On the accuracy of Hari-Zimmerman Method for solving the positive definite generalized eigenvalue problem

Vjeran Hari
Department of Mathematics

Josip Matejaš
Faculty of Economics and Business

University of Zagreb

Subtle Error Analysis

We have derived error estimates

- without neglecting the terms of higher order (in machine precision) of the errors

and

- by taking into account the signs of the errors.

Using such an aproach we can achieve:

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

Assumptions

Let u denote the round-off unit, i.e.

$$u \in \{\underbrace{2^{-23}, 2^{-24}}, \ \underbrace{2^{-52}, 2^{-53}}, \ 2^{-64}, \ 10^{-12}\}.$$

single $1.2 \cdot 10^{-7}$

double $2.2 \cdot 10^{-16}$

extended $5.4 \cdot 10^{-20}$

calculator

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

Standard model of aritmetic

We use the standard model of arithmetic 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

To obtain the exact expressions for the erros, we use

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

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

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

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,

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

Now, we have

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

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

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 HZ method

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

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

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

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

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

CP1

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

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

$$\boxed{\begin{array}{ll} c1 = (\rho \cdot cs - \xi \cdot sn) / \tau; & s1 = (\rho \cdot sn + \xi \cdot cs) / \tau \\ c2 = (\rho \cdot cs + \xi \cdot sn) / \tau; & s2 = (\rho \cdot sn - \xi \cdot cs) / \tau \end{array}}$$

CP2

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

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

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

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

% update the rest of rows and columns i and j

....

until convergence

Critical point 1

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

Let $k = \dfrac{a_{ij}}{a_{ii} + a_{jj}} : \dfrac{b_{ij}}{1+1}$, $\quad \mu = \max\left\{ \left|\dfrac{1}{1-k}\right|, \left|\dfrac{1}{1/k-1}\right| \right\}.$

If $\mu \le \dfrac{1}{\sqrt{u}}$ then $\mathrm{fl}(t2) = (1 + \varepsilon_{t2})t2,$ where

$$\begin{aligned} |\varepsilon_{t2}| &\le (2\mu + 6.5)u + (14.00001\mu + 40.37728)u^2 \\ &\le (2\mu + 6.50484)u. \end{aligned}$$

μ is large if $k \approx 1$ which means that the 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

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

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

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

Critical point 2

$$c1, c2 = \frac{\rho \cdot cs \mp \xi \cdot sn}{\tau}, \quad s1, s2 = \frac{\rho \cdot sn \pm \xi \cdot cs}{\tau},$$

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

Let $\nu = \max\left\{ \left| \frac{1}{1 \pm \frac{\rho}{\xi} \cdot t} \right|, \left| \frac{1}{1 \pm \frac{\xi}{\rho} \cdot \frac{1}{t}} \right| \right\}, \quad t = \frac{sn}{cs}.$

If $\mathrm{fl}(z) = (1 + \varepsilon_z) z, \; z \in \{c1, c2, s1, s2\},$ then

$$\max\{|\varepsilon_{c1}|, |\varepsilon_{c2}|, |\varepsilon_{s1}|, |\varepsilon_{s2}|\} \le \left[\nu(2.0058\mu + 13.3118) + 9.0032\right] \cdot u$$

$$\text{Large } \nu \quad \Leftrightarrow \quad t \approx \pm \frac{\xi}{\rho} = \pm \frac{b_{ij}}{1 + \sqrt{1 - b_{ij}^2}}$$

$$\Leftrightarrow \quad t2 \approx \pm \frac{b_{ij}}{\sqrt{1 - b_{ij}^2}}$$

$$\Leftrightarrow \quad \left\{ \frac{a_{ij}}{a_{ii}} \approx \frac{b_{ij}}{1} \quad or \quad \frac{a_{ij}}{a_{jj}} \approx \frac{b_{ij}}{1} \right\}$$

$$\left(\text{note that } 1 = b_{ii} = b_{jj}\right),$$

which is proportionality for first or second row and column.