On the Accuracy of Element-wise Jacobi Methods for PGEP

Josip Matejaš
Faculty of Economics and Business

Vjeran Hari
Department of Mathematics

University of Zagreb, CROATIA

This work has been fully supported by Croatian Science Foundation under the project IP–2014–09–3670.

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

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

Falk-Langemeyer Method (FL)

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

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

where F is elementary plane matrix with unit diagonal.

On the level of 2 by 2 pivot submatrices,

$$\begin{pmatrix} a'_{ii} & 0 \\ 0 & a'_{jj} \end{pmatrix} = \begin{pmatrix} 1 & -\beta \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix} \begin{pmatrix} 1 & \alpha \\ -\beta & 1 \end{pmatrix}$$

$$\begin{pmatrix} b'_{ii} & 0 \\ 0 & b'_{jj} \end{pmatrix} = \begin{pmatrix} 1 & -\beta \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} \begin{pmatrix} 1 & \alpha \\ -\beta & 1 \end{pmatrix}$$

FL Method

Advantages

- it solves PGEP where *(A,B)* is a definite matrix pair, i.e. *tA+sB* is positive definite for some *t,s*
- transformations are cheap to apply (saxpy operations only)

Shortcomings

- the norms of iteration matrices and matrix of accumulated transformations gradually increase, occasionally they have to be normalized
- the stopping criterion uses the normalized matrices.

Hari-Zimmermann Method (HZ)

is the normalized version of the FL method. It uses the simplified iteration matrices

$$B^{(0)}, B^{(1)}, B^{(2)}, B^{(3)}, \ldots$$

with unit diagonal. It has the initial step

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

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

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

and it maintains the unit diagonal of B.

HZ Method

On the level of 2 by 2 pivot submatrices,

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

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

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

diagonalizes $\tilde{B}$	makes unit diagonal of $\tilde{B}$	diagonalizes updated $\tilde{A}$

HZ Method

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

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

$\tilde{Z}$ can be stated in the form

$$\tilde{Z} = \begin{bmatrix} \mathrm{c}1 & -\mathrm{s}1 \\ \mathrm{s}2 & \mathrm{c}2 \end{bmatrix} = \frac{1}{\tau}\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\psi & \cos\psi \end{bmatrix},$$

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 Method

Advantages

- it is a proper generalization of the standard Jacobi method for the matrix *A* (reduces to it if *B*=*I*)
- it does not require renormalization of the iterated matrices
- It has been shown that it is an excellent choice for the kernel algorithm for the block Jacobi method for GSVD

Shortcomings

- one of the matrices, A or B, has to be positive definite
- transformations are somewhat more expensive

Cholesky-Jacobi Method (CJ)

On the level of 2 by 2 submatrices, instead of

$$\tilde{Z} = \tilde{J}_B \tilde{D}_B \tilde{J}_A \quad \text{(as in HZ)}$$

we have

$$\tilde{Z} = \tilde{L}_B^{-T} \tilde{J}_A \quad (LL^T J \ \text{ algorithm})$$

or

$$\tilde{Z} = \tilde{R}_B^{-T} \tilde{J}_A \quad (RR^T J \ \text{ algorithm}),$$

where $\tilde{B} = \tilde{L}_B \tilde{L}_B^T$ or $\tilde{B} = \tilde{R}_B \tilde{R}_B^T$.

In each step *CJ* chooses $LL^T J$ or $RR^T J$ algorithm.

Advantages and shortcomings as in HZ method.

Subtle Error Analysis

We have derived error estimates

- without neglecting the terms of higher order (in machine precision) of the errors
- by taking into account the signs of the errors.

Using such an aproach we can detect:

- suppression of the initial and intermediate errors
- cancellation of the initial and intermediate errors

Assumptions

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

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

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

Note that $u \le 2^{-23} < 1.1920929 \cdot 10^{-7}$.

Standard Model of Aritmetic

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

Proper Error Estimates

We use the exact expressions for the errors:

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

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

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

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 an approximation of x,

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

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

$$\begin{aligned}
\mathrm{fl}(v) &= (1+\varepsilon_1)\left[1 + (1+\varepsilon_2)(1+\varepsilon_x)^2 x^2\right] \\
&= \left[1 + \boxed{\varepsilon_1 + \varepsilon_2 \frac{x^2}{1+x^2}} + \boxed{2\varepsilon_x \frac{x^2}{1+x^2}} + \boxed{\eta \frac{x^2}{1+x^2}}\right] \cdot v ,
\end{aligned}$$

