CERTIFIED AND SYMBOLIC-NUMERIC COMPUTATION **2023, Lyon, May 22-26**

Centro Nacional de la Jubilaccion Scientifica

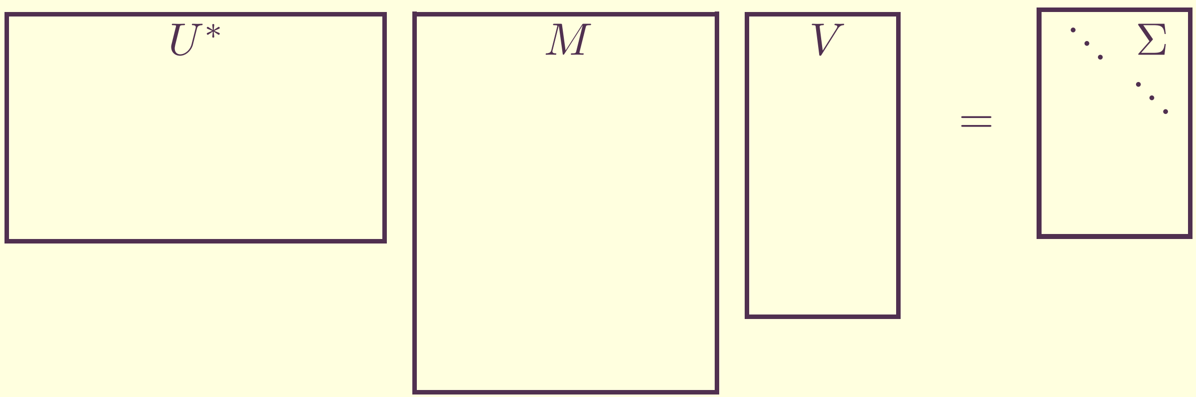
High Order Methods For The SVD

Jean-Claude Yakoubsohn

Work in collaboration with **Diego Armentano** (Universidad de la Republica,Uruguay)

Let $M \in \mathbb{C}^{m \times n}$, $m \geqslant n$, be a matrix.

We know that there exists two Stiefel matrices $U \in \mathbb{C}^{m \times \ell}$,$V \in \mathbb{C}^{n \times q}$ and a diagonal matrix $\Sigma \in \mathbb{C}^{\ell \times q}$ such that



A wonderfull paper on the applications of the SVD :

Martin, C. D., and Porter, M. A. The extraordinary svd. *The American Mathematical Monthly 119*, 10 (2012), 838–851.

The background of this talk comes from

1. Davies, P. I., and Smith, M. I. Updating the singular value decomposition. *Journal of computational and applied mathematics 170*, 1 (2004), 145–167.

2. Joris van der Hoeven, and Jean-Claude Yakoubsohn. Certified Singular Value Decomposition, 2018, HAL 01941987.

3. Rima Khouja, Bernard Mourrain and Jean-Claude Yakoubsohn. Newton-type methods for simultaneous matrix diagonalization. *Calcolo*, 2022, 59:38.

We denote

$$\begin{aligned} \Delta &= U^*MV - \Sigma \\ E(U) &= U^*U - I \end{aligned}$$

We consider the system

$$f(U, V, \Sigma) = \begin{pmatrix} E(U) \\ E(V) \\ U^*MV - \Sigma \end{pmatrix}$$

We will use

$$\Delta + \Sigma = U^*MV$$

A map $f: E \to F$ where E and F are two Banach spaces define a method of order $p+1$ if for all convergent sequence $x_{k+1} = f(x_k)$, $k \geqslant 0$, one has

$$\|x_k - x^*\| \leqslant c2^{-(p+1)^k+1}\|x_0 - x^*\|$$

where $x^*$is the limit of the sequence $(x_k)_{k \geqslant 0}$ and c a positive constant.

In the Stiefel space we consider multiplicative perturbations such type :

1. $U(I+\Omega)$ where W is an Hermitian matrix.

2. $U(I+\Theta)$ where Θ is a skew Hermitian matrix.

