On the Accuracy of the Element-wise Jacobi Methods for PGEP

Josip Matejaš[1]* and **Vjeran Hari**[2]**

[1] Faculty of Economy, University of Zagreb, Kennedyjev trg 6, 10000 Zagreb, Croatia
[2] Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia

We analyze the relative accuracy of two new element-wise Jacobi-type methods for the positive definite generalized eigenvalue problem $Ax = \lambda Bx$, where A and B are symmetric matrices and B is positive definite. A detailed error analysis is used, and the appropriate numerical tests are performed. If A and B are well-behaved positive definite matrices then the transformation parameters will have small relative errors and numerical tests indicate the high relative accuracy of the methods.

Copyright line will be provided by the publisher

1 Introduction

We consider accuracy properties of two Jacobi methods for the positive definite generalized eigenvalue problem (PGEP), $Ax = \lambda Bx$, $x \neq 0$, where A and B are symmetric matrices of order n and B is positive definite. If only A is positive definite then we consider the eigenproblem $Bx = \lambda Ax$, $x \neq 0$. Note that these two eigenproblems have the same eigenvectors and the reciprocal eigenvalues.

The most well known Jacobi method for solving PGEP is the Falk-Langemeyer (FL) method. It was introduced in [3] and its asymptotic convergence and accuracy were studied in [7] and [5], respectively. A single FL-step annihilates the pivot element at position (i, j), $i < j$ by the congruence transformation $A' = F^T AF$, $B' = F^T BF$, where F is elementary plane matrix with unit diagonal. The FL method is well defined for a wider class of definite pairs of symmetric matrices [7] and the transformations can be efficiently applied by using BLAS1 saxpy operations. Nonetheless, the iteration matrices occasionally need normalization during the process. When the FL method is used as a kernel algorithm for the block Jacobi method, this can be a demanding task on contemporary CPU and GPU parallel computing machines [6]. Fortunately, one can use other Jacobi methods as kernel algorithms, in particular the Hari-Zimmermann (HZ) and the Cholesky-Jacobi (CJ) method (see [1, 8]) which can be seen as normalized versions of the FL method.

In this short communication, we shall concentrate on the HZ method. Unlike the FL method, it initially scales the matrix B symmetrically, and then in each step, it maintains the unit diagonal of B. The method uses congruence transformations. In each step, the transformation matrix F is obtained as a product of three matrices: Jacobi rotation for B, the diagonal matrix which restores the unit diagonal of B and Jacobi rotation for the updated A. On the level of two by two pivot submatrices, the transformation matrix has the form

$$\tilde{F} = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1+b_{ij}}} & 0 \\ 0 & \frac{1}{\sqrt{1-b_{ij}}} \end{bmatrix} \begin{bmatrix} \cos(\theta - \frac{\pi}{4}) & -\sin(\theta - \frac{\pi}{4}) \\ \sin(\theta - \frac{\pi}{4}) & \cos(\theta - \frac{\pi}{4}) \end{bmatrix} = \frac{1}{\tau} \begin{bmatrix} \cos\phi & -\sin\phi \\ sin\psi & \cos\psi \end{bmatrix}. \quad (1)$$

In [2] it has been shown that the angles are defined as follows

$$t2 = \tan(2\theta) = \frac{2a_{ij} - (a_{ii} + a_{jj})b_{ij}}{(a_{ii} - a_{jj})\tau}, \quad \tau = \sqrt{(1 - b_{ij})(1 + b_{ij})}, \quad -\frac{\pi}{4} \leq \theta \leq \frac{\pi}{4}, \quad (2)$$

$$\cos\phi = \rho\cos\theta - \xi\sin\theta, \ \sin\phi = \rho\sin\theta + \xi\cos\theta, \ \cos\psi = \rho\cos\theta + \xi\sin\theta, \ \sin\psi = \rho\sin\theta - \xi\cos\theta, \quad (3)$$

where $\rho = 0.5(\sqrt{1 - b_{ij}} + \sqrt{1 + b_{ij}})$, $\xi = b_{ij}/(2\rho)$. Here $A = (a_{rt})$ and $B = (b_{rt})$ are the current iteration matrices and i, j, $i < j$ are the pivot indices. Now, if we use in each step, instead of Jacobi rotation, inverse of the (lower or upper) Cholesky factor of $\tilde{B}$, we obtain the Cholesky-Jacobi (CJ) method. In each step we can choose the most appropriate algorithm and the choice mostly depends on how accurately we can compute the elements of $\tilde{F}$. In this way, we obtain a hybrid method with the best properties. That is the reason for our detailed accuracy analysis of the methods. The global convergence of the HZ and CJ methods has been proved in [2].

