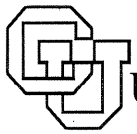
On Principal Angles Between Subspaces of Euclidean Space *

Zlatko Drmac

CU-CS-838-97



University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

* This work was supported by National Science Foundation grants ACS-9357812 and ASC-9625912 and Department of Energy grant DE-FG03-94ER25215.

On Principal Angles Between Subspaces of Euclidean Space*

Zlatko Drmač†

March 31, 1997

Abstract

Let $X \in \mathbf{R}^{m\times p}$, $Y \in \mathbf{R}^{m\times q}$ be of full column rank, and let $\mathcal{X} = \mathrm{span}(X)$, $\mathcal{Y} = \mathrm{span}(Y)$. This paper addresses the issues of accurate floating–point computation of the principal angles between $\mathcal{X}$ and $\mathcal{Y}$. The cosines of the principal angles are usually computed using the Björck–Golub algorithm as the singular values of $Q_x^T Q_y$, where Q_x and Q_y are orthonormal matrices computed by the QR factorizations of X and Y, respectively. This paper shows that the Björck–Golub algorithm is mixed stable in the following sense: The computed cosines of the principal angles approximate with small relative error the exact cosines of the principal angles between $\tilde{\mathcal{X}} \equiv \mathrm{span}(X + \Delta X)$ and $\tilde{\mathcal{Y}} \equiv \mathrm{span}(Y + \Delta Y)$, where $\epsilon_x \equiv \max_{1\le i\le p} \|\Delta X e_i\|_2/\|Xe_i\|_2$ and $\epsilon_y \equiv \max_{1\le j\le q} \|\Delta Y e_j\|_2/\|Ye_j\|_2$ are bounded by modest polynomials of m, p, q times machine precision. Hence, the backward error is bounded in the angle metric by $\sin\angle(\mathcal{X},\tilde{\mathcal{X}}) \le \sqrt{p}\epsilon_x\|X_c^\dagger\|_2$, $\sin\angle(\mathcal{Y},\tilde{\mathcal{Y}}) \le \sqrt{q}\epsilon_y\|Y_c^\dagger\|_2$, where $X = X_c\mathrm{diag}(\|Xe_i\|_2)$, $Y = Y_c\mathrm{diag}(\|Ye_i\|_2)$. This paper also recommends that in the Björck–Golub algorithm the QR factorizations are computed with complete pivoting of Powell and Reid. It is shown that in that case the Björck–Golub algorithm achieves high accuracy if X and Y can be written as $\Pi_1^{(x)} X \Pi_2^{(x)} = D_1 X_s D_2$, $\Pi_1^{(y)} Y \Pi_2^{(y)} = D_3 Y_s D_4$, with (arbitrarily ill–conditioned) diagonal matrices D_1, D_2, D_3, D_4 with descending diagonal entries, permutation matrices $\Pi_1^{(x)}$, $\Pi_2^{(x)}$, $\Pi_1^{(y)}$, $\Pi_2^{(y)}$, and with well conditioned X_s and Y_s. Further, this paper proposes a new mixed stable algorithm, based on Gaussian elimination with pivoting. The new algorithm approximates with small relative error the exact cosines of the principal angles between $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$, where $\sin\angle(\mathcal{X},\hat{\mathcal{X}})$ and $\sin\angle(\mathcal{Y},\hat{\mathcal{Y}})$ depend on the accuracy of the pivoted LU factorizations of X and Y. The new algorithm is more accurate than the Björck–Golub algorithm in cases when the LU factorization with pivoting is more accurate than the QR factorization.

*Technical Report CU–CS–838–97, Department of Computer Science, University of Colorado at Boulder.

†Department of Computer Science, Engineering Center ECOT 7–7, University of Colorado at Boulder, Boulder, CO 80309–0430, (zlatko@cs.colorado.edu; http://www.cs.colorado.edu/~zlatko/) This research was supported by National Science Foundation grants ACS–9357812 and ASC–9625912 and Department of Energy grant DE–FG03–94ER25215.

1 Introduction

Let $X \in \mathbf{R}^{m\times p}$, $Y \in \mathbf{R}^{m\times q}$ be full column rank matrices with $p \geq q$ and let $\mathcal{X} = \mathrm{span}(X)$, $\mathcal{Y} = \mathrm{span}(Y)$, $k = \dim(\mathcal{X} \bigcap \mathcal{Y})$. Let $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{Y}$ be the corresponding orthogonal projectors, and let $k + r = \mathrm{rank}(\mathcal{P}_\mathcal{X}\mathcal{P}_\mathcal{Y})$. In a suitably chosen orthonormal basis, the matrix representations $[\mathcal{P}_\mathcal{X}]$ and $[\mathcal{P}_\mathcal{Y}]$ of $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{Y}$, respectively, read (cf. [45])

$$[\mathcal{P}_\mathcal{X}] = I_k \oplus \left(\oplus_{i=k+1}^{k+r} \Xi_\mathcal{X}^{(i)}\right) \oplus \Omega_\mathcal{X}, \quad [\mathcal{P}_\mathcal{Y}] = I_k \oplus \left(\oplus_{i=k+1}^{k+r} \Xi_\mathcal{Y}^{(i)}\right) \oplus \Omega_\mathcal{Y}, \tag{1}$$

where $\Omega_\mathcal{X}$ and $\Omega_\mathcal{Y}$ are diagonal matrices with diagonal entries zero or one, and such that $\Omega_\mathcal{X}\Omega_\mathcal{Y} = \mathbf{O}$ and

$$\Xi_\mathcal{X}^{(i)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} [1, 0], \quad \Xi_\mathcal{Y}^{(i)} = \begin{bmatrix} \cos\vartheta_i \\ \sin\vartheta_i \end{bmatrix} [\cos\vartheta_i, \sin\vartheta_i], \quad 0 < \vartheta_i < \frac{\pi}{2}, \quad k+1 \leq i \leq k+r. \tag{2}$$

(The matrices $\Omega_\mathcal{X}$ and $\Omega_\mathcal{Y}$ may be void.) Principal angles between $\mathcal{X}$ and $\mathcal{Y}$ are, by definition, $\vartheta_1 \equiv \cdots \equiv \vartheta_k \equiv 0$, $\vartheta_{k+1}, \ldots, \vartheta_{k+r}$, $\vartheta_{k+r+1} \equiv \cdots \equiv \vartheta_q \equiv \pi/2$. Obviously, $\sigma_i \equiv \cos\vartheta_i$, $i = 1, \ldots, k+r$, are the non–zero singular values of $\mathcal{P}_\mathcal{X}\mathcal{P}_\mathcal{Y}$. (Note that $\Xi_\mathcal{X}^{(i)}\Xi_\mathcal{Y}^{(i)} = \begin{bmatrix} \cos\vartheta_i & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\vartheta_i & \sin\vartheta_i \\ -\sin\vartheta_i & \cos\vartheta_i \end{bmatrix}$.)

