We will consider the effect of a block diagonalization on matrices which are diagonally dominant i.e.

$$||D^{-1}(A-D)|| < 1$$

where D is the diagonal part of A as well as on the scaled diagonally dominant matrices for which

$$|||D|^{-1/2}(A-D)|D|^{-1/2}|| < 1.$$

(Here $\|\cdot\|$ is the spectral norm. As it is known, on symmetric matrices the scaled diagonal dominance is a weaker property than the usual scaled dominance).

In Section 2 we prove that Jacobi or Gauss - Seidel like iterations solve the Riccati equations with a number of steps independent of the dimension. This allows to find few extremal eigenpairs in some n^2 operations, both in the diagonally dominant and the scaled diagonally dominant case.

In Section 3 we study the effect of block diagonalization and show that the (scaled) diagonal dominance is preserved after the block diagonalization step. This opens the prospective to use our block transformations as an iterative correction on matrices whose eigensolution was poor because of their high condition number (see [3]).

In Section 4 we present some simple illustrative numerical examples.

2 Block diagonalization

Let $A \in C^{n\times n}$,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \tag{1}$$

where $a = [a_{ij}]$, $d = [_{ij}]$ are square of order m, $n-m$, respectively.[2]

We will consider two types of block transformations

---

[1]Block rotations with the corresponding Riccati equations were also used in [5] as iterative improvement to a computed eigensolution.

[2]We use lower case letters for matrix blocks to remind the connection with the classical Jacobi algorithm.

a)

$$A \to A_1 = V^{-1}AV \ , \qquad A_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \tag{2}$$

$$V = V(u,t) = \begin{bmatrix} I & u \\ -t & I \end{bmatrix} \begin{bmatrix} (I+ut)^{-1/2} & 0 \\ 0 & (I+tu)^{-1/2} \end{bmatrix} \tag{3}$$

$$V^{-1} = \begin{bmatrix} (I+ut)^{-1/2} & 0 \\ 0 & (I+tu)^{-1/2} \end{bmatrix} \begin{bmatrix} I & -u \\ t & I \end{bmatrix} \tag{4}$$

Here the square root is taken as an analytic continuation of the binomial series (if it exists, see [9]).

b)

$$A \to A_2 = U^{-1}AU \ , \qquad A_2 = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} \tag{5}$$

$$U = U(u,t) = \begin{bmatrix} I & u \\ -t & I \end{bmatrix} \begin{bmatrix} L_1^{-1} & 0 \\ 0 & L_2^{-1} \end{bmatrix} \tag{6}$$

$$U^{-1} = \begin{bmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{bmatrix} \begin{bmatrix} I & -u \\ t & I \end{bmatrix} \tag{7}$$

where

$$I + ut = R_1 L_1 \quad , \quad I + tu = R_2 L_2 \ , \tag{8}$$

with L_1, L_2 upper triangular, R_1, R_2 lower triangular and $\operatorname{diag} L_i = \operatorname{diag} R_i > 0, \quad i = 1,2.$

These two types are the simplest block transfomations which tend to the identity, if the matrix A is nearly block diagonal. The second transformation avoids the use of matrix square roots. It should, however, be noted that even in the first case the complexity in computing the square roots is tolerable, if $m \ll n$, because $I + ut, \ \ I + tu$ are low - rank perturbations of identity matrices.

For instance, if $m = 1$ then in (3) ut is just a number and

$$(I + tu)^{-1/2} = I - \frac{tu}{1 + ut + (1 + ut)^{1/2}}$$

as is immediately seen by using the binomial formula. The variables u, t can be understood as some kind of "tangents". In [9] it was shown how to use "$tan\frac{x}{2}$" substitution to get rid of all square roots in (3) .

In case a) we have

$$\begin{aligned} a_1 &= (I + ut)^{-1/2}(a - uc - bt + udt)(I + ut)^{-1/2} \\ b_1 &= (I + ut)^{-1/2}(au - ud + b - ucu)(I + tu)^{-1/2} \\ c_1 &= (I + tu)^{-1/2}(ta - dt + c - tbt)(I + ut)^{-1/2} \\ d_1 &= (I + tu)^{-1/2}(tau + cu + tb + d)(I + tu)^{-1/2} \end{aligned} \tag{9}$$

and in case b):

$$\begin{aligned} a_2 &= R_1^{-1}(a - uc - bt + udt)L_1^{-1} \\ b_2 &= R_1^{-1}(au - ud + b - ucu)L_2^{-1} \\ c_2 &= R_2^{-1}(ta - dt + c - tbt)L_1^{-1} \\ d_2 &= R_2^{-1}(tua + cu + tb + d)L_2^{-1} \end{aligned} \tag{10}$$

Block diagonalization means to solve the Riccati equations

$$ta - dt + c - tbt = 0 \tag{11}$$

$$au - ud + b - ucu = 0 \ . \tag{12}$$

Note that both

$$V(t^H, t) \quad \text{and} \quad U(t^H, t)$$

are unitary, in this case (8) are Cholesky decompositions. If A is Hermitian then with t solving (11) the matrix $u = t^H$ solves (12) and vice versa.