2 Accuracy of the HZ and CJ methods

To obtain better accuracy estimates we use the subtle error analysis from [4, 5] which does not neglect nonlinear parts of the errors and takes into account the signs of the errors. By such approach, the error bounds can be greatly improved. The

* Corresponding author: e-mail jmatejas@efzg.hr,
** Coauthor: e-mail hari@math.hr,

Copyright line will be provided by the publisher

suppression and cancellation of the errors can occur In numerical computation and that error analysis can detect it. We assume that the unit round-off (machine epsilon) $\mathbf{u}$ complies with the IEEE standard. We use the standard model of arithmetic where the relative errors in the basic operations $(+, -, \cdot, /, \sqrt{})$ are not greater than the unit round-off. In the analysis we use the exact expressions for the errors, like $(1+\varepsilon_1)(1+\varepsilon_2) = 1+\varepsilon_1+\varepsilon_2+\varepsilon_1\varepsilon_2$, $\frac{1+\varepsilon_1}{1+\varepsilon_2} = 1+\varepsilon_1-\varepsilon_2+\frac{\varepsilon_2(\varepsilon_2-\varepsilon_1)}{1+\varepsilon_2}$, etc., where the linear and nonlinear parts are separated. Such expressions shed more light on the structure of the errors and enable obtaining sharper error bounds. In particular, one can detect good situations when the initial errors are suppressed (multiplied by some small factor) or canceled to a certain extent, and use such information to derive much smaller error bounds. In addition, one can also detect the critical points in the algorithm where the errors can rapidly grow up.

Let us consider the HZ algorithm. One critical point in the algorithm comes from the subtraction in the numerator of the quotient which defines 2θ in the relation (2). Using the subtle error analysis, for the computed value of $t2$ we obtain $\mathrm{fl}(t2) = (1+\varepsilon_{t2})t2$ with $|\varepsilon_{t2}| \leq (2\mu + 6.505)\mathbf{u}$ provided that $\mu = \max\left\{\left|\frac{2a_{ij}}{2a_{ij}-(a_{ii}+a_{jj})b_{ij}}\right|, \left|\frac{(a_{ii}+a_{jj})b_{ij}}{2a_{ij}-(a_{ii}+a_{jj})b_{ij}}\right|\right\} \leq 1/\sqrt{\mathbf{u}}$. Thus, inaccuracy can occur if μ is large which means that the pivot submatrices are in some sense close to proportionality. Using the computed $t2$ we compute $t = \tan\theta$, $sn = \sin\theta$, $cs = \cos\theta$ and obtain $\mathrm{fl}(t) = (1+\varepsilon_t)t$, $\mathrm{fl}(sn) = (1+\varepsilon_{sn})sn$, $\mathrm{fl}(cs) = (1+\varepsilon_{cs})cs$, where

$$\varepsilon_t = \omega_t + \frac{1}{\sqrt{1+t2^2}}\,\varepsilon_{t2} + \eta_t\,, \quad \varepsilon_{cs} = \omega_{sn} + \frac{t^2}{(1+t^2)\sqrt{1+t2^2}}\,\varepsilon_{t2} + \eta_{sn}\,, \quad \varepsilon_{cs} = \omega_{cs} + \frac{1}{(1+t^2)\sqrt{1+t2^2}}\,\varepsilon_{t2} + \eta_{cs}\,. \tag{4}$$

Here $\omega_t, \omega_{sn}, \omega_{cs}$ are produced by current operations and $\eta_t, \eta_{sn}, \eta_{cs}$ are the nonlinear parts of the errors. We obtained $|\omega_t|, |\omega_{sn}|, |\omega_{cs}| < 7\mathbf{u}$ and $|\eta_t|, |\eta_{sn}|, |\eta_{cs}| \ll \mathbf{u}$.