$$\begin{aligned}
\eta &= \varepsilon_1\varepsilon_2 + 2(\varepsilon_1 + \varepsilon_2)\varepsilon_x + \varepsilon_x^2 \\
&\quad + 2\varepsilon_1\varepsilon_2\varepsilon_x + (\varepsilon_1 + \varepsilon_2)\varepsilon_x^2 \\
&\quad + \varepsilon_1\varepsilon_2\varepsilon_x^2
\end{aligned}$$

suppression of ε_x

Cancellation of the Initial Errors

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

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

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

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

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

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

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

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

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

Application in PGEP Methods

If $x = \tan^2 \varphi$ then

$$\sin^2 \varphi = \frac{x}{1+x}, \qquad \cos^2 \varphi = \frac{1}{1+x}.$$

If $x = \cot^2 \varphi$ then

$$\sin^2 \varphi = \frac{1}{1+x}, \qquad \cos^2 \varphi = \frac{x}{1+x}.$$

HZ Algorithm

repeat

select the pivot pair (i, j) with $i < j$

% compute $t2 = \tan(2\theta)$, $t = \tan\theta$, $cs = \cos\theta$, $sn = \sin\theta$

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

$$t2 = \left[2a_{ij} - (a_{ii}+a_{jj})b_{ij}\right] / \left[\tau(a_{ii}-a_{jj})\right]$$ **CP**

$$t = t2 / \left(1+\sqrt{1+t2^2}\right); \quad cs = 1/\sqrt{1+t^2}; \quad sn = t/\sqrt{1+t^2}$$

% compute $c1 = \cos\phi$, $s1 = \sin\phi$, $c2 = \cos\psi$, $s2 = \sin\psi$

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

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

$$y = (a_{ii}-a_{jj})/\tau; \quad z = (1-t)/(1+t);$$

% update the pivot submatrices $\{\delta = \mathrm{sgn}(b_{ij})\}$

$$x = \left(a_{ii} - 2\delta a_{ij} + a_{jj}\right) / \left(1-\delta b_{ij}\right); \quad a_{ii} = x + yz; \quad a_{jj} = x - y/z$$

$$a_{ij} = 0; \quad a_{ji} = 0; \quad b_{ij} = 0; \quad b_{ji} = 0$$

% update the rest of rows and columns i and j

.....

until convergence

Critical Points

$$\tilde{A} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} 1 & b_{ij} \\ b_{ij} & 1 \end{pmatrix}, \quad t2 = \frac{2a_{ij} - (a_{ii} + a_{jj})b_{ij}}{(a_{ii} - a_{jj})\sqrt{1 - b_{ij}^2}}.$$

Let $\mu = \max\left\{ \left| \frac{2a_{ij}}{2a_{ij} - (a_{ii} + a_{jj})b_{ij}} \right|, \left| \frac{(a_{ii} + a_{jj})b_{ij}}{2a_{ij} - (a_{ii} + a_{jj})b_{ij}} \right| \right\}$

$$= \max\left\{ \left| \frac{1}{1-k} \right|, \left| \frac{1}{1/k - 1} \right| \right\}, \text{ where } k = \frac{a_{ij}}{a_{ii} + a_{jj}} \ : \ \frac{b_{ij}}{1+1}.$$

Let $\mathrm{fl}(t2) = (1 + \varepsilon_{t2})t2.$

If $\mu \le \frac{1}{\sqrt{u}}$ then $|\varepsilon_{t2}| \le (2\mu + 6.505)u$.

If $|\varepsilon_{t2}|$ is large $\Rightarrow \mu$ is large $\Rightarrow k \approx 1 \Rightarrow$ pivot submatrices are close to proportionality with respect to trace.

Errors in *t*, *cs*, *sn*

$$t = \frac{t2}{1+\sqrt{1+t2^2}}, \quad cs = \frac{1}{\sqrt{1+t^2}}, \quad sn = \frac{t}{\sqrt{1+t^2}}.$$

If $\mathrm{fl}(z) = (1+\varepsilon_z)z, \; z \in \{t, cs, sn\}$, then

$$\varepsilon_t = \boxed{\theta} + \frac{1}{\sqrt{1+t2^2}}\varepsilon_{t2} + \boxed{\eta_t}, \qquad |\theta| \le 4u,$$

$$\varepsilon_{cs} = \boxed{\phi - \frac{t^2}{1+t^2}\theta} - \frac{t^2}{(1+t^2)\sqrt{1+t2^2}}\varepsilon_{t2} + \boxed{\eta_{cs}},$$

$$\varepsilon_{sn} = \boxed{\phi + \frac{1}{1+t^2}\theta} + \frac{1}{(1+t^2)\sqrt{1+t2^2}}\varepsilon_{t2} + \boxed{\eta_{sn}}, \qquad |\phi| \le 3u.$$

Well Behaved Matrix Pairs

Let A and B are positive definite matrices which can be well scaled symmetrically, i.e. the condition number of

