Block-Jacobi methods with Newton-steps and non-unitary joint matrix diagonalization

Martin Kleinsteuber and Hao Shen

Department of Electrical and Computer Engineering
Technische Universität München, Arcisstr. 21, 80333 Munich, Germany
{kleinsteuber,hao.shen}@tum.de
http://www.gol.ei.tum.de

Abstract. In this work, we consider block-Jacobi methods with Newton steps in each subspace search and prove their local quadratic convergence to a local minimum with non-degenerate Hessian under some orthogonality assumptions on the search directions. Moreover, such a method is exemplified for non-unitary joint matrix diagonalization, where we present a block-Jacobi-type method on the oblique manifold with guaranteed local quadratic convergence.

Keywords: Jacobi algorithms, local convergence properties, manifold optimization, signal separation.

1 Introduction

Jacobi-type methods have been very successful in numerical linear algebra for the task of diagonalizing matrices, dating back to the seminal work of Carl Gustaf Jacob Jacobi from 1846 [1], where a scheme is presented to iteratively find the eigenvalue decomposition of a Hermitian matrix. A characteristic feature of many Jacobi-type methods is that they act to minimize the distance to diagonality while preserving some predefined constraints, cf. [2] and the references therein. Furthermore, their inherent parallelizability makes them also useful for large scale matrix computations.

In the context of jointly diagonalizing a set of Hermitian matrices, eg. JADE [3, 4] is a very prominent example that borrows the idea of iteratively minimizing the distance to joint diagonality by means of unitary similarity transforms. Moreover, Jacobi-type algorithms have also demonstrated their promising performance in solving the problem of blind source separation, such as [5, 6].

Block Jacobi-type procedures were developed as a generalization of standard Jacobi method in terms of grouped variables for solving symmetric eigenvalue problems or singular value problems [7]. In each Jacobi sweep, it is required to solve a sequence of sub-optimization problems, which can be computationally expensive or infeasible. To cope with this problem, we propose to employ a Newton-step that approximates a solution to this sub problem. We show that under this setting, if the predefined search directions are orthogonal with respect to the non degenerate Hessian of the minimum, then the Jacobi method converges locally to that minimum with quadratic, hence superlinear rate. We exemplify these insights in order to develop an efficient method for the problem of non-unitary joint matrix diagonalization (JMD), which arises in the context of Blind Source Separation.

2 Block-Jacobi-type methods on manifolds

From a viewpoint of geometric optimization, Jacobi-type methods can be considered as a generalization of coordinate descent methods to the manifold setting. Given some point on a manifold, Jacobi-type methods optimize a cost function along some predefined directions in the tangent space in order to find the next iterate on the manifold. In practice, this requires two more algorithmic ingredients: i) a map from the tangent space to the manifold, which is traditionally done via the Riemannian exponential, but also more general concepts like retraction have been introduced for general line search method adaptions to the manifold setting; (ii) a practical step-size selection rule that approximates the search for a minimizer of the restricted cost function. In the following, we provide a formal setup and introduce some notations that help to make these concepts more concrete.

Let M be an n-dimensional smooth manifold and consider the problem of minimizing a smooth function $f\colon M \to \mathbb{R}$. Let the map $\mu\colon \mathbb{R}^n \times M \to M$ be smooth and fulfill the property that $\mu_x\colon \mathbb{R}^n \to M, \quad v \mapsto \mu(v, x)$ is a local parametrization around x with $\mu_x(0) = x$.[1] Actually, it would suffice for μ_x to be defined only in an appropriate neighbourhood of 0, but we omit this detail for the sake of readability. In order to explain the predefined directions on the manifold for the purpose of optimization, let $\mathbb{R}^n = \oplus_i V^{(i)}$ be a vector space decomposition of $\mathbb{R}^n$ into a direct sum of N subspaces V_i, with $\dim V_i = \ell_i$. By a slight abuse of notation, we denote by $V_i \in \mathbb{R}^{n \times \ell_i}$ also basis of these vector spaces. Since μ_x is a local diffeomorphism, its differential map at 0

$$T_0\mu_x\colon \mathbb{R}^n \to T_xM \tag{1}$$

is bijective, hence the images of V_i under this map, namely $\mathcal{V}_x^{(i)} := T_0\mu_x(V_i)$, form a direct vector space decomposition of the tangent space T_xM, i.e.

$$T_xM = \oplus_i \mathcal{V}_x^{(i)}. \tag{2}$$

Note, that the restrictions of μ_x to subspace V_i, i.e. $\mu_x(V_i)$ for all $i = 1, \dots, N$, are often referred to as basic transformations. Our main result in this paper states that Jacobi-type methods are locally quadratically convergent to a local minimum of f, if the Hessian at this minimum is non-degenerate and if the above decomposition of the tangent space is orthogonal w.r.t. this Hessian at the local minimum.