3. $U(I+\Omega)(I+\Theta)$

In the space of diagonal matrices we consider the additive perturbations such type :

1. $\Sigma + S$

The ingredients are :

1. Approximation of the Stiefel group at the order $p+1$. Find Ω such that $U_1 = (I+\Omega)U$ verifies

$$\|\boldsymbol{E}(\boldsymbol{U_1})\| \;=\; \boldsymbol{O}(\|\boldsymbol{E}(\boldsymbol{U})\|^{\boldsymbol{p+1}}).$$

2. Computation of a triplet (X, Y, S) so that

$$\Delta_1 \;=\; (I+X)^*(\Delta+\Sigma)(I+Y)V-\Sigma-S$$

satisfies

$$\|\boldsymbol{\Delta_1}\| \;=\; \boldsymbol{O}(\|\boldsymbol{\Delta}\|^{\boldsymbol{p+1}}).$$

3. The map $\boldsymbol{H_p}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\Sigma}) = \begin{pmatrix} (\boldsymbol{I}+\boldsymbol{\Theta})(\boldsymbol{I}+\boldsymbol{\Omega})\boldsymbol{U} \\ (\boldsymbol{I}+\boldsymbol{\Psi})(\boldsymbol{I}+\boldsymbol{\Lambda})\boldsymbol{V} \\ \boldsymbol{\Sigma}+\boldsymbol{S} \end{pmatrix}$ defines a method of order $p+1$.

Remarks

1. Newton's method is based on the cancellation of $f(x) + Df(x)(x - x_0)$.

2. Here we cancel $f(x) + L(x)(x - x_0)$ where $L(x)$ is a part of $Df(x)$.

3. There is no matrix to invert.

4. We perform only additions and multiplications of matrices.

$$\boldsymbol{E(U + U\,\Omega) \;=\; E(U) + 2\,\Omega + \Omega\,E(U) + E(U)\,\Omega + \Omega^2 + \Omega\,E(U)\,\Omega}$$

If $\boldsymbol{\Omega = -I + (I + E(U))^{-1/2}}$ then $\boldsymbol{E(U(I + \Omega)) = 0}$.

Let us consider the truncated Taylor series of $-1 + (1+u)^{-1/2}$ at $u = 0$ at the order p :

$$s_p(u) \;=\; -\frac{1}{2}u + \frac{3}{8}u^2 + \cdots + (-1)^p c_p\, u^p \quad \text{with} \quad c_k = \frac{1}{2^k\, k!}\prod_{i=1}^{k-1}(2\,i+1).$$

Theorem 1. *Let $p \geqslant 1$. Let U_0 be such that $\|\boldsymbol{E(U_0)}\| \leqslant \boldsymbol{\varepsilon < 1/2}$. Then the sequence*

$$\boldsymbol{U_{i+1} \;=\; U_i(I + E(s_p(U_i))) \qquad i \geqslant 0,}$$

converges at the order $p+1$ to an Stiefel matrix U_∞.

See

Bjork Bowie AN ITERATIVE ALGORITHM FOR COMPUTING THE BEST ESTIMATE OF AN ORTHOGONAL MATRIX

SIAM J. NUMER. ANAL. Vol. 8, No. 2, June 1971

Let $X^* = -X$ and $Y^* = -Y$ and $S = \mathrm{diag}(\Delta)$.

$$\begin{aligned} \boldsymbol{\Delta_1} &= (\boldsymbol{I}+\boldsymbol{X})^*(\boldsymbol{\Delta}+\boldsymbol{\Sigma})(\boldsymbol{I}+\boldsymbol{Y}) - \boldsymbol{\Sigma} - \boldsymbol{S} \\ &= \underline{\boldsymbol{\Delta} - \boldsymbol{S} - \boldsymbol{X}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Delta} + \boldsymbol{Y}\boldsymbol{\Delta}} - \boldsymbol{X}(\boldsymbol{\Delta}+\boldsymbol{\Sigma})\boldsymbol{Y} \end{aligned}$$

Lemma 2. *Let* (X, Y, S) *such that* $\boldsymbol{\Delta} - \boldsymbol{S} - \boldsymbol{X}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{Y} = \boldsymbol{0}$.

1. *Then* (X, Y, S) *is given by explicit formulas.*

2. $\|\boldsymbol{S}\| \leqslant \|\boldsymbol{\Delta}\|$, $\|\boldsymbol{X}\|, \|\boldsymbol{Y}\| \leqslant \boldsymbol{\kappa}\,\|\boldsymbol{\Delta}\|$ *where*

$$\boldsymbol{\kappa} := \boldsymbol{\kappa}(\boldsymbol{\Sigma}) = \max\left(1, \quad \max_i \frac{1}{\boldsymbol{\sigma_i}}, \quad \max_{i \neq j}\left(\frac{1}{|\boldsymbol{\sigma_i} - \boldsymbol{\sigma_j}|} + \frac{1}{\boldsymbol{\sigma_i} + \boldsymbol{\sigma_j}}\right)\right).$$

3. $\|\boldsymbol{\Delta_1}\| = \boldsymbol{O}(\|\boldsymbol{\Delta}\|^2)$.

