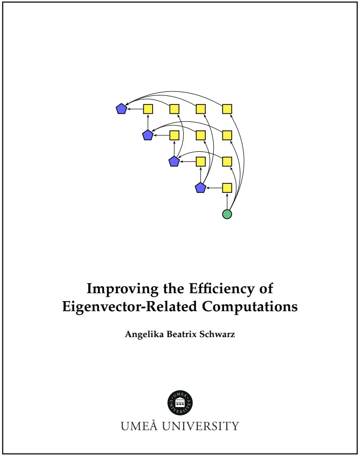
Towards BLAS-3 Robust Solvers in LAPACK

Angelika Schwarz

BLIS Retreat 2022

Disclaimer: All results are from the author's PhD thesis or the ongoing integration into reference LAPACK conducted in the author's free time.

In particular, the work and the results are not related to the author's current or previous employer.



Motivation 1

An eigenvector corresponding to λ can be obtained from the Schur form

$$\begin{bmatrix} T_{11} & t_{12} & T_{13} \\ 0 & \lambda & t_{23} \\ 0 & 0 & T_{33} \end{bmatrix}$$

via solving $(T_{11} - \lambda I)x = -t_{12}$.

> *[...] it appears that one could just use the Level 2 BLAS routine [trsv] for solving triangular systems [...]. Unfortunately we can not, because [...] we anticipate solving ill-conditioned systems which could lead to overflow. In the case of condition estimation, we want a condition estimate as a warning if overflow is possible, since overflow is generally fatal and to be avoided.*
>
> *[Demmel, 1992, p.13]*

Motivation 2 ([Kjelgaard Mikkelsen et al., 2019, Sec.8])

$$\underbrace{\begin{bmatrix} 1 & -2 & & & \\ & 1 & -2 & & \\ & & \ddots & \ddots & \\ & & & 1 & -2 \\ & & & & 1 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{b}$$

The exact solution is

$$y = \begin{bmatrix} 2^{m-1} \\ 2^{m-2} \\ \vdots \\ 2^1 \\ 2^0 \end{bmatrix}.$$

- `dtrsv` introduces `inf` when $m \geq 1025$.
- The system is well-conditioned

$$\frac{\| |T^{-1}| \, |T| \, |y| \|_\infty}{\|y\|_\infty} = 2m - 1$$

To deal with potential overflow, we had to write new versions of all the triangular solvers in LAPACK which scaled in the innermost loop to avoid overflow.

[Demmel, 1992, p.13]

Instead of solving $Ty = b$, solve $Tx = \alpha b$, $\alpha \leq 1$, representing $y = \alpha^{-1}x$.

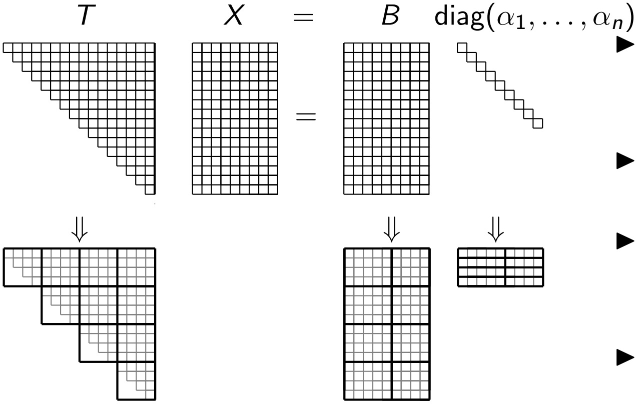
Robust Solvers in LAPACK

Purpose	LAPACK 3.10	Plan/Status
triangular solve $Tx = \alpha b$	`LATRS`	PR up for `LATRS3`
triangular banded solve	`LATBS`	TODO
eigenvectors from Schur matrix $(S - \lambda I)x = \alpha b$	`TREVC(3)`	algorithm update, WIP
eigenvectors from Hessenberg matrix $(H - \lambda I)x^{(1)} = \alpha x^{(0)}$	`HSEIN`	RQ-based replacement, WIP
generalized eigenvectors $(S - \lambda T)x = \alpha b$	`TGEVC`	TODO
triangular Sylvester equation $S_1 X + X S_2 = \alpha C$	`TRSYL`	PR up for `TRSYL3`

- LAPACK 3.10 solvers are at most BLAS-2
- Called from condition number estimators `[ge,tr,po]con`
- Used to solve the general Sylvester equation

Robust triangular solve LATRS ⟶ LATRS3 (PR)

