On the High Relative Accuracy of the **HZ method**

Josip Matejaš
Faculty of Economics and Business

Vjeran Hari
Faculty of Science

University of Zagreb, CROATIA

Sveučilište u Zagrebu
Ekonomski fakultet

Positive Definite Generalized Eigenvalue Problem (PGEP)

$$Ax = \lambda Bx, \qquad x \neq 0,$$

where A and B are symmetric matrices of order n and B is positive definite.

If A is positive definite and B is not, then we consider

$$Bx = \mu Ax, \quad x \neq 0 \quad \Rightarrow \quad \mu = \frac{1}{\lambda}.$$

Hari-Zimmermann Method (HZ)

is the normalized version of the Falk-Langemeyer method. It generates the sequence of matrices

$$A^{(k+1)} = Z_k^T A^{(k)} Z_k, \quad B^{(k+1)} = Z_k^T B^{(k)} Z_k, \quad k = 0,1,2,...$$

where each $B^{(k)}$ has unit diagonal. It has a preliminary step

$$A^{(0)} = DAD, \; B^{(0)} = DBD, \; D = diag\left(b_{11}^{-1/2}, b_{22}^{-1/2}, ..., b_{nn}^{-1/2}\right).$$

A single HZ-step annihilates the pivot elements at position (i,j), $i < j$ by the congruence transformation (k is omitted)

$$A' = Z^T A Z, \; B' = Z^T B Z,$$

and it maintains the unit diagonal of B.

HZ Method

On the level of 2 by 2 pivot submatrices,

$\hat{A}' = \hat{Z}^T \hat{A}\hat{Z}$, $\hat{B}' = \hat{Z}^T \hat{B}\hat{Z}$ where

$$\hat{Z} = \hat{J}_B \hat{D}_B \hat{J}_A$$

$$= \begin{pmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{1+b_{ij}}} & 0 \\ 0 & \frac{1}{\sqrt{1-b_{ij}}} \end{pmatrix} \begin{pmatrix} \cos(\theta-\frac{\pi}{4}) & -\sin(\theta-\frac{\pi}{4}) \\ \sin(\theta-\frac{\pi}{4}) & \cos(\theta-\frac{\pi}{4}) \end{pmatrix}.$$

↓ diagonalizes $\hat{B}$

↓ makes unit diagonal of $\hat{B}$

↓ diagonalizes updated $\hat{A}$

HZ Method

$$\rho = \frac{\sqrt{1+b_{ij}} + \sqrt{1-b_{ij}}}{2}, \quad \xi = \frac{b_{ij}}{2\rho}, \quad \tau = \sqrt{(1-b_{ij})(1+b_{ij})},$$

$$\tan(2\theta) = \frac{2a_{ij} - (a_{ii} + a_{jj})b_{ij}}{(a_{ii} - a_{jj}) \cdot \tau}, \quad -\frac{\pi}{4} \le \theta \le \frac{\pi}{4}.$$

$\hat{Z}$ can be written in the form

$$\hat{Z} = \begin{bmatrix} c1 & -s1 \\ s2 & c2 \end{bmatrix} = \frac{1}{\tau}\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\psi & \cos\psi \end{bmatrix},$$

where

$$\cos\phi = \rho\cos\theta - \xi\sin\theta, \quad \sin\phi = \rho\sin\theta + \xi\cos\theta,$$
$$\cos\psi = \rho\cos\theta + \xi\sin\theta, \quad \sin\psi = \rho\sin\theta - \xi\cos\theta.$$

HZ Algorithm (first part)