In the following, we first consider the case where the $\mathcal{V}_x^{(i)}$ are one-dimensional, leading to (one-dimensional) predefined directions in the tangent-space along

[1] That is, μ_x^{-1} is a coordinate chart around x.

which optimization is performed in each step. We then consider the generalization to higher dimensions, leading to a manifold adaption of block-coordinate descent methods on manifolds.

2.1 Coordinate descent on Manifolds

Let us consider the case where the dimension of the V_i's is equal to one, i.e. $\ell_i = 1$ for all $i = 1, \dots n$. A straightforward adaption of coordinate descent methods to manifolds for minimizing a smooth cost function f is now as follows. It consists of iterating *sweeps*, where one sweep sequentially works off all the initially predetermined directions V_i. That is, starting from some point $x \in M$, we determine the local minimum[2] that is closest to zero of the restricted cost function $t \mapsto f \circ \mu_x(V_1 t)$. This minimum t^* then delivers the initial point $x_{\text{new}} = \mu_x(V_1 t^*)$ for a subsequent minimization along the next predetermined direction $\mu_{x_{\text{new}}}(V_2 t)$. This procedure is repeated until all directions V_i are worked off. The Jacobi-sweep is visualized in Figure 1 and concretized in Algorithm 1.

Algorithm 1 *Jacobi-Sweep on a manifold M*

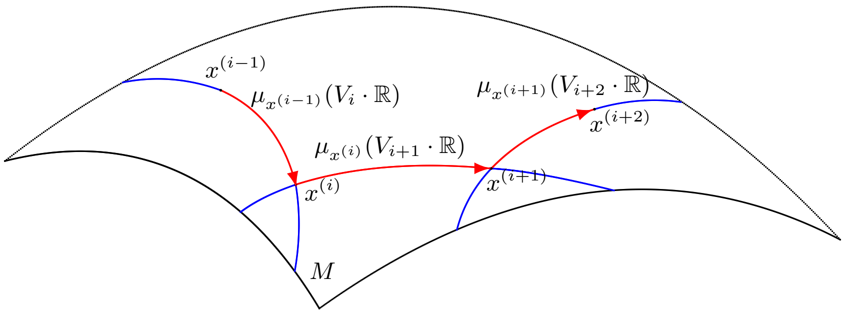
```
INPUT: initial point $x^{(0)} \in M$ and and directions $V_i$, $i = 1...n$
FOR $i = 1, \dots, n$ DO
    STEP 1. Compute the local minumum $t^*$ with smallest absolute value of
        $$\varphi\colon \mathbb{R} \to \mathbb{R}, \quad t \mapsto f \circ \mu_{x^{(i-1)}}(V_i t) \qquad (3)$$
    STEP 2. Set $x^{(i)} := \mu_{x^{(i-1)}}(V_i t^*)$.
    STEP 3. Increase $i$.
```

It can be shown that a *Jacobi method*, that is iterating Algorithm 1, leads to a locally quadratic convergent algorithm, if the descent directions are orthogonal with respect to the non-degenerated Hessian at a local minimum of f.

Theorem 1 ([8]). *Let M be an n-dimensional manifold and let x^* be a local minimum of the smooth cost function $f\colon M \to \mathbb{R}$ with nondegenerate Hessian $\mathsf{H}_f(x^*)$. If the $\mathcal{V}_i := T_0\mu_{x^*}(V_i) \in T_{x^*}M$ are orthogonal with respect to $\mathsf{H}_f(x^*)$, then the Jacobi method is locally quadratic convergent to x^*.*

In practice, the search for a local minimum in STEP 1 of Algorithm 1 is often infeasible. We therefore follow a different approach that is based on a one dimensional Newton optimisation step. Similar approximations of the optimal

[2] The reason why we choose a local and not a global minimum here is that for the convergence analysis, this choice is needed to be smooth around a minimizer of the cost function. This can only be guaranteed by choosing the nearest local minimum along basic transformations.