Extend LATRS solving $Tx = \alpha b$ [Anderson, 1991] to multiple right-hand sides using BLAS-3 [Kjelgaard Mikkelsen et al., 2019]:



- LATRS3 structurally identical to blocked triangular solve
- Small triangular solve through LATRS
- Preprocess linear updated to allow usage of GEMM
- Local scale factors for each column segment

Linear Update in LATRS3 (PR)

Compute with overflow
$B_{ik}\mathrm{diag}(\tilde{\alpha}_{ik_1}^{-1}, \ldots, \tilde{\alpha}_{ik_n}^{-1}) \leftarrow B_{ik}\mathrm{diag}(\alpha_{ik_1}^{-1}, \ldots, \alpha_{ik_n}^{-1}) - T_{ij}(X_{jk}\mathrm{diag}(\alpha_{jk_1}^{-1}, \ldots, \alpha_{jk_n}^{-1}))$

for $\ell \leftarrow k_1 : k_n$ **do**
- Consistent scaling $\gamma_\ell \leftarrow \min\{\alpha_{i\ell}, \alpha_{j\ell}\}$
- Scale $b_{i\ell} \leftarrow \frac{\alpha_{i\ell}}{\gamma_\ell} x_{i\ell}$; $x_{j\ell} \leftarrow \frac{\alpha_{j\ell}}{\gamma_\ell} x_{j\ell}$
- Compute $\xi_\ell \leq 1$ such that $\|\xi_\ell b_{i\ell}\|_\infty + \|T_{ij}\|_\infty \|\xi_\ell x_{j\ell}\|_\infty \leq \Omega$
- Scale $b_{i\ell} \leftarrow \xi_\ell x_{i\ell}$; $x_{j\ell} \leftarrow \xi_\ell x_{j\ell}$
- Update local scale factor $\tilde{\alpha}_{i,\ell} \leftarrow \gamma_\ell \xi_\ell$

$B_{ik} \leftarrow B_{ik} - T_{ij} X_{jk}$ (GEMM)

Linear Update in LATRS3 (PR)

Compute with overflow

$B_{ik}\mathrm{diag}(\tilde{\alpha}_{ik_1}^{-1},\dots,\tilde{\alpha}_{ik_n}^{-1}) \leftarrow B_{ik}\mathrm{diag}(\alpha_{ik_1}^{-1},\dots,\alpha_{ik_n}^{-1}) - T_{ij}(X_{jk}\mathrm{diag}(\alpha_{jk_1}^{-1},\dots,\alpha_{jk_n}^{-1}))$

for $\ell \leftarrow k_1 : k_n$ **do**

- Consistent scaling $\gamma_\ell \leftarrow \min\{\alpha_{i\ell},\alpha_{j\ell}\}$
- Scale $b_{i\ell} \leftarrow \frac{\alpha_{i\ell}}{\gamma_\ell} x_{i\ell}$; $x_{j\ell} \leftarrow \frac{\alpha_{j\ell}}{\gamma_\ell} x_{j\ell}$
- Compute $\xi_\ell \leq 1$ such that $\|\xi_\ell b_{i\ell}\|_\infty + \|T_{ij}\|_\infty \|\xi_\ell x_{j\ell}\|_\infty \leq \Omega$
- Scale $b_{i\ell} \leftarrow \xi_\ell x_{i\ell}$; $x_{j\ell} \leftarrow \xi_\ell x_{j\ell}$
- Update local scale factor $\tilde{\alpha}_{i,\ell} \leftarrow \gamma_\ell \xi_\ell$

$B_{ik} \leftarrow B_{ik} - T_{ij} X_{jk}$ (GEMM)

Example with 2 columns:

$$\left[\left(\tfrac{1}{2}\right)^{-1}\left[\; b_{i1} \;\right] \;\middle|\; \left(\tfrac{1}{8}\right)^{-1}\left[\; b_{i2} \;\right]\right] - T_{ij}\left[\left(\tfrac{1}{4}\right)^{-1}\left[\; x_{j1} \;\right] \;\middle|\; \left(\tfrac{1}{1}\right)^{-1}\left[\; x_{j2} \;\right]\right]$$

Linear Update in LATRS3 (PR)