Furthermore the Riccati equation (11) determines the matrix

$$S = S(t) = \begin{bmatrix} I & 0 \\ -t & I \end{bmatrix} \qquad \text{with} \qquad S^{-1} = \begin{bmatrix} I & 0 \\ t & I \end{bmatrix}$$

such that

$$S^{-1}AS = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}.$$

The obtained upper block triangular matrix is is then block diagonalized as

$$\begin{bmatrix} I & -u \\ 0 & I \end{bmatrix} \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} \begin{bmatrix} I & u \\ 0 & I \end{bmatrix} = \begin{bmatrix} * & 0 \\ 0 & * \end{bmatrix} \ ;$$

the corresponding Riccati equation is now linear.

Next we will be solving Riccati equations (11) and (12) in the case where the matrices b, c are small in some sense.

Let

$$a = \Delta_a + a^u + a^l \qquad , \qquad d = \Delta_d + d^u + d^l \tag{13}$$

where Δ_a, a^u, a^l is the diagonal, strict upper triangular, strict lower triangular part of a , respectively (and similarly for d).

Gauss-Seidel iterates t^j for the Riccati equation (11) are defined by

$$t^{j+1}(\Delta_a + a^u) - (\Delta_d + d^l)t^{j+1} + t^j a^l - d^u t^j + c - t^j b t^j = 0, \tag{14}$$

t^0 arbitrary. Similarly, the iteration for (12) reads

$$(\Delta_a + a^u)u^{j+1} - u^{j+1}(\Delta_d + d^l) + a^l u^j - u^j d^u + b - u^j c u^j = 0, \tag{15}$$

u^0 arbitrary.

Besides (14) there are other iterations, defined by the leading terms

$$\begin{array}{lcl} t^{j+1}(\Delta_a + a^u) & - & (\Delta_d + d^u)t^{j+1} \\ t^{j+1}(\Delta_a + a^l) & - & (\Delta_d + d^l)t^{j+1} \\ t^{j+1}(\Delta_a + a^l) & - & (\Delta_d + d^u)t^{j+1} \\ t^{j+1}(\Delta_a + a^l) & - & \Delta_d t^{j+1} \\ t^{j+1}(\Delta_a + a^u) & - & \Delta_d t^{j+1} \\ t^{j+1}\Delta_a & - & (\Delta_d + d^u)t^{j+1} \\ t^{j+1}\Delta_a & - & (\Delta_d + d^l)t^{j+1} \\ t^{j+1}\Delta_a & - & \Delta_d t^{j+1} \end{array} \tag{16}$$

(and similarly for the iteration (15) with u). The last choice is just the Jacobi iteration. They are all treated in the same way. We will now state a convergence result for the iteration (14).

We first consider the case in which the matrix A is nearly block-diagonal in the standard sense.

Theorem 1 *Let A be as in (1) , (13). Assume that*

$$2\sqrt{\|b\|_E\,\|c\|_E} + \|a^u\| + \|a^l\| + \|d^u\| + \|d^l\| < \beta \tag{17}$$

where[3]

$$\beta = \min_{i,j} |a_{jj} - d_{ii}| \ . \tag{18}$$

Then the Riccati equation (11) has a unique solution t^* in the ball $K = K(0,r)$ with

$$r = \frac{2\|c\|_E}{\beta - \|a^u\| - \|a^l\| - \|d^u\| - \|d^l\|} \ .$$

Moreover, for the sequence t^j from (14) we have

$$\|t^* - t^j\|_E \leq \frac{k^j}{1-k}\|t^1 - t^0\|_E \ ,$$

with

$$k = \frac{\|a^l\| + \|d^u\| + 2r\|b\|_E}{\beta - \|a^u\| - \|d^l\|} < 1 \ .$$

The estimates for the u - variables are obtained just by interchanging the roles of b and c . Moreover, the same conditions imply the convergence of any of the iterations in (16) , and similarly for u^j.

We omit the proof of this theorem. It can be easily reconstructed from the harder, scaled diagonally case below. The conditions (17), (18) in fact imply that the spectra of a and d are sufficiently distant.

[3] Here $\|\cdot\|_E$ is the Euclidean norm on matrices and $\|\cdot\|$ is the operator norm on matrices or matrix operators, induced by the Euclidean matrix norm.

