Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

A SMOOTH VERSION OF SYLVESTER'S LAW OF INERTIA AND ITS NUMERICAL REALIZATION*

PETER KUNKEL†

Abstract. A smooth version of Sylvester's law of inertia is presented for symmetric matrix functions of constant rank. The techniques used in the proof are constructive but the resulting numerical approaches are unstable, and therefore require stabilization. Two different stabilization techniques are suggested, one based on a descent method and one based on Newton's method. Some numerical tests are included to demonstrate the applicability of the obtained numerical methods.

Key words. Matrix functions, Sylvester's law of inertia, Numerical realization, Stabilization.

AMS subject classifications. 15A21, 65F30, 65K10.

1. Introduction. Dealing with matrices both theoretically and numerically, a common tool is the use of factorizations and canonical forms under certain classes of transformations. Common factorizations are the LR and QR decomposition in the context of the solution of linear equations, and Schur form and the singular value decomposition in the context of eigenvalue problems, see [5]. Given additional structural properties of the matrices under consideration, the used factorization should maintain these properties since they may reflect physical properties or lead to more efficient algorithms. In the case of symmetric matrices, typical factorizations are the Cholesky decomposition and its rational (i.e., root-free) modification also called LDLT decomposition as well as Sylvester's law of inertia, see again [5]. Further important structural classes are skew-symmetric matrices, Hamiltonian matrices, and symplectic matrices which occur, e.g., in the context of geometric integration of structured ordinary differential equations, see [6].

Turning to smooth matrix functions, which play a central role for example in the theoretical and numerical treatment of linear differential-algebraic equations with variable coefficients, see [8], it is then a natural question whether we can obtain similar factorizations in a pointwise manner such that all factors inherit the smoothness of the given matrix function. It is well-known that this is sometimes only possible locally or only with loss of smoothness even if we assume that the pointwise rank of the matrix function is constant, see [3]. In the cases where such smooth factorizations exist globally, we can use them to derive suitable canonical forms without losing smoothness or being forced to restrict the domain to a sufficiently small open set. An example in this respect can be found in [9] where pointwise skew-symmetric matrices were treated and it was shown that under constant rank assumptions we can smoothly transform a given self-adjoint differential-algebraic equation by appropriately chosen changes of bases in such a way that the dynamics are described by a Hamiltonian system implying that the dynamics are described by a symplectic flow.

Recently another structural class of differential-algebraic equations has been considered in [10, 11] containing pointwise symmetric matrix function. To get a canonical form under smooth pointwise congruence transformations in the spirit of [9], a global and smooth version of Sylvester's law of inertia is needed. The purpose of this paper is to present such a result. Since it turns out that the constructive algorithms which are used in the proof yield unstable transformations the result is not satisfactory from the numerical point

*Received by the editors on March 24, 2020. Accepted for publication on June 16, 2020. Handling Editor: Heike Fassbender.
†Mathematisches Institut, Universität Leipzig, Augustusplatz 10, D-04109 Leipzig, Germany (kunkel@math.uni-leipzig.de).

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

of view. We therefore also include techniques to get stable numerical procedures.

The paper is organized as follows. In Section 2, we introduce the notation and provide some basic results we need in the further course of the paper. We then state and prove a smooth version of Sylvester's law of inertia in Section 3. Since it turns out that from the numerical point of view, the constructive techniques used so far are not satisfactory we develop stabilizations of the approach in Section 4. We then present some numerical experiments to illustrate the obtained results in Section 5 and close with a conclusion in Section 6.

2. Preliminaries. Let C^k for $k \in \mathbb{N}_0$ denote the class of k-times continuously differentiable functions, let C^∞ denote the class of infinitely often continuously differentiable functions, and let C^ω denote the class of (real) analytic functions. Furthermore, let $\mathrm{GL}(n)$ be the general linear group containing all invertible matrices in $\mathbb{R}^{n,n}$ and let $\mathrm{O}(n)$ be the orthogonal group containing all orthogonal matrices in $\mathbb{R}^{n,n}$.

Consider a matrix function $E \in C^k(\mathbb{I}, \mathbb{R}^{m,n})$ with $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$, where $\mathbb{I} \subset \mathbb{R}$ is a compact interval with non-empty interior. Throughout this paper, we will assume that E has constant rank in the sense that there is an $r \in \mathbb{N}_0$ with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$. Additionally, all relations between matrix functions are to be understood to hold pointwise. In the case $k \neq 0$, we write $\dot{E}$ for $\frac{d}{dt}E$. Finally, we write I_n for the identity matrix in $\mathbb{R}^{n,n}$.

A classical result for such a matrix function, which is also needed in what follows, is due to [4].

THEOREM 2.1. *Let $A \in C^k(\mathbb{I}, \mathbb{R}^{m,n})$ with $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$ have constant rank. Then there exist matrix functions $U \in C^k(\mathbb{I}, \mathrm{O}(m))$ and $V \in C^k(\mathbb{I}, \mathrm{O}(n))$ such that*

$$U^T A V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \tag{2.1}$$

with $\Sigma \in C^k(\mathbb{I}, \mathrm{GL}(r))$.

In particular, Theorem 2.1 guarantees the existence of smooth bases of kernel, cokernel, range, and corange of A of the same class of smoothness as A being defined globally on the whole interval $\mathbb{I}$. Note that a similar global result in the case when A depends on multiple parameters cannot hold in view of the hairy ball theorem, sometimes also called the theorem of the combed hedgehog. For a local version of Theorem 2.1 in the case of multiple parameters, see, e.g., [8].

Since we deal with pointwise congruence in the symmetric case, we need the following immediate consequence of Theorem 2.1, see, e.g., [12]

COROLLARY 2.2. *Let $A \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$ with $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$ have constant rank and let $A^T = A$, i.e., let A be pointwise symmetric. Then there exists a matrix function $Q \in C^k(\mathbb{I}, \mathrm{O}(n))$ such that*

$$Q^T A Q = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

with $\Sigma \in C^k(\mathbb{I}, \mathrm{GL}(r))$.

Proof. Taking $Q = U$ from Theorem 2.1, we get

$$Q^T A = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad B = \begin{bmatrix} \Sigma & 0 \end{bmatrix} V^T,$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

where B possesses pointwise full row rank. By symmetry of A, it follows that

$$A^TQ = AQ = \begin{bmatrix} B^T & 0 \end{bmatrix}.$$

Hence,

$$Q^TAQ = \begin{bmatrix} BQ \\ 0 \end{bmatrix} = \begin{bmatrix} Q^TB^T & 0 \end{bmatrix}$$

and the claim holds. ∎

The first aim of the present paper is to obtain a smooth version of Sylvester's law of inertia. Although Sylvester's law of inertia is a well-known textbook result we state it here together with a proof for later reference.

THEOREM 2.3. *Let $E \in \mathbb{R}^{n,n}$ and let $E^T = E$. Then there exists a matrix $W \in \mathrm{GL}(n)$ with*

$$W^TEW = \begin{bmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In particular, p and q are the number of positive and negative eigenvalues of E.

Proof. Since E is symmetric, the spectral theorem guarantees the existence of a $Q \in \mathrm{O}(n)$ such that

$$Q^TEQ = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

with all eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Without loss of generality, we may assume that

$$\lambda_1, \ldots, \lambda_p > 0, \quad \lambda_{p+1}, \ldots, \lambda_{p+q} < 0, \quad \lambda_{p+q+1}, \ldots, \lambda_n = 0.$$

Defining $D = \mathrm{diag}(d_1, \ldots, d_n)$ by $d_i = 1/\sqrt{|\lambda_i|}$ for $i = 1, \ldots, p+q$ and $d_i = 1$ otherwise yields the desired result with $W = QD$. ∎

As we will explain in the next section, the proof of Theorem 2.3 cannot be transferred to the case of matrix functions.

3. A smooth version of Sylvester's law of inertia. In order to prove a smooth version of Sylvester's law of inertia, an obvious idea would be to follow the proof of Theorem 2.3 and use a smooth version of the spectral theorem. But this is only possible under loss of smoothness. Even for matrix functions of class C^∞ the eigenvalues may only be of class C^1, see [3] for details. This, however, does not mean that there is no smooth version of Sylvester's law of inertia.