$tol = 4u;$ select the pivot pair (i, j) with $i < j$

% *** STOPPING CRITERION ***

if $|a_{ij}| / \sqrt{a_{ii} \cdot a_{jj}} \le tol$ **then** $a_{ij} = 0;\ a_{ji} = 0;$ **endif**

if $|b_{ij}| \le tol$ **then** $b_{ij} = 0;\ b_{ji} = 0;$ **endif**

if $a_{ij} \ne 0$ **or** $b_{ij} \ne 0$ **then**

% *** INITIAL PARAMETERS AND TANGENT OF ROTATION ANGLE ***

$$\rho = (\sqrt{1 + b_{ij}} + \sqrt{1 - b_{ij}}) / 2; \quad \xi = b_{ij} / (2\rho); \quad \tau = \sqrt{(1 - b_{ij})(1 + b_{ij})}$$

if $2 \cdot a_{ij} = (a_{ii} + a_{jj}) \cdot b_{ij}$ **then** $t = 0;$ **else**

if $a_{ii} = a_{jj}$ **then** $t = 1;$ **else**

$$t2 = (2 \cdot a_{ij} - (a_{ii} + a_{jj}) \cdot b_{ij}) / (\tau \cdot (a_{ii} - a_{jj})); \quad t = t2 / (1 + \sqrt{1 + t2^2});$$

endif

% *** ELEMENTS OF TRANSFORMATION MATRIX ***

$cs = 1 / \sqrt{1 + t^2}; \quad sn = t / \sqrt{1 + t^2};$

$c1 = (\rho \cdot cs - \xi \cdot sn) / \tau; \quad s1 = (\rho \cdot sn + \xi \cdot cs) / \tau;$

$c2 = (\rho \cdot cs + \xi \cdot sn) / \tau; \quad s2 = (\rho \cdot sn - \xi \cdot cs) / \tau;$

Subtle Error Analysis

We have derived error estimates

- without neglecting the terms of higher order (in machine precision) of the errors,
- without neglecting the signs of the errors and variables.

So, we use exact expressions for the errors.

Using such an approach we can detect:

- suppression of the initial and intermediate errors
- cancellation of the initial and intermediate errors
- weak parts of the algorithm (where the error(s) can explode)

First Assumption

Let u denote the round-off unit (machine epsilon) according to IEEE standard, i.e.

$$u \in \{\underbrace{2^{-23}, 2^{-24}}_{\text{single}}, \quad \underbrace{2^{-52}, 2^{-53}}_{\text{double}}, \quad \underbrace{2^{-63}, 2^{-64}}_{\text{extended}}, \quad \underbrace{2^{-112}, 2^{-113}}_{\text{quadruple}}, \quad \underbrace{10^{-12}}_{\text{calculator}}\}$$

single	double	extended	quadruple	calculator
$6 \cdot 10^{-8}$	10^{-16}	$5 \cdot 10^{-20}$	10^{-34}	

Note that $u \le 2^{-23} < 1.1921 \cdot 10^{-7}$.

Second Assumption: Standard Model of Arithmetic

We use the model where the floating point result of the basic operations is given by

$$\mathrm{fl}(a \pm b) = (1 + \varepsilon_1)(a \pm b)$$

$$\mathrm{fl}(a \cdot b) = (1 + \varepsilon_2)(a \cdot b)$$

$$\mathrm{fl}(a / b) = (1 + \varepsilon_3)(a / b)$$

$$\mathrm{fl}(\sqrt{a}) = (1 + \varepsilon_4)(\sqrt{a})$$

where $|\varepsilon_i| \le u, \; i = 1,2,3,4.$

Some Examples of Error Expressions

We use the exact expressions for the errors:

$$(1+\varepsilon_1)(1+\varepsilon_2) = 1 + \boxed{\varepsilon_1 + \varepsilon_2} + \boxed{\varepsilon_1\varepsilon_2}$$

$$\frac{1+\varepsilon_1}{1+\varepsilon_2} = 1 + \boxed{\varepsilon_1 - \varepsilon_2} + \boxed{\frac{\varepsilon_2(\varepsilon_2 - \varepsilon_1)}{1+\varepsilon_2}}$$

$$\sqrt{1+\varepsilon_1} = 1 + \boxed{\frac{\varepsilon_1}{2}} - \boxed{\frac{\varepsilon_1^2}{4 + 2\varepsilon_1 + 4\sqrt{1+\varepsilon_1}}}$$

Linear and nonlinear parts of the error.