We now consider matrices with the scaled diagonal dominance i.e. we set

$$A = DA_0D = \begin{bmatrix} D_a & 0 \\ 0 & D_d \end{bmatrix} \begin{bmatrix} a_0 & b_0 \\ c_0 & d_0 \end{bmatrix} \begin{bmatrix} D_a & 0 \\ 0 & D_d \end{bmatrix} \tag{19}$$

where D is diagonal, diagonal entries of A_0 have the absolute value one and A_0 is diagonally dominant in the usual sense.
We use the substitution

$$t = D_d\, \tau D_a^{-1} \quad , \quad u = D_a^{-1} \nu D_d \tag{20}$$

such that (11) , (12) are equivalent to

$$\tau a_0 - d_0 D_d^2\, \tau D_a^{-2} + c_0 - \tau b_0 D_d^2\, \tau D_a^{-2} = 0, \tag{21}$$

$$a_0 \nu - D_a^{-2} \nu D_d^2 d_0 + b_0 - D_a^{-2} \nu D_d^2 c_o \nu = 0. \tag{22}$$

Let

$$a_0 = E_a + a_0^u + a_0^l \quad , \quad d_0 = E_d + d_0^u + d_0^l \tag{23}$$

where E_a, a_0^u, a_0^l is the diagonal, strict upper triangular, strict lower triangular part of a_0 , respectively (and similarly for d_0). The equations (14) , (15) are equivalent to

$$\tau^{j+1}(E_a + a_0^u) - (E_d + d_0^l) D_d^2\, \tau^{j+1} D_a^{-2} + \tau^j a_0^l - d_0^u D_d^2\, \tau^j D_a^{-2} + \tag{24}$$
$$c_0 - \tau^j b_0 D_d^2\, \tau^j D_a^{-2} = 0 \, ,$$

τ_0 arbitrary,

$$(E_a + a_0^u)\nu^{j+1} - D_a^{-2} \nu^{j+1} D_d^2 (E_d + d_0^l) + a_0^l\, \nu^j - D_a^{-2} \nu^j D_d^2\, d_0^u + \tag{25}$$
$$b_0 - D_a^{-2} \nu^j D_d^2\, c_0 \nu^j = 0$$

ν_0 arbitrary.

Theorem 2 *Let A be as in (19) , (23). Assume that*

$$\alpha = \frac{\max_i |d_{ii}|}{\min_j |a_{jj}|} \leq 1 \tag{26}$$

and

$$2\sqrt{\alpha\|b_0\|_E\,\|c_0\|_E} + \|a_0^u\| + \|a_0^l\| + \alpha(\|d_0^u\| + \|d_0^l\|) < \beta_s \tag{27}$$

where

$$\beta_s = \min_{i,j}(1 - \frac{d_{ii}}{a_{jj}}) \ . \tag{28}$$

Then the equation (21) has a unique solution τ^* *in the ball* $K = K(0, r)$ *with*

$$r = \frac{2\gamma||c_0||_E}{1 - \gamma(||a_0^l|| + \alpha\|d_0^u\|)} \ , \tag{29}$$

$$\gamma = \frac{1}{\beta_s - ||a_0^u|| - \alpha||d_0^l||} \ . \tag{30}$$

Moreover, for the sequence τ^j *from (24) we have*

$$\|\tau^* - \tau^j\|_E \leq \frac{k^j}{1-k}\|\tau^1 - \tau^0\|_E \ ,$$

with

$$k = \gamma(||a_0^l|| + \alpha||d_0^u|| + 2\alpha r||b_0||_E) < l \ . \tag{31}$$

Again (27) and (26) ensure a sufficient distance of the spectra of a and d but now this distance is measured in the "relative sense".

Proof. We can write the equation (21) as

$$\tau(E_a + a_0^u + a_0^l) - (E_d + d_0^u + d_0^l)D_d^2\,\tau D_a^{-2} + c_0 - \tau b_0 D_d^2\,\tau D_a^{-2} = 0. \tag{32}$$

Let $\mathcal{L}$ be the linear operator defined by

$$\mathcal{L}\tau = \tau(E_a + a_0^u) - (E_d + d_0^l)D_d^2\,\tau D_a^{-2} \ . \tag{33}$$

The equation (32) is equivalent to

$$\mathcal{L}\tau = -\tau a_0^l + d_0^u\ D_d^2\,\tau D_a^{-2} - c_0 + \tau b_0 D_d^2\,\tau D_a^{-2} \ .$$

If $\mathcal{L}$ is invertible then this equation is written as

$$\mathcal{P}(\tau) = \tau$$

with

$$\mathcal{P}(\tau) = \mathcal{L}^{-1}(-\tau a_0^l + d_0^u D_d^2\,\tau D_a^{-2} - c_0 + \tau b_0 D_d^2\,\tau D_a^{-2})\ .$$

Set $\mathcal{L} = \mathcal{L}_0 + \mathcal{L}_1$, where

$$\mathcal{L}_0\tau = \tau E_a - E_d D_d^2\,\tau D_a^{-2} \quad , \qquad \mathcal{L}_1\tau = \tau a_0^u - d_0^l D_d^2\,\tau D_a^{-2}\ .$$

Then $\mathcal{L}_0$ is obviously invertible and

$$||\mathcal{L}_0^{-1}|| = \frac{1}{\beta_s}\ .$$

Thus

$$\mathcal{L} = (\mathcal{I} + \mathcal{L}_1\mathcal{L}_0^{-1})\mathcal{L}_0\ .$$