THEOREM 3.1. *Let $E \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$ with $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$ have constant rank r and let $E^T = E$. Then there exists a matrix function $W \in C^k(\mathbb{I}, \mathrm{GL}(n))$ such that*

$$W^TEW = \begin{bmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.2}$$

with $p, q \in \{0, \ldots, n\}$ independent of $t \in \mathbb{I}$ and $p + q = r$.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

Proof. Applying Corollary 2.2 we immediately get the third block row and column in (3.2). Thus, we are allowed to assume without loss of generality that $r = n$. Since the eigenvalues depend continuously on the matrix, see [3], they cannot change sign. Hence, p and q are constant on $\mathbb{I}$. For convenience, we then write $S = \operatorname{diag}(I_p, -I_q)$. Furthermore, let $\hat{t} \in \mathbb{I}$ and $\hat{W} \in \mathbb{R}^{n,n}$ with $\hat{W}^T E(\hat{t})\hat{W} = S$ according to Theorem 2.3. As it is typical for proofs in this area we need to distinguish two cases.

For $k \neq 0$, the initial value problem

$$\dot{W} = -\tfrac{1}{2}E^{-1}\dot{E}W, \quad W(\hat{t}) = \hat{W}$$

consisting of a matricial linear ordinary differential equation possesses a solution $W \in C^k(\mathbb{I}, GL(n))$ satisfying

$$\begin{aligned}\tfrac{d}{dt}(W^TEW) &= \dot{W}^TEW + W^T\dot{E}W + W^TE\dot{W}\\ &= -\tfrac{1}{2}(E^{-1}\dot{E}W)^TEW + W^T\dot{E}W - \tfrac{1}{2}W^TE(E^{-1}\dot{E}W)\\ &= -\tfrac{1}{2}W^T\dot{E}W + W^T\dot{E}W - \tfrac{1}{2}W^T\dot{E}W = 0\end{aligned}$$

due to the symmetry of E, $\dot{E}$, and E^{-1}. Hence, W^TEW is constant. Because of $W(\hat{t})^TE(\hat{t})W(\hat{t}) = \hat{W}^TE(\hat{t})\hat{W} = S$ the constant is given by S and $W^TE\,W = S$ holds.

For $k = 0$, let

$$\hat{W}^TE\hat{W} = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

such that $E_{11} \in C^0(\mathbb{I}, \mathbb{R}^{p,p})$, $E_{12} = E_{21}^T \in C^0(\mathbb{I}, \mathbb{R}^{p,q})$, and $E_{22} \in C^0(\mathbb{I}, \mathbb{R}^{q,q})$. Moreover, we have that $E_{11}(\hat{t}) = I_p$, $E_{12}(\hat{t}) = E_{21}(\hat{t})^T = 0$, and $E_{22}(\hat{t}) = -I_q$. Thus, there exists a sufficiently small (relatively) open neighborhood $\hat{\mathbb{I}} \subseteq \mathbb{I}$ of $\hat{t}$ such that E_{11} is pointwise symmetric positive definite, E_{22} is pointwise symmetric negative definite, and E_{12} is arbitrarily small in norm if we restrict all functions to $\hat{\mathbb{I}}$. Since Cholesky decomposition is a smooth process there is an $L_{11} \in C^0(\hat{\mathbb{I}}, \mathbb{R}^{p,p})$ with

$$E_{11} = L_{11}L_{11}^T.$$

Hence,

$$\begin{bmatrix} L_{11}^{-1} & 0 \\ 0 & I_q \end{bmatrix}\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}\begin{bmatrix} L_{11}^{-T} & 0 \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & L_{11}^{-1}E_{12} \\ E_{12}^TL_{11}^{-T} & E_{22} \end{bmatrix}$$

and

$$\begin{bmatrix} I_p & 0 \\ -E_{12}^TL_{11}^{-T} & I_q \end{bmatrix}\begin{bmatrix} I_p & L_{11}^{-1}E_{12} \\ E_{12}^TL_{11}^{-T} & E_{22} \end{bmatrix}\begin{bmatrix} I_p & -L_{11}^{-1}E_{12}^T \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & \tilde{E}_{22} \end{bmatrix}$$

with the Schur complement $\tilde{E}_{22} = E_{22} - E_{12}^TE_{11}^{-1}E_{12}$. For sufficiently small $\hat{\mathbb{I}}$, the matrix function $\tilde{E}_{22} \in C^0(\hat{\mathbb{I}}, \mathbb{R}^{q,q})$ as perturbation of E_{22} is still pointwise symmetric negative definite. Accordingly, there is an $L_{22} \in C^0(\hat{\mathbb{I}}, \mathbb{R}^{q,q})$ with

$$-\tilde{E}_{22} = L_{22}L_{22}^T$$

and

$$\begin{bmatrix} I_p & 0 \\ 0 & L_{22}^{-1} \end{bmatrix}\begin{bmatrix} I_p & 0 \\ 0 & \tilde{E}_{22} \end{bmatrix}\begin{bmatrix} I_p & 0 \\ 0 & L_{22}^{-T} \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} = S.$$

Thus,

$$W = \hat{W}\begin{bmatrix} L_{11}^{-T} & 0 \\ 0 & I_q \end{bmatrix}\begin{bmatrix} I_p & -L_{11}^{-1}E_{12}^T \\ 0 & I_q \end{bmatrix}\begin{bmatrix} I_p & 0 \\ 0 & L_{22}^{-T} \end{bmatrix}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

gives the desired property $W^TEW = S$ on $\hat{\mathbb{I}}$. Since this can be done for every $\hat{t} \in \mathbb{I}$ we obtain an open covering of $\mathbb{I}$. Due to the compactness of $\mathbb{I}$ it contains a finite open covering of $\mathbb{I}$. From this finite open covering we can deduce finitely many points

$$t_0 < t_1 < \cdots < t_{m-1} < t_m$$

such that $\mathbb{I} = [t_0, t_m]$ and

$$W_i^T E W_i = S \quad \text{on} \ \ [t_i, t_{i+1}]$$

for $i = 0, \ldots, m-1$, where $W_i \in C^0([t_i, t_{i+1}], \mathbb{R}^{n,n})$ from the above construction. In general, the pieces W_i do not combine to a continuous function on $\mathbb{I}$. Hence, we need to modify the pieces to fit them together. For this, consider $i \in \{1, \ldots, m-1\}$. At the point t_i, we have

$$W_{i-1}(t_i)^T E(t_i) W_{i-1}(t_i) = S, \quad W_i(t_i)^T E(t_i) W_i(t_i) = S$$

implying that

$$W_{i-1}(t_i)^{-T} S W_{i-1}(t_i)^{-1} = W_i(t_i)^{-T} S W_i(t_i)^{-1}.$$

Defining

$$\tilde{W}_i = W_i \, W_i(t_i)^{-1} W_{i-1}(t_i),$$

yields

$$\tilde{W}_i(t_i) = W_i(t_i) W_i(t_i)^{-1} W_{i-1}(t_i) = W_{i-1}(t_i)$$

and

$$\begin{aligned} \tilde{W}_i^T E \tilde{W}_i &= W_{i-1}(t_i)^T W_i(t_i)^{-T} W_i^T E \, W_i \, W_i(t_i)^{-1} W_{i-1}(t_i), \\ &= W_{i-1}(t_i)^T W_i(t_i)^{-T} S \, W_i(t_i)^{-1} W_{i-1}(t_i) = S, \end{aligned}$$

hence continuity on $[t_{i-1}, t_{i+1}]$. Proceeding in this way from left to right we obtain a matrix function $W \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ satisfying $W^T E \, W = S$. ∎

Remark 3.2. The previous proof for the case $k = 0$ can also be based on a generalization of the Cholesky decomposition for indefinite matrices of the form

$$\Pi^T A \Pi = L^T D L \tag{3.3}$$

with a permutation matrix Π describing the necessary pivoting, a unit lower triangular matrix L, and a block diagonal matrix D consisting of 1-by-1- and 2-by-2-blocks, see [1]. For the construction of the reference transformation $\hat{W}$, we need to blockwise diagonalize D, transform the so obtained diagonal entries to ± 1 as in the proof of Theorem 2.3, and reorder to finally get S. For the locally smooth construction, we can then use (3.3) for $\hat{W}^T E \hat{W}$ but with no permutation and only allowing for 1-by-1-blocks in the diagonal matrix.

From the theoretical point of view, the posed problem is solved and we have proven a smooth version of Sylvester's law of inertia. From the numerical point of view, the positive aspect of the above proof is that it is constructive, i.e., it directly proposes a method to determine a possible smooth transformation W satisfying (3.2). Numerical tests, however, show that even if there exits a possible nicely bounded W the accordingly computed W may blow up in norm.