The cosines of the principal angles are also known as canonical correlations and have important applications in statistic, econometric, geology. Golub and Zha [26] discuss various equivalent characterizations of the principal angles. For instance, they show that the SVD of $Q_x^\tau Q_y$ can be used to solve constrained optimization problems such as

$$\max_{A,B} \mathrm{Trace}(A^\tau X^\tau Y B), \quad \text{where} \quad A^\tau X^\tau X A = I_p, \quad B^\tau Y^\tau Y B = I_q,$$

and the orthogonal Procrustes problem $\min_{U^\tau U = I} \|Q_x - Q_y U\|_F$. (For a thorough description of the principal angles see [37].)

In this paper, we analyze numerical computation of the principal angles. Björck and Golub [10] have shown that the principal angles can be computed via the SVD of $Q_x^\tau Q_y$, where $X = Q_x R_x$ and $Y = Q_y R_y$ are the QR factorizations of X and Y, respectively. (Taking $\mathcal{P}_\mathcal{X} \equiv Q_x Q_x^\tau$, $\mathcal{P}_\mathcal{Y} \equiv Q_y Q_y^\tau$ and writing the SVD of $Q_x^\tau Q_y$ as $Q_x^\tau Q_y = W\Sigma V^\tau$ we obtain $\mathcal{P}_\mathcal{X}\mathcal{P}_\mathcal{Y} = (Q_x W)\Sigma(Q_y V)^\tau$.) In § 2, we show that the Björck–Golub algorithm for numerical computation of the principal angles is mixed stable: the computed approximations of the singular values of $\mathcal{P}_\mathcal{X}\mathcal{P}_\mathcal{Y}$ approximate with small relative error the singular values of $\mathcal{P}_{\tilde{\mathcal{X}}}\mathcal{P}_{\tilde{\mathcal{Y}}}$, where $\tilde{\mathcal{X}} = \mathrm{span}(X + \Delta X)$, $\tilde{\mathcal{Y}} = \mathrm{span}(Y + \Delta Y)$ and $\max_{1\leq i\leq p} \|\Delta X e_i\|_2/\|X e_i\|_2$, $\max_{1\leq i\leq q} \|\Delta Y e_i\|_2/\|Y e_i\|_2$ are, up to factors of the dimensions, of the order of the machine precision. From this estimate, it follows that the Björck–Golub algorithm has equally small backward error angles $\angle(\mathcal{X}, \tilde{\mathcal{X}})$, $\angle(\mathcal{Y}, \tilde{\mathcal{Y}})$ for all bases XD_1, YD_2 of $\mathcal{X}$, $\mathcal{Y}$, where D_1, D_2 are arbitrary diagonal nonsingular matrices. We also show that the Björck–Golub algorithm achieves high accuracy on a wider class of problems if the QR factorizations of X and Y are computed with complete pivoting of Powell and Reid [32].

Further, we propose a new algorithm for accurate computation of the principal angles. Detailed description and analysis of the new algorithm are given in § 3. As in the Björck–Golub algorithm, our algorithm computes the SVD of $Q_x^\tau Q_y$, where Q_x and Q_y are orthonormal bases of $\mathcal{X}$ and $\mathcal{Y}$, respectively. The main novelty in the new algorithm is that Q_x and Q_y are not computed using the QR factorizations of X and Y. Instead of that, Q_x and Q_y are certain permuted lower trapezoidal orthonormal bases of $\mathcal{X}$ and $\mathcal{Y}$. The matrices Q_x and Q_y are computed by a variant of the modified Gram–Schmidt algorithm applied to permuted unit lower trapezoidal LU factors $\Pi_1 L_x$, $\Pi_2 L_y$ (Π_1, Π_2 permutations) of permuted X and Y, respectively. Error analysis shows that the new algorithm computes with small relative error the singular values of $\mathcal{P}_{\hat{\mathcal{X}}}\mathcal{P}_{\hat{\mathcal{Y}}}$, where $\hat{\mathcal{X}} = \mathrm{span}(\Pi_1(L_x + \Delta L_x))$, $\hat{\mathcal{Y}} = \mathrm{span}(\Pi_2(L_y + \Delta L_y))$, and $\max_{1\leq i\leq p} \|\Delta L_x e_i\|_2/\|L_x e_i\|_2$, $\max_{1\leq i\leq q} \|\Delta L_y e_i\|_2/\|L_y e_i\|_2$ depend on the accuracy of Gaussian elimination with pivoting. Moreover, $\sin\angle(\mathcal{X}, \hat{\mathcal{X}})$ and $\sin\angle(\mathcal{Y}, \hat{\mathcal{Y}})$ are correspondingly small. Another interesting feature of the new algorithm is that it computes with

nearly the same accuracy the principal angles of all pairs $(\mathrm{span}(D_1 X D_2), \mathrm{span}(D_3 Y D_4))$, D_1, D_2, D_3, D_4 arbitrary diagonal nonsingular matrices. We also explain (using the analysis of the new algorithm) similar high accuracy of the Björck–Golub algorithm with complete pivoting of Powell and Reid. However, since the accuracy of the new algorithm depends on backward perturbation in the bases computed by Gaussian elimination (rather than backward perturbation from the QR factorization) the new algorithm is generally more accurate than the Björck–Golub algorithm with complete pivoting. In § 4, we give numerical examples that illustrate this difference between the two approaches.

2 Analysis of the Björck–Golub algorithm

Golub and Zha [25] have shown that the Björck–Golub algorithm has the same forward error bounds as a backward stable algorithm. However, they did not show that the algorithm itself is backward stable. In this section, we prove that the Björck–Golub algorithm is mixed stable: the computed canonical correlations are close approximations of the exact canonical correlations of certain matrices $\tilde{X} \approx X$ and $\tilde{Y} \approx Y$. Detailed analysis, presented in § 2.2, shows that the backward and the forward errors are independent of column scalings of X and Y. In § 2.3, we show that the Björck–Golub algorithm achieves much higher accuracy if the QR factorization is computed with complete pivoting of Powell and Reid [32].

2.1 Preliminaries

Before we start with detailed analysis of the computation of the principal angles, let us recall some elementary results about floating–point computation of the QR factorization, floating–point matrix product, and two fundamental perturbation results for the singular values. We use the standard model of floating–point arithmetic,

$$\boldsymbol{fl}(a \odot b) = (a \odot b)(1+\xi), \quad \boldsymbol{fl}(\sqrt{c}) = \sqrt{c}(1+\zeta), \quad |\xi|, |\zeta| \leq \boldsymbol{\epsilon}, \tag{3}$$

where a, b and c are floating–point numbers, $\odot$ denotes any of the four elementary operations $+$, $-$, $\cdot$ and $\div$, and $\boldsymbol{\epsilon}$ is the machine precision (round–off unit). The following two propositions can be proved using (3). (For a proof of Proposition 2.1 see e.g. [16], [27]. For a proof of Proposition 2.2 see [24].)