Suppression of the Initial Errors

Let us estimate the error in evaluation of the expression

$$v = 1 + x^2 .$$

Suppose that we have at disposal only an approximation of x,

$$\mathrm{fl}(x) = (1 + \varepsilon_x)x .$$

For the computed value $\mathrm{fl}(v)$ we have

$$\mathrm{fl}(v) = (1+\varepsilon_1)\left[1 + (1+\varepsilon_2)(1+\varepsilon_x)^2 x^2\right]$$

$$= \left[1 + \boxed{\varepsilon_1 + \varepsilon_2 \frac{x^2}{1+x^2}} + \boxed{2\varepsilon_x \frac{x^2}{1+x^2}} + \boxed{\eta \frac{x^2}{1+x^2}}\right] \cdot v ,$$

$$\eta = \varepsilon_1\varepsilon_2 + 2(\varepsilon_1 + \varepsilon_2)\varepsilon_x + \varepsilon_x^2 + 2\varepsilon_1\varepsilon_2\varepsilon_x + (\varepsilon_1 + \varepsilon_2)\varepsilon_x^2 + \varepsilon_1\varepsilon_2\varepsilon_x^2$$

suppression of ε_x

Linear and nonlinear parts of the error.

Cancellation of the Initial Errors

$$z = \frac{x}{1+x}, \quad x > 0 \qquad \ldots \qquad \mathrm{fl}(x) = (1+\varepsilon_x)x$$

$$\mathrm{fl}(z) = \frac{1+\varepsilon_1}{1+\varepsilon_2} \cdot \frac{(1+\varepsilon_x)x}{1+(1+\varepsilon_x)x} = \frac{1+\varepsilon_1}{1+\varepsilon_2} \cdot \frac{1+\varepsilon_x}{1+\dfrac{\varepsilon_x x}{1+x}} \cdot z$$

$$\frac{1+\varepsilon_1}{1+\varepsilon_2} = 1 + \boxed{\varepsilon_1 - \varepsilon_2} + \boxed{\frac{\varepsilon_2(\varepsilon_2 - \varepsilon_1)}{1+\varepsilon_2}},$$

$$\frac{\varepsilon_x}{1+x}$$

$$\frac{1+\varepsilon_x}{1+\dfrac{\varepsilon_x x}{1+x}} = 1 + \varepsilon_x - \frac{\varepsilon_x x}{1+x} - \boxed{\frac{\varepsilon_x^2 x}{(1+x)(1+x+\varepsilon_x x)}}.$$

Linear and nonlinear parts of the error.

Cancellation of the Initial Errors

For $z = \dfrac{x}{1+x}$, $x > 0$ ... $\mathrm{fl}(x) = (1+\varepsilon_x)x$

the final error is

$$\mathrm{fl}(z) = \left(1 + \boxed{\varepsilon_1 - \varepsilon_2} + \boxed{\frac{\varepsilon_x}{1+x}} + \boxed{\eta}\right) \cdot z\,,$$

where $|\varepsilon_1|, |\varepsilon_2| \le u$, and

$$\eta = \frac{\varepsilon_x(\varepsilon_1 - \varepsilon_2)}{1+x} + \left(1 + \frac{\varepsilon_x}{1+x}\right)\frac{\varepsilon_2^2 - \varepsilon_1\varepsilon_2}{1+\varepsilon_2}$$

$$- \left(1 + \varepsilon_1 - \varepsilon_2 + \frac{\varepsilon_2^2 - \varepsilon_1\varepsilon_2}{1+\varepsilon_2}\right) \cdot \frac{\varepsilon_x^2 x}{(1+x)(1+x+\varepsilon_x x)}\,.$$

Linear and nonlinear parts of the error.

