ROUNDING ERROR ANALYSIS OF INDEFINITE ORTHOGONALIZATION

Miro Rozložník

joint work with Alicja Smoktunowicz and Felicja Okulicka-Dłużewska

Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic

GAMM-Matheon Workshop "Matrix Computations for Sparse Recovery"
Berlin, April 9-11, 2014

Orthogonalization with the standard inner product

$$A = (a_1, \dots, a_n) \in \mathcal{R}^{m,n},\ m \geq n = rank(A)$$

orthogonal basis Q of $span(A)$:

$$Q = (q_1, \dots, q_n) \in \mathcal{R}^{m,n},\ Q^T Q = I_n$$

$A = QR$, $R \in \mathcal{R}^{n,n}$ upper triangular with positive diagonal

$$C = A^T A = R^T R$$

Orthogonalization with a non-standard inner product

$B \in \mathcal{R}^{m,m}$ symmetric positive definite, inner product $\langle \cdot, \cdot \rangle_B$

$$A = (a_1, \ldots, a_n) \in \mathcal{R}^{m,n},\ m \geq n = rank(A)$$

B-orthonormal basis of $span(A)$:

$$Q = (q_1, \ldots, q_n) \in \mathcal{R}^{m,n},\ Q^T B Q = I_n$$

$A = QR$, $R \in \mathcal{R}^{n,n}$ upper triangular with positive diagonal

$$C = A^T B A = R^T R$$

Indefinite orthogonalization with a symmetric bilinear form

$B \in \mathcal{R}^{m,m}$ symmetric indefinite and nonsingular, bilinear form

$A = (a_1, \dots, a_n) \in \mathcal{R}^{m,n}$, $m \geq n = rank(A)$

B-orthonormal basis of $span(A)$:

$Q = (q_1, \dots, q_n) \in \mathcal{R}^{m,n}$, $Q^T B Q = \Omega \in \mathrm{diag}(\pm 1)$

$A = QR$, $R \in \mathcal{R}^{n,n}$ upper triangular with positive diagonal

if no principal minor of C vanishes (if C is strongly nonsingular)

$$C = A^T B A = R^T \Omega R$$

Signed Cholesky factorization of an indefinite matrix

$$C_j = A_j^T B A_j = \begin{pmatrix} C_{j-1} & c_{1:j-1,j} \\ c_{1:j-1,j}^T & c_{j,j} \end{pmatrix} =$$

$$\begin{pmatrix} R_{j-1}^T & 0 \\ r_{1:j-1,j}^T & r_{j,j} \end{pmatrix} \begin{pmatrix} \Omega_{j-1} & 0 \\ 0 & \omega_j \end{pmatrix} \begin{pmatrix} R_{j-1} & r_{1:j-1,j} \\ 0 & r_{j,j} \end{pmatrix}$$

$$r_{1:j-1,j} = \Omega_{j-1}^{-1} R_{j-1}^{-T} c_{1:j-1,j}$$

$$r_{j,j}^2 \omega_j = c_{j,j} - r_{1:j-1,j}^T \Omega_{j-1} r_{1:j-1,j} = c_{j,j} - c_{1:j-1,j}^T C_{j-1}^{-1} c_{1:j-1,j} = s_j$$

$$R_j = \begin{pmatrix} R_{j-1} & r_{1:j-1,j} \\ 0 & r_{j,j} \end{pmatrix} = \begin{pmatrix} R_{j-1} & R_{j-1} C_{j-1}^{-1} c_{1:j-1,j} \\ 0 & \sqrt{|s_j|} \end{pmatrix}$$