Proposition 3. *Let* $\Sigma = \mathrm{diag}(\sigma_1, \ldots \sigma_n) \in \mathbb{D}^{m\times n}$ *and* $\Delta = (\delta_{i,j}) \in \mathbb{C}^{m\times n}$*. Consider the diagonal matrix* $S \in \mathbb{R}^{m\times n}$ *and the two skew Hermitian matrices* $X = (x_{i,j}) \in \mathbb{C}^{m\times m}$ *and* $Y = (y_{i,j}) \in \mathbb{C}^{n\times n}$ *that are defined by the following formulas :*

- *For* $1 \leqslant i \leqslant n$*, we take* $S_{i,i} = \mathrm{Re}\, \delta_{i,i}$ *and* $x_{i,i} - y_{i,i} = \frac{\mathrm{Im}\, \delta_{i,i}}{2\, \sigma_i}\,\mathrm{i}$.
- *For* $1 \leqslant i < j \leqslant n$*, we take*

$$x_{i,j} = \frac{1}{2}\left(\frac{\delta_{i,j} + \overline{\delta_{i,j}}}{\sigma_j - \sigma_i} + \frac{\delta_{i,j} - \overline{\delta_{i,j}}}{\sigma_j + \sigma_i}\right),$$

$$y_{i,j} = \frac{1}{2}\left(\frac{\delta_{i,j} - \overline{\delta_{i,j}}}{\sigma_j - \sigma_i} - \frac{\delta_{i,j} - \overline{\delta_{i,j}}}{\sigma_j + \sigma_i}\right).$$

- *For* $n+1 \leqslant i \leqslant m$ *and* $1 \leqslant j \leqslant n$*, we take* $x_{i,j} = \frac{1}{\sigma_j}\delta_{i,j}$
- *For* $n+1 \leqslant i \leqslant m$ *and* $n+1 \leqslant j \leqslant m$*, we take* $x_{i,j} = 0$*.*

Then we have

$$\Delta - S - X\Sigma + \Sigma Y = 0.$$

We ned to generalize $(1-u)(1+u)-1=u^2$.

We remark that

$$\left(1+\sqrt{1+u^2}-u-1\right)\left(1+\sqrt{1+u^2}+u-1\right) \;=\; 1$$

Considering the function $c_p(u)$ the truncated Taylor series of $\sqrt{1+u^2}+u-1$ at $u=0$ at the order p we have

$$(1+c_p(-u))(1+c_p(u))-1 \;=\; O(u^{p+1}).$$

One has

$$c_p(u) \;=\; u+\sum_{k=1}^{\max(k:2k\leqslant p)}(-1)^{k+1}\frac{(2k)!}{4^k(k!)^2(2k-1)}u^{2k}=u+\frac{1}{2}u^2-\frac{1}{8}u^4-\frac{5}{128}u^6+\cdots$$

Let X_1, X_2 two shew Hermitian matrices and $\Delta_1 = (I + c_2(X_1))^*(\Delta + \Sigma)(I + Y) - \Sigma - S_1$

$$\begin{aligned} \Delta_2 &= (I + c_2(X_1 + X_2))^*(\Delta + \Sigma)(I + c_2(Y_1 + Y_2)) - \Sigma - S_1 - S_2 \\ &= \Delta_1 - S_2 - X_2\Sigma + \Sigma Y_2 + \cdots + \end{aligned}$$

Lemma 4. *If* $\Delta - S_1 - X_1\Sigma + \Sigma Y_1 = 0$ *and* $\Delta_1 - S_2 - X_2\Sigma + \Sigma Y_2 = 0$ *one has*

$$\|\Delta_2\| \leqslant O(\|\Delta\|^3)$$

Now with $p = 2$, $X = X_1 + X_2$ and $U_1 = (I_m + c_p(X))U$ then

$$\begin{aligned} E(U_1) &= (I + c_p(-X))E(U)(I + c_p(X)) + (I + c_p(-X))(I + c_p(X)) - I \\ &= (I + c_p(-X))E(U)(I + c_p(X)) + O(X^{p+1}) \end{aligned}$$