HZ Algorithm (first part)

$tol = 4u;$ select the pivot pair (i, j) with $i < j$

% *** STOPPING CRITERION ***

if $|a_{ij}| / \sqrt{a_{ii} \cdot a_{jj}} \le tol$ **then** $a_{ij} = 0;$ $a_{ji} = 0;$ **endif**

if $|b_{ij}| \le tol$ **then** $b_{ij} = 0;$ $b_{ji} = 0;$ **endif**

if $a_{ij} \ne 0$ **or** $b_{ij} \ne 0$ **then**

% *** INITIAL PARAMETERS AND TANGENT OF ROTATION ANGLE ***

$$\rho = (\sqrt{1 + b_{ij}} + \sqrt{1 - b_{ij}}) / 2; \quad \xi = b_{ij} / (2\rho); \quad \tau = \sqrt{(1 - b_{ij})(1 + b_{ij})}$$

if $2 \cdot a_{ij} = (a_{ii} + a_{jj}) \cdot b_{ij}$ **then** $t = 0;$ **else**

if $a_{ii} = a_{jj}$ **then** $t = 1;$ **else**

$$t2 = (2 \cdot a_{ij} - (a_{ii} + a_{jj}) \cdot b_{ij}) / (\tau \cdot (a_{ii} - a_{jj})); \quad t = t2 / (1 + \sqrt{1 + t2^2});$$

endif

% *** ELEMENTS OF TRANSFORMATION MATRIX ***

$cs = 1 / \sqrt{1 + t^2};$ $sn = t / \sqrt{1 + t^2};$

$c1 = (\rho \cdot cs - \xi \cdot sn) / \tau;$ $s1 = (\rho \cdot sn + \xi \cdot cs) / \tau;$

$c2 = (\rho \cdot cs + \xi \cdot sn) / \tau;$ $s2 = (\rho \cdot sn - \xi \cdot cs) / \tau;$

HZ Algorithm (second part)

% * * * CORRECT THE TRANSFORMATION MATRIX IF NECESSARY * * *

$d1 = c1^2 + s2^2 + 2 \cdot c1 \cdot s2 \cdot b_{ij};$ $\qquad d2 = c2^2 + s1^2 - 2 \cdot c2 \cdot s1 \cdot b_{ij};$

if $|1 - d1| / d1 > tol$ **then** $d = \sqrt{d1};\ c1 = c1/d;\ s2 = s2/d;$ **endif**

if $|1 - d2| / d2 > tol$ **then** $d = \sqrt{d2};\ c2 = c2/d;\ s1 = s1/d;$ **endif**

Note that (A, B), (A', B') and $(\tilde{A}, \tilde{B})$, where

$A' = Z^T A Z, \quad B' = Z^T B Z,$

$\tilde{A} = \mathrm{fl}(Z)^T \cdot A \cdot \mathrm{fl}(Z), \quad \tilde{B} = \mathrm{fl}(Z)^T \cdot B \cdot \mathrm{fl}(Z),$

have the same eigenvalues,

but we must also maintain the unit diagonal of B.

HZ Algorithm (third part)

% * * * UPDATING THE PIVOT SUBMATRICES ***

$a_{ii'} = c1^2 \cdot a_{ii} + s2^2 \cdot a_{jj} + 2 \cdot c1 \cdot s2 \cdot a_{ij}$; $a_{jj'} = s1^2 \cdot a_{ii} + c2^2 \cdot a_{jj} - 2 \cdot c2 \cdot s1 \cdot a_{ij}$;

$a_{ij'} = (c1 \cdot c2 - s1 \cdot s2) \cdot a_{ij} + (c2 \cdot s2 \cdot a_{jj} - c1 \cdot s1 \cdot a_{ii})$; $a_{ji'} = a_{ij'}$;

$b_{ij'} = (c1 \cdot c2 - s1 \cdot s2) \cdot b_{ij} + (c2 \cdot s2 - c1 \cdot s1)$; $b_{ji'} = b_{ij'}$;

% * * * UPDATING THE REST OF PIVOT ROWS AND COLUMNS ***

for $k = 1, \ldots, n,\ \ k \neq i, j$ **do**

$a_{ki'} = c1 \cdot a_{ki} + s2 \cdot a_{kj}$; $b_{ki'} = c1 \cdot b_{ki} + s2 \cdot b_{kj}$; $a_{ik'} = a_{ki'}$; $b_{ik'} = b_{ki'}$;