By (27) and

$$||\mathcal{L}_1|| \leq ||a_0^u|| + \alpha||d_0^l||$$

we obtain

$$\|\mathcal{L}_1\|\|\mathcal{L}_0^{-1}\| \leq \frac{1}{\beta_s}(\|a_0^u\| + \alpha\|d_0^l\|) < 1\ .$$

Thus, $\mathcal{L}$ is invertible and

$$\|\mathcal{L}^{-1}\| \leq \frac{\|\mathcal{L}_0^{-1}\|}{1 - \|\mathcal{L}_1\|\|\mathcal{L}_0^{-1}\|} \leq \gamma \tag{34}$$

$$\gamma = \frac{1}{\beta_s - ||a_0^u|| - \alpha||d_0^l||}\ .$$

We will find r such that $\mathcal{P} : K \longrightarrow K \quad K = K(0,r)$. Let $\tau \in K$. Then

$$\|\mathcal{P}(\tau)\|_E \leq \gamma\left[(\|a_0^l\| + \alpha\|d_0^u\|)\|\tau\|_E + \alpha\|b_0\|_E\,\|\tau\|_E^2 + \|c_0\|_E\right]\ .$$

The inclusion $\mathcal{P}(K) \subseteq K$ will follow, if

$$\alpha\gamma\ \|b_0\|_E\, r^2 - \left[1 - \gamma(\|a_0^l\| + \alpha\|d_0^u\|)\right]\ r + \gamma\ \|c_0\|_E \leq 0. \tag{35}$$

The assumption (27) implies the positivity of the discriminant of the quadratic function above. Consequently, the solutions of (35) make the closed interval $[r_{min}, r_{max}]$ with

$$r_{min} = \frac{2\gamma\|c_0\|_E}{g_s + \sqrt{g_s^2 - 4\alpha\gamma^2\|b_0\|_E\,\|c_0\|_E}}\ ,$$

$$r_{max} = \frac{2\gamma \|c_0\|_E}{g_s - \sqrt{g_s^2 - 4\alpha\gamma^2 \|b_0\|_E \, \|c_0\|_E}} \ ,$$

with

$$g_s = 1 - \gamma(\|a_0^l\| + \alpha \|d_0^u\|) \ .$$

For simplicity we take

$$r = \frac{2\gamma \ \|c_0\|_E}{g_s} \in (r_{min}, r_{max}) \ . \tag{36}$$

We prove the contractivity. Let $\bar{\tau}, \tau \in K$, then

$$\mathcal{P}(\bar{\tau}) - \mathcal{P}(\tau) =$$

$$\mathcal{L}^{-1}[-(\bar{\tau} - \tau)a_0^l + d_0^u \ D_d^2(\bar{\tau} - \tau)D_a^{-2} + \bar{\tau} b_0 D_d^2(\bar{\tau} - \tau)D_a^{-2} + (\bar{\tau} - \tau) b_0 D_d^2 \ \tau D_a^{-2}] \ ,$$

and

$$\|\mathcal{P}(\bar{\tau}) - \mathcal{P}(\tau)\|_E \leq k \|\bar{\tau} - \tau\|_E \ .$$

with k from (31). The condition (27) implies $k < 1$.

Let $\tau^1 = \mathcal{P}(\tau^0)$ for any $\tau^0 \in K$,

$$\tau^j = \mathcal{P}(\tau^{j-1}) = \mathcal{L}^{-1}(-\tau^{j-1} a_0^l + d_0^u \ D_d^2 \tau^{j-1} D_a^{-2} - c_0 + \tau^{j-1} b_0 D_d^2 \tau^{j-1} D_a^{-2}) \ .$$

By Banach fixed-point theorem [8] there is a unique solution $\tau^* \in K$ of the equation (21) satisfying

$$\|\tau^* - \tau^j\|_E \leq \frac{k^j}{1-k} \|\tau^1 - \tau^0\|_E \ .$$

Q.E.D.

The estimates for ν are again obtained by interchanging the roles of b_0 and c_0 .[4]

Analogous results are valid for all iterations listed in (16), more precisely, same assumptions lead to same results, although optimal constants, like respective r_{min} and k actually vary from case to case.

[4]In fact our assumption (27) is made a bit stronger to accomodate both iterations

Example 3 Let A in (1)

$$A = \begin{bmatrix} 1 & 2\varepsilon\sqrt{\varepsilon} & \varepsilon\sqrt{\varepsilon} \\ 2\varepsilon\sqrt{\varepsilon} & 4\varepsilon^2 & 2\varepsilon^2\sqrt{\varepsilon} \\ \varepsilon\sqrt{\varepsilon} & 2\varepsilon^2\sqrt{\varepsilon} & \varepsilon^2 \end{bmatrix} \tag{37}$$

and

$$a = \begin{bmatrix} 1 & 2\varepsilon\sqrt{\varepsilon} \\ 2\varepsilon\sqrt{\varepsilon} & 4\varepsilon^2 \end{bmatrix}$$

with $0 < \varepsilon < 1$.
Here the assumption (17) of Theorem 1 is not fullfilled.

