Computing the Polar Decomposition in Matrix Groups

Nick Higham
Department of Mathematics
University of Manchester

higham@ma.man.ac.uk
http://www.ma.man.ac.uk/~higham/

Joint work with Niloufer Mackey, D. Steven Mackey, and Françoise Tisseur.

THE UNIVERSITY of MANCHESTER

Group Background

Given nonsingular M and $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$,

$$\langle x, y \rangle_{\mathrm{M}} = \begin{cases} x^T M y, & \text{real or complex bilinear forms,} \\ x^* M y, & \text{sesquilinear forms.} \end{cases}$$

Define automorphism group

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n \times n} : \langle Ax, Ay \rangle_{\mathrm{M}} = \langle x, y \rangle_{\mathrm{M}},\ \forall x, y \in \mathbb{K}^n \,\}.$$

Recall adjoint $A^\star$ of $A \in \mathbb{K}^{n \times n}$ wrt $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ defined by

$$\langle Ax, y \rangle_{\mathrm{M}} = \langle x, A^\star y \rangle_{\mathrm{M}} \quad \forall x, y \in \mathbb{K}^{n \times n}.$$

Can show: $$A^\star = \begin{cases} M^{-1} A^T M, & \text{for bilinear forms,} \\ M^{-1} A^* M, & \text{for sesquilinear forms,} \end{cases}$$

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n \times n} : A^\star = A^{-1} \,\}.$$

Some Automorphism Groups

Space	M	$A^\star$	**Automorphism group**, $\mathbb{G}$
Groups corresponding to a bilinear form			
$\mathbb{R}^n$	I	A^T	Real orthogonals
$\mathbb{C}^n$	I	A^T	Complex orthogonals
$\mathbb{R}^n$	$\Sigma_{p,q}$	$\Sigma_{p,q}A^T\Sigma_{p,q}$	Pseudo-orthogonals
$\mathbb{R}^n$	R	RA^TR	Real perplectics
$\mathbb{R}^{2n}$	J	$-JA^TJ$	Real symplectics
$\mathbb{C}^{2n}$	J	$-JA^TJ$	Complex symplectics
Groups corresponding to a sesquilinear form			
$\mathbb{C}^n$	I	A^*	Unitaries
$\mathbb{C}^n$	$\Sigma_{p,q}$	$\Sigma_{p,q}A^*\Sigma_{p,q}$	Pseudo-unitaries
$\mathbb{C}^{2n}$	J	$-JA^*J$	Conjugate symplectics

$$R=\begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \qquad J=\begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}, \qquad \Sigma_{p,q}=\begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$$

Questions

Recall polar decomposition of $A \in \mathbb{C}^{n\times n}$:

$$A = UH, \qquad U^*U = I, \quad H = H^* \geq 0.$$

If $A \in \mathbb{G}$

- When do its polar factors lie in the group?
- How can we exploit any group structure when computing U and H?

Structure of Polar Factors

Let $\mathfrak{U}$ denote set of autom. groups for which M is unitary.

Two results of Mackey, Mackey & Tisseur, 2003:

Theorem 1 *Let* $\mathbb{G} \in \mathfrak{U}$ *and* $A \in \mathbb{G}$. *Then in* $A = UH$ *the polar factors* U *and* H *also belong to* $\mathbb{G}$.

Theorem 2 *Let* $\mathbb{G} \in \mathfrak{U}$ *and* $A \in \mathbb{G}$. *The singular values of* A *occur in reciprocal pairs* σ *and* $1/\sigma$, *with the same multiplicity.*

Structure-Preserving Iterations

Theorem 3 *Consider*

$$Z_{k+1} = Z_k\, P_{mm}(I - Z_k^* Z_k)\, Q_{mm}(I - Z_k^* Z_k)^{-1}, \quad Z_0 = A,$$

where $P_{mm}(t)/Q_{mm}(t)$ *is the* $[m/m]$ *Padé approximant to* $(1-t)^{-1/2}$ *and* $m \geq 1$. *If* $\mathbb{G} \in \mathfrak{U}$ *and* $A \in \mathbb{G}$ *then*

- $Z_k \in \mathbb{G}$ *for all* k,
- $Z_k \to U$ *at order* $2m+1$.

Structure-Preserving Iterations

Theorem 4 *Consider*

$$Z_{k+1} = Z_k\, P_{mm}(I - Z_k^* Z_k)\, Q_{mm}(I - Z_k^* Z_k)^{-1}, \quad Z_0 = A,$$

where $P_{mm}(t)/Q_{mm}(t)$ *is the* $[m/m]$ *Padé approximant to* $(1-t)^{-1/2}$ *and* $m \geq 1$. *If* $\mathbb{G} \in \mathfrak{U}$ *and* $A \in \mathbb{G}$ *then*

- $Z_k \in \mathbb{G}$ *for all* k,
- $Z_k \to U$ *at order* $2m+1$.