Proposition 2.1 *Let $X \in \mathbf{R}^{m\times p}$, $m \geq p$, and let the QR factorization of X be computed using Givens rotations or Householder reflections. If the computation is performed in floating–point arithmetic, then there exist backward error δX and an orthonormal matrix $\tilde{Q}$ such that*

$$X + \delta X = \tilde{Q}\tilde{R}, \quad \|\delta X e_i\|_2 \leq \boldsymbol{\epsilon}_{QR}(m,p)\|Xe_i\|_2, \quad 1 \leq i \leq p, \tag{4}$$

where $\boldsymbol{\epsilon}_{QR}(m,p)$ is bounded by machine precision $\boldsymbol{\epsilon}$ times a modestly growing polynomial of m and p. Furthermore, it holds that

$$|\delta X| \leq \boldsymbol{\epsilon}_{QR}(m,p) E |X|, \quad E = ee^T \ \ (e = (1,\ldots,1)^T). \tag{5}$$

Proposition 2.2 *The floating–point product Z of an $p \times m$ matrix A and an $m \times q$ matrix B satisfies*

$$Z = AB + E, \quad |E| \leq \boldsymbol{\epsilon}_{MM}(m)|A| \cdot |B|, \quad 0 \leq \boldsymbol{\epsilon}_{MM}(m) \leq (1+\boldsymbol{\epsilon})^{m+1} - 1, \tag{6}$$

where the absolute value and the inequality are taken element–wise. Generally, $\boldsymbol{\epsilon}_{MM}(m)$ depends on the details of implementations. For instance, using double precision accumulation, the bound for $\boldsymbol{\epsilon}_{MM}(m)$ can be reduced to $O(1)\boldsymbol{\epsilon}$ for all $m < 1/\boldsymbol{\epsilon}$.

The following theorem estimates the absolute and the relative size of the perturbations of the singular values of a matrix.

Theorem 2.1 *Let $\tilde{S} = (I+E)S(I+F)$, and let $\sigma_1 \geq \cdots \geq \sigma_n$ and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_n$ be the singular values of S and $\tilde{S}$, respectively. Then $\max_{1\leq i\leq n} |\tilde{\sigma}_i - \sigma_i| \leq ||\tilde{S}-S||_2$. Furthermore, if $\max\{||E||_2, ||F||_2\} < 1$, then*

$$\max_{1\leq i\leq n} \frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq (1+||E||_2)(1+||F||_2) - 1. \tag{7}$$

For a proof of relation (7) in Theorem 2.1 see e.g. [22, Lemma 6.4 and Corollary 6.1], [28, Problem 12 in § 3.3], [16], [21], [29].

2.2 Mixed stability

The Björck–Golub algorithm follows a three step scheme: *(i)* compute the orthonormal QR factors Q_x, Q_y of the data matrices X, Y; *(ii)* compute the matrix product $S = Q_x^\tau Q_y$; *(iii)* compute the SVD of S. By following these steps in floating–point computation and by using the results from § 2.1, we obtain the following theorem.

Theorem 2.2 *Let $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_q$ be the singular values computed by the Björck–Golub algorithm. Then there exist $\tilde{X} = X + \Delta X \in \mathbf{R}^{m\times p}$, $\tilde{Y} = Y + \Delta Y \in \mathbf{R}^{m\times q}$ with the following two properties:*

(i) The values $\max_{1\leq i\leq p} ||\Delta X e_i||_2/||Xe_i||_2$ and $\max_{1\leq i\leq q} ||\Delta Y e_i||_2/||Ye_i||_2$ are of the order of machine precision times a moderate polynomial of the corresponding matrix dimensions.

(ii) If $\sigma'_1 \geq \cdots \geq \sigma'_q$ are the exact cosines of the principal angles between $\mathrm{span}(\tilde{X})$ and $\mathrm{span}(\tilde{Y})$, then, for all i, either $\tilde{\sigma}_i = \sigma'_i = 0$ or $|\tilde{\sigma}_i - \sigma'_i|/\sigma'_i$ is less than machine precision times a moderate polynomial of the matrix dimensions.

Proof: To simplify the notation, we use η_i, $i = 1, 2, \ldots$ to denote small non–negative values bounded by machine precision times a moderate function of matrix dimensions. For all reasonable dimensions, the values of η_i are much less than one.
Let $\tilde{Q}_x$, $\tilde{Q}_y$, $\tilde{R}_x$, $\tilde{R}_y$ be the computed approximations of Q_x, Q_y, R_x, R_y, respectively. Then there exist backward perturbations δX, δY and an $\eta_1 \ll 1$ such that (cf. Proposition 2.1)

$$X + \delta X = \tilde{Q}_x \tilde{R}_x, \quad Y + \delta Y = \tilde{Q}_y \tilde{R}_y, \quad \max\left\{ \max_{1\leq i\leq p} \frac{||\delta X e_i||_2}{||Xe_i||_2}, \max_{1\leq i\leq q} \frac{||\delta Y e_i||_2}{||Ye_i||_2} \right\} \leq \eta_1. \tag{8}$$

Note that computation of the orthogonal factors is generally not backward stable unless the computed matrices $\tilde{Q}_x$ and $\tilde{Q}_y$ are exactly orthonormal. (We generally cannot say that the computed matrix $\tilde{Q}_x$ is exact orthogonal factor of some $\tilde{X} \approx X$.) The best we can prove is mixed stability: $\tilde{Q}_x$ and $\tilde{Q}_y$ are close to exact orthogonal factors of $X + \delta X$ and $Y + \delta Y$, respectively. This is ensured since there exists an $\eta_2 \ll 1$ such that

$$\max\{||\tilde{Q}_x^\tau \tilde{Q}_x - I||_F, ||\tilde{Q}_y^\tau \tilde{Q}_y - I||_F\} \leq \eta_2.$$

(Here we assume that we use a QR factorization algorithm that ensures almost orthogonality of the computed matrices $\tilde{Q}_x$ and $\tilde{Q}_y$.) So, for example, if $\tilde{Q}_x = Q'_x(I+T'_x)$ is the (exact) QR factorization of $\tilde{Q}_x$, then the upper triangular matrix T'_x satisfies $||T'_x||_2 \leq \eta_2$ and the mixed stability of the computation of the orthonormal QR factor follows from the relation

$$\tilde{X} = X + \Delta X \equiv X + \delta X = Q'_x \left((I + T'_x)\tilde{R}_x \right), \quad ||\tilde{Q}_x - Q'_x||_2 \leq ||T'_x||_2.$$