$x^{(i-1)}$
$\mu_{x^{(i-1)}}(V_i \cdot \mathbb{R})$
$\mu_{x^{(i+1)}}(V_{i+2} \cdot \mathbb{R})$
$x^{(i+2)}$
$\mu_{x^{(i)}}(V_{i+1} \cdot \mathbb{R})$
$x^{(i)}$
$x^{(i+1)}$
M

Fig. 1. Illustration of Jacobi-type method on smooth manifolds.

step size have already been used in [9, 10]. The idea is to replace STEP 1 by the approximation

$$t^* := -\left(\tfrac{\mathrm{d}^2}{\mathrm{d}\,t^2}\varphi(t)\big|_{t=0}\right)^{-1} \tfrac{\mathrm{d}}{\mathrm{d}\,t}\varphi(t)\big|_{t=0}. \tag{4}$$

We show in a more general result on block-coordinate descent methods in the next section that this approximation maintains the local quadratic convergence property.

2.2 Approximate Block-Coordinate descent on manifolds

With the established setting, it is quite easy to generalize the above line-search methods to so-called block-Jacobi-type methods. The idea is to extend the search for a new iterate to higher dimensional subspaces, i.e. we drop the assumption that $\ell_i = 1$. To search for a new iterate, we propose to extend the one-dimensional Newton step (4) to higher dimensions. More precisely, for the i-th iteration within one sweep, we consider the restricted cost function

$$\varphi \colon \mathbb{R}^{\ell_i} \to \mathbb{R}, \quad t \mapsto f \circ \mu_{x^{(i-1)}}(V_i t). \tag{5}$$

Note, that now $t \in \mathbb{R}^{\ell_i}$ and $V_i \in \mathbb{R}^{n \times \ell_i}$. We denote the Hessian matrix of φ at 0 as $\mathsf{H}_\varphi(0)$ and the standard gradient as $\nabla_\varphi(0)$. The ℓ_i-dimensional Newton step

$$t^* := -\mathsf{H}_\varphi(0)^{-1}\nabla_\varphi(0) \tag{6}$$

is then a straightforward generalization of (4) to higher dimensions. The sweep for a block-Jacobi method with Newton-step size is summarized in Algorithm 2.

Algorithm 2 *Block-Jacobi-Sweep with Newton-step size on a manifold M*

```
INPUT: initial point $x^{(0)} \in M$ and and matrices $V_i \in \mathbb{R}^{n \times \ell_i}$, $i = 1...N$
FOR $i = 1, \dots, N$ DO
    STEP 1. Compute the $t^* \in \mathbb{R}_i^{\ell}$ according to (6).
    STEP 2. Set $x^{(i)} := \mu_{x^{(i-1)}}(V_i t^*)$.
```

We study local convergence properties of the Block-Jacobi-Sweep on a manifold M with Newton-step size.

Theorem 2. *Let M be an n-dimensional manifold and let x^* be a local minimum of the smooth cost function $f\colon M \to \mathbb{R}$ with nondegenerate Hessian $\mathsf{H}_f(x^*)$. If the subspaces $\mathcal{V}_i := T_0\mu_{x^*}(V_i) \subset T_{x^*}M$ are orthogonal with respect to $\mathsf{H}_f(x^*)$, then the Block-Jacobi method which consists of iterating Algorithm 2 is locally quadratic convergent to x^*.*

Sketch of the proof. The proof is inspired by the local quadratic convergence result for block-Jacobi methods on manifolds with an exact step-size selection rule in [11]. It consists of essentially three steps. First, one has to show that a local minimum x^* is a fixed-point of the algorithm. Second, we formulate one sweep as a map $s\colon M \mapsto M$ and compute its derivative at x^*. Third, we show that this derivative vanishes if the subspaces $\mathcal{V}_i$ are orthogonal with respect to the Hessian at x^*. Finally, using the Taylor Series expansion, we can then conclude that

$$\|s(x) - x^*\| \leq C\|x - x^*\|^2 \tag{7}$$

for all x being close enough to x^*, which ensures local quadratic convergence of the algorithm.