The matrix (37) can be written in form (19) as

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2\varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix} \begin{bmatrix} 1 & \sqrt{\varepsilon} & \sqrt{\varepsilon} \\ \sqrt{\varepsilon} & 1 & \sqrt{\varepsilon} \\ \sqrt{\varepsilon} & \sqrt{\varepsilon} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2\varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix} \tag{38}$$

with

$$a_0 = \begin{bmatrix} 1 & \sqrt{\varepsilon} \\ \sqrt{\varepsilon} & 1 \end{bmatrix} .$$

Here are the assumptions of Theorem 2 satisfied for ε small enough. In fact, we obtain

$$\alpha = \frac{1}{4}, \quad \|b_0\|_E = \|c_0\|_E = \sqrt{2\varepsilon}, \quad \|a_0^u\| = \|a_0^l\| = \sqrt{\varepsilon}, \quad \beta_s = \frac{3}{4} ,$$

(d_0^u, d_0^l are not existent) and for (27)

$$(\sqrt{2} + 2)\sqrt{\varepsilon} < 3/4 .$$

3 Preserving of block diagonality

The rest of the paper gives estimates of the changes of the diagonal blocks of A after the similarity transformations (3) and (6) . Due to the special type of these block transformations this change will be quadratically small thus generalizing similar properties of Jacobi - like diagonalization in the 2×2 case. Similar results hold for scaled matrices.

In case a) by (9) we obtain

$$a - a_1 = (I + ut)^{-1/2}[w + (uc + bt - udt)(I + ut)^{-1/2}]$$

with

$$w = (I + ut)^{1/2}a - a(I + ut)^{-1/2} \ .$$

For $\|ut\| < 1$ using binomial expansion we obtain

$$\|(I + ut)^{-1/2}\| \leq \frac{1}{\sqrt{1 - \|ut\|}} \tag{39}$$

and

$$\|w\|_E \leq \frac{\|ut\|\|a\|_E}{\sqrt{1 - \|ut\|}} \ .$$

Then

$$\|a - a_1\|_E \leq \frac{1}{1 - \|ut\|}(\|ut\|\|a\|_E + \|uc + bt - udt\|_E) \ . \tag{40}$$

Similarly, for $\|tu\| < 1$

$$\|d - d_1\|_E \leq \frac{1}{1 - \|tu\|}(\|tu\|\|d\|_E + \|cu + tb + tua\|_E) \ . \tag{41}$$

In case b) we need the following lemma.

Lemma 4 *Let* $\|X\|_E \leq 1/4$. *Then* $I + X$ *can be decomposed as*

$$I + X = RL \tag{42}$$

$R,\ L^T$ *upper triangular with positive diagonal and*

$$R = I + Y \qquad , \qquad L = I + Z \tag{43}$$

$$\|Y\|_E, \|Z\|_E \leq \rho_{min} = \rho(X) = \frac{2\|X\|_E}{1 + \sqrt{1 - 4\|X\|_E}} \ . \tag{44}$$

For the proof see [7], Th. 2.1.

By (10) we obtain

$$a - a_2 = R_1^{-1}[R_1 a - aL_1^{-1} + (uc + bt - udt)L_1^{-1}] \ .$$

Since

$$R_1 L_1 = I + ut$$

by Lemma 4

$$R_1 = I + Y_1 \qquad \text{and} \qquad L_1^{-1} = I + Z_1$$

with

$$\|Y_1\|_E \le \rho(ut) \le 2\|ut\|_E \quad \text{and} \quad \|Z_1\|_E \le \frac{\rho(ut)}{1-\rho(ut)} \le \frac{2\|ut\|_E}{1-2\|ut\|_E}$$

it follows

$$\|a - a_2\|_E \le \|R_1^{-1}\|[(\|Y_1\| + \|Z_1\|)\|a\|_E + \|uc + bt - udt\|_E \|L_1^{-1}\|] \ . \qquad (45)$$

Similarly,

$$\|d - d_2\|_E \le \|R_2^{-1}\|[(\|Y_2\| + \|Z_2\|)\|d\|_E + \|cu + tb + tua\|_E \, \|L_2^{-1}\|] \qquad (46)$$

with

$$\|Y_2\|_E \le 2\|tu\|_E \qquad \text{and} \qquad \|Z_2\|_E \le \frac{2\|tu\|_E}{1-2\|tu\|_E} \ .$$

As we see in all these formulae the change of diagonal blocks is quadratical in r from Theorem 1.