Example 3.3. Let

$$E(t) = \begin{bmatrix} \cos(2\pi t) & \sin(2\pi t) \\ -\sin(2\pi t) & \cos(2\pi t) \end{bmatrix}^T \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix} \begin{bmatrix} \cos(2\pi t) & \sin(2\pi t) \\ -\sin(2\pi t) & \cos(2\pi t) \end{bmatrix}.$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

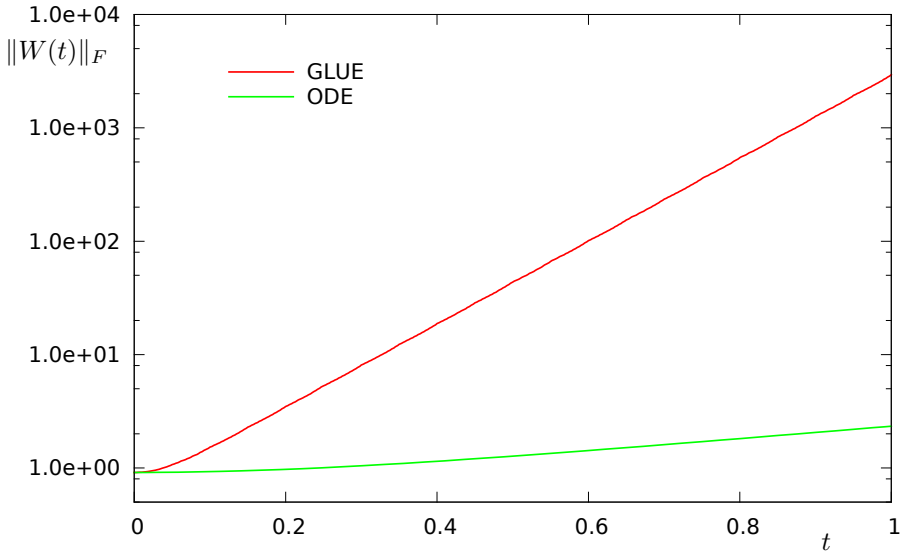

FIGURE 1. *Norm of W as function of t for Example 3.3 constructed according to the proof of Theorem 3.1.*

Figure 1 shows the Frobenius norm of W as a function of t obtained by the procedures from the proof of Theorem 3.1, where `ODE` refers to the case $k \neq 0$ and `GLUE` refers to the case $k = 0$. Obviously, we have exponential growth in both approaches which is very drastic in the second case.

4. Numerical realization. In order to stabilize the procedures from the proof of Theorem 3.1 we consider

$$W^T E W = S, \quad S = \operatorname{diag}(I_p, -I_q) \tag{4.4}$$

with $E \in \mathbb{R}^{n,n}$ and nonsingular $W \in \mathbb{R}^{n,n}$ and try to transform it according to

$$Z^T W^T E W Z = Z^T S Z$$

with nonsingular $Z \in \mathbb{R}^{n,n}$ such that $Z^T S Z = S$ and WZ is smaller in some appropriate norm. Using the Frobenius norm this leads us to an optimization problem of the form

$$\|WZ\|_F^2 = \min \quad \text{subject to} \quad Z^T S Z = S \tag{4.5}$$

for $Z \in \mathbb{R}^{n,n}$. The solution set of the constraint is given by the (quadratic) Lie group $\mathrm{O}(p,q)$, the so-called indefinite orthogonal group of signature (p,q), see, e.g., [7]. It is known that $\mathrm{O}(p,q)$ is closed with $\dim \mathrm{O}(p,q) = \frac{1}{2}n(n+1)$. Furthermore, $\mathrm{O}(p,q)$ coincides with $\mathrm{O}(n)$ for $pq = 0$ and is thus bounded in this case. However, $\mathrm{O}(p,q)$ is unbounded for $pq \neq 0$ which actually leads to the problems observed for Example 3.3.

THEOREM 4.1. *Let $W \in \mathbb{R}^{n,n}$ be nonsingular. Then the optimization problem (4.5) possesses a solution.*

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

Proof. The relation $\|Z\|_W = \|WZ\|_F$ defines a new matrix norm. The problem (4.5) can then be written as

$$\|Z\|_W^2 = \min \quad \text{subject to} \quad Z \in \mathrm{O}(p,q).$$

Choosing some fixed $\hat{Z} \in \mathrm{O}(p,q)$ this optimization problem is equivalent to

$$\|Z\|_W^2 = \min \quad \text{subject to} \quad Z \in \mathrm{O}(p,q),\ \|Z\|_W \le \|\hat{Z}\|_W,$$

where now the solution set of the constraints is not only closed but also bounded and thus compact. The claim follows then from the solvability of the best approximation problem with respect to compact sets, see, e.g., [2]. ∎

In order to solve (4.5) we first need to treat the constraint. For this, we look for a suitable parametrization of $\mathrm{O}(p,q)$. The quantities of the form $A^{1/2}$ used there denote the square root of the symmetric positive definite matrix A, i.e., the unique symmetric positive definite matrix $A^{1/2}$ with the property $A^{1/2}A^{1/2} = A$. Furthermore, the inverse of $A^{1/2}$ is written as $A^{-1/2}$.

LEMMA 4.2. *For every $A \in \mathbb{R}^{q,p}$, the relation*

$$(I_p + A^TA)^{-1/2}A^T(I_q + AA^T)^{1/2} = A^T \tag{4.6}$$

holds.

Proof. Let

$$U^TAV = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r),$$

with U, V orthogonal, $\sigma_1 \ge \cdots \ge \sigma_r > 0$, $r = \operatorname{rank} A$, be a singular value decomposition of A. Then

$$\begin{aligned}
&(I_p + A^TA)^{-1/2}A^T(I_q + AA^T)^{1/2} \\
&= \left(V \begin{bmatrix} I + \Sigma^2 & 0 \\ 0 & I \end{bmatrix} V^T\right)^{-1/2} V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^T \left(U \begin{bmatrix} I + \Sigma^2 & 0 \\ 0 & I \end{bmatrix} U^T\right)^{1/2} \\
&= V \begin{bmatrix} (I + \Sigma^2)^{-1/2}\Sigma(I + \Sigma^2)^{1/2} & 0 \\ 0 & 0 \end{bmatrix} U^T = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^T = A^T. \qquad \blacksquare
\end{aligned}$$

THEOREM 4.3. *The Lie group $\mathrm{O}(p,q)$ can be parametrized according to*

$$\mathrm{O}(p,q) = \left\{ \left[\begin{matrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{matrix} \right] \middle| \begin{array}{l} Z_{21} \in \mathbb{R}^{q,p} \text{ arbitrary} \\ Z_{11} = Q_{11}(I_p + Z_{21}^TZ_{21})^{1/2},\ Q_{11} \in \mathrm{O}(p), \\ Z_{22} = (I_q + Z_{21}Z_{21}^T)^{1/2}Q_{22},\ Q_{22} \in \mathrm{O}(q), \\ Z_{12} = Z_{11}^{-T}Z_{21}^TZ_{22} = Q_{11}Z_{21}^TQ_{22} \end{array} \right\}. \tag{4.7}$$

Proof. The property $Z^TSZ = S$ reads

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}^T \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix},$$

or

$$Z_{11}^TZ_{11} - Z_{21}^TZ_{21} = I_p, \quad Z_{11}^TZ_{12} - Z_{21}^TZ_{22} = 0, \quad Z_{12}^TZ_{12} - Z_{22}^TZ_{22} = -I_q.$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

Let $Z \in \mathrm{O}(p,q)$. Then

$$Z_{11}^T Z_{11} = I_p + Z_{21}^T Z_{21} = (I_p + Z_{21}^T Z_{21})^{1/2}(I_p + Z_{21}^T Z_{21})^{1/2},$$

or

$$(I_p + Z_{21}^T Z_{21})^{-1/2} Z_{11}^T Z_{11} (I_p + Z_{21}^T Z_{21})^{-1/2} = I_p,$$

and hence,

$$Q_{11} = Z_{11}(I_p + Z_{21}^T Z_{21})^{-1/2} \in \mathrm{O}(p).$$

Moreover,

$$Z_{12} = Z_{11}^{-T} Z_{21}^T Z_{22}.$$

Furthermore,

$$\begin{aligned} Z_{22}^T Z_{22} &= I_q + Z_{12}^T Z_{12} = I_q + Z_{22}^T Z_{21} Z_{11}^{-1} Z_{11}^{-T} Z_{21}^T Z_{22} \\ &= I_q + Z_{22}^T Z_{21}(I_p + Z_{21}^T Z_{21})^{-1} Z_{21}^T Z_{22}, \end{aligned}$$

or

$$Z_{22}^T (I_q - Z_{21}(I_p + Z_{21}^T Z_{21})^{-1} Z_{21}^T) Z_{22} = I_q.$$

Using the Sherman-Morrison formula, see, e.g., [5], this is the same as

$$Z_{22}^T (I_q + Z_{21} Z_{21}^T)^{-1} Z_{22} = I_q,$$

or

$$Z_{22}^T (I_q + Z_{21} Z_{21}^T)^{-1/2} (I_q + Z_{21} Z_{21}^T)^{-1/2} Z_{22} = I_q,$$

and hence,

$$Q_{22} = (I_q + Z_{21} Z_{21}^T)^{-1/2} Z_{22} \in \mathrm{O}(q).$$

Finally, we have

$$Z_{11}^{-T} Z_{21}^T Z_{22} = Q_{11}(I_p + Z_{21}^T Z_{21})^{-1/2} Z_{21}^T (I_q + Z_{21} Z_{21}^T)^{1/2} Q_{22} = Q_{11} Z_{21}^T Q_{22}$$

due to Lemma 4.2.