Cholesky factorization and singular value decomposition

$$R_j^T R_j =$$
$$\begin{pmatrix} I & 0 \\ c_{1:j-1,j}^T C_{j-1}^{-1} & 1 \end{pmatrix} \begin{pmatrix} R_{j-1}^T R_{j-1} & 0 \\ 0 & \omega_j s_j \end{pmatrix} \begin{pmatrix} I & C_{j-1}^{-1} c_{1:j-1,j} \\ 0 & 1 \end{pmatrix}$$

$$C_j = \begin{pmatrix} I & 0 \\ c_{1:j-1,j}^T C_{j-1}^{-1} & 1 \end{pmatrix} \begin{pmatrix} C_{j-1} & 0 \\ 0 & s_j \end{pmatrix} \begin{pmatrix} I & C_{j-1}^{-1} c_{1:j-1,j} \\ 0 & 1 \end{pmatrix}$$

The norm of the triangular factor

$$R_j^T R_j = \omega_1 C_j + 2 \sum_{i=1,\ldots,j-1;\ \omega_{i+1} \neq \omega_i} \begin{pmatrix} 0 & 0 \\ 0 & C_j \backslash C_i \end{pmatrix}$$

$$\|R_j\|^2 \leq \|C_j\| + 2 \sum_{i=1,\ldots,j-1;\ \omega_{i+1} \neq \omega_i} \|C_j \backslash C_i\|,$$

The norm of the inverse of the triangular factor

$$(R_j^T R_j)^{-1} = \begin{pmatrix} (R_{j-1}^T R_{j-1})^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \omega_j \left[C_j^{-1} - \begin{pmatrix} C_{j-1}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right]$$

$$\|R_j^{-1}\|^2 \leq \|C_j^{-1}\| + 2 \sum_{i=1,\ldots,j-1;\ \omega_{i+1} \neq \omega_i} \|C_i^{-1}\|$$

Condition number of factors R and Q

$$\|R\| \leq \|C\| \|R^{-1}\|$$

$$\kappa(R) \leq \|C\| \left(\|C^{-1}\| + 2 \sum_{j;\ \omega_{j+1} \neq \omega_j} \|C_j^{-1}\| \right)$$

$$\|Q\| \leq \|A\| \|R^{-1}\|, \quad \sigma_{min}(Q) \geq \frac{\sigma_{min}(A)}{\|R\|}$$

$$\kappa(Q) \leq \kappa(A) \kappa(R)$$

Example with $\kappa(R) \approx \kappa^{1/2}(B)$

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \; B = \begin{pmatrix} 1 & \sqrt{\varepsilon} \\ \sqrt{\varepsilon} & -\varepsilon \end{pmatrix}$$

$$Q = R^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & \frac{1}{\sqrt{\varepsilon}} \end{pmatrix}, \quad R = Q^{-1} = \begin{pmatrix} 1 & \sqrt{\varepsilon} \\ 0 & \sqrt{\varepsilon} \end{pmatrix}, \quad \Omega = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\|B\| \approx 1 + \varepsilon \text{ and } \sigma_{min}(B) = 2\varepsilon$$

$$\|R\| \approx \sqrt{1+\varepsilon}, \;\; \sigma_{min}(R) \approx \sqrt{\varepsilon}, \;\; \kappa(R) = \kappa(Q) \approx \frac{1}{\sqrt{\varepsilon}}$$

Example with $\kappa(R) \gg \kappa^{1/2}(B)$

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \; B = \begin{pmatrix} \varepsilon & 1 \\ 1 & -\varepsilon \end{pmatrix}$$

$$Q = R^{-1} = \begin{pmatrix} \frac{1}{\sqrt{\varepsilon}} & -\frac{1}{\sqrt{\varepsilon(1+\varepsilon^2)}} \\ 0 & \frac{\sqrt{\varepsilon}}{\sqrt{1+\varepsilon^2}} \end{pmatrix}, \quad R = Q^{-1} = \begin{pmatrix} \sqrt{\varepsilon} & \frac{1}{\sqrt{\varepsilon}} \\ 0 & \frac{\sqrt{1+\varepsilon^2}}{\sqrt{\varepsilon}} \end{pmatrix}, \quad \Omega = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\|B\| = \sigma_{min}(B) = \sqrt{1+\varepsilon^2}$$

$$\|R\| \approx \frac{\sqrt{2}}{\sqrt{\varepsilon}}, \; \sigma_{min}(R) \approx \frac{\sqrt{\varepsilon}}{\sqrt{2}}, \quad \kappa(R) = \kappa(Q) \approx \frac{2}{\varepsilon}$$