Now we consider the case when u, t are the solutions of the Riccati equations in case a). From (12) we have

$$udt = aut + bt - ucut \ . \qquad (47)$$

Inserting (47) in (9) we obtain

$$a_1 = (I + ut)^{-1/2}(a - uc)(I + ut)^{1/2} \ . \qquad (48)$$

Similarly, from (11)

$$tau = dtu - cu + tbtu \ . \tag{49}$$

Inserting (49) in (9) we obtain

$$d_1 = (I + tu)^{-1/2}(d + tb)(I + tu)^{1/2} \ . \tag{50}$$

Now

$$a - a_1 = (I + ut)^{-1/2}[w_1 + uc(I + ut)^{1/2}]$$

with

$$w_1 = (I + ut)^{1/2}a - a(I + ut)^{1/2} \ .$$

For $\|ut\| < 1$ we obtain

$$\|(I + ut)^{1/2}\| \leq 1 + \frac{\|ut\|}{1 + \sqrt{1 - \|ut\|}} \leq 1 + \|ut\| \tag{51}$$

and

$$\|w_1\|_E \leq \frac{2\|ut\|\|a\|_E}{1 + \sqrt{1 + \|ut\|}} \ . \tag{52}$$

Using (39) , (52) and (51) we obtain

$$\|a - a_1\|_E \leq \frac{1}{\sqrt{1 - \|ut\|}} \left[\frac{2\|ut\|\|a\|_E}{1 + \sqrt{1 + \|ut\|}} + (1 + \|ut\|)\|uc\|_E \right] \tag{53}$$

and for $\|tu\| < 1$

$$\|d - d_1\|_E \leq \frac{1}{\sqrt{1 - \|tu\|}} \left[\frac{2\|tu\|\|d\|_E}{1 + \sqrt{1 + \|tu\|}} + (1 + \|tu\|)\|tb\|_E \right] \ . \tag{54}$$

In case b) we have $b_2 = 0$ and $c_2 = 0$ with (8) we obtain

$$a_2 = R_1^{-1}(a - uc)R_1, \tag{55}$$

$$d_2 = R_2^{-1}(d + tb)R_2 \tag{56}$$

By (8) and Lemma 4 we have

$$R_1 = I + Y_1 \qquad , \qquad R_2 = I + Y_2$$

with

$$\|Y_1\|_E \le 2\|ut\|_E \qquad , \qquad \|Y_2\|_E \le 2\|tu\|_E \ .$$

It follows

$$\|a - a_2\|_E \le \|R_1^{-1}\|(2\|Y_1\|\|a\|_E + \|uc\|_E \, \|R_1\|) \ , \tag{57}$$

$$\|d - d_2\|_E \le \|R_2^{-1}\|(2\|Y_2\|\|d\|_E + \|tb\|_E \, \|R_2\|) \ . \tag{58}$$

Now as before we consider matrices with scaled diagonal dominance (19). By substituting (20) in (3) , (4) , (6) , (7) for

$$A_1 = V^{-1} D A_0 D V$$

and

$$A_2 = U^{-1} D A_0 D U$$

we obtain

$$\begin{aligned}
a_1 &= D_a[(I + D_a^{-2}\nu D_d^2\tau)^{-1/2}\bar{a}_0(I + \nu D_d^2\tau D_a^{-2})^{-1/2}]D_a \\
b_1 &= D_a[(I + D_a^{-2}\nu D_d^2\tau)^{-1/2}\bar{b}_0(I + D_d^2\tau D_a^{-2}\nu)^{-1/2}]D_d \\
c_1 &= D_d[(I + \tau D_a^{-2}\nu D_d^2)^{-1/2}\bar{c}_0(I + \nu D_d^2\tau D_a^{-2})^{-1/2}]D_a \\
d_1 &= D_d[(I + \tau D_a^{-2}\nu D_d^2)^{-1/2}\bar{d}_0(I + D_d^2\tau D_a^{-2}\nu)^{-1/2}]D_d
\end{aligned} \tag{59}$$

and

$$\begin{aligned}
a_2 &= D_a(D_a^{-1}R_1^{-1}D_a\bar{a}_0D_aL_1^{-1}D_a^{-1})D_a \\
b_2 &= D_a(D_a^{-1}R_1^{-1}D_a\bar{b}_0D_dL_2^{-1}D_d^{-1})D_d \\
c_2 &= D_d(D_d^{-1}R_2^{-1}D_d\bar{c}_0D_aL_1^{-1}D_a^{-1})D_a \\
d_2 &= D_d(D_d^{-1}R_2^{-1}D_d\bar{d}_0D_dL_2^{-1}D_d^{-1})D_d.
\end{aligned} \tag{60}$$

with

$$\begin{aligned}
\bar{a}_0 &= a_0 - D_a^{-2}\nu D_d^2 c_0 - b_0 D_d^2\tau D_a^{-2} + D_a^{-2}\nu D_d^2 d_0 D_d^2\tau D_a^{-2} \\
\bar{b}_0 &= a_0\nu - D_a^{-2}\nu D_d^2 c_0\nu + b_0 - D_a^{-2}\nu D_d^2 d_0 \\
\bar{c}_0 &= \tau a_0 + c_0 - \tau b_0 D_d^2\tau D_a^{-2} - d_0 D_d^2\tau D_a^{-2} \\
\bar{d}_0 &= \tau a_0\nu + c_0\nu + \tau b_0 + d_0 \ .
\end{aligned} \tag{61}$$