$a_{kj'} = c2 \cdot a_{kj} - s1 \cdot a_{ki}$; $b_{kj'} = c2 \cdot b_{kj} - s1 \cdot b_{ki}$; $a_{jk'} = a_{kj'}$; $b_{jk'} = b_{kj'}$;

endfor

endif % ($a_{ij} \neq 0$ or $b_{ij} \neq 0$)

Well Behaved Matrix Pairs

Let A and B be positive definite matrices that can be well-scaled symmetrically, i.e. the spectral condition numbers $\kappa_2(A_S)$ and $\kappa_2(B_S)$, where

$$A_S = D_A^{-1/2} A D_A^{-1/2}, \quad D_A = diag(A),$$
$$B_S = D_B^{-1/2} B D_B^{-1/2}, \quad D_B = diag(B),$$

are small. We have $A_S = (a_{lm}^{(S)}), \ B_S = (b_{lm}^{(S)}),$

$$a_{lm}^{(S)} = \frac{a_{lm}}{\sqrt{a_{ll} a_{mm}}}, \quad b_{lm}^{(S)} = \frac{b_{lm}}{\sqrt{b_{ll} b_{mm}}} = \frac{b_{lm}}{\sqrt{1 \cdot 1}} = b_{lm} \Rightarrow B_S = B.$$

We assume that

(A1) $\quad \left|a_{lm}^{(S)}\right| \le \alpha_0 \le 0.5, \quad \left|b_{lm}\right| \le \beta_0 \le 0.5, \quad l \ne m.$

Perturbation Theorem (Drmač)

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ $(\lambda_1' \geq \lambda_2' \geq \ldots \geq \lambda_n')$ be the eigenvalues of the pair (A, B) $((A+\delta A, B+\delta B))$ where A, B are positive definite and δA, δB are symmetric perturbations.

Let $(\delta A)_S = D_A^{-1/2} \delta A D_A^{-1/2}$, $\quad \varepsilon_{A_S} = \|(\delta A)_S\|_2 / \|A_S\|_2$

and $(\delta B)_S = D_B^{-1/2} \delta B D_B^{-1/2}$, $\quad \varepsilon_{B_S} = \|(\delta B)_S\|_2 / \|B_S\|_2$

If

$$\|(\delta A)_S\|_2 \cdot \|A_S^{-1}\|_2 < 1 \quad \text{and} \quad \|(\delta B)_S\|_2 \cdot \|B_S^{-1}\|_2 < 1,$$

then

$$\max_{1 \leq i \leq n} \frac{|\lambda_i' - \lambda_i|}{\lambda_i} \leq \frac{\varepsilon_{A_S} \kappa_2(A_S) + \varepsilon_{B_S} \kappa_2(B_S)}{1 - \varepsilon_{B_S} \kappa_2(B_S)}.$$

Application of Perturbation Theorem

Note again that (A,B), (A',B') and $(\tilde{A},\tilde{B})$, where

$$A' = Z^T AZ, \quad B' = Z^T BZ,$$

$$\tilde{A} = \mathrm{fl}(Z)^T \cdot A \cdot \mathrm{fl}(Z), \quad \tilde{B} = \mathrm{fl}(Z)^T \cdot B \cdot \mathrm{fl}(Z),$$

have the same eigenvalues.

We apply the perturbation theorem to the matrix pairs $(\tilde{A},\tilde{B})$ and $(\mathrm{fl}(\tilde{A}),\mathrm{fl}(\tilde{B}))$, where

$$\mathrm{fl}(\tilde{A}) = \tilde{A} + \delta\tilde{A}, \quad \mathrm{fl}(\tilde{B}) = \tilde{B} + \delta\tilde{B}.$$

We need to estimate $\left\|(\delta\tilde{A})_S\right\|_2$ and $\left\|(\delta\tilde{B})_S\right\|_2$.