Triangular factor from classical Gram-Schmidt vs. indefinite Cholesky factor

Exact arithmetic:

$$r_{i,j} = \omega_i^{-1} \left(a_j, q_i\right)_B = \left(a_j, \frac{a_i - \sum_{k=1}^{i-1} r_{k,i} q_k}{\omega_i r_{i,i}}\right)_B = \frac{(a_j, a_i)_B - \sum_{k=1}^{i-1} r_{k,i}\omega_k r_{k,j}}{\omega_i r_{i,i}}$$

Finite precision arithmetic:

$$\bar{r}_{1:j,k}^T \bar{\Omega}_j \bar{r}_{1:j,j} = c_{k,j} + \Delta r_{1:j-1,k}^T \bar{\Omega}_j \bar{r}_{1:j,j} + a_k^T B \Delta a_j$$
$$\bar{\omega}_j \bar{r}_{j,j}^2 = c_{j,j} - \bar{r}_{1:j-1,j}^T \bar{\Omega}_{j-1} \bar{r}_{1:j-1,j} + \Delta c_{j,j}$$

Classical Gram-Schmidt computes a stable Cholesky factor of $C = A^T B A$

Indefinite Cholesky B-QR factorization:

assuming $\mathcal{O}(u)\kappa(C)\|A\|^2\|B\| \max_{j,\ \bar{\omega}_{j+1}\neq\bar{\omega}_j} \|C_j^{-1}\| < 1$

$$C + \Delta C = \bar{R}^T \bar{\Omega} \bar{R},$$
$$\|\Delta C\| \leq \mathcal{O}(u)[\|\bar{R}\|^2 + \|B\|\|A\|^2]$$

Classical Gram-Schmidt (B-CGS) process :

$$A + \Delta A = \bar{Q}\bar{R},\ \|\Delta A\| \leq \mathcal{O}(u)\|\bar{Q}\|\|\bar{R}\|,$$
$$C + \Delta C = \bar{R}^T \bar{\Omega} \bar{R},$$
$$\|\Delta C\| \leq \mathcal{O}(u)[\|\bar{R}\|^2 + \|B\|\|A\|\|\bar{Q}\|\|\bar{R}\| + \|B\|\|A\|^2]$$

Indefinite Cholesky QR factorization: factorization error and loss of orthogonality

$$\bar{Q} = \text{fl}(A\bar{R}^{-1})$$

$$\|\bar{Q}^T B\bar{Q} - \bar{\Omega}\| \quad \leq \quad \mathcal{O}(u)[\kappa^2(\bar{R}) + \|\bar{R}^{-1}\|^2\|A\|^2\|B\| + 2\|B\bar{Q}\|\|\bar{Q}\|\kappa(\bar{R})]$$

Classical Gram-Schmidt (B-CGS) process :

$$\|\bar{Q}^T B\bar{Q} - \bar{\Omega}\| \leq \\ \mathcal{O}(u)\left[\kappa^2(\bar{R}) + \|\bar{R}^{-1}\|^2\|A\|^2\|B\| + 3\|BA\|\|\bar{R}^{-1}\|\|\bar{Q}\|\kappa(\bar{R})\right]$$