Conversely, let $Z \in \mathbb{R}^{n,n}$ be in the set of the right-hand side of (4.7). Then

$$I_p + Z_{21}^T Z_{21} - Z_{11}^T Z_{11} = I_p + Z_{21}^T Z_{21} - (I_p + Z_{21}^T Z_{21})^{1/2} Q_{11}^T Q_{11} (I_p + Z_{21}^T Z_{21})^{1/2} = 0$$

and

$$Z_{11}^T Z_{12} - Z_{21}^T Z_{22} = Z_{11}^T Z_{11}^{-T} Z_{21}^T Z_{22} - Z_{21}^T Z_{22} = 0$$

as well as

$$\begin{aligned} &I_q + Z_{12}^T Z_{12} - Z_{22}^T Z_{22} \\ &\quad = I_q + Z_{22}^T Z_{21} Z_{11}^{-1} Z_{11}^{-T} Z_{21}^T Z_{22} - Z_{22}^T Z_{22} \\ &\quad = I_q + Z_{22}^T Z_{21}(I_p + Z_{21}^T Z_{21})^{-1} Z_{21}^T Z_{22} - Z_{22}^T Z_{22} \\ &\quad = I_q - Z_{22}^T (I_q - Z_{21}(I_p + Z_{21}^T Z_{21})^{-1} Z_{21}^T) Z_{22} \\ &\quad = I_q - Z_{22}^T (I_q + Z_{21} Z_{21}^T)^{-1} Z_{22} \\ &\quad = I_q - Q_{22}^T (I_q + Z_{21} Z_{21}^T)^{1/2} (I_q + Z_{21} Z_{21}^T)^{-1} (I_q + Z_{21} Z_{21}^T)^{1/2} Q_{22} = 0 \end{aligned}$$

implying that $Z \in \mathrm{O}(p,q)$. ∎

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

An immediate consequence of the parametrization (4.7) is that $\mathrm{O}(p,q)$ is one-to-one to the product $\mathrm{O}(p) \times \mathbb{R}^{q,p} \times \mathrm{O}(q)$ reflecting the property

$$\dim \mathrm{O}(p,q) = \dim \mathrm{O}(p) + \dim \mathbb{R}^{q,p} + \dim \mathrm{O}(q) = \tfrac{1}{2}p(p+1) + pq + \tfrac{1}{2}q(q+1) = \tfrac{1}{2}n(n+1).$$

This can also formulated in the following way.

COROLLARY 4.4. *Every $Z \in \mathrm{O}(p,q)$ can be factorized according to*

$$Z = \begin{bmatrix} Q_{11} & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} (I_p + Z_{21}^T Z_{21})^{1/2} & Z_{21}^T \\ Z_{21} & (I_q + Z_{21} Z_{21}^T)^{1/2} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & Q_{22} \end{bmatrix}.$$

On the basis of the preceding structural properties of $\mathrm{O}(p,q)$ we will now discuss two possibilities to stabilize a given smooth factorization (4.4).

4.1. Stabilization by a descent method. Turning back to the optimization problem (4.5) we observe that the Frobenius norm is invariant under orthogonal transformations. Hence, we concentrate on matrices $Z \in \mathrm{O}(p,q)$ for which $Q_{11} = I_p$ and $Q_{22} = I_q$. In the spirit of Jacobi's method for the determination of all eigenvalues of a symmetric matrix, see, e.g., [5], we construct an iterative procedure by choosing the remaining parameter Z_{21} as zero matrix with the exception of a parameter $\sigma \in \mathbb{R}$ at the position (l,k) of the resulting Z with suitably chosen $k \in \{1,\ldots,q\}$ and $l \in \{q+1,\ldots,n\}$. Denoting $W = [\, w_1 \ \cdots \ w_n \,]$, building WZ for such a special Z only alters the two columns given by k and l. For convenience, we therefore use the shorthand notation $x = w_k$ and $y = w_l$. Note that $x \neq 0$ and $y \neq 0$ due to the nonsingularity of W. Application of the special Z then reads

$$[\, x \ y \,] \begin{bmatrix} \sqrt{1+\sigma^2} & \sigma \\ \sigma & \sqrt{1+\sigma^2} \end{bmatrix} = \begin{bmatrix} \sqrt{1+\sigma^2}x + \sigma y & \sigma x + \sqrt{1+\sigma^2} y \end{bmatrix}.$$

Since the squared Frobenius norm is the sum of the squared Euclidean norms of the columns, the optimization problem (4.5) reduces to

$$\left\| \sqrt{1+\sigma^2}x + \sigma y \right\|_2^2 + \left\| \sigma x + \sqrt{1+\sigma^2} y \right\|_2^2 = \min \tag{4.8}$$

and thus to

$$(1+2\sigma^2)(\|x\|_2^2 + \|y\|_2^2) + 4\sigma\sqrt{1+\sigma^2}\, x^T y = \min.$$

Omitting the additive constant, scaling appropriately, and setting

$$c = \frac{2x^T y}{\|x\|_2^2 + \|y\|_2^2},$$

we finally arrive at the problem

$$f(\sigma) = \min, \quad f(\sigma) = \sigma^2 + c\,\sigma\sqrt{1+\sigma^2}. \tag{4.9}$$

Observing

$$f'(\sigma) = 2\sigma + c\left(\sqrt{1+\sigma^2} + \sigma^2/\sqrt{1+\sigma^2}\right),$$

the requirement $f'(\sigma) = 0$ is equivalent to

$$2\sigma\sqrt{1+\sigma^2} + c\,(1+2\sigma^2) = 0, \tag{4.10}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

or

$$f(\sigma) = -c, \quad f(\sigma) = \frac{2\sigma\sqrt{1+\sigma^2}}{1+2\sigma^2}. \tag{4.11}$$

A short computation shows that f is monotone increasing with

$$\lim_{\sigma\to\pm\infty} f(\sigma) = \pm 1.$$

If $b = 0$ corresponding to $x^T y = 0$, the solution of (4.9) is given by $\sigma = 0$. If $b \neq 0$ corresponding to $x^T y \neq 0$, we observe that x and y are linearly independent due to the nonsingularity of W implying that

$$(x \pm y)^T(x \pm y) > 0 \iff \|x\|_2^2 \pm 2x^T y + \|y\|_2^2 > 0 \iff -1 < \frac{2x^T y}{\|x\|_2^2 + \|y\|_2^2} < 1.$$

Hence,

$$|c| = \frac{2|x^T y|}{\|x\|_2^2 + \|y\|_2^2} < 1$$

and (4.11) possesses a unique solution $\sigma \in \mathbb{R}$.

The above discussion leads to the following algorithm for the solution of (4.5).

ALGORITHM 4.5. *Starting with $C^{(0)} = W$ we define $C^{(\nu+1)}$ for given $C^{(\nu)}$ by choosing $k_\nu \in \{1, \ldots, q\}$ and $l_\nu \in \{q+1, \ldots, n\}$ such that $x^{(\nu)T} y^{(\nu)} \neq 0$ for $x^{(\nu)}$ being the k_ν-th column and $y^{(\nu)}$ being the l_ν-th column of $C^{(\nu)}$ and solving the corresponding scalar nonlinear problem (4.8). With the resulting special transformation denoted here by $Z^{(\nu)}$ we then set $C^{(\nu+1)} = C^{(\nu)} Z^{(\nu)}$.*

The question now is whether the sequence $\{C^{(\nu)}\}_{\nu\in\mathbb{N}_0}$ defined by Algorithm 4.5 converges. We start the examination of Algorithm 4.5 by looking at its descent property.