Bounds for the Perturbation Matrices

Let $\delta\tilde{A}$, $\delta\tilde{B}$ be the perturbations caused by finite arithmetic computation in HZ algorithm. If (A1) holds then

$$\begin{aligned}\left\|(\delta\tilde{A})_S\right\|_2 &\le \left(8\alpha_{ij} + 4\sqrt{2}\,\alpha'_{ij'}\sqrt{n-2}\right)u + \left(12.002\alpha_{ij} + 2\sqrt{2}\,\alpha'_{ij'}\sqrt{n-2}\right)u^2 \\ &\le \left(8.001\alpha_{ij} + 5.657\alpha'_{ij'}\sqrt{n-2}\right)u,\end{aligned}$$

$$\begin{aligned}\left\|(\delta\tilde{B})_S\right\|_2 &\le \left(11\beta_{ij} + 4\sqrt{2}\,\beta'_{ij'}\sqrt{n-2}\right)u + \left(227.033\beta_{ij} + 2\sqrt{2}\,\beta'_{ij'}\sqrt{n-2}\right)u^2 \\ &\le \left(11.001\beta_{ij} + 5.657\beta'_{ij'}\sqrt{n-2}\right)u,\end{aligned}$$

where
$$\alpha_{ij} = \frac{1+|a_{ij}^{(S)}|}{1-|a_{ij}^{(S)}|} \le 3, \quad \alpha'_{ij'} = \frac{\alpha_0}{\sqrt{1-|a_{ij}^{(S)}|}} \le \frac{\sqrt{2}}{2},$$

$$\beta_{ij} = \frac{1+|b_{ij}|}{1-|b_{ij}|} \le 3, \quad \beta'_{ij'} = \frac{\beta_0}{\sqrt{1-|b_{ij}|}} \le \frac{\sqrt{2}}{2}.$$

Note that $\alpha_{ij}, \beta_{ij} \to 1$, $\alpha'_{ij'}\beta'_{ij'} \to 0$ as the process advances.

Numerical experiments

Numerical experiments are made using

- MATLAB R2019b on a 64-bit PC
- *symbolic toolbox* and *variable precision arithmetic* to compute the actual relative errors of the main variables
- sample of 1186 well-scaled matrix pairs of order 100
- MATLAB eig function for comparison

Numerical experiments

Two versions of HZ method, with two different pairs of formulas for computing diagonal elements of *A*, were tested:

HZ

$$a'_{ii'} = c1^2 \cdot a_{ii} + s2^2 \cdot a_{jj} + 2 \cdot c1 \cdot s2 \cdot a_{ij}$$