Compute with overflow
$B_{ik}\mathrm{diag}(\tilde{\alpha}_{ik_1}^{-1}, \ldots, \tilde{\alpha}_{ik_n}^{-1}) \leftarrow B_{ik}\mathrm{diag}(\alpha_{ik_1}^{-1}, \ldots, \alpha_{ik_n}^{-1}) - T_{ij}(X_{jk}\mathrm{diag}(\alpha_{jk_1}^{-1}, \ldots, \alpha_{jk_n}^{-1}))$

for $\ell \leftarrow k_1 : k_n$ **do**
- Consistent scaling $\gamma_\ell \leftarrow \min\{\alpha_{i\ell}, \alpha_{j\ell}\}$
- Scale $b_{i\ell} \leftarrow \frac{\alpha_{i\ell}}{\gamma_\ell} x_{i\ell}$; $x_{j\ell} \leftarrow \frac{\alpha_{j\ell}}{\gamma_\ell} x_{j\ell}$
- Compute $\xi_\ell \leq 1$ such that $\|\xi_\ell b_{i\ell}\|_\infty + \|T_{ij}\|_\infty \|\xi_\ell x_{j\ell}\|_\infty \leq \Omega$
- Scale $b_{i\ell} \leftarrow \xi_\ell x_{i\ell}$; $x_{j\ell} \leftarrow \xi_\ell x_{j\ell}$
- Update local scale factor $\tilde{\alpha}_{i,\ell} \leftarrow \gamma_\ell \xi_\ell$

$B_{ik} \leftarrow B_{ik} - T_{ij} X_{jk}$ (GEMM)

Example with 2 columns:

$$\left[\left(\tfrac{1}{4}\right)^{-1} \left[\tfrac{1}{2} b_{i1} \right] \;\middle|\; \left(\tfrac{1}{8}\right)^{-1} \left[b_{i2} \right] \right] - T_{ij} \left[\left(\tfrac{1}{4}\right)^{-1} \left[x_{j1} \right] \;\middle|\; \left(\tfrac{1}{8}\right)^{-1} \left[\tfrac{1}{8} x_{j2} \right] \right]$$

Linear Update in LATRS3 (PR)

Compute with overflow
$B_{ik}\mathrm{diag}(\tilde{\alpha}_{ik_1}^{-1}, \dots, \tilde{\alpha}_{ik_n}^{-1}) \leftarrow B_{ik}\mathrm{diag}(\alpha_{ik_1}^{-1}, \dots, \alpha_{ik_n}^{-1}) - T_{ij}(X_{jk}\mathrm{diag}(\alpha_{jk_1}^{-1}, \dots, \alpha_{jk_n}^{-1}))$

for $\ell \leftarrow k_1 : k_n$ **do**
- Consistent scaling $\gamma_\ell \leftarrow \min\{\alpha_{i\ell}, \alpha_{j\ell}\}$
- Scale $b_{i\ell} \leftarrow \frac{\alpha_{i\ell}}{\gamma_\ell} x_{i\ell}$; $x_{j\ell} \leftarrow \frac{\alpha_{j\ell}}{\gamma_\ell} x_{j\ell}$
- Compute $\xi_\ell \leq 1$ such that $\|\xi_\ell b_{i\ell}\|_\infty + \|T_{ij}\|_\infty \|\xi_\ell x_{j\ell}\|_\infty \leq \Omega$
- Scale $b_{i\ell} \leftarrow \xi_\ell x_{i\ell}$; $x_{j\ell} \leftarrow \xi_\ell x_{j\ell}$
- Update local scale factor $\tilde{\alpha}_{i,\ell} \leftarrow \gamma_\ell \xi_\ell$

$B_{ik} \leftarrow B_{ik} - T_{ij}X_{jk}$ (GEMM)

Example with 2 columns:

$$\left(\tfrac{1}{4}\right)^{-1}\left[\ \tfrac{1}{2} b_{i1}\ \right] \ \left(\tfrac{1}{8}\right)^{-1}\left[\ b_{i2}\ \right] - T_{ij} \left(\tfrac{1}{4}\right)^{-1}\left[\ x_{j1}\ \right] \ \left(\tfrac{1}{8}\right)^{-1}\left[\ \tfrac{1}{8} x_{j2}\ \right]$$

Linear Update in LATRS3 (PR)

Compute with overflow
$B_{ik}\mathrm{diag}(\tilde{\alpha}_{ik_1}^{-1}, \ldots, \tilde{\alpha}_{ik_n}^{-1}) \leftarrow B_{ik}\mathrm{diag}(\alpha_{ik_1}^{-1}, \ldots, \alpha_{ik_n}^{-1}) - T_{ij}(X_{jk}\mathrm{diag}(\alpha_{jk_1}^{-1}, \ldots, \alpha_{jk_n}^{-1}))$