Similarly, $Y + \delta Y = Q'_y(I + T'_y)\tilde{R}_y$, $||T'_y||_2 \leq \eta_2$. Let $\tilde{S} = \boldsymbol{fl}(\tilde{Q}_x^\tau \tilde{Q}_y)$. Then $\tilde{S} = \tilde{Q}_x^\tau \tilde{Q}_y + E_S$, where $||E_S||_2 \leq \eta_3$ (cf. Proposition 2.2). The computed singular values $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_q$ are the exact

singular values of $\tilde{S} + \delta\tilde{S}$, where $\|\delta\tilde{S}\|_2 \leq \eta_4$. Here the values of $\eta_3 \ll 1$, $\eta_4 \ll 1$ depend on the details of computation (cf. [24]). We can write the matrix $\tilde{S} + \delta\tilde{S}$ as

$$\begin{aligned} \tilde{S} + \delta\tilde{S} &= \tilde{Q}_x^\tau \tilde{Q}_y + E_S + \delta\tilde{S} = \tilde{Q}_x^\tau (\tilde{Q}_y + (\tilde{Q}_x^\tau)^\dagger (E_S + \delta\tilde{S})) \\ &= (I + T_x')^\tau \left((Q_x')^\tau Q_y'' \right) (I + T_y''), \end{aligned}$$

where $(\tilde{Q}_x^\tau)^\dagger = Q_x'(I + T_x')^{-\tau}$ and $\tilde{Q}_y + (\tilde{Q}_x^\tau)^\dagger (E_S + \delta\tilde{S}) = Q_y''(I + T_y'')$ is the QR factorization of an almost orthonormal matrix with $\|T_y''\|_2 \leq \eta_5 \ll 1$. Note that

$$\tilde{Y} = Y + \Delta Y \equiv Y + \delta Y + (\tilde{Q}_x^\tau)^\dagger (E_S + \delta\tilde{S})\tilde{R}_y = Q_y''(I + T_y'')\tilde{R}_y,$$

and that, for all $1 \leq i \leq q$,

$$\begin{aligned} \| \left((\tilde{Q}_x^\tau)^\dagger (E_S + \delta\tilde{S})\tilde{R}_y \right) e_i \|_2 &\leq \frac{\|E_S\|_2 + \|\delta\tilde{S}\|_2}{1 - \|T_x'\|_2} \frac{1 + \eta_1}{1 - \|T_y'\|_2} \|Y e_i\|_2 \\ &\leq (1 + \eta_1) \frac{\eta_3 + \eta_4}{(1 - \eta_2)^2} \|Y e_i\|_2. \end{aligned}$$

The proof completes by noting that the singular values of $(Q_x')^\tau Q_y''$ are the cosines of the principal angles between $\mathrm{span}(\tilde{X})$ and $\mathrm{span}(\tilde{Y})$ and by invoking Theorem 2.1. Q.E.D.

The backward errors in Theorem 2.2 are small norm–wise relative errors in the columns of the bases X and Y of $\mathcal{X}$, $\mathcal{Y}$, respectively. These estimates, however, do not guarantee that the backward perturbations of $\mathcal{X}$ and $\mathcal{Y}$ are small in the *angle metric*. In this paper we use the following angle metric introduced by Wedin [45]. For any two subspaces $\mathcal{S}_1$, $\mathcal{S}_2$ with corresponding orthogonal projections $\mathcal{P}_{\mathcal{S}_1}$, $\mathcal{P}_{\mathcal{S}_2}$ (possibly of different ranks), the angle $\angle(\mathcal{S}_1, \mathcal{S}_2)$ is defined by

$$\angle(\mathcal{S}_1, \mathcal{S}_2) = \min \left\{ \arcsin \left(\|(I - \mathcal{P}_{\mathcal{S}_2})\mathcal{P}_{\mathcal{S}_1}\|_2 \right), \arcsin \left(\|(I - \mathcal{P}_{\mathcal{S}_1})\mathcal{P}_{\mathcal{S}_2}\|_2 \right) \right\}.$$

(If $\dim(\mathcal{S}_1) = \dim(\mathcal{S}_2)$, $\sin\angle(\mathcal{S}_1, \mathcal{S}_2) = \|\mathcal{P}_{\mathcal{S}_1} - \mathcal{P}_{\mathcal{S}_2}\|_2$.) Consider now the angle $\angle(\mathcal{X}, \tilde{\mathcal{X}})$, where $\mathcal{X} = \mathrm{span}(X)$, $\tilde{\mathcal{X}} = \mathrm{span}(X + \Delta X)$, and X, ΔX are as in Theorem 2.2. If $D_X = \mathrm{diag}(\|Xe_i\|_2)$, $X = X_c D_X$, $\Delta X_c = \Delta X D_X^{-1}$, then $\mathcal{X} = \mathrm{span}(X_c)$, $\tilde{\mathcal{X}} = \mathrm{span}(X_c + \Delta X_c)$ and an estimate of Wedin [45] yields

$$\sin\angle(\mathcal{X}, \tilde{\mathcal{X}}) \leq \|X_c^\dagger\|_2 \sqrt{p} \max_{1 \leq i \leq p} \frac{\|\Delta X e_i\|_2}{\|Xe_i\|_2}.$$

Hence, we have the following corollary of Theorem 2.2.

Corollary 2.1 *Let the assumptions of Theorem 2.2 hold. Then there exist subspaces $\tilde{\mathcal{X}}$, $\tilde{\mathcal{Y}}$ and a modest polynomial $f(m, p, q)$ such that*

$$\sin\angle(\mathcal{X}, \tilde{\mathcal{X}}) \leq \|X_c^\dagger\|_2 \sqrt{p} \max_{1 \leq i \leq p} \frac{\|\Delta X e_i\|_2}{\|Xe_i\|_2}, \quad \sin\angle(\mathcal{Y}, \tilde{\mathcal{Y}}) \leq \|Y_c^\dagger\|_2 \sqrt{q} \max_{1 \leq i \leq q} \frac{\|\Delta Y e_i\|_2}{\|Ye_i\|_2},$$

and such that the computed values $(\tilde{\sigma}_i)_{i=1}^q$ are up to a relative error of order $f(m, p, q)\epsilon$ the exact singular values of $\mathcal{P}_{\tilde{\mathcal{X}}}\mathcal{P}_{\tilde{\mathcal{Y}}}$.

In the principal angle computation, the angles $\angle(\mathcal{X}, \tilde{\mathcal{X}})$, $\angle(\mathcal{Y}, \tilde{\mathcal{Y}})$ seem to be very natural metric to measure backward errors. The following perturbation estimate illustrates this observation.

Theorem 2.3 *Let $\mathcal{X}$, $\tilde{\mathcal{X}}$, $\mathcal{Y}$, $\tilde{\mathcal{Y}}$ be subspaces of $\mathbf{R}^m$ (or $\mathbf{C}^m$) with $\dim(\mathcal{X}) = \dim(\tilde{\mathcal{X}})$, $\dim(\mathcal{Y}) = \dim(\tilde{\mathcal{Y}})$, and let*

$$\eta = \sin\angle(\mathcal{X}, \tilde{\mathcal{X}}) + \sin\angle(\mathcal{Y}, \tilde{\mathcal{Y}}) + \sin\angle(\mathcal{X}, \tilde{\mathcal{X}}) \cdot \sin\angle(\mathcal{Y}, \tilde{\mathcal{Y}}).$$

Let $\Sigma = \mathrm{diag}(\sigma_i)$, $\Xi = \mathrm{diag}(\xi_j)$, $\tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_i)$, $\tilde{\Xi} = \mathrm{diag}(\tilde{\xi}_j)$ be the singular values of $\mathcal{P}_{\mathcal{X}}\mathcal{P}_{\mathcal{Y}}$, $(I - \mathcal{P}_{\mathcal{X}})\mathcal{P}_{\mathcal{Y}}$, $\mathcal{P}_{\tilde{\mathcal{X}}}\mathcal{P}_{\tilde{\mathcal{Y}}}$, $(I - \mathcal{P}_{\tilde{\mathcal{X}}})\mathcal{P}_{\tilde{\mathcal{Y}}}$, respectively. Then

$$\|\Sigma - \tilde{\Sigma}\|_2 \leq \eta, \quad \|\Xi - \tilde{\Xi}\|_2 \leq \eta. \tag{9}$$