Iterations $z_{k+1} = f(z_k)$:

m	$f(x)$
1	$\dfrac{x(3+x^2)}{1+3x^2}$
2	$\dfrac{x(5+10x^2+x^4)}{1+10x^2+5x^4}$

Iterations (all with $X_0 = A$)

Cubic (structure-preserving):

$$X_{k+1} = \frac{1}{3} X_k [I + 8(I + 3X_k^* X_k)^{-1}].$$

Quintic (structure-preserving):

$$x_{k+1} = x_k \left[\frac{1}{5} + \frac{8}{5x_k^2 + 7 - \dfrac{16}{5x_k^2 + 3}} \right].$$

Scaled Newton iteration (not structure-preserving):

$$X_{k+1} = \frac{1}{2} \left[\gamma^{(k)} X_k + \frac{1}{\gamma^{(k)}} X_k^{-*} \right], \quad \gamma^{(k)} = \left(\frac{\|X_k^{-1}\|_F}{\|X_k\|_F} \right)^{1/2}.$$

$$X_0 \in \mathbb{G} \in \mathfrak{U} \quad \Rightarrow \quad \gamma^{(0)} = 1.$$

Experiment

Random symplectic $A \in \mathbb{R}^{12\times 12}$, $\|A\|_2 = 310 = \|A^{-1}\|_2$.

$$\mu_{\mathbb{O}}(A) = \frac{\|A^*A - I\|_2}{\|A\|_2^2}, \qquad \mu_{\mathbb{G}}(A) = \frac{\|A^\star A - I\|_2}{\|A\|_2^2}.$$

k	Newton (scaled)		Cubic	
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$
0	1.0e+0	7.0e-18	1.0e+0	7.0e-18
1	1.0e+0	1.0e+0	1.0e+0	8.9e-17
2	8.6e-01	8.6e-01	1.0e+0	8.1e-16
3	2.0e-01	2.0e-01	9.9e-01	6.3e-15
4	3.2e-03	3.2e-03	9.4e-01	5.0e-14
5	9.0e-07	9.0e-07	5.7e-01	2.8e-13
6	6.0e-14	1.3e-13	3.6e-02	5.2e-13
7	4.3e-16	1.1e-13	3.2e-06	5.3e-13
8			3.8e-16	5.3e-13

Newton Behaviour

Theorem 5 *Let $\mathbb{G} \in \mathfrak{U}$, $A \in \mathbb{G}$, and X_k be the Newton iterates, either unscaled or with Frobenius scaling. Then $X_k^\star = X_k^*$, $k \geq 1$. Moreover,*

$$M X_k = X_k M, \quad \textit{real bilinear, complex sesquilinear forms},$$
$$M X_k = \overline{X_k} M, \quad \textit{complex bilinear forms}.$$

Implications:

- ★ Tethering.
- ★ Structure in X_k:

$$A \text{ pseudo-orthogonal} \Rightarrow X_k \text{ block-diagonal},$$
$$A \text{ symplectic} \Rightarrow X_k = \begin{bmatrix} E_k & F_k \\ -F_k & E_k \end{bmatrix}.$$

Convergence Tests

Padé iteration function f_{mm} satisfies

$$f_{mm}(\sigma^{-1}) = f_{mm}(\sigma)^{-1}, \qquad f_{mm}(1) = 1,$$
$$1 < \sigma \;\Rightarrow\; 1 < f_{mm}(\sigma) < \sigma,$$
$$1 \le \mu < \sigma \;\Rightarrow\; f_{mm}(\mu) < f_{mm}(\sigma).$$

Can show that for $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$:

$$\|U - Z_k\|_2 = f_{mm}^{(k)}(\sigma_1) - 1\,.$$

Moreover, $\|Z_k\|_F$ decreases monotonically. Stop when

$$\frac{\|\widehat{Z}_{k+1}\|_F}{\|\widehat{Z}_k\|_F} \ge 1 - \delta\,.$$

Scaled Newton: latter test applicable for $k \ge 1$.

Conclusions

Structured iteration (cubic, quintic) versus *scaled* Newton

- ★ Newton has slightly better observed numerical stability.
- ★ Newton usually requires the fewest flops.
- ★ Convergence prediction possible with structured iterations.
- ★ Which is best depends on matrix A, group $\mathbb{G}$, and user's accuracy requirements.

Have analogous results for matrix sign function, with *no restrictions on* $\mathbb{G}$.

- ▶ `http://www.ma.man.ac.uk/~nareports/narep426.pdf`