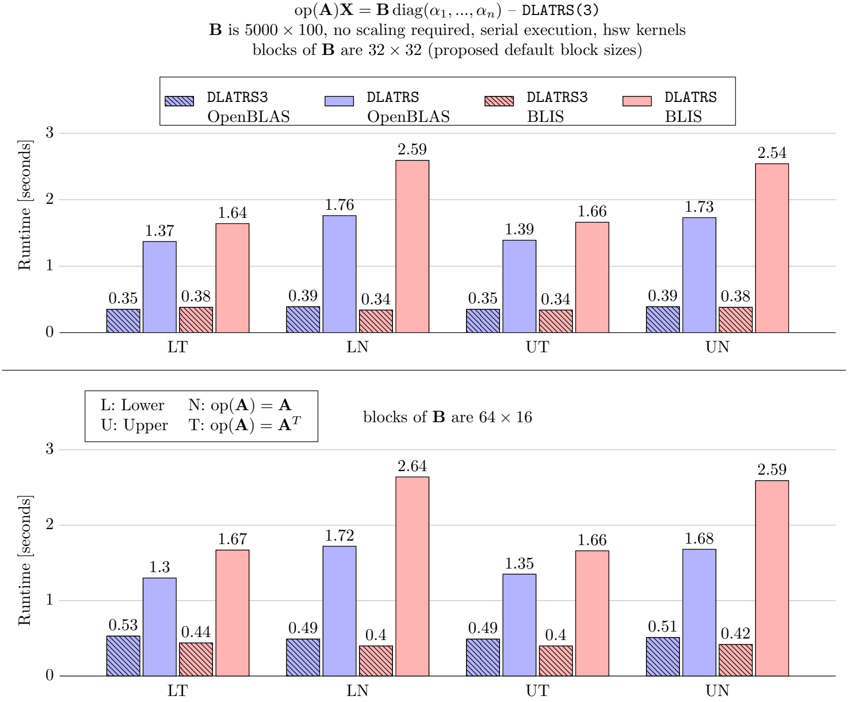
for $\ell \leftarrow k_1 : k_n$ **do**

- Consistent scaling $\gamma_\ell \leftarrow \min\{\alpha_{i\ell}, \alpha_{j\ell}\}$
- Scale $b_{i\ell} \leftarrow \frac{\alpha_{i\ell}}{\gamma_\ell} x_{i\ell}$; $x_{j\ell} \leftarrow \frac{\alpha_{j\ell}}{\gamma_\ell} x_{j\ell}$
- Compute $\xi_\ell \leq 1$ such that $\|\xi_\ell b_{i\ell}\|_\infty + \|T_{ij}\|_\infty \|\xi_\ell x_{j\ell}\|_\infty \leq \Omega$
- Scale $b_{i\ell} \leftarrow \frac{\alpha_{i\ell}}{\gamma_\ell} x_{i\ell}$; $x_{j\ell} \leftarrow \frac{\alpha_{j\ell}}{\gamma_\ell} x_{j\ell}$
- Update local scale factor $\tilde{\alpha}_{i,\ell} \leftarrow \gamma_\ell \xi_\ell$

$B_{ik} \leftarrow B_{ik} - T_{ij} X_{jk}$ (GEMM)

Example with 2 columns:

$$\left[\left(\xi_\ell \tfrac{1}{4}\right)^{-1} \left[\; \xi_\ell\left(\tfrac{1}{2} b_{i1}\right) \; \right] \;\middle|\; \left(\tfrac{1}{8}\right)^{-1} \left[\; b_{i2} \; \right] \right] - T_{ij} \left[\left(\xi_\ell \tfrac{1}{4}\right)^{-1} \left[\; \xi_\ell x_{j1} \; \right] \;\middle|\; \left(\tfrac{1}{8}\right)^{-1} \left[\; \tfrac{1}{8} x_{j2} \; \right] \right]$$



op(**A**)**X** = **B** diag($\alpha_1, ..., \alpha_n$) – DLATRS(3)
B is 5000 × 100, no scaling required, serial execution, hsw kernels
blocks of **B** are 32 × 32 (proposed default block sizes)
DLATRS3 OpenBLAS
DLATRS OpenBLAS
DLATRS3 BLIS
DLATRS BLIS
Runtime [seconds]
0
1
2
3
0.35
1.37
0.38
1.64
0.39
1.76
0.34
2.59
0.35
1.39
0.34
1.66
0.39
1.73
0.38
2.54
LT
LN
UT
UN
L: Lower
N: op(**A**) = **A**
U: Upper
T: op(**A**) = $\mathbf{A}^T$
blocks of **B** are 64 × 16
Runtime [seconds]
0
1
2
3
0.53
1.3
0.44
1.67
0.49
1.72
0.4
2.64
0.49
1.35
0.4
1.66
0.51
1.68
0.42
2.59
LT
LN
UT
UN

Issues LATRS $Tx = \alpha b$ [Anderson, 1991]

- Cause of $\alpha = 0$ not clear: $a_{jj} = 0$ or *badly scaled*?