Classical Gram-Schmidt process with reorthogonalization (B-CGS2)

$$u_j^{(1)} = a_j - Q_{j-1} r_{1:j-1,j}^{(1)} = (I - Q_{j-1}\Omega_{j-1}^{-1} Q_{j-1}^T B) a_j,$$
$$u_j^{(2)} = u_j^{(1)} - Q_{j-1} r_{1:j-1,j}^{(2)} = (I - Q_{j-1}\Omega_{j-1}^{-1} Q_{j-1}^T B)^2 a_j = u_j^{(1)}$$

$$\|\bar{Q}_{j-1}^T B \left(\frac{\bar{u}_j^{(2)}}{\bar{r}_{j,j}} \right) \| \lessapprox \|\bar{\Omega}_{j-1} - \bar{Q}_{j-1}^T B \bar{Q}_{j-1}\|^2 \| \frac{\bar{r}_{1:j-1,j}}{\bar{r}_{j,j}} \|$$

$$1/r_{j,j} = |s_j|^{-1/2} \leq \|C_j^{-1}\|^{1/2}, \ \|r_{1:j-1,j}\|/r_{j,j} \leq \|R_j\| \|C_j^{-1}\|^{1/2}$$

Cholesky QR factorization with iterative refinement and classical Gram-Schmidt with reorthogonalization: loss of orthogonality

$$A^T BA = (R^{(1)})^T \Omega^{(1)} R^{(1)},\ Q^{(1)} = A(R^{(1)})^{-1}$$

$$(Q^{(1)})^T BQ^{(1)} = (R^{(2)})^T \Omega^{(2)} R^{(2)},\ Q^{(2)} = Q^{(1)}(R^{(2)})^{-1}$$

$$Q = Q^{(2)},\ R = R^{(2)} R^{(1)}$$

$$\|(\bar{Q}^{(2)})^T B\bar{Q}^{(2)} - \bar{\Omega}^{(2)}\| \leq \mathcal{O}(u) \left[\|B\|\|\bar{Q}^{(1)}\|^2 + \|B\bar{Q}^{(2)}\|\|\bar{Q}^{(2)}\|\right]$$

CGS with reorthogonalization (B-CGS2):

$$\mathcal{O}(u)\kappa(A)\|A\|^2\|B\|\|C\|(\|C^{-1}\| + \max_{j,\ \bar{\omega}_{j+1}\neq\bar{\omega}_j} \|C_j^{-1}\|)^2 < 1$$

$$\|\bar{Q}^T B\bar{Q} - \bar{\Omega}\| \leq \mathcal{O}(u)\|B\|\|\bar{Q}\|^2$$

Numerical experiments - model examples

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} R_{11}^T & 0 \\ R_{12}^T & R_{22}^T \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix},$$

1. $\kappa(C_{11}) = 100 \ll \kappa(C) \approx 10^{2i}$, $\kappa(C_{12}) = 10^i$ for $i = 0, \ldots, 8$; $C_{22} = 0$ ($\|C_{11}\| = \|C_{12}\| = 1$)

2. $\kappa(C_{11}) = 10^i \gg \kappa(C) = 1$ for $i = 0, \ldots, 16$; $C_{11}^2 + C_{12}^2 = I$ $C_{22} = -C_{11}$ ($\|C_{11}\| = 1/2$)

The spectral properties of computed factors with respect to the conditioning of the submatrix C_{12} for Problem 1.

$\|C_{12}^{-1}\|$	$\|C^{-1}\|$	$\|S_{22}\|$	$\|\bar{R}\| = \|\bar{Q}^{-1}\|$	$\|\bar{R}^{-1}\| = \|\bar{Q}\|$
10^0	1.6180e+00	1.0000e+02	1.4142e+01	1.4142e+01
10^1	1.0099e+02	1.0000e+02	1.4142e+01	1.4142e+01
10^2	1.0001e+04	1.0000e+02	1.4142e+01	1.0001e+02
10^3	1.0000e+06	1.0000e+02	1.4142e+01	1.0000e+03
10^4	1.0000e+08	1.0000e+02	1.4142e+01	1.0000e+04
10^5	1.0000e+10	1.0000e+02	1.4142e+01	1.0000e+05
10^6	1.0000e+12	1.0000e+02	1.4142e+01	1.0000e+06
10^7	9.9808e+13	1.0000e+02	1.4142e+01	1.0000e+07
10^8	1.8925e+16	1.0000e+02	1.4142e+01	1.0000e+08

The factorization error $\|A - \bar{Q}\bar{R}\|$ with respect to the conditioning of the submatrix C_{12} for Problem 1.