Furthermore, let $\Theta = \mathrm{diag}(\vartheta_i)$, $\tilde{\Theta} = \mathrm{diag}(\tilde{\vartheta}_i)$ *be the principal angles between* $\mathcal{X}$ *and* $\mathcal{Y}$, *and between* $\tilde{\mathcal{X}}$ *and* $\tilde{\mathcal{Y}}$, *respectively. Then* $\eta < (\sqrt{2}-1)/\sqrt{2}$ *implies*

$$\|\Theta - \tilde{\Theta}\|_2 \leq \frac{\eta}{\sqrt{1-(\eta+1/\sqrt{2})^2}}. \tag{10}$$

Proof: The first inequality in (9) follows from

$$\begin{aligned} \|\Sigma - \tilde{\Sigma}\|_2 &\leq \|\mathcal{P}_{\mathcal{X}}\mathcal{P}_{\mathcal{Y}} - \mathcal{P}_{\tilde{\mathcal{X}}}\mathcal{P}_{\tilde{\mathcal{Y}}}\|_2 = \|(\mathcal{P}_{\mathcal{X}} - \mathcal{P}_{\tilde{\mathcal{X}}})\mathcal{P}_{\mathcal{Y}} + \mathcal{P}_{\mathcal{X}}(\mathcal{P}_{\mathcal{Y}} - \mathcal{P}_{\tilde{\mathcal{Y}}}) - (\mathcal{P}_{\mathcal{X}} - \mathcal{P}_{\tilde{\mathcal{X}}})(\mathcal{P}_{\mathcal{Y}} - \mathcal{P}_{\tilde{\mathcal{Y}}})\|_2 \\ &\leq \sin\angle(\mathcal{X},\tilde{\mathcal{X}}) + \sin\angle(\mathcal{Y},\tilde{\mathcal{Y}}) + \sin\angle(\mathcal{X},\tilde{\mathcal{X}})\cdot\sin\angle(\mathcal{Y},\tilde{\mathcal{Y}}). \end{aligned}$$

Similarly, the second inequality in (9) follows from $\|\Xi - \tilde{\Xi}\|_2 \leq \|(I-\mathcal{P}_{\mathcal{X}})\mathcal{P}_{\mathcal{Y}} - (I-\mathcal{P}_{\tilde{\mathcal{X}}})\mathcal{P}_{\tilde{\mathcal{Y}}}\|_2$ and from the identity $\angle(\mathcal{X},\tilde{\mathcal{X}}) = \angle(\mathcal{X}^{\perp},\tilde{\mathcal{X}}^{\perp})$. To prove (10), we first note that

$$|\vartheta_i - \tilde{\vartheta}_i| = \int_{\min\{\sigma_i,\tilde{\sigma}_i\}}^{\max\{\sigma_i,\tilde{\sigma}_i\}} \frac{dt}{\sqrt{1-t^2}} = \int_{\min\{\xi_i,\tilde{\xi}_i\}}^{\max\{\xi_i,\tilde{\xi}_i\}} \frac{dt}{\sqrt{1-t^2}}, \tag{11}$$

and that $\min\{\max\{\sigma_i,\tilde{\sigma}_i\},\max\{\xi_i,\tilde{\xi}_i\}\} \leq 1/\sqrt{2}+\eta$. Then we estimate the integrals in (11). Q.E.D.

Corollary 2.1 shows that the backward error angle in the Björck–Golub algorithm is independent of column scalings of the bases X and Y, and that this angle might be large only if $\min_{D=\mathrm{diag}}\kappa_2(XD)$ and $\min_{D=\mathrm{diag}}\kappa_2(YD)$ are large. (Here we recall the near optimality of $\kappa_2(X_c)$, $\kappa_2(X_c) \leq \sqrt{p}\min_{D=\mathrm{diag}}\kappa_2(XD)$; see [42].) In that case, certain, even very small, norm–wise relative changes of the columns of the ill–conditioned basis X might cause arbitrarily large flutter of the corresponding subspace. The following example will illustrate. Let

$$X = \begin{bmatrix} 1 & 1 \\ \epsilon & -\epsilon \\ \epsilon & \epsilon \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} 1 & 1 \\ \epsilon & -\epsilon \\ \epsilon & -\epsilon \end{bmatrix}, \quad |\epsilon| \ll 1, \tag{12}$$

and let $\mathcal{X} = \mathrm{span}(X)$, $\tilde{\mathcal{X}} = \mathrm{span}(\tilde{X})$, $\mathcal{Y} = \mathrm{span}(Y)$. Note that the angle between $\mathcal{X}$ and $\mathcal{Y}$ is fairly large ($\mathcal{X}$ is close to $\mathrm{span}(e_1+e_2)$) and that the corresponding columns of X and $\tilde{X}$ differ by small ($O(\epsilon)$) angles. However, it holds that $\mathcal{Y} \subset \tilde{\mathcal{X}}$. Using MATLAB with $\epsilon = 1000 * \mathrm{eps} \approx 2.22\cdot 10^{-13}$, we compute the orthogonal factors of X and $\tilde{X}$, respectively, as

$$\tilde{Q}_x \approx \begin{bmatrix} -1.00 & 2.22\cdot 10^{-13} \\ -2.22\cdot 10^{-13} & -1.00 \\ -2.22\cdot 10^{-13} & 2.46\cdot 10^{-26} \end{bmatrix}, \quad \widetilde{\tilde{Q}_x} \approx \begin{bmatrix} -1.00 & 3.14\cdot 10^{-13} \\ -2.22\cdot 10^{-13} & -7.07\cdot 10^{-1} \\ -2.22\cdot 10^{-13} & -7.07\cdot 10^{-1} \end{bmatrix}.$$

Hence, the principal angles between $\mathcal{X}$ and $\mathcal{Y}$ are poorly determined in the presence of such errors. This behavior is also captured by the following theorem of Golub and Zha [25].