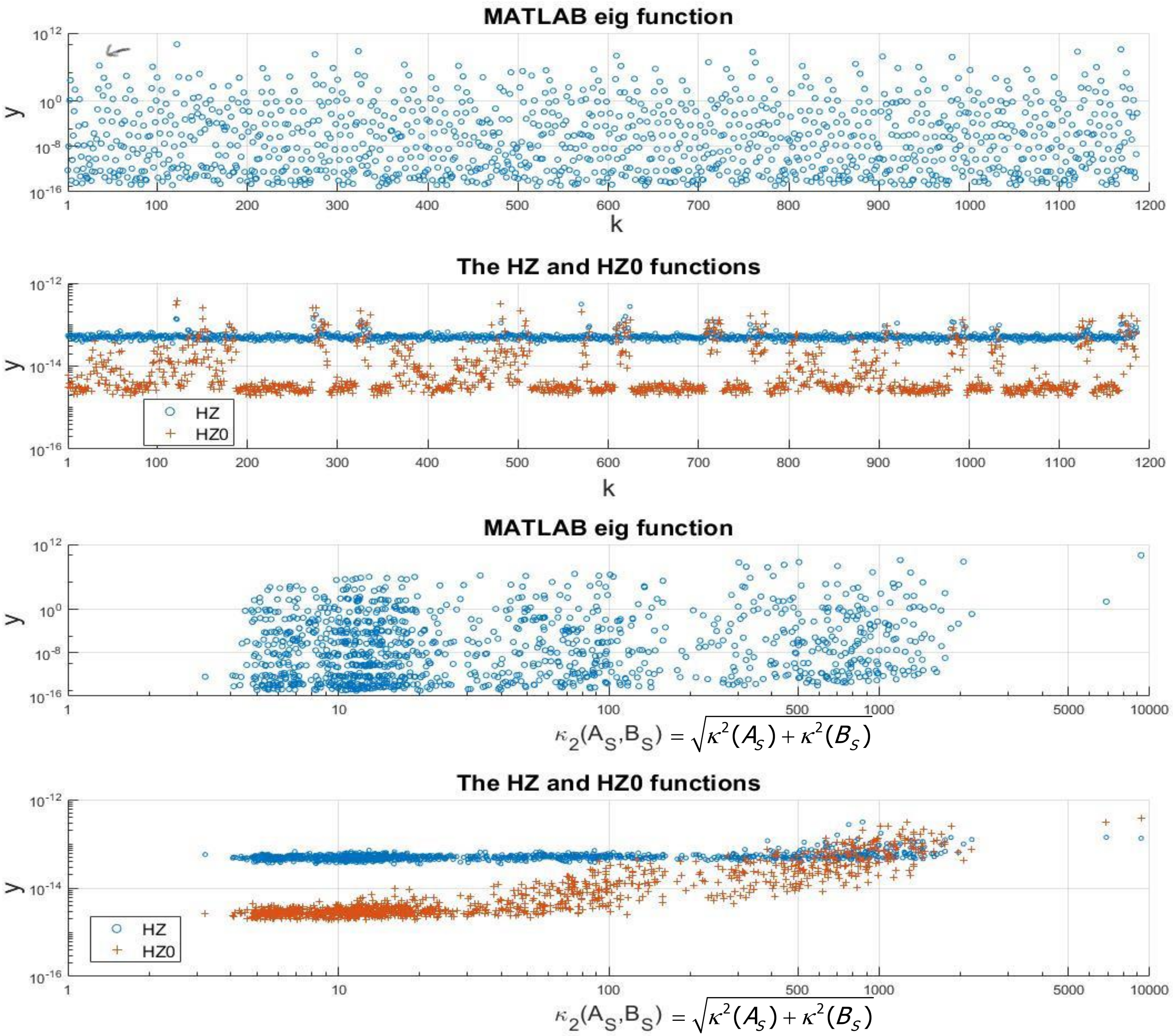
$$a'_{jj'} = s1^2 \cdot a_{ii} + c2^2 \cdot a_{jj} - 2 \cdot c2 \cdot s1 \cdot a_{ij}$$

HZ0

$$a'_{ii'} = a_{ii} + \left[(b_{ij} / \tau - s1)(b_{ij} / \tau + s1) \cdot a_{ii} + (2 \cdot c1 \cdot a_{ij} + s2 \cdot a_{jj}) \cdot s2\right]$$

$$a'_{jj'} = a_{jj} - \left[(s2 - b_{ij} / \tau)(s2 + b_{ij} / \tau) \cdot a_{jj} + (2 \cdot c2 \cdot a_{ij} - s1 \cdot a_{ii}) \cdot s1\right]$$

Numerical experiments – results



References

V. Hari: *Globally convergent Jacobi methods for positive definite matrix pairs*, Numer. Alg. 79 (1) (2018), pp. 221-249.

V. Hari: *On the global convergence of the complex HZ method*, SIAM J. Matrix Anal. Appl. 40 (4) (2019), pp. 1291-1310.

V. Hari: *On the global convergence of the block Jacobi method for the positive definite generalized eigenvalue problem*, Calcolo 58 (24) (2021).

V. Hari: *On the quadratic convergence of the complex HZ method for the positive definite generalized eigenvalue problem*, Linear Algebra and Appl. 632 (1) (2022), pp.153-192.

J. Matejaš, V. Hari: *On the high relative accuracy of the HZ method*, under second round of revision, Applied Mathematics and Comp., pp. 1-25.

Thanks

for your

attention.