Now by (19) and (59) after a lengthy calculation we obtain

$$\|D_a^{-1}(a - a_1)D_a^{-1}\|_E \le \alpha s \tag{62}$$

and

$$\|D_d^{-1}(d-d_1)D_d^{-1}\|_E \le s \tag{63}$$

where α is as in (26) and $\alpha\|\nu\|\|\tau\|<1$ and

$$s = \frac{\|\nu\|\|\tau\|(\|a_0\|_E + \alpha\|d_0\|_E) + \|\tau\|\|b_0\|_E + \|\nu\|\|c_0\|_E}{1-\alpha\|\nu\|\|\tau\|} \ . \tag{64}$$

Similarly, by (19) and (60) we obtain

$$\|D_a^{-1}(a-a_2)D_a^{-1}\|_E \le$$

$$\|R_1^{-1}\|[(\|Y_1\|+\|Z_1\|)\|a_0\|_E+\alpha(\alpha\|\nu\|\|\tau\|\|d_0\|_E+\|\tau\|\|b_0\|_E+\|\nu\|\|c_0\|_E)\|L_1^{-1}\|] \ , \tag{65}$$

$$\|D_d^{-1}(d-d_2)D_d^{-1}\|_E \le$$

$$\|R_2^{-1}\|[(\|Y_2\|+\|Z_2\|)\|d_0\|_E + (\|\nu\|\|\tau\|\|a_0\|_E + \|\tau\|\|b_0\|_E + \|\nu\|\|c_0\|_E)\|L_2^{-1}\|] \tag{66}$$

where

$$\|Y_1\|_E, \|Y_2\|_E \le 2\alpha\|\nu\|_E\,\|\tau\|_E \qquad \text{and} \qquad \|Z_1\|_E, \|Z_2\|_E \le \frac{2\alpha\|\nu\|_E\,\|\tau\|_E}{1-2\alpha\|\nu\|_E\,\|\tau\|_E} \ .$$

The estimates (62) , (63) are "scaled" analogs of (53) , (54) and similarly (65) , (66) are analogs of (45), (46), respectively. If $b_1=0$ and $c_1=0$ then ν, τ is the solution of (22) , (21) respectively. Inserting (19) and (20) in (48) and (50) we obtain

$$a_1 = D_a(I + D_a^{-2}\nu D_d^2\tau)^{-1/2}(a_0 - D_a^{-2}\nu D_d^2 c_0)(I+\nu D_d^2\tau D_a^{-2})^{1/2}D_a \ , \tag{67}$$

and

$$d_1 = D_d(I+\tau D_a^{-2}\nu D_d^2)^{-1/2}(a_0+\tau b_0)(I+D_d^2\tau D_a^{-2}\nu)^{1/2}D_d \ . \tag{68}$$

For $\alpha\|\nu\|\|\tau\|<1$,

$$\|D_a^{-1}(a-a_1)D_a^{-1}\|_E \le$$

$$\leq \frac{\alpha\|\nu\|}{\sqrt{1-\alpha\|\nu\|\|\tau\|}}\left[\frac{2\|\tau\|}{1+\sqrt{1+\alpha\|\nu\|\|\tau\|}}\|a_0\|_E+(1+\alpha\|\nu\|\|\tau\|)\|c_o\|_E\right] \ , \tag{69}$$

$$\|D_d^{-1}(d-d_1)D_d^{-1}\|_E \leq$$
$$\leq \frac{\|\tau\|}{\sqrt{1-\alpha\|\nu\|\|\tau\|}}\left[\frac{2\alpha\|\nu\|}{1+\sqrt{1+\alpha\|\nu\|\|\tau\|}}\|d_0\|_E+(1+\alpha\|\nu\|\|\tau\|)\|b_0\|_E\right] . \tag{70}$$

Similarly, for $b_2=0$ and $c_2=0$ we obtain

$$a_2 = D_a[D_a^{-1}R_1^{-1}D_a(a_0-D_a^{-2}\nu D_d^2 c_0)(I+\nu D_d^2\tau D_a^{-2})D_a L_1^{-1}D_a^{-1}]D_a \ , \tag{71}$$

$$d_2 = D_d[D_d^{-1}R_2^{-1}D_d(d_0+\tau b_0)(I+D_d^2\tau D_a^{-2}\nu)D_d L_2^{-1}D_d^{-1}]D_d \ . \tag{72}$$

Then

$$\|D_a^{-1}(a-a_1)D_a^{-1}\|_E \leq$$
$$\|R_1^{-1}\|[(\|Y_1\|+\|S_1\|+\|Y_1\|\|S_1\|+\alpha\|\nu\|\|\tau\|)\|a_0\|_E+\alpha\|\nu\|(1+\alpha\|\tau\|)\|c_0\|_E]\|L_1^{-1}\| \ , \tag{73}$$

$$\|D_d^{-1}(d-d_1)D_d^{-1}\|_E \leq$$
$$\leq \|R_2^{-1}\|[(\|Y_2\|+\|S_2\|+\|Y_2\|\|S_2\|+\alpha\|\nu\|\|\tau\|)\|d_0\|_E+\|\tau\|(1+\alpha\|\nu\|)\|b_0\|_E]\|L_2^{-1}\| \tag{74}$$

where by Lemma 4

$$R_1 = I+Y_1 \quad , \quad R_2 = I+Y_2 \quad , \quad L_1 = I+S_1 \quad , \quad L_2 = I+S_2$$

and

$$\|Y_1\|_E, \|Y_2\|_E, \|S_1\|_E, \|S_2\|_E \leq 2\alpha\|\nu\|_E\,\|\tau\|_E \ .$$