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

are small. We have

$$a_{ij}^{(S)} = \frac{a_{ij}}{\sqrt{a_{ii} a_{jj}}}, \quad b_{ij}^{(S)} = \frac{b_{ij}}{\sqrt{b_{ii} b_{jj}}} = \frac{b_{ij}}{\sqrt{1 \cdot 1}} = b_{ij} \Rightarrow B_S = B.$$

Accuracy in the Critical Points

If $\mu = \max\left\{ \left| \frac{2a_{ij}}{2a_{ij} - (a_{ii} + a_{jj})b_{ij}} \right|, \left| \frac{(a_{ii} + a_{jj})b_{ij}}{2a_{ij} - (a_{ii} + a_{jj})b_{ij}} \right| \right\}$,

then $|\varepsilon_{t2}| \leq (2\mu + 6.505)u$

and $|t2| = \frac{2\sqrt{a_{ii}a_{jj}}}{a_{ii} + a_{jj}} \cdot \left| \frac{a_{ij}^{(S)}}{\gamma \cdot \tau} \right| \cdot \frac{1}{\mu} = \left| \frac{b_{ij}^{(S)}}{\gamma \cdot \tau} \right| \cdot \frac{1}{\mu}$,

where $\tau = \sqrt{1 - b_{ij}^2}$ and $\gamma = \frac{a_{ii} - a_{jj}}{a_{ii} + a_{jj}}$ is the relative gap.

Thus, if $|\varepsilon_{t2}|$ is large $\Rightarrow \mu$ is large $\Rightarrow |t2| = O(1/\mu)$.

Accuracy of *t*, *sn*, *cs*

For large μ we have $|\varepsilon_{t2}| = O(2\mu u)$ and $|t2| = O(1/\mu) \Rightarrow$

$$t \approx \frac{t2}{2}\left(1 - \frac{t2^2}{4}\right), \quad sn \approx t\left(1 - \frac{t^2}{2}\right), \quad cs \approx 1 - \frac{t^2}{2} \quad \Rightarrow$$

$$t,\ sn = O(1/2\mu), \quad cs \approx 1.$$

$$\varepsilon_t = \theta + \frac{1}{\sqrt{1+t2^2}}\varepsilon_{t2} + \eta_t,$$

$$\varepsilon_{cs} = \phi - \frac{t^2}{1+t^2}\theta - \frac{t^2}{(1+t^2)\sqrt{1+t2^2}}\varepsilon_{t2} + \eta_{cs} \Rightarrow |\varepsilon_{cs}| < 4u + u$$

$$\varepsilon_{sn} = \phi + \frac{1}{1+t^2}\theta + \frac{1}{(1+t^2)\sqrt{1+t2^2}}\varepsilon_{t2} + \eta_{sn}, \quad |\phi|, |\theta| \le 4u.$$

$$\frac{1}{4\mu^2} \qquad 2\mu u$$

Accuracy of transformation matrix

For large μ: $|\varepsilon_{t2}|$, $|\varepsilon_t|$, $|\varepsilon_{sn}|$ are large and $|\varepsilon_{cs}|$ is small,

but $|t2|$, $|t|$ and $|sn|$ are small and $cs \approx 1$,

In the expressions for the errors in the elements of transformation matrix, $c1, s1, c2$ and $s2$

$$\varepsilon_{cs} \quad \text{and} \quad t \cdot \varepsilon_{sn} \quad \text{appear!}$$

Since $\varepsilon_{cs} = O(u)$ and $t \cdot \varepsilon_{sn} = O\left(\frac{1}{2\mu} \cdot 2\mu u\right) = O(u)$,