Theorem 2.4 *Let* $X \in \mathbf{R}^{m\times p}$ *and* $Y \in \mathbf{R}^{m\times q}$ *be full column rank matrices and let* $X = X_c D_X$, $Y = Y_c D_Y$, *where* $D_X = \mathrm{diag}(\|Xe_i\|_2)$, $D_Y = \mathrm{diag}(\|Ye_i\|_2)$. *Let* $\tilde{X} = X + \Delta X$, $\tilde{Y} = Y + \Delta Y$ *be also full column rank matrices such that* $|\Delta X| \leq \epsilon G_X|X|$, $|\Delta Y| \leq \epsilon G_Y|Y|$, *where* $0 \leq \epsilon \ll 1$ *and* G_X, G_Y *are matrices with nonnegative elements. Let* $\mathcal{X} = \mathrm{span}(X)$, $\mathcal{Y} = \mathrm{span}(Y)$, $\tilde{\mathcal{X}} = \mathrm{span}(\tilde{X})$, $\tilde{\mathcal{Y}} = \mathrm{span}(\tilde{Y})$. *Let* $\mathcal{C}(\mathcal{X},\tilde{\mathcal{X}})$ *be the orthogonal complement of* $\mathcal{X}\bigcap\tilde{\mathcal{X}}$ *in* $\mathcal{X}+\tilde{\mathcal{X}}$, *and let* ξ *be the minimal angle between* $\mathcal{C}(\mathcal{X},\tilde{\mathcal{X}})$ *and* $\mathcal{Y}$. *Similarly, let* ζ *be defined as minimal angle between* $\mathcal{C}(\mathcal{Y},\tilde{\mathcal{Y}})$ *and* $\tilde{\mathcal{X}}$. *If* Σ, $\tilde{\Sigma}$ *are the diagonal matrices of the cosines of the principal angles between* $\mathcal{X}$ *and* $\mathcal{Y}$, *and between* $\tilde{\mathcal{X}}$ *and* $\tilde{\mathcal{Y}}$, *respectively, then*

$$\|\tilde{\Sigma} - \Sigma\|_2 \leq \epsilon\left(\sqrt{p(m-p)}\|G_X\|_2\kappa_2(X_c)\cos\xi + \sqrt{q(m-q)}\|G_Y\|_2\kappa_2(Y_c)\cos\zeta\right).$$

This theorem shows that the accuracy of the singular values of $\mathcal{P}_{\mathcal{X}}\mathcal{P}_{\mathcal{Y}}$ depends on the condition number of column scaled matrices X_c and Y_c.

Remark 2.1 It can be easily shown that the error bound in Theorem 2.4 can be improved to

$$||\tilde{\Sigma} - \Sigma||_2 \leq \epsilon\,(\sqrt{p}||G_X||_2\kappa_2(X_c)\cos\xi + \sqrt{q}||G_Y||_2\kappa_2(Y_c)\cos\zeta)\,.$$

We conclude our analysis of the Björck–Golub algorithm with an experimental illustration of the bounds in Theorem 2.2 and Corollary 2.1.

Example 2.1 We generate test pair (X, Y) as follows. We write X, Y as $X = X_s D_X$, $Y = Y_s D_Y$, where $D_X = \mathrm{diag}(||Xe_i||_2)$, $D_Y = \mathrm{diag}(||Ye_i||_2)$, and $\kappa_2(X_s), \kappa_2(Y_s) \in \{10^i,\ i = 2, \ldots, 6\}$, $\kappa_2(D_X) \in \{10^8, 10^{12}, 10^{16}\}$, $\kappa_2(D_Y) \in \{10^9, 10^{13}, 10^{15}\}$. For fixed values of the condition numbers $(\kappa_2(X_s), \kappa_2(D_X), \kappa_2(Y_s), \kappa_2(D_Y))$ we generate X_s, D_X, Y_s, D_Y with different distributions of singular values. We use the procedure `DLATM1()` from [14], and we choose the values of the parameter `MODE` so that the distributions of the singular values of X_s, D_X, Y_s, D_Y are from the set $\{5, 3\} \times \{5\} \times \{5, -4\} \times \{5\}$. In this way, we generate 900 test pairs (X, Y), divided into 25 classes, where the pairs from the same class $\mathcal{C}_{ij}$ have nearly the same value of $(\kappa_2(X_c), \kappa_2(Y_c)) \approx (10^i, 10^j)$, $2 \leq i, j \leq 6$. We measure the backward error angles in the following way. We use single precision floating–point arithmetic to find approximate orthonormal bases $\tilde{Q}_x$ and $\tilde{Q}_x^{\perp}$ for $\mathcal{X}$ and $\mathcal{X}^{\perp}$, respectively. Then, we use double precision computation to compute the sine of the angle between $\mathrm{span}(\tilde{Q}_x)$ and $\mathcal{X}$. This is accomplished by an application of the Björck–Golub algorithm to the matrices $\tilde{Q}_x^{\perp}$ and X. The same procedure is applied to $\tilde{Q}_y$ and Y. (It is clear from the proof of Theorem 2.2 that the computation of the orthonormal bases introduces the major part of the error. Hence, this experiment gives a useful insight into the overall accuracy of the algorithm.) The QR factorizations are computed using the LAPACK [1] procedure `SGEQRF()`. The results of the test with $m = 200$, $p = 100$, $q = 50$ are given in Figure 1, where

$$\mathbf{e}_{ij} = \max_{(X,Y)\in\mathcal{C}_{ij}} \max\{\sin\angle(\mathrm{span}(\tilde{Q}_x), \mathcal{X}), \sin\angle(\mathrm{span}(\tilde{Q}_y), \mathcal{Y})\}.$$

Note that Figure 1 indicates that the error bounds in Theorem 2.2 and Corollary 2.1 are almost attainable.

2.3 Modified algorithm

In some situations, large $\kappa_2(X_c)$ and $\kappa_2(Y_c)$ represent artificial ill–conditioning and Björck–Golub algorithm can be modified to be more accurate than the above theory predicts. For example, let $X = X_s D$, where X_s is well–conditioned. Then replacing X with $D'X$, where D' is ill–conditioned diagonal matrix, produces an ill–conditioned problem in the framework of the analyses of Theorem 2.2 and Theorem 2.4. Numerical experiments clearly show that the algorithm indeed becomes unstable and that it computes with large backward (and forward) errors. Such large errors are caused by known stability problem of the QR factorization of the matrix with widely differing row norms. To overcome this problem, we use the QR factorization with column and row pivoting, as suggested by Powell and Reid [32]. The results of an experiment are reported in the next example.

Example 2.2 We follow a similar test procedure as in Example 2.1. The only difference is that the generated matrices $X = X_s D_X$, $Y = Y_s D_Y$ are updated as follows: $X := D'_X X$, $Y := D'_Y Y$, where D'_X, D'_Y are randomly generated in the same way (with the same parameters) as D_X, D_Y, respectively. In this way, $\kappa_2(X_c)$ and $\kappa_2(Y_c)$ are increased, but in a very structured way. The QR factorizations are computed by a modification of the LAPACK procedure `SGEQPF()`. The results of the test are shown in Figure 2. In Figure 3, we show the values of $\mathbf{e}_x = \sin\angle(\tilde{Q}_x, \mathcal{X})$, $\kappa_2(X_c)$, $\mathbf{e}_x/\kappa_2(X_c)$, where X_c is obtained from X by column equilibration.