The fixed-point condition holds because $Df(x^*) = 0$ implies by use of the chain rule, that $\nabla_\varphi(0) = 0$, and hence $t^*(x^*) = 0$ for all directions. We now consider one step within a sweep given by

$$r\colon M \to M, \quad x \mapsto \mu(V_i t^*(x), x). \tag{8}$$

Using $t^*(x^*) = 0$, the derivative of r at x^* applied to a tangent vector $\xi \in T_{x^*}M$ is

$$\begin{aligned} Dr(x)|_{x=x^*}\xi &= D_1\mu(V_i(Dt^*(x)|_{x=x^*}\xi), x^*) + D_2\mu(0, x^*)\xi && (9)\\ &= D_1\mu(V_i(Dt^*(x)|_{x=x^*}\xi), x^*) + \xi && (10) \end{aligned}$$

where D_l denotes the derivative w.r.t. the l-th argument. So the next step is to calculate $Dt^*(x)|_{x=x^*}\xi$. Let $e_k \in \mathbb{R}^{\ell_i}$ be the k-th standard basis vector and denote

$$\xi_k := D_1\mu(V_i e_k, x) \in T_xM. \tag{11}$$

Then the (k,l) entry of $\mathsf{H}_\varphi(0)$ is $\mathsf{H}_f(x^*)(\xi_k,\xi_l)$ and the k-th entry of $\nabla_\varphi(0)$ is $Df(x)\xi_k$. Using the product rule, we have

$$Dt^*(x)|_{x=x^*}\xi = -\big(D\mathsf{H}_\varphi(0)^{-1}\xi\big)\nabla_\varphi(0)|_{x=x^*} - \mathsf{H}_\varphi(0)^{-1}\big(D\nabla_\varphi(0)|_{x=x^*}\xi\big) \quad (12)$$

The first summand vanishes since $\nabla_\varphi(0)|_{x=x^*} = 0$. The entries of the vector on the right-hand side are $D(Df(x)|_{x=x^*}\xi_k)\xi = \mathsf{H}_f(x^*)(\xi_k,\xi)$. It follows that if $\xi^{(i)} \in \mathcal{V}_i$, i.e. if $\xi^{(i)} = \sum_k h_k\xi_k$ is a linear combination of the ξ_k, then

$$Dt^*(x)|_{x=x^*}\xi^{(i)} = -\mathsf{H}_\varphi(0)^{-1}\mathsf{H}_\varphi(0)|_{x=x^*}\begin{bmatrix} h_1, \\ \vdots \\ h_{\ell_i} \end{bmatrix} = -\begin{bmatrix} h_1, \\ \vdots \\ h_{\ell_i} \end{bmatrix}, \quad (13)$$

so that, by using (10) and the fact that $D_1\mu(V_ih,x^*) = \xi^{(i)}$, the derivative $Dr(x)|_{x=x^*}$ annihilates the $\mathcal{V}_i$-component of ξ. On the other hand, if $\xi^{(i),\perp}$ is such that $\mathsf{H}_f(x^*)(\xi_k,\xi^{(i),\perp}) = 0$ for all k, then $Dt^*(x)|_{x=x^*}\xi^{(i),\perp} = 0$ and thus

$$Dr(x)|_{x=x^*}\xi^{(i),\perp} = \xi^{(i),\perp}. \quad (14)$$

This shows that the derivative of the i-th step in the Jacobi-Sweep is an orthogonal projection with respect to $\mathsf{H}_f(x^*)$ onto $\mathcal{V}_i^\perp$. Therefore, using the fixed-point property of x^* and the chain rule, we conclude that

$$Ds(x)|_{x=x^*}\xi = 0, \quad (15)$$

which concludes the proof of Theorem 2.

Remark 1. It is worthwhile to notice that convergence of Jacobi-type algorithms is strongly dependent on the construction of basic transformations. Local quadratic convergence can only be attained, when the subspaces in the tangent space specified by the basic transformations are orthogonal with respect to the Hessian at a critical point. Unfortunately, both characterization of the Hessian at critical points and construction of computationally light basic transformations are nontrivial tasks in general.

3 Applications in Signal Separation

In order to investigate performance of the theoretical results presented in the last section, we employ the problem of joint matrix diagonalization as an illustrative and important example. Given a set of $m\times m$ real symmetric matrices $\{C_i\}_{i=1}^n$, constructed by $C_i = A\Lambda_iA^\top$, for $i = 1,\ldots,n$, where $\Lambda_i = \operatorname{diag}\big(\lambda_{i1},\ldots,\lambda_{im}\big) \in \mathbb{R}^{m\times m}$ with $\lambda_{ij} \neq 0$ for $j = 1,\ldots,m$ and $A \in Gl(m)$. The problem of estimating the matrix A given only the set $\{C_i\}_{i=1}^n$ leads to finding an $X \in Gl(m)$ such that the matrices $Y_i = X^\top C_iX$ are simultaneously diagonalised. In a generic situation, a joint diagonalizer X can only be determined up to column-wise permutation and scaling, i.e. if X is a diagonalizer, so is any XDP where D

is an $m \times m$ invertible diagonal matrix and P an $m \times m$ permutation matrix, cf. [12] for a uniqueness analysis of non-unitary JMD. To deal with the scaling ambiguity, we restrict the solutions to the oblique manifold, i.e.