LEMMA 4.6. *Let*

$$\hat{x} = \sqrt{1+\sigma^2}x + \sigma y, \quad \hat{y} = \sigma x + \sqrt{1+\sigma^2}y$$

with σ satisfying

$$\sigma\sqrt{1+\sigma^2}(\|x\|_2^2 + \|y\|_2^2) + (1+2\sigma^2)x^T y = 0$$

according to (4.10). Then

$$\|\hat{x}\|_2^2 + \|\hat{y}\|_2^2 = \frac{1}{1+2\sigma^2}(\|x\|_2^2 + \|y\|_2^2).$$

Proof. We have

$$\begin{aligned}
\|\hat{x}\|_2^2 + \|\hat{y}\|_2^2 &= (1+2\sigma^2)(\|x\|_2^2 + \|y\|_2^2) + 4\sigma\sqrt{1+\sigma^2}x^T y \\
&= (1+2\sigma^2)(\|x\|_2^2 + \|y\|_2^2) - 4\sigma\sqrt{1+\sigma^2}\frac{\sigma\sqrt{1+\sigma^2}}{1+2\sigma^2}(\|x\|_2^2 + \|y\|_2^2) \\
&= \left[(1+2\sigma^2) - \frac{4\sigma^2(1+\sigma^2)}{1+2\sigma^2}\right](\|x\|_2^2 + \|y\|_2^2) \\
&= \frac{1}{1+2\sigma^2}(\|x\|_2^2 + \|y\|_2^2). \qquad \square
\end{aligned}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

LEMMA 4.7. *Let $\hat{W} \in \mathbb{R}^{n,n}$ be the result of one step of Algorithm 4.5 applied to the nonsingular matrix $W \in \mathbb{R}^{n,n}$. Then*

$$\|\hat{W}\|_F^2 - \|W\|_F^2 = -\frac{2\sigma^2}{1+2\sigma^2}(\|x\|_2^2 + \|y\|_2^2) \tag{4.12}$$

in the notation of (4.10).

Proof. Let $W = [\, w_1 \ \cdots \ w_n \,]$ and $\hat{W} = [\, \hat{w}_1 \ \cdots \ \hat{w}_n \,]$. Algorithm 4.5 yields

$$\hat{w}_i = w_i \quad \text{for} \quad i \neq k, l$$

and

$$\hat{w}_k = \sqrt{1+\sigma^2} w_k + \sigma w_l, \quad \hat{w}_l = \sigma w_k + \sqrt{1+\sigma^2} w_l.$$

According to Lemma 4.6, we then have

$$\|\hat{w}_k\|_2^2 + \|\hat{w}_l\|_2^2 = \frac{1}{1+2\sigma^2}(\|w_k\|_2^2 + \|w_l\|_2^2)$$

implying

$$\begin{aligned}
\|\hat{W}\|_F^2 - \|W\|_F^2 &= \sum_{i=1}^{n} \|\hat{w}_i\|_2^2 - \sum_{i=1}^{n} \|w_i\|_2^2 \\
&= (\|\hat{w}_k\|_2^2 + \|\hat{w}_l\|_2^2) - (\|w_k\|_2^2 + \|w_l\|_2^2) \\
&= \left(\frac{1}{1+2\sigma^2} - 1 \right) (\|w_k\|_2^2 + \|w_l\|_2^2) \\
&= -\frac{2\sigma^2}{1+2\sigma_2}(\|x\|_2^2 + \|y\|_2^2). \qquad \square
\end{aligned}$$

A first consequence of Algorithm 4.5 is then that we can characterize optimal solutions by a suitable algebraic property.

THEOREM 4.8. *Let $Z \in \mathrm{O}(p,q)$ be a solution of (4.5) and let $C = WZ$ with $C = [\, C_1 \ C_2 \,]$ split according to the block structure of Z in (4.7). Then*

$$C_2^T C_1 = 0 \tag{4.13}$$

and the solution set of (4.5) is given by

$$\mathbb{L} = \left\{ Y \in \mathrm{O}(p,q) \,\middle|\, Y = Z \begin{bmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{bmatrix}, \ Q_{11}, Q_{22} \text{ orthogonal} \right\}. \tag{4.14}$$

Proof. Applying Algorithm 4.5 to $C = WZ$ the choice of k and l selects a column x from C_1 and a column y from C_2. If $C_2^T C_1 \neq 0$, we can therefore choose x and y such that $x^T y \neq 0$ implying $\sigma \neq 0$. In view of (4.12) the given Z then cannot be optimal. Hence, (4.13) holds.

Let Z_1, Z_2 be two solutions of (4.5) with $Z_1^T W^T W Z_1$ and $Z_2^T W^T W Z_2$ being blockdiagonal according to (4.13). Writing $C = WZ_1$ and $\hat{C} = WZ_2$, we have

$$\hat{C} = CZ, \quad Z = Z_1^{-1} Z_2, \quad Z^T S Z = S$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

with

$$C = [\, C_1 \;\; C_2 \,], \quad C_1^T C_2 = 0, \quad \hat{C} = [\, \hat{C}_1 \;\; \hat{C}_2 \,], \quad \hat{C}_1^T \hat{C}_2 = 0.$$

Using (4.7) we get

$$\hat{C}_1 = C_1 Z_{11} + C_2 Z_{21}, \quad \hat{C}_2 = C_1 Z_{12} + C_2 Z_{22},$$

and thus,

$$\begin{aligned} 0 &= (C_1 Z_{12} + C_2 Z_{22})^T (C_1 Z_{11} + C_2 Z_{21}) \\ &= Z_{12}^T C_1^T C_1 Z_{11} + Z_{22}^T C_2^T C_2 Z_{21} \\ &= Z_{22}^T Z_{21} Z_{11}^{-1} C_1^T C_1 Z_{11} + Z_{22}^T C_2^T C_2 Z_{21}. \end{aligned}$$

Since Z_{22} is nonsingular, this is equivalent to the (homogeneous) Sylvester equation

$$Z_{21}(Z_{11}^{-1} C_1^T C_1 Z_{11}) + (C_2^T C_2) Z_{21} = 0.$$

Observing that the matrices in parantheses are both symmetric positive definite due to the nonsingularity of C and Z_{11}, the only possible solution is given by $Z_{21} = 0$ implying $Z_{12} = 0$ and Z_{11}, Z_{22} being orthogonal. Hence, Z is orthogonal, and therefore,

$$\|WZ_2\|_F = \|WZ_1 Z_1^{-1} Z_2\|_F = \|WZ_1 Z\|_F = \|WZ_1\|_F. \qquad \square$$

Before we can actually show convergence we need an estimate which guarantees a lower bound for the descent away from the solution.

LEMMA 4.9. *Given the compact set*

$$\mathbb{D} = \{C = WZ \mid Z \in O(p,q),\ \|C\|_F \le \|W\|_F,\ W \text{ nonsingular}\},$$

then there exists a $\delta > 0$ so that

$$\min_{\substack{C \in \mathbb{D} \\ i=1,\ldots,n}} \|Ce_i\|_2 \ge \delta, \tag{4.15}$$

with e_i, $i = 1, \ldots, n$, being the canonical basis vectors in $\mathbb{R}^n$.