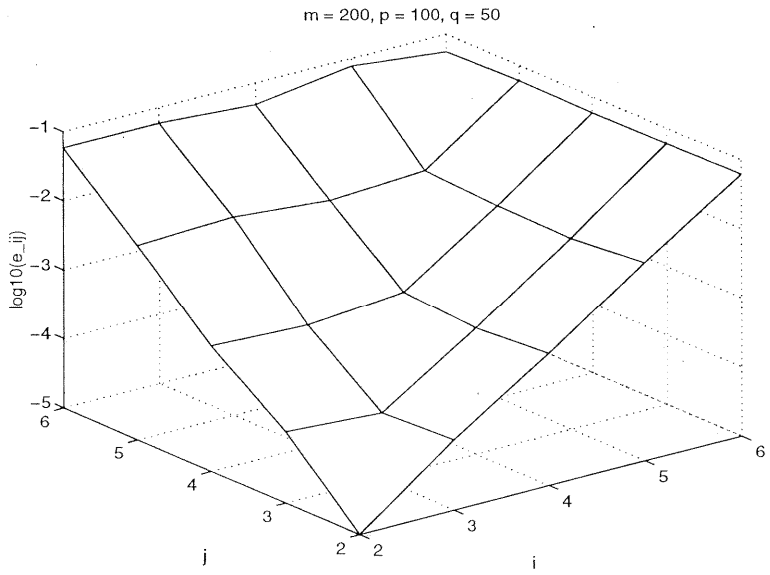


Figure 1: The values of $\log_{10} \mathbf{e}_{ij}$, $2 \leq i, j \leq 6$ in Example 2.1. Note that $\log_{10} \mathbf{e}_{ij} \approx 10^{7-\max\{i,j\}}$, as predicted by the theory.

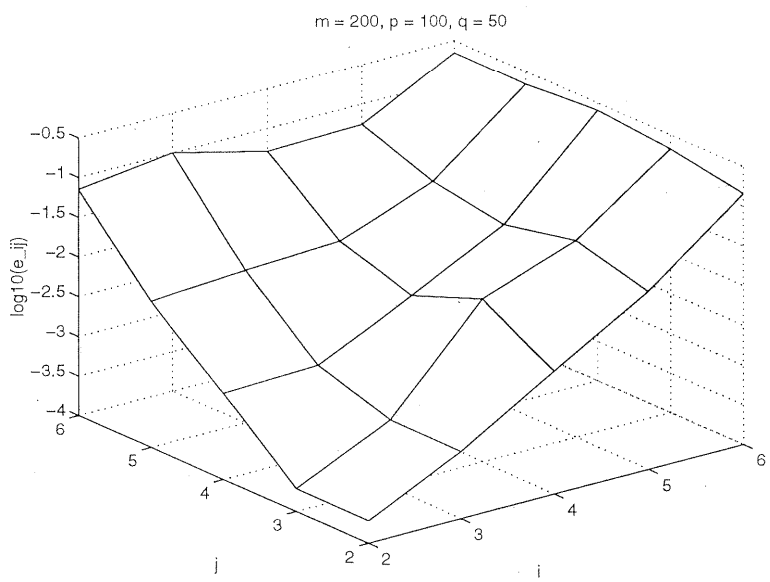


Figure 2: The values of $\log_{10} \mathbf{e}_{ij}$, $2 \leq i, j \leq 6$ in Example 2.2.

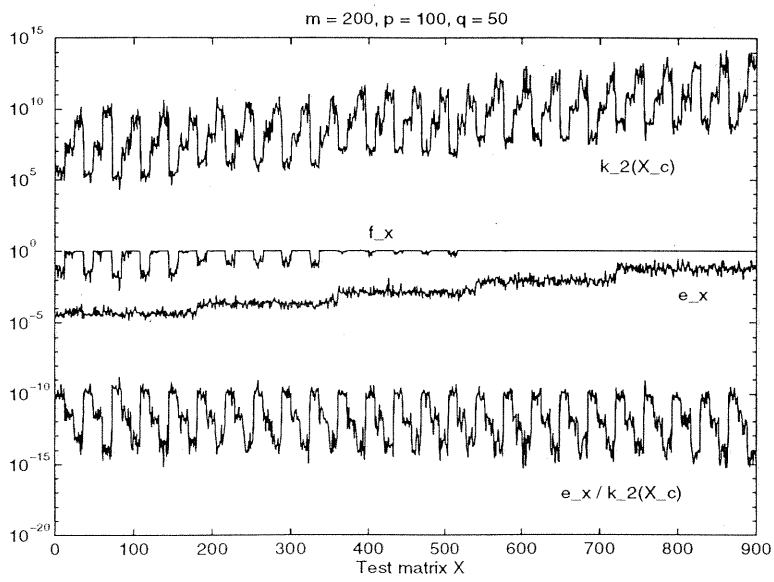


Figure 3: The values of $\mathbf{e}_x = \sin\angle(\tilde{Q}_x, \mathcal{X})$, $\kappa_2(X_c)$, $\mathbf{e}_x/\kappa_2(X_c)$ for all 900 examples of the basis X and with the modified Björck–Golub algorithm (Example 2.2). The values of $\mathbf{f}_x$ represent the backward error angles $\sin\angle(\tilde{Q}_x, \mathcal{X})$, where $\tilde{Q}_x$ is computed with QR factorization without Powell–Reid pivoting.

Note the similarity of the error behavior in Figure 1 and Figure 2. (In this example, we have observed similar accuracy if the Powell–Reid pivoting is replaced by simple reordering of the rows of the matrix so that their Euclidean norms are in descending order.)

In the next two examples, we show the difference in the forward errors for the two variants of the Björck–Golub algorithm.

Example 2.3 In this example we show that QR based computation of the orthogonal bases can introduce large error and it can fail to detect that, for example, one of the principal angles is close to $\pi/2$. We take the bases X and Y to be

$$X \approx \begin{bmatrix} 0.57378941\cdot 10^{17} & -0.74737239\cdot 10^{09} & -0.10439621\cdot 10^{02} \\ -0.75415686\cdot 10^{29} & 0.25173789\cdot 10^{22} & -0.11089462\cdot 10^{14} \\ -0.52912208\cdot 10^{19} & 0.51559708\cdot 10^{12} & -0.63842515\cdot 10^{04} \\ 0.26020839\cdot 10^{26} & -0.72667785\cdot 10^{18} & 0.14745371\cdot 10^{10} \\ 0.21463361\cdot 10^{22} & -0.76107815\cdot 10^{14} & 0.39906168\cdot 10^{06} \\ 0.13388386\cdot 10^{26} & -0.48858418\cdot 10^{19} & 0.75605997\cdot 10^{11} \\ -0.43084490\cdot 10^{20} & 0.33985776\cdot 10^{13} & -0.38962076\cdot 10^{05} \end{bmatrix}$$

$$Y \approx \begin{bmatrix} 0.12378225\cdot 10^{+00} & -0.17331250\cdot 10^{+13} \\ 0.84008590\cdot 10^{-09} & 0.17773952\cdot 10^{+05} \\ -0.26428604\cdot 10^{-14} & -0.98536731\cdot 10^{-01} \\ 0.13059467\cdot 10^{-12} & -0.80072369\cdot 10^{+00} \\ 0.18943973\cdot 10^{-11} & -0.20708348\cdot 10^{+01} \\ -0.16178360\cdot 10^{+01} & -0.33048027\cdot 10^{+13} \\ 0.40286435\cdot 10^{-06} & 0.10409793\cdot 10^{+09} \end{bmatrix},$$

where the entries of the corresponding double precision arrays are shown to eight decimal places. In Table 13, we show the computed approximations of the cosines of the principal angles. (π–Björck-Golub refers to the Björck-Golub algorithm with pivoting suggested by Powell and Reid.