Now, let us assume that A and B belong to the class of positive definite matrices which can be well scaled symmetrically, i.e. the condition numbers of $A_S = (\mathrm{diag}(A))^{-1/2}A(\mathrm{diag}(A))^{-1/2}$ and $B_S = (\mathrm{diag}(B))^{-1/2}B(\mathrm{diag}(B))^{-1/2}$ are small. Note that $B_S = B$, because B has unit diagonal. For such matrices each b_{ij} will be small, hence $\tau \approx 1$. If we express $|t2|$ from (2) in terms of scaled elements $a_{ij}^{(S)} = a_{ij}/\sqrt{a_{ii}a_{jj}}$ and $b_{ij}^{(S)} = b_{ij}$, we obtain $|t2| = \frac{2\sqrt{a_{ii}a_{jj}}}{a_{ii}+a_{jj}}\left|\frac{a_{ij}^{(S)}}{\gamma\cdot\tau}\right| \cdot \frac{1}{\mu} = \left|\frac{b_{ij}^{(S)}}{\gamma\cdot\tau}\right| \cdot \frac{1}{\mu}$, where $\gamma = \frac{a_{ii}-a_{jj}}{a_{ii}+a_{jj}}$ is the relative gap. The error $|\varepsilon_{t2}|$ can be large if μ is large, $|\varepsilon_{t2}| = O(2\mu\mathbf{u})$, and then, for that class of matrices, $|t2|$ is small, $|t2| = O(1/\mu)$. Hence $t \approx \frac{t2}{2}(1-\frac{t2^2}{2}) = O(\frac{1}{2\mu})$, $sn \approx t(1-\frac{t^2}{2}) = O(\frac{1}{2\mu})$, $cs \approx 1-\frac{t^2}{2} \approx 1$. From (4) we see that large $|\varepsilon_{t2}|$ can cause large $|\varepsilon_t|$ and $|\varepsilon_{sn}|$ but $|\varepsilon_{cs}|$ is reduced by the factor t^2. Indeed, we have $t^2|\varepsilon_{t2}| = O(\frac{1}{4\mu^2}) \cdot O(2\mu\mathbf{u}) = O(\frac{\mathbf{u}}{2\mu})$. So, the cosine is always computed accurately, $|\varepsilon_{cs}| = O(\mathbf{u})$. Next, we apply the subtle error analysis to the elements of the transformation matrix (1). In the expressions defining their errors the factors ε_{cs} and $t\varepsilon_{sn}$ appear. Since $|\varepsilon_{cs}| = O(\mathbf{u})$ and $|t\varepsilon_{sn}| = O(\frac{1}{2\mu}\cdot 2\mu\mathbf{u}) = O(\mathbf{u})$, we see that in spite of possible significant inaccuracy in $\mathrm{fl}(t2)$, the transformation matrix will be computed to high relative accuracy.

The same result has been obtained for the CJ algorithm in which the numerical expressions are similar to those in HZ.

3 Future Work

The goal of this research that uses such kind of subtle error analysis is to prove the following theorem.

Theorem 3.1 *On the class of pairs of symmetric positive definite matrices that can be well scaled symmetrically, the HZ and CJ methods compute the eigenvalues of PGEP to high relative accuracy.*

This conclusion is strongly supported by the numerical experiments (cf. [2]). Afterwards our next goal will be to prove the high relative accuracy of the block Jacobi method which uses the HZ or CJ method as the kernel algorithm.

Acknowledgements This work has been fully supported by Croatian Science Foundation under the project: IP_09_2014_3670.

References

[1] V. Hari, On Cyclic Jacobi Methods for the Positive Definite Generalized Eigenvalue Problem. Ph.D. thesis, University of Hagen (1984).
[2] V. Hari, Globally convergent Jacobi methods for positive definite matrix pairs, to appear in Numerical Algorithms, https://doi.org/10.1007/s11075-017-0435-5.
[3] F. S. Langemeyer, Das Jacobische Rotations-Verfahren fũr reel symmetrische Matrizen-Paare I, II, Elektronische Datenverarbeitung, 30-43 (1960).
[4] J. Matejaš, Accuracy of the Jacobi Method on Scaled Diagonally Dominant Symmetric Matrices. SIAM. J. Matrix Anal. Appl. **31(1)**, 133-153 (2009).
[5] J. Matejaš, Accuracy of one step of the Falk-Langemeyer method. Numerical Algorithms **68(4)**, 645-670 (2015).
[6] V. Novaković, S. Singer, S Singer, Blocking and Parallelization of the Hari – Zimmermann Variant of the Falk – Langemeyer Algorithm for the Generalized SVD, Parallel Comput. **49**, 136-152 (2015).
[7] I. Slapničar, V. Hari, On the Quadratic Convergence of the Falk-Langemeyer Method for Definite Matrix Pairs. SIAM J. Matrix Anal. Appl. **12(1)**, 84-114 (1991).
[8] K. Zimmermann, On the Convergence of the Jacobi Process for Ordinary and Generalized Eigenvalue Problems. Ph.D. Thesis, Dissertation No. 4305 ETH, Zũrich (1965).

Copyright line will be provided by the publisher