Hence

$$\text{order of } \|E(U_1)\| = \min(\text{order of } \|E(U)\|, p + 1)$$

Let

$$(U,V,\Sigma) \;\to\; H_p(U,V,\Sigma) = \begin{pmatrix} U\,(I+\Omega_p)\,(I+\Theta_p) \\ V(I+\Lambda_p)(I+\Psi_p) \\ \Sigma + S \end{pmatrix}$$

where :

1. $\Omega_p = s_p(E(U)\,)$ and $\Lambda = s_p(E(V)\,)$.
2. $\Theta_k = c_p(X_1 + \cdots + X_k)$ and $\Psi_k = c_p(Y_1 + \cdots + Y_k)$, $\qquad 1 \leqslant k \leqslant p$.
3. The X_k's and Y_k's are skew Hermitian matrices satisfying

$$S_k = \mathrm{diag}(\Delta_k), \qquad \Delta_k - S_k - X_k\Sigma + \Sigma Y_k \;=\; 0, \qquad 1 \leqslant k \leqslant p$$

where the Δ_k's are defined as

$$\Delta_1 \;=\; (I+\Omega_p)(\Delta+\Sigma)(I+\Lambda_p) - \Sigma$$

$$\Delta_k \;=\; (I+\Theta^*_{k-1})(\Delta_1+\Sigma)(I+\Psi_{k-1}) - \Sigma - \sum_{k=1}^{k-1} S_k, \qquad 2 \leqslant k \leqslant p+1$$

The computation of $H_p(U, V, \Sigma)$ only requires matrix additions and multiplications without resolution of linear systems. The following table gives the number of addittion and multiplications to evaluate $H_p(U, V, \Sigma)$.

	$E_m(U)$	$s_p(E_m(U))$	$c_p(X)$	$\Delta_k - S_k - X_k\Sigma + \Sigma Y_k$	S	Δ_k
matrix additions	1	p	p^2		p	
matrix multiplications	1	p	p^2			
additions				$10np$		$(m+4n)\,p$
multiplications				$(m-n+8)\,np$		$(m+n)\,mnp$

This implies $2(p^2 + p + 1)(m^2 + n^2) + (m + 14n)p$ additions

and $2\,(p^2 + p + 1)(m^3 + n^3) + (m^2 + mn + m - n + 8)np$ multiplications.

From $\Sigma_0 := \mathrm{diag}(\sigma_1, \ldots, \sigma_\ell)$ we define

1. $K := K(\Sigma_0) = \max_i (1, |\sigma_i|)$.

2. $\kappa := \kappa(\Sigma_0) = \max\left(1, \quad \max_i \frac{1}{|\sigma_i|}, \quad \max_{i \neq j}\left(\frac{1}{|\sigma_i - \sigma_j|} + \frac{1}{|\sigma_i + \sigma_j|}\right)\right)$

Theorem 5. *Let $p \geqslant 1$ and M a complex matrix. From (U_0, V_0, Σ_0), let us define the sequence*

$$(U_{i+1}, V_{i+1}, \Sigma_{i+1}) = H_p(U_i, V_i, \Sigma_i), \quad i \geqslant 0.$$

We consider the constants defined by

	$p=1$	$p=2$	$p \geqslant 3$
a	2	$4/3$	$4/3$
u_0	0.0289	0.046	0.0297

If

$$\mathbf{max}\,\left(\quad \boldsymbol{\kappa^a K^a}\,||\boldsymbol{E(U_0)}||, \quad \boldsymbol{\kappa^a K^a}||\boldsymbol{E(V_0)}||, \quad \boldsymbol{\kappa^a K^{a-1}}\,||\boldsymbol{\Delta_0}|| \quad \right) \leqslant \boldsymbol{u_0}$$

then the sequence $(U_i, V_i, \Sigma_i)_{i \geqslant 0}$ converges to a solution $(U_\infty, V_\infty, \Sigma_\infty)$ of SVD system with an order of convergence equal to $p+1$.

Our numerical experiments are done with the `Julia Programming Language` coupled with the library `ArbNumerics` of Jeffrey Sarnoff.

To intialize our method we start with a triplet (U_0, V_0, Σ_0) computed by the function *svd* of *Julia*.

We consider for $i \geqslant 0$ the quantities

$$\boldsymbol{\varepsilon_i = \max\left(\quad \kappa_i^a\, K_i^a\, \|E(U_i)\|, \quad \kappa_i^a\, K_i^a\, \|E(V_i)\|, \quad \kappa_i^a K_i^{a-1}\|\Delta_i\|\right)}.$$

We show the behaviour of $\boldsymbol{e_i = -\left\lfloor \log_2\left(\frac{\varepsilon_i}{u_0}\right)\right\rfloor}$.