$$\mathcal{OB}(m) := \left\{X \in \mathbb{R}^{m \times m} \middle| \operatorname{ddiag}(X^\top X) = I_m, \operatorname{rank} X = m\right\}, \tag{16}$$

where $\operatorname{ddiag}(Z)$ forms a diagonal matrix, whose diagonal entries are just those of Z, and I_m is the $m \times m$ identity matrix.

We employ the popular off-norm function for measuring the diagonality of matrices, i.e.

$$f \colon \mathcal{OB}(m) \to \mathbb{R}, \quad X \mapsto \frac{1}{4} \sum_{i=1}^{n} \left\| \operatorname{off}(X^\top C_i X) \right\|_{\mathrm{F}}^2, \tag{17}$$

where $\operatorname{off}(Z) = Z - \operatorname{ddiag}(Z)$ is a matrix by setting the diagonal entries of Z to zero, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. In order to develop a block Jacobi algorithm to minimise the cost function f, we recall firstly a local parameterisation on $\mathcal{OB}(m)$. Let us denote the set of all $m \times m$ matrices with all diagonal entries equal to zero by

$$\mathfrak{off}(m) = \left\{Z \in \mathbb{R}^{m \times m} | z_{ii} = 0, \text{ for } i = 1, \ldots, m\right\}, \tag{18}$$

then, for every point $X \in \mathcal{OB}(m)$, the following map

$$\mu_X \colon \mathfrak{off}(m) \to \mathcal{OB}(m), \quad Z \mapsto X(I_m + Z) \operatorname{diag}\left\{\tfrac{1}{\|X(e_1+z_1)\|}, \ldots, \tfrac{1}{\|X(e_m+z_m)\|}\right\}, \tag{19}$$

where $Z = [z_1, \ldots, z_m] \in \mathfrak{off}(m)$ and e_i is the i-th standard basis vector of $\mathbb{R}^m$, is a local and smooth parameterisation around X. Let us define the set of matrices, whose entries are all zero except the (i,j) and (j,i) position, as

$$V_{ij} := \left\{Z = (z_{ij}) \in \mathbb{R}^{m \times m} | z_{pq} = 0, \text{ for } (p,q) \notin \{(i,j),(j,i)\}\right\}, \tag{20}$$

with $\oplus_{i \neq j} V_{ij} = \mathfrak{off}(m)$. We denote

$$\mathcal{V}_{ij}(X) := \{\tfrac{\mathrm{d}}{\mathrm{d}\,t}\, \mu_X(t \cdot Z)|_{t=0} \,| Z \in V_{ij}\}, \tag{21}$$

being a predefined vector space decomposition of the tangent space $T_X\mathcal{OB}(m)$, i.e. $T_X\mathcal{OB}(m) = \oplus_{i \neq j} \mathcal{V}_{ij}(X)$.

The results in [4] have shown that the subspaces $\mathcal{V}_{ij}(X^*)$ are orthogonal with respect to $\mathsf{H}_f(X^*)$, hence validate the feasibility of construction of a block Jacobi algorithm with Newton step size selection that is locally quadratically convergent to an exact joint diagonalizer.

The task of our experiment is to jointly diagonalize a set of symmetric matrices $\{\widetilde{C}_i\}_{i=1}^n$, constructed by

$$\widetilde{C}_i = A \Lambda_i A^\top + \varepsilon E_i, \qquad i = 1, \ldots, n, \tag{22}$$

where $A \in \mathbb{R}^{m \times m}$ is a randomly picked matrix in $\mathcal{OB}(m)$, diagonal entries of