In concluding this section we can say that in all cases the change of the diagonal elements is quadratically small in the measure of the (scaled) diagonal dominance and the same kind of diagonal dominance is shared by the transformed matrix as well. In view of the results in [1] and [2] this strongly suggests that all these transformations, at least in the Hermitian case, are candidates for an eigenvalue algorithm with high relative accuracy. The basis of our last estimates are the formulae (48), (50), (55) and (56) for the transformed diagonal blocks. These formulae are of independent interest. To illustrate this, assume that A is Hermitian and V unitary and that block diagonalization took place, i.e. $c_1 = b_1^H = 0$. Then using binomial expansion and $\|t\| = \mathcal{O}(\|c\|)$ the formula (48) yields

$$a_1 = a - t^H c + \frac{at^H t - t^H ta}{2} + \mathcal{O}(\|c\|^4).$$

Since a_1 has to be Hermitian we obtain

$$a_1 = a - \frac{t^H c + (t^H c)^H}{2} + \mathcal{O}(\|c\|^4),$$

i.e.

$$a_1 = \text{ Hermitian part of } (a - t^H c) \tag{75}$$

up to a quartic error. The same is true of d_1. A scaled variant of this approximation could be derived as well.

4 Some numerical examples

We have tried some typical iterations with a MATLAB - based code which just yields the number of iterations which, in turn allows to estimate the real efficiency. As predicted by our theory the number of iterations is independent of the dimensions n, m, which allows to compute few eigenpairs with the cost of some n^2 operations. Our first example were (in the MATLAB notation)

```
A = diag(1:300)+ rand(300)/80
```

which looks "diagonally dominant" but is well-beyond the Gershgorin separation. For $m = 3,\ 5,\ 20,\ 150$ we obtained block diagonality to working accuracy after 10 Jacobi iterations or $6 - 8$ Gauss-Seidel iterations (14), (15). The second example

```
A = diag(200:-1:1)(eye(200) + rand(200)/10000)diag(200:-1:1)
```

is only scaled diagonally dominant according to [1]. For $m = 2,\ 5,\ 20,\ 100$ we needed $10-13$ Jacobi iterations and $9-12$ Gauss-Seidel iterations.

To illustrate high relative accuracy of the algorithm consider the matrix

$$A = \begin{bmatrix} 1\cdot 10^{20} & 2 & 3 & 4 \\ 2 & 4\cdot 10^{20} & 5 & 6 \\ 3 & 5 & 7 & 8 \\ 4 & 6 & 8 & 9 \end{bmatrix}$$

whose small eigenvalues are up to a 10^{-20} error equal to those of

$$\begin{bmatrix} 7 & 8 \\ 8 & 9 \end{bmatrix}.$$

Both eigenvalues are correctly reproduced by our method (standard eigensolver from MATLAB yielded completely wrong values 0 and 10^4).

A rigorous justification of the above illustrated accuracy as well as extensive numerical experiments shall be presented elsewhere.

References

[1] J. Barlow, J. Demmel, Computing accurate eigensystems of scaled diagonally dominant matrices, SIAM J. Numer. Anal. **27** (1988) 762-791.

[2] Demmel, J., Veselić, K., Jacobi's method is more accurate than QR, SIAM J. Matrix Anal. Appl., **13** (1992) 1204–1245.

[3] Z. Drmač, K. Veselić, On the spectra of nearly unitarily similar Hermitian matrices, LAA **309** (2000) 191 -215.

[4] C.G.J. Jacobi, Über ein leichtes Verfahren die in der Theorie der Säculärstörungen vorkomenden Gleichungen numerisch aufzulösen, Crelle's J. **330** (1846) 51-94.

[5] A.N. Malyshev, On iterative refinement for the spectral decomposition of Symmetric Matrices, IRISA, preprint No 628, (1992) 1-25.

[6] G.L.G. Sleipen and H.A. van der Vorst, A Jacobi-Davison iteration method for linear eigenvalue problems, SIAM J. Matrix Anal. Appl. **17** (1996) 401-425.

[7] G. W. Stewart, On the perturbation of LU, Cholesky, and QR factorizations, SIAM J. Matrix Anal. Appl. **14** (1993) 551-566

[8] J. Stoer, R. Bulirsch, Introduction to numerical analysis, Spinger, Berlin 1980

[9] K. Veselić, On a new class of elementary matrices, Numer. Math. **33** (1979) 173-180.