Proof. Take a fixed $i = 1, \ldots, n$ and consider

$$\min_{C \in \mathbb{D}} \|Ce_i\|_2.$$

The minimum, say δ_i, exists since we have a continuous function on a compact domain and is nonnegative since the function is nonnegative. Assume that $\delta_i = 0$. Then there is a $C \in \mathbb{D}$ with $\|Ce_i\|_2 = 0$, hence $Ce_i = 0$ implying that C is singular in contradiction to $C = WZ$ with $Z \in O(p,q)$ and W nonsingular. Thus, $\delta_i > 0$ and we can choose $\delta = \min\{\delta_1, \ldots, \delta_n\} > 0$. $\square$

THEOREM 4.10. *Let the sequence $C_{\nu \in \mathbb{N}_0}^{(\nu)}$ be generated by Algorithm 4.5 with intermediate quantities $x^{(\nu)}$, $y^{(\nu)}$, and σ_ν. Furthermore, let $\hat{Z} \in \mathbb{L}$ and $\hat{C} = W\hat{Z}$. Then $C^{(\nu)} \in \mathbb{D}$ for all $\nu \in \mathbb{N}_0$ satisfying*

$$\|\hat{C}\|_F \le \|C^{(\nu+1)}\|_F \le \|C^{(\nu)}\|_F \le \|W\|_F. \tag{4.16}$$

Moreover,

$$\sum_{\nu=0}^{\infty} \frac{4\sigma_\nu^2}{1 + 2\sigma_\nu^2} \le \frac{\|W\|_F - \|\hat{C}\|_F}{\delta^2} \tag{4.17}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

and

$$\lim_{\nu\to\infty}\sigma_\nu = 0, \quad \lim_{\nu\to\infty} x^{(\nu)T}y^{(\nu)} = 0. \tag{4.18}$$

Proof. The claim (4.16) follows by induction directly from (4.12) and from $\hat{Z}$ being an optimal solution of (4.5). In more detail we have

$$\|C^{(\nu+1)}\|_F^2 - \|C^{(\nu)}\|_F^2 = -\frac{2\sigma_\nu^2}{1+2\sigma_\nu^2}(\|x^{(\nu)}\|_2^2 + \|y^{(\nu)}\|_2^2) \le -\frac{4\sigma_\nu^2}{1+2\sigma_\nu^2}\delta^2,$$

where we utilized (4.15). Summing up then yields

$$\sum_{\nu=0}^{\mu}\frac{4\sigma_\nu^2}{1+2\sigma_\nu^2} \le \frac{\|C^{(0)}\|_F^2 - \|C^{(\mu+1)}\|_F^2}{\delta^2} \le \frac{\|C^{(0)}\|_F^2 - \|\hat{C}\|_F^2}{\delta^2},$$

and the limit of the sum for $\mu\to\infty$ exists and is bounded as claimed in (4.17). It immediately follows that

$$\lim_{\nu\to\infty}\frac{4\sigma_\nu^2}{1+2\sigma_\nu^2} = 0,$$

which is equivalent to the first part of (4.18). For the second part, we observe that

$$f(\sigma_\nu) = \frac{x^{(\nu)T}y^{(\nu)}}{\|x^{(\nu)}\|_2^2 + \|y^{(\nu)}\|_2^2}$$

according to (4.11). Since $|f(\sigma_\nu)| \le |\sigma_\nu|$ due to a simple calculation, we finally obtain

$$|\sigma_\nu| \ge \frac{|x^{(\nu)T}y^{(\nu)}|}{2\|C^{(\nu)}\|_F} \ge \frac{|x^{(\nu)T}y^{(\nu)}|}{2\|W\|_F}$$

showing the second part of (4.18). ∎

The preceeding theorem shows that we approach the set of optimal solutions of the optimization problem (4.5) as long as (4.18) implies

$$\lim_{\nu\to\infty} C_2^{(\nu)T}C_1^{(\nu)} = 0, \quad C^{(\nu)} = \begin{bmatrix} C_1^{(\nu)} & C_2^{(\nu)} \end{bmatrix}.$$

Actually this is a property of the pivot strategy, that is, of the choice of k_ν, l_ν in Algorithm 4.5. One possibility is to choose k_ν, l_ν in such a way that

$$\left|x^{(\nu)T}y^{(\nu)}\right| = \left\|\operatorname{vec}(C_2^{(\nu)T}C_1^{(\nu)})\right\|_\infty$$

holds, with vec stacking all entries of a matrix into a vector. Compare this with the classical form of Jacobi's method, see [5]. Another possibility is to guarantee that (k_ν, l_ν) equals (k,l) for each possible pair with $k = 1,\dots,p$, $l = p+1,\dots,n$ for infinitely many $\nu\in\mathbb{N}_0$. In this case, it is also allowed that some of the σ_ν vanish.

In order to get a smooth process that stabilizes the approach of Section 3, we can take a fixed sequence of pairs (k,l) so that every possiblity actually occurs, for example

$$(1,q+1),\dots,(1,n),(2,q+1),\dots,(2,n),\dots,(q,q+1),\dots,(q,n) \tag{4.19}$$

and repeat it a certain number of times. Compare this with the so-called cyclic form of Jacobi's method, see again [5]. Since all steps in the procedure are analytic and only a finite number of steps are performed the overall process is analytic maintaining the smoothness of the original matrix function W.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

4.2. Stabilization by Newton's method. In view of Theorem 4.8 we can replace the optimization problem (4.5) by the solution of a system of nonlinear equations suggested by (4.13). In order to fix a unique solution we are looking for a solution of a special structure.

LEMMA 4.11. *The solution set* $\mathbb{L}$ *contains a matrix* Z *of the form*

$$Z = \begin{bmatrix} (I_p + Z_{21}^T Z_{21})^{1/2} & Z_{21}^T \\ Z_{21} & (I_q + Z_{21} Z_{21}^T)^{1/2} \end{bmatrix}.$$

Proof. Let $\tilde{Z} \in \mathbb{L}$. Due to Theorem 4.3, it has the form

$$\tilde{Z} = \begin{bmatrix} Q_{11}(I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} & Q_{11} \tilde{Z}_{21}^T Q_{22} \\ \tilde{Z}_{21} & (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{1/2} Q_{22} \end{bmatrix},$$

which can be written as

$$\tilde{Z} = \begin{bmatrix} (I_p + Z_{21}^T \tilde{Z}_{21})^{1/2} & Z_{21}^T \\ Z_{21} & (I_q + Z_{21} Z_{21}^T)^{1/2} \end{bmatrix} \begin{bmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{bmatrix}$$

with $Z_{21} = \tilde{Z}_{21} Q_{11}^T$. The claim follows from the structure of $\mathbb{L}$ according to (4.14). ∎

Thus, we are left with the determination of a suitable matrix $Z_{21} \in \mathbb{R}^{q,p}$. We therefore consider the system of nonlinear equations

$$F(Z_{21}) = 0, \tag{4.20}$$

with $F : \mathbb{R}^{q,p} \to \mathbb{R}^{q,p}$ defined by

$$\begin{aligned} &\text{(a)} \quad F(Z_{21}) = C_2^T C_1, \\ &\text{(b)} \quad C = [\, C_1 \;\; C_2 \,] = WZ, \\ &\text{(c)} \quad Z = \begin{bmatrix} (I_p + Z_{21}^T Z_{21})^{1/2} & Z_{21}^T \\ Z_{21} & (I_q + Z_{21} Z_{21}^T)^{1/2} \end{bmatrix}. \end{aligned} \tag{4.21}$$

Let Z_{21}^* be a solution of (4.20) with Z^* according to (4.21c) and $W^* = WZ^*$. Then (4.20) is equivalent to

$$\tilde{F}(\tilde{Z}_{21}) = 0 \tag{4.22}$$

with $\tilde{F} : \mathbb{R}^{q,p} \to \mathbb{R}^{q,p}$ defined by

$$\begin{aligned} &\text{(a)} \quad \tilde{F}(\tilde{Z}_{21}) = \tilde{C}_2^T \tilde{C}_1, \\ &\text{(b)} \quad \tilde{C} = [\, \tilde{C}_1 \;\; \tilde{C}_2 \,] = W^* \tilde{Z}, \\ &\text{(c)} \quad \tilde{Z} = \begin{bmatrix} (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} & \tilde{Z}_{21}^T \\ \tilde{Z}_{21} & (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{1/2} \end{bmatrix}. \end{aligned}$$

In particular, (4.22) possesses the solution $\tilde{Z}_{21}^* = 0$. Without loss of generality, we are thus allowed to assume that $Z_{21}^* = 0$ is a solution of (4.20).

In the following we use the notion of differentials replacing the more familiar d by Δ thus mimicking the notation known in the context of Newton's method.