  SCALE is DOUBLE PRECISION
  *The scaling factor s for the triangular system A * x = s*b [...]If SCALE = 0, the matrix A is singular or badly scaled, and the vector x is an exact or approximate solution to A*x = 0.*

  DLATRS documentation (LAPACK 3.10)

- Upper bounds based on columns norms of T, x and b can overestimate growth. Entries can be flushed unnecessarily.

$$\alpha^{-1}x = \left(\frac{1}{2^{951}}\right)^{-1} \begin{bmatrix} 2^{30} \\ \vdots \\ 2^{-1074} \\ 0 \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \textit{flushed} \\ \textit{flushed} \end{matrix}$$

Discussion

- Should the new routines be a drop-in replacement of the existing solvers? Should they produce identical scaled representations?
- `LATRS`: $\alpha = 0$ signals either that $a_{jj} = 0$ or *badly scaled* matrices. It is impossible to tell what the cause is. Should `INFO = J` be used to signal $a_{jj} = 0$?
- Upcoming change with `NaN`/`Inf` propagation: Should α be an integer representing the scale factor $\alpha = \frac{1}{2^i}$? This would de facto guarantee $\alpha > 0$ for all non-singular problems.

References

Anderson, E. (1991).
Robust triangular solves for use in condition estimation.
LAPACK Working Note 36, USA.

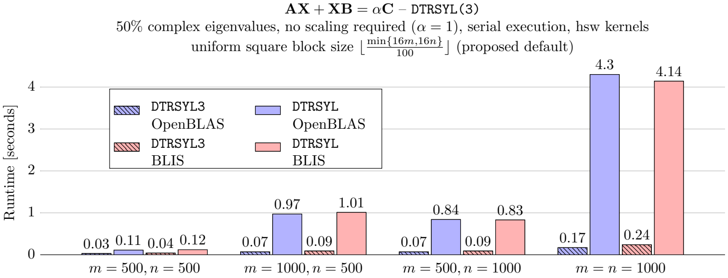
Demmel, J. (1992).
Open Problems in Numerical Linear Algebra.
LAPACK Working Note 47.

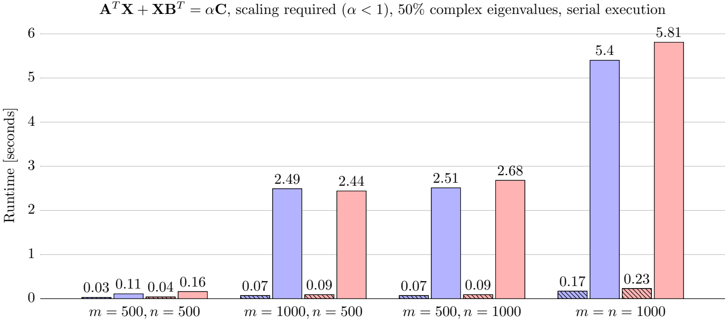
Kjelgaard Mikkelsen, C. C., Schwarz, A. B., and Karlsson, L. (2019).
Parallel robust solution of triangular linear systems.
Concurrency and Computation: Practice and Experience, 31(19):e5064.

Schwarz, A. (2022).
Robust Level-3 BLAS Inverse Iteration from the Hessenberg Matrix.
ACM Trans. Math. Softw., 48(3).

Schwarz, A. and Kjelgaard Mikkelsen, C. C. (2020).
Robust task-parallel solution of the triangular sylvester equation.
In Wyrzykowski, R., Deelman, E., Dongarra, J., and Karczewski, K., editors, *Parallel Processing and Applied Mathematics*, pages 82–92, Cham. Springer International Publishing.

TRSYL → TRSYL3 [Schwarz and Kjelgaard Mikkelsen, 2020]





Preliminary results DHSEIN [Schwarz, 2022]

Computation of a single (right) eigenvectors by inverse iteration from the Hessenberg matrix:

$$(H - \lambda I)x^{(1)} = \alpha x^{(0)}$$

- 25% selected eigenvalues
- first start vector always leads to convergence in a single iteration
- only right eigenvectors
- Change algorithm from LU (LAPACK 3.10) to RQ factorization
- No drop-in replacement: difference workspace requirements

	LAPACK 3.10	proposed
$n = 1000$	0.11s	0.78s
$n = 2000$	0.45s	10.20s