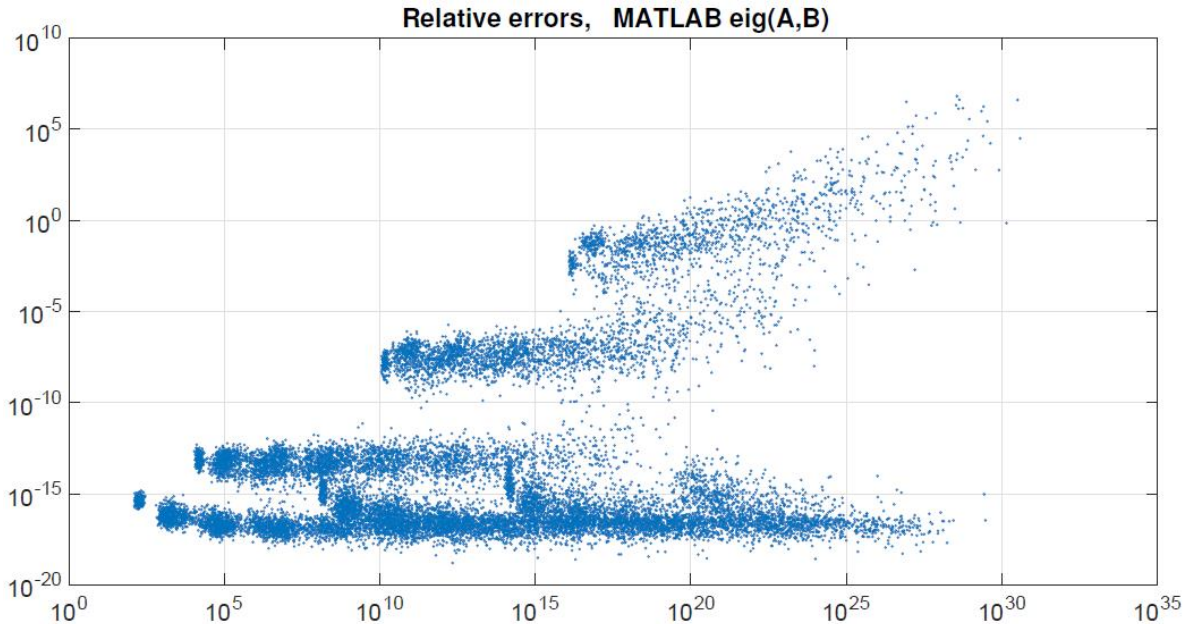
we have $\varepsilon_{c1}, \varepsilon_{s1}, \varepsilon_{c2}, \varepsilon_{s2} = O(u)$.

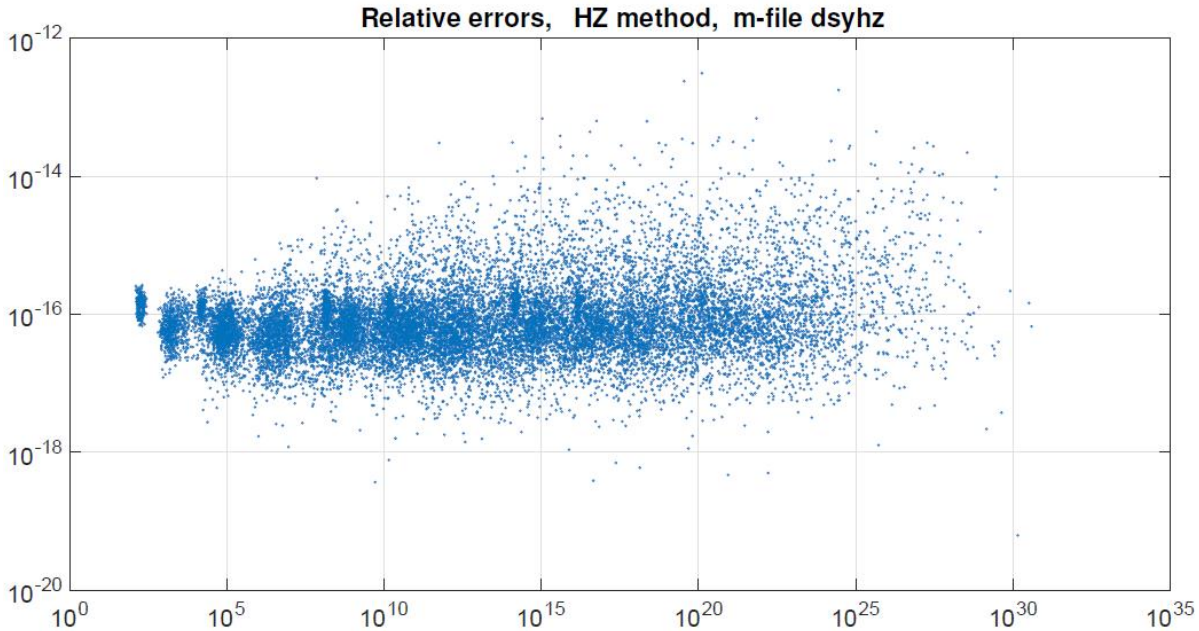
Accuracy of the Method

On the class of positive definite matrices which can be well scaled symmetrically, i.e. the condition numbers of the scaled matrices are small, HZ method computes the eigenvalues of

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

to high relative accuracy.


Relative errors, MATLAB eig(A,B)
10^{10}
10^{5}
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{0}
10^{5}
10^{10}
10^{15}
10^{20}
10^{25}
10^{30}
10^{35}


Relative errors, HZ method, m-file dsyhz
10^{-12}
10^{-14}
10^{-16}
10^{-18}
10^{-20}
10^{0}
10^{5}
10^{10}
10^{15}
10^{20}
10^{25}
10^{30}
10^{35}

Thank you

for your

attention.