In order to show that $Z_{21}^* = 0$ is a regular solution we need to show that the Newton correction ΔZ_{21} defined by

$$\Delta F(Z_{21}) = F'(Z_{21}) \Delta Z_{21} = -F(Z_{21}) \tag{4.23}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

vanishes at the solution Z_{21}^*, i.e., for vanishing right-hand side in (4.23). Note that $A^{1/2}A^{1/2} = A$ implies $\Delta A^{1/2}A^{1/2} + A^{1/2}\Delta A^{1/2} = \Delta A$, which is a uniquely solvable Sylvester equation for $\Delta A^{1/2}$. Taking $A = I_p + Z_{21}^T Z_{21}$, we get $\Delta A = \Delta Z_{21}^T Z_{21} + Z_{21}^T \Delta Z_{21}$. Hence, $A = I_p$ and $\Delta A = 0$ for $Z_{21} = 0$ resulting in $\Delta(I_p + Z_{21}^T Z_{21})^{1/2} = 0$ at the solution. Similarly, we have $\Delta(I_q + Z_{21} Z_{21}^T)^{1/2} = 0$ at the solution. Thus, the Newton correction at the solution is given by

$$\begin{aligned} &\text{(a)} && \Delta C_2^T C_1 + C_2^T \Delta C_1 = 0, \\ &\text{(b)} && \Delta C = [\, \Delta C_1 \;\; \Delta C_2 \,] = W \Delta Z, \\ &\text{(c)} && \Delta Z = \begin{bmatrix} 0 & \Delta Z_{21}^T \\ \Delta Z_{21} & 0 \end{bmatrix}. \end{aligned}$$

Writing $W = [\, W_1 \;\; W_2 \,]$ yields $\Delta C = [\, W_2 \Delta Z_{21} \;\; W_1 \Delta Z_{21}^T \,]$. Observing that $Z = I$ for $Z_{21} = 0$ and thus $C = W$, we finally arrive at

$$\Delta Z_{21} W_1^T W_1 + W_2^T W_2 \Delta Z_{21} = 0. \tag{4.24}$$

Since W is nonsingular, the matrices $W_1^T W_1$ and $W_2^T W_2$ are symmetric positive definite so that the Sylvester equation (4.24) possesses the unique solution $\Delta Z_{21} = 0$.

THEOREM 4.12. *The problem (4.20) possesses a unique solution which is regular in the sense that the linearization at the solution is invertible.*

Proof. Regularity was shown by the preceeding discussion. Uniqueness follows from (4.14) with Z from Lemma 4.11 since we require $Q_{11} = I_p$ and $Q_{22} = I_q$ in (4.14). ∎

Taking the dependence on W into account, we write (4.20) as

$$F(Z_{21}, W) = 0.$$

Due to the implicit function theorem, this can locally be solved for Z_{21} according to

$$Z_{21} = G(W)$$

with a function $G : \mathrm{GL}(n) \to \mathbb{R}^{q,p}$. Since F is analytic, also G is analytic. In the case of a path $W \in C^k(\mathbb{I}, \mathrm{GL}(n))$, as in Theorem 3.1, we get a path $Z \in C^k(\mathbb{I}, \mathrm{O}(p,q))$ stabilizing the approach of Section 3.

As a by-product of the numerical considerations we have shown the following stronger version for Theorem 3.1.

THEOREM 4.13. *Let $E \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$ with $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$ have constant rank and let $E^T = E$. Then there exists a matrix function $W \in C^k(\mathbb{I}, \mathrm{GL}(n))$ satisfying (3.2) and being pointwise of minimal norm in the sense of (4.5).*

4.3. Alternative parametrization. We actually based the approach of Section 4.2 on a parametrization of $\mathrm{O}(p,q)$ slightly different from (4.7).

THEOREM 4.14. *The Lie group $\mathrm{O}(p,q)$ can be parametrized according to*

$$\mathrm{O}(p,q) = \left\{ \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \middle| \begin{array}{l} Z_{21} \in \mathbb{R}^{q,p} \text{ arbitrary} \\ Z_{11} = Q_{11} L_{11}^T, \;\; Q_{11} \in \mathrm{O}(p), \\ Z_{22} = L_{22} Q_{22}, \;\; Q_{22} \in \mathrm{O}(q), \\ Z_{12} = Z_{11}^{-T} Z_{21}^T Z_{22} \end{array} \right\}, \tag{4.25}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

where

$$L_{11}L_{11}^T = I_p + Z_{21}^T Z_{21}, \quad L_{22}L_{22}^T = I_q + Z_{21}Z_{21}^T \tag{4.26}$$

by means of the Cholesky factorization.

Proof. The proof follows the lines of those for Theorem 4.3. Let $Z \in \mathrm{O}(p,q)$ according to (4.25) with block Z_{21}. Then

$$Z_{11}^T Z_{11} = I_p + Z_{21}^T Z_{21} = L_{11}L_{11}^T,$$

or

$$L_{11}^{-1} Z_{11}^T Z_{11} L_{11}^{-T} = I_p,$$

and hence,

$$Q_{11} = Z_{11}L_{11}^{-T} \in \mathrm{O}(p).$$

Moreover,

$$Z_{12} = Z_{11}^{-T} Z_{21}^T Z_{22}.$$

Furthermore,

$$Z_{22}^T Z_{22} = I_q + Z_{12}^T Z_{12} = I_q + Z_{22}^T Z_{21} Z_{11}^{-1} Z_{11}^{-T} Z_{21}^T Z_{22} = I_q + Z_{22}^T Z_{21}(I_p + Z_{21}^T Z_{21})^{-1} Z_{21}^T Z_{22},$$

or

$$Z_{22}^T (I_q - Z_{21}(I_p + Z_{21}^T Z_{21})^{-1} Z_{21}^T) Z_{22} = I_q.$$

Using the Sherman-Morrison formula, this is the same as

$$Z_{22}^T (I_q + Z_{21}Z_{21}^T)^{-1} Z_{22} = I_q,$$

or

$$Z_{22}^T (L_{22}L_{22}^T)^{-1} Z_{22} = I_q,$$

and hence,

$$Q_{22} = L_{22}^{-1} Z_{22} \in \mathrm{O}(q).$$

Conversely, let $Z \in \mathbb{R}^{n,n}$ be in the set of the right-hand side of (4.25). Then

$$I_p + Z_{21}^T Z_{21} - Z_{11}^T Z_{11} = I_p + Z_{21}^T Z_{21} - L_{11}Q_{11}^T Q_{11} L_{11}^T = 0$$

as well as

$$\begin{aligned} I_q + Z_{12}^T Z_{12} - Z_{22}^T Z_{22} &= I_q + Z_{22}^T Z_{21} Z_{11}^{-1} Z_{11}^{-T} Z_{21}^T Z_{22} - Z_{22}^T Z_{22} \\ &= I_q - Z_{22}^T (I_q - Z_{21}(I_p + Z_{21}Z_{21}^T)^{-1} Z_{21}^T) Z_{22} \\ &= I_q - Z_{22}^T (I_q + Z_{21}^T Z_{21})^{-1} Z_{22} \\ &= I_q - Z_{22}^T L_{22}^T (L_{22}L_{22}^T)^{-1} L_{22} Q_{22} = 0, \end{aligned}$$

implying that $Z \in \mathrm{O}(p,q)$. ∎

All results of Section 4.2 carry over to the parametrization (4.25). While the parametrization (4.7) has the advantage to exhibit more symmetry leading to a simpler mathematical discussion, the parametrization (4.25) has the advantage to use Cholesky factorization which is a finite process instead of the determination of the square root of matrices which requires an iterative process.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

LEMMA 4.15. *The solution set* $\mathbb{L}$ *contains a matrix* Z *of the form*

$$Z = \begin{bmatrix} L_{11}^T & L_{11}^{-1} Z_{21}^T L_{22} \\ Z_{21} & L_{22} \end{bmatrix}$$

with L_{11}, L_{22} *according to (4.26).*

Proof. Due to Lemma 4.11, the solution set $\mathbb{L}$ contains a matrix $\tilde{Z}$ of the form

$$\tilde{Z} = \begin{bmatrix} (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} & \tilde{Z}_{21}^T \\ \tilde{Z}_{21} & (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{1/2} \end{bmatrix}.$$

Let L_{11}, L_{22} nonsingular lower triangular and Q_{11}, Q_{22} orthogonal be given by QR factorization according to

$$L_{11}^T Q_{11} = (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2}, \quad L_{22} Q_{22} = (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{1/2}$$

and set $Z_{21} = \tilde{Z}_{21} Q_{11}^T$. Then

$$\tilde{Z} = \begin{bmatrix} L_{11}^T Q_{11} & Q_{11}^T Z_{21}^T \\ Z_{21} Q_{11} & L_{22} Q_{22} \end{bmatrix} = \begin{bmatrix} L_{11}^T & Q_{11}^T Z_{21}^T Q_{22}^T \\ Z_{21} & L_{22} \end{bmatrix} \begin{bmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{bmatrix}$$

satisfying

$$\begin{aligned} L_{11} L_{11}^T &= Q_{11} (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} Q_{11}^T \\ &= Q_{11} (I_p + Q_{11}^T Z_{21}^T Z_{21} Q_{11}) Q_{11} = I_p + Z_{21}^T Z_{21} \end{aligned}$$

and

$$\begin{aligned} L_{22} L_{22}^T &= (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{1/2} Q_{22}^T Q_{22} (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{1/2} \\ &= I_q + Z_{21} Q_{11} Q_{11}^T Z_{21}^T = I_q + Z_{21} Z_{21}^T. \end{aligned}$$