$\|C_{12}^{-1}\|$	Cholesky B-QR	Cholesky B-QR2	B-CGS	B-CGS2
10^0	9.0448e-16	4.0019e-14	3.5544e-15	1.1411e-14
10^1	3.7826e-15	1.7094e-14	2.5165e-15	9.4835e-15
10^2	2.0509e-15	1.4189e-14	2.9717e-16	1.1512e-14
10^3	1.5382e-15	1.3225e-14	4.4431e-16	5.9412e-15
10^4	7.9169e-16	1.4906e-14	2.4825e-16	1.3652e-14
10^5	1.2152e-15	1.5119e-14	2.6803e-16	7.8625e-15
10^6	1.1653e-15	8.8771e-15	4.5776e-16	9.0056e-15
10^7	1.7904e-15	2.2160e-14	1.2413e-16	6.5767e-15
10^8	1.8611e-15	2.5766e-14	1.0175e-15	1.1846e-14

The loss of B-orthogonality $\|\bar{\Omega} - \bar{Q}^T B \bar{Q}\|$ with respect to the conditioning of the submatrix C_{12} for Problem 1.

$\|C_{12}^{-1}\|$	Cholesky B-QR	Cholesky B-QR2	B-CGS	B-CGS2
10^0	6.9767e-15	3.1373e-15	4.5838e-15	3.1956e-15
10^1	8.5940e-14	6.6516e-15	5.1740e-14	7.1550e-15
10^2	1.8989e-12	5.6400e-14	4.4021e-12	5.1951e-14
10^3	4.8268e-10	3.2421e-13	1.5760e-10	4.4188e-13
10^4	2.9594e-08	4.9631e-12	1.1656e-08	2.6936e-12
10^5	1.5621e-06	3.7820e-11	1.8274e-06	2.9007e-11
10^6	2.4082e-05	2.0335e-10	2.3673e-04	2.8010e-10
10^7	3.7036e-02	2.5207e-09	9.6352e-03	2.9913e-09
10^8	6.5241e-01	2.0603e-08	4.1306e-01	2.4907e-08

The spectral properties of computed factors with respect to the conditioning of the submatrix C_{11} for Problem 2.

$\|C_{11}^{-1}\|$	$\|C^{-1}\|$	$\|S_{22}\|$	$\|\bar{R}\| = \|\bar{Q}^{-1}\|$	$\|\bar{R}^{-1}\| = \|\bar{Q}\|$
10^{0}	1.0000e+00	2.0000e+00	1.9319e+00	1.9319e+00
10^{1}	1.0000e+00	2.0000e+01	6.3226e+00	6.3226e+00
10^{2}	1.0000e+00	2.0000e+02	2.0000e+01	2.0000e+01
10^{3}	1.0000e+00	2.0000e+03	6.3246e+01	6.3246e+01
10^{4}	1.0000e+00	2.0000e+04	2.0000e+02	2.0000e+02
10^{5}	1.0000e+00	2.0000e+05	6.3246e+02	6.3246e+02
10^{6}	1.0000e+00	2.0000e+06	2.0000e+03	2.0000e+03
10^{7}	1.0000e+00	2.0000e+07	6.3246e+03	6.3246e+03
10^{8}	1.0000e+00	2.0000e+08	2.0000e+04	2.0000e+04
10^{9}	1.0000e+00	2.0000e+09	6.3246e+04	6.3246e+04
10^{10}	1.0000e+00	2.0000e+10	2.0000e+05	2.0000e+05
10^{11}	1.0000e+00	2.0000e+11	6.3246e+05	6.3246e+05
10^{12}	1.0000e+00	2.0000e+12	2.0000e+06	2.0000e+06
10^{13}	1.0000e+00	1.9999e+13	6.3245e+06	6.3245e+06
10^{14}	1.0000e+00	2.0004e+14	2.0188e+07	2.0520e+07
10^{15}	1.0000e+00	2.0011e+15	6.6349e+07	5.2040e+07

The factorization error $\|A - \bar{Q}\bar{R}\|$ with respect to the conditioning of the principal submatrix C_{11} for Problem 2.