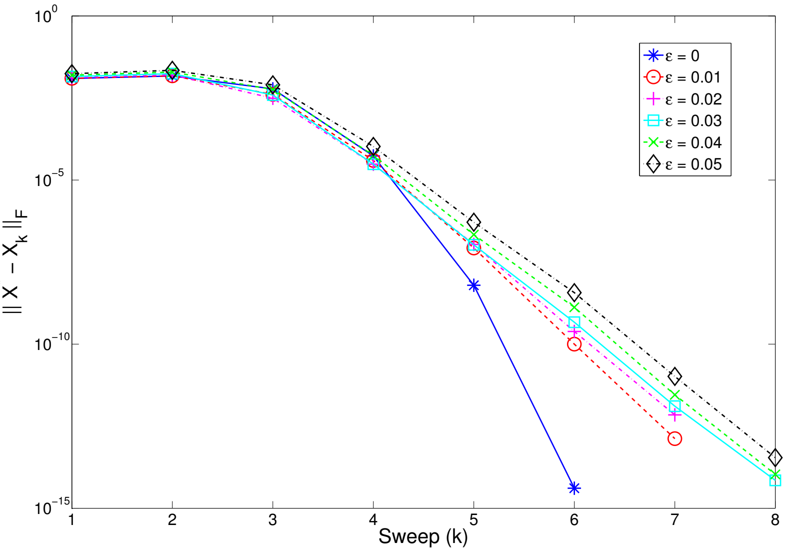


Fig. 2. Convergence properties of the proposed block Jacobi algorithm.

Λ_i are drawn from a uniform distribution on the interval $(9, 11)$, $E_i \in \mathbb{R}^{m \times m}$ is the symmetric part of an $m \times m$ matrix, whose entries are generated from a uniform distribution on the unit interval $(-0.5, 0.5)$, representing additive noise, and $\varepsilon \in \mathbb{R}$ is the noise level. We set $m = 5$, $n = 20$, and run six tests in accordance with increasing noise, by using $\varepsilon = d \times 10^{-2}$ where $d = 0, \ldots, 5$.

The convergence of algorithms is measured by the distance of the accumulation point $X^* \in \mathcal{OB}(m)$ to the current iterate $X_k \in \mathcal{OB}(m)$, i.e. by $\|X_k - X^*\|_{\mathrm{F}}$. According to Fig. 2, it is clear that our proposed algorithm converges locally quadratically fast to a joint diagonalizer under the exact nonunitary JMD setting, i.e. $\varepsilon = 0$, while with presence of noise, the algorithm seems to converge only linearly.

References

1. Jacobi, C.G.J.: Über ein leichtes verfahren, die in der theorie der säcularstörungen vorkommenden gleichungen numerisch aufzulösen. Crelle's J. für die reine und angewandte Mathematik **30** (1846) 51–94
2. Kleinsteuber, M.: A sort-Jacobi algorithm for semisimple Lie algebras. Linear Algebra and its Applications **430**(1) (2009) 155–173
3. Cardoso, J.F.: High-order contrasts for independent component analysis. Neural Computation **11**(1) (1999) 157–192
4. Shen, H., Hüper, K.: Block Jacobi-type methods for non-orthogonal joint diagonalisation. In: Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing. (2009) 3285–3288

5. Shen, H.: Geometric Algorithms for Whitened Linear Independent Component Analysis. PhD thesis, The Australian National University, Australia (2008)
6. Shen, H., Kleinsteuber, M., Hüper, K.: Block Jacobi-type methods for log-likelihood based linear independent subspace analysis. In: Proceedings of the 17^{th} IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2007), Thessaloniki, Greece (2007) 133–138
7. Golub, G., van Loan, C.F.: Matrix Computations. 4th edn. The Johns Hopkins University Press, Baltimore (2013)
8. Kleinsteuber, M.: Jacobi-Type Methods on Semisimple Lie Algebras – A Lie Algebraic Approach to the Symmetric Eigenvalue Problem. PhD thesis, Bayerische Julius-Maximilians-Universität Würzburg, Germany (2006)
9. Götze, J., Paul, S., Sauer, M.: An efficient Jacobi-like algorithm for parallel eigenvalue computation. IEEE Transactions on Computers **42**(9) (1993) 1058–1065
10. Modi, J.J., Pryce, J.D.: Efficient implementation of Jacobi's diagonalization method on the DAP. Numerische Mathematik **46**(3) (1985) 443–454
11. Hüper, K.: A Calculus Approach to Matrix Eigenvalue Algorithms. PhD thesis, University of Würzburg, Germany (2002)
12. Kleinsteuber, M., Shen, H.: Uniqueness analysis of non-unitary matrix joint diagonalization. IEEE Transactions on Signal Processing **61**(7) (2013) 1786–1796