It remains to show that $Q_{11}^T Z_{21}^T Q_{22}^T = L_{11}^{-1} Z_{21}^T L_{22}$ which follows from

$$\begin{aligned} &L_{11} Q_{11}^T Z_{21}^T Q_{22}^T L_{22}^{-1} \\ &\quad = Q_{11} (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} Q_{11}^T Q_{11} \tilde{Z}_{21}^T Q_{22}^T Q_{22} (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{-1/2} \\ &\quad = Q_{11} (I_p + \tilde{Z}_{21}^T \tilde{Z}_{21})^{1/2} \tilde{Z}_{21}^T (I_q + \tilde{Z}_{21} \tilde{Z}_{21}^T)^{-1/2} = Q_{11} \tilde{Z}_{21}^T = Z_{21}^T \end{aligned}$$

with the help of (4.6). ∎

One can also think of other parametrizations. Using the Cayley transform parametrizations of the Lie group $\mathrm{O}(p, q)$ correspond to parametrizations of its Lie algebra and vice versa. Thus, we can also think of parametrizations of the Lie algebra and transfer them to $\mathrm{O}(p, q)$. At this point, it is not clear whether there is an optimal parametrization with respect to efficiency and robustness of the representation of the elements in $\mathrm{O}(p, q)$ and in the subset needed for Newton's method and of Newton's method itself.

5. Numerical illustration. We implemented the constructions for a smooth version of Sylvester's law of inertia from the proof of Theorem 3.1 as well as both approaches for their stabilization as described in Section 4.1 and in Section 4.2, the latter together with the alternative parametrization described in Section 4.3.

EXAMPLE 5.1. Figure 2 shows the Frobenius norm of W for E given in Example 3.3 as function of t obtained by stabilization in comparison with the original norms as shown in Figure 1. Again `ODE` refers to the case $k \neq 0$ and `GLUE` refers to the case $k = 0$ while `STABLE` refers to the stabilization of the construction in

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

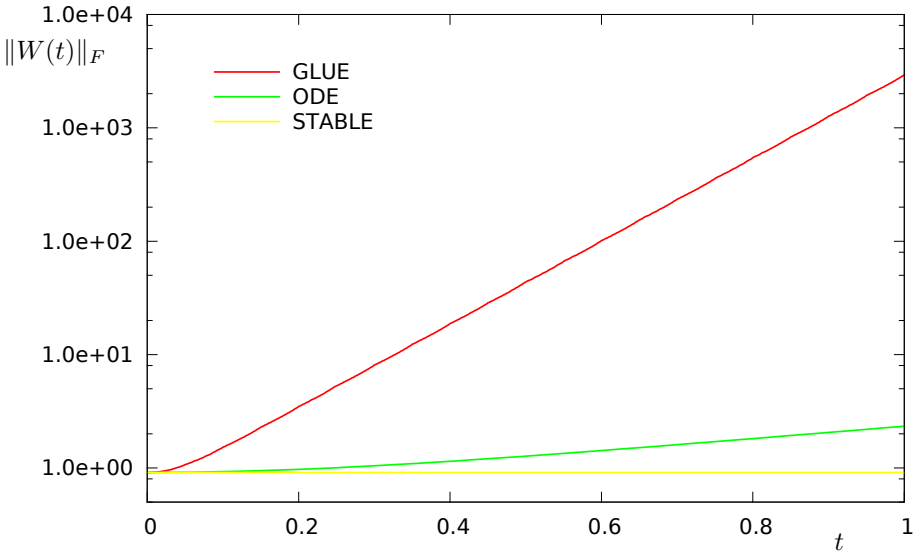

FIGURE 2. *Norm of W as function of t for Example 5.1 using stabilization.*

the case $k = 0$. Note that for this example both approaches of Section 4 yield undistinguishable results when we use (4.19) for at least three times. In particular, both approaches yield nicely bounded transformations for the smooth version of Sylvester's law of inertia.

EXAMPLE 5.2. Choosing

$$E = U^T \operatorname{diag}(2, -3, 3, -2)\, U$$

with

$$U(t) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \sin(2\pi t) & 1 & 0 & 0 \\ 0 & \cos(2\pi t) & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Figure 3 shows again the Frobenius norm of W as function of t obtained by stabilization in comparison with the norm obtained using the construction for $k = 0$ labeled by `GLUE`. The used stabilization techniques were the descent method with going through (4.19) once labeled by `DESCENT1` and twice labeled by `DESCENT2` and Newton's method labeled by `NEWTON`. It can be seen that the results of the descent method approach the result of Newton's method quickly with increasing number of applications of (4.19).

6. Conclusions. We presented a smooth version of Sylvester's law of inertia maintaing the smoothness of the given matrix-valued symmetric matrix function of constant rank. The proof distinguishes between the case of continuous functions and continuously differentiable functions. The proof is constructive but the resulting numerical procedures are unstable. We therefore developed two possible stabilization techniques. Numerical examples verified the obtained approaches.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 542-560, August 2020.

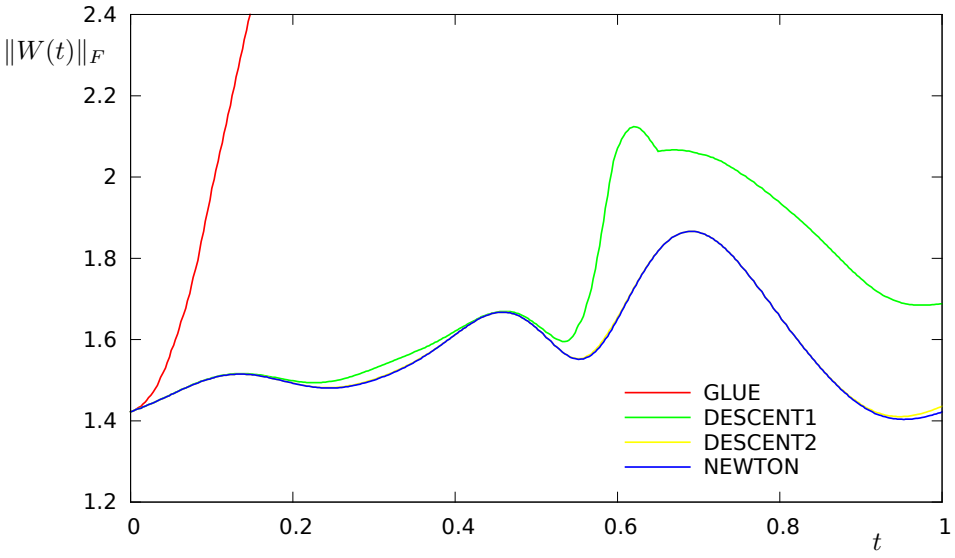

FIGURE 3. *Norm of W as function of t for Example 5.2 using stabilization.*

REFERENCES

[1] J.R. Bunch and L. Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Math. Comp.*, 31:163–179, 1977.
[2] E.W. Cheney. *Introduction to Approximation Theory.* McGraw-Hill, New York, 1966.
[3] L. Dieci and T. Eirola. On smooth decompositions of matrices. *SIAM J. Matrix Anal. Appl.*, 20:800–819, 1999.
[4] V. Doležal. The existence of a continuous basis of a certain subspace of E_r which depends on a parameter. *Cas. Pro. Pest. Mat.*, 89:466–468, 1964.
[5] G.H. Golub and C.F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 1983.
[6] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer, Berlin, 2002.
[7] J. Hilgert and K.H. Neeb. *Structure and Geometry of Lie Groups.* Springer, New York, 2012.
[8] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations. Analysis and Numerical Solution.* EMS Publishing House, Zürich, 2006.
[9] P. Kunkel, V. Mehrmann, and L. Scholz. Self-adjoint differential-algebraic equations. *Math. Control Signals Systems*, 26:47–76, 2014.
[10] L. Scholz. Condensed forms for linear port-Hamiltonian descriptor systems. Preprint 09-2017, Institut für Mathematik, Technische Universität Berlin, 2017.
[11] L. Scholz. Condensed forms for linear port-Hamiltonian descriptor systems. *Electron. J. Linear Algebra*, 35:65–89, 2019.
[12] L. Weiss and P.L. Falb. Doležal's theorem, linear algebra with continuously parametrized elements, and time-varying systems. *Mathematical Systems Theory*, 3:67–75, 1969.