$\|C_{11}^{-1}\|$	Cholesky B-QR	Cholesky B-QR2	B-CGS	B-CGS2
10^0	2.2204e-16	3.4158e-31	2.2204e-16	2.2204e-16
10^1	1.8577e-15	4.4404e-15	7.6343e-16	2.5796e-15
10^2	9.5582e-15	2.5418e-14	3.5531e-15	2.8651e-14
10^3	8.1635e-14	5.6963e-13	1.6381e-14	2.8060e-13
10^4	4.8395e-13	2.7736e-12	6.3553e-14	1.8356e-12
10^5	6.7123e-12	3.4801e-11	1.8190e-12	3.3911e-11
10^6	4.3895e-11	2.8659e-10	7.2760e-12	1.7619e-10
10^7	3.2539e-11	5.1621e-09	1.1732e-10	2.4764e-09
10^8	1.9919e-09	3.8291e-08	5.8208e-11	1.1369e-08
10^9	1.9037e-08	4.7511e-07	3.4298e-08	3.1724e-07
10^{10}	9.4905e-08	3.1411e-06	2.9802e-08	1.5431e-06
10^{11}	2.0371e-06	3.1822e-05	2.3842e-07	2.0807e-05
10^{12}	4.6287e-06	2.6973e-04	1.7481e-05	3.7244e-04
10^{13}	3.4565e-04	4.3527e-03	6.1035e-05	1.6198e-03
10^{14}	2.3032e-03	8.4629e-02	3.0518e-05	1.8111e-02
10^{15}	7.8736e-03	8.4428e-01	8.9503e-03	1.5765e-01

The loss of B-orthogonality $\|\bar{\Omega} - \bar{Q}^T B \bar{Q}\|$ with respect to the conditioning of the principal submatrix C_{11} for Problem 2.

$\|C_{11}^{-1}\|$	Cholesky B-QR	Cholesky B-QR2	B-CGS	B-CGS2
10^0	5.0322e-16	3.2067e-16	5.3413e-16	3.9373e-16
10^1	1.2883e-15	8.7715e-16	1.5521e-15	1.2610e-15
10^2	4.5583e-15	3.5957e-15	4.6097e-15	3.2657e-15
10^3	1.9874e-14	1.6704e-14	2.6765e-14	2.2026e-14
10^4	1.5159e-13	1.2480e-13	1.4222e-13	1.3054e-13
10^5	1.0447e-12	8.1751e-13	1.1241e-12	1.2374e-12
10^6	1.0511e-11	7.1311e-12	1.6597e-11	6.4763e-12
10^7	5.8440e-11	5.0812e-11	2.1037e-10	5.1101e-11
10^8	3.5174e-10	2.3857e-10	6.4724e-10	5.8383e-10
10^9	5.6336e-09	4.7359e-09	8.5080e-09	3.2390e-09
10^{10}	6.4206e-08	4.7271e-08	1.8162e-07	4.7073e-08
10^{11}	3.3127e-07	2.8293e-07	1.0061e-06	4.2164e-07
10^{12}	3.4508e-06	2.6920e-06	7.6409e-06	6.0936e-06
10^{13}	2.2361e-05	5.5208e-05	1.3357e-04	4.7861e-03
10^{14}	5.4077e-04	3.6470e-04	6.8111e-04	2.1676e+00
10^{15}	5.4339e-03	2.9211e-03	1.0174e-02	4.1463e+00

Thank you for your attention!!!

References:
R, F. Okulicka-Dluzewska, A. Smoktunowicz: Indefinite orthogonalization with rounding errors, submitted 2013.
R, J. Kopal, M. Tuma, A. Smoktunowicz: Numerical stability of orthogonalization methods with a non-standard inner product, BIT Numerical Mathematics (2012) 52:1035–1058.