Iterations / Order	2	3	4	5	6	7
0	7	8	9	8	8	8
1	18	35	47	59	69	85
2	44	112	194	311	427	604
3	92	346	787	1571	2580	4353

We determine an index q such that :

1. $\Sigma_0 = \begin{pmatrix} \Sigma_{0,q} & \\ & \Sigma_{0,n-q} \\ 0 & 0 \end{pmatrix}$
2. $\kappa(\Sigma_q)^a \, K(\Sigma)^{a-1} \|\Delta_0\| \leqslant u_0$

We then approximate the thin SVD associated to $\Sigma_{0,q}$.

$M = \left(\frac{1}{i+j}\right)_{1 \leqslant i,j \leqslant n}$.

Lemma 6. $\sigma_{1+k} \leqslant 4\exp\left(\frac{\pi^2}{2\mathrm{Log}(4n)}\right)^{-2k} \sigma_1$

B. Beckermann, A. Townsend, On the singular values of matrices with displacement structure, SIAM Journal on Matrix Analysis and Applications Vol. 38, 4, 2017.

The table gives the value of q with respect n and $p+1$.

$p+1=2$	$n=2:7, q=n$	$n=8:9, q=7$	$n=10:16, q=8$	$n=17:30, q=9$	$n=34:40, q=10$
$p+1\geqslant 3$	$n=2:10, q=n$	$n=11:12, q=10$	$n=13:19, q=11$	$n=20:29, q=12$	$n=30:40, q=13$

$$(U,V,\Sigma) \;\to\; \mathrm{DS}(U,V,\Sigma) = \begin{pmatrix} U\Big(I+X_1+X_2+\frac{1}{2}X_1^2\Big) \\ V\Big(I+Y_1+Y_2+\frac{1}{2}Y_1^2\Big) \\ \Sigma+S_1+S_2 \end{pmatrix}$$

where

$$\begin{aligned} X_1\Sigma - \Sigma Y_1 + S_1 &= \Delta_1 := \Delta = U^*MV - \Sigma \\ X_2\Sigma - \Sigma Y_2 + S_2 &= \Delta_2 := -\tfrac{1}{2}X_1(\Delta+S_1) + \tfrac{1}{2}(\Delta+S_1)Y_1 \end{aligned}$$

This differs from the map $H_2(U,V,\Sigma) = \begin{pmatrix} U(I+\Omega_p)\Big(I+X_1+X_2+\frac{1}{2}(X_1+X_2)^2\Big) \\ V(I+\Lambda_p)\Big(I+Y_1+Y_2+\frac{1}{2}(Y_1+Y_2)^2\Big) \\ \Sigma+S_1+S_2 \end{pmatrix}$

Remember $$\mathrm{DS}(U,V,\Sigma)=\begin{pmatrix} U\left(I+X_1+X_2+\frac{1}{2}X_1^2\right) \\ V\left(I+Y_1+Y_2+\frac{1}{2}Y_1^2\right) \\ \Sigma+S_1+S_2 \end{pmatrix} := \begin{pmatrix} U_1 \\ V_1 \\ \Sigma_1 \end{pmatrix}$$

Let us define $$\overline{\mathrm{DS}}(U,V,\Sigma)=\begin{pmatrix} U\left(I+X_1+X_2+\frac{1}{2}(X_1+X_2)^2\right) \\ V\left(I+Y_1+Y_2+\frac{1}{2}(Y_1+Y_2)^2\right) \\ \Sigma+S_1+S_2 \end{pmatrix} := \begin{pmatrix} \overline{U}_1 \\ \bar{V}_1 \\ \bar{\Sigma}_1 \end{pmatrix}$$

Theorem 7.

1. *If* $\boldsymbol{\kappa^{5/4}K^{2/5}\|\Delta\| \leqslant \varepsilon \leqslant 0.1}$ *then*

$$\|U_1^* M V_1 - \Sigma_1\| \leqslant (8+18\varepsilon+33\varepsilon^2)\varepsilon^3.$$

2. *If* $\boldsymbol{\kappa^{6/5}K^{3/10}\varepsilon_1 \leqslant \varepsilon \leqslant 0.1}$ *then*

$$\|\bar{U}_1^* M \bar{V}_1 - \bar{\Sigma}_1\| \leqslant (6+21\varepsilon+54\varepsilon^2)\varepsilon^3$$

Thanks for your attention