Algorithm 3.1 is described in § 3. At this point, the only purpose of the last row in Table 13 is to give a second set of double precision reference values (singular value approximations).)

$\tilde{\sigma}_i$	Björck-Golub (single)	Björck-Golub (double)	π–Björck-Golub (single)
$\tilde{\sigma}_1$	$0.10000002 \cdot 10^{1}$	0.9999999910693745	$0.10000000 \cdot 10^{1}$
$\tilde{\sigma}_2$	0.91120803	$0.2269574724944604 \cdot 10^{-6}$	$0.21987161 \cdot 10^{-6}$
Algorithm 3.1 (double): $\tilde{\sigma}_1 \approx 0.9999999910693748$, $\tilde{\sigma}_2 \approx 0.2219392787298458 \cdot 10^{-6}$.			

(13)

Since $\tilde{\sigma}_2 \approx 3.7\epsilon$, it is determined only to an absolute uncertainty of order ϵ. To illustrate this, we multiply the entries of X and Y by randomly chosen numbers $1 \pm \epsilon_{ij}$ with $|\epsilon_{ij}| \leq 10^{-4}$. The single precision Björck–Golub algorithm and π–Björck-Golub algorithm compute, respectively, $\tilde{\sigma}_2 \approx 0.99112201$ and $\tilde{\sigma}_2 \approx 0.75409122 \cdot 10^{-6}$. The double precision computation gives $\tilde{\sigma}_2 \approx 0.7685475597770073 \cdot 10^{-6}$. The maximal principal angle is not sensitive to this change since $\tilde{\vartheta}_2 \approx \arccos(0.21987161 \cdot 10^{-6})$ and $\tilde{\tilde{\vartheta}}_2 \approx \arccos(0.75409122 \cdot 10^{-6})$ satisfy $\tilde{\vartheta}_2/\tilde{\tilde{\vartheta}}_2 \approx 1.0000003$ and $(\pi/2)/\min\{\tilde{\vartheta}_2, \tilde{\tilde{\vartheta}}_2\} \approx 1.0000005$. This is obvious from the formula

$$\vartheta_i = \frac{\pi}{2} - \int_0^{\sigma_i} \frac{dt}{\sqrt{1-t^2}}.$$

Remark 2.2 The value of $\tilde{\sigma}_1 = 0.10000002 \cdot 10^{1}$ in Table 13 shows that mixed stability is the right framework for the numerical analysis of principal angle computation. No backward perturbation can in exact arithmetic lead to $0.10000002 \cdot 10^{1}$ as the cosine of a principal angle. (Strictly speaking, the Björck–Golub algorithm (and Algorithm 3.1 in § 3) are not backward stable.)

Example 2.4 The most critical part of the principal angle computation is the computation of the orthonormal bases of the given spaces. If that computation introduces large errors that "rotate" the initial spaces, there is no way to tell which singular value of $\mathcal{P}_{\mathcal{X}}\mathcal{P}_{\mathcal{Y}}$ will suffer the largest perturbation. In this example, we show that the largest error might be in the largest singular value, while the smallest one is computed very accurately. Let

$$X \approx \begin{bmatrix} 0.81909804 \cdot 10^{01} & -0.85610022 \cdot 10^{02} & -0.19108842 \cdot 10^{12} \\ -0.31793150 \cdot 10^{11} & 0.15111104 \cdot 10^{13} & 0.26747300 \cdot 10^{22} \\ -0.51921289 \cdot 10^{12} & 0.32394455 \cdot 10^{13} & 0.74985519 \cdot 10^{22} \\ -0.12806811 \cdot 10^{16} & 0.32962115 \cdot 10^{16} & 0.11506216 \cdot 10^{26} \\ 0.11302525 \cdot 10^{03} & -0.85968597 \cdot 10^{03} & -0.16852694 \cdot 10^{13} \\ 0.85886880 \cdot 10^{16} & -0.89292760 \cdot 10^{17} & -0.17015941 \cdot 10^{27} \\ 0.14028936 \cdot 10^{05} & -0.69895642 \cdot 10^{06} & -0.11412105 \cdot 10^{16} \end{bmatrix}$$

$$Y \approx \begin{bmatrix} -0.77654567 \cdot 10^{-4} & -0.42605337 \cdot 10^{-06} \\ -0.52320495 \cdot 10^{-7} & -0.42627118 \cdot 10^{-09} \\ -0.12184166 \cdot 10^{-6} & -0.47657759 \cdot 10^{-09} \\ 0.34901023 \cdot 10^{-6} & 0.19476305 \cdot 10^{-08} \\ 0.22741771 \cdot 10^{+4} & 0.86991999 \cdot 10^{+01} \\ 0.15964494 \cdot 10^{-8} & 0.15686126 \cdot 10^{-10} \\ 0.75523679 \cdot 10^{-9} & 0.46711879 \cdot 10^{-11} \end{bmatrix}$$

The computed singular values are given in Table 14. (As in Example 2.3, the last row in Table 14 and in Table 15 is used only as a second set of reference values.)

$\tilde{\sigma}_i$	Björck-Golub (single)	Björck-Golub (double)	π–Björck-Golub (single)
$\tilde{\sigma}_1$	0.99059296	$0.5015345317976148 \cdot 10^{-2}$	$0.48222207 \cdot 10^{-2}$
$\tilde{\sigma}_2$	$0.25136729 \cdot 10^{-9}$	$0.2510846255712600 \cdot 10^{-9}$	$0.22261035 \cdot 10^{-9}$
Algorithm 3.1 (double): $\tilde{\sigma}_1 \approx 0.5015345505648627 \cdot 10^{-2}$, $\tilde{\sigma}_2 \approx 0.2510846257369576 \cdot 10^{-9}$.			

(14)

To illustrate how well σ_1 and σ_2 are determined by the data, we introduce random rounding errors of order 10^{-4} into the entries of X and Y and we run the test again. The results are shown in Table 15.

$\tilde{\sigma}_i$	Björck-Golub (single)	Björck-Golub (double)	π–Björck-Golub (single)
$\tilde{\sigma}_1$	0.99053305	$0.5013070874085607 \cdot 10^{-2}$	$0.48202435 \cdot 10^{-2}$
$\tilde{\sigma}_2$	$0.24991473 \cdot 10^{-9}$	$0.2502365149198476 \cdot 10^{-9}$	$0.22178702 \cdot 10^{-9}$
Algorithm 3.1 (double): $\tilde{\sigma}_1 \approx 0.5013070984780922 \cdot 10^{-2}$, $\tilde{\sigma}_2 \approx 0.2502365153154889 \cdot 10^{-9}$.			

(15)

The benefits of row pivoting in the QR factorization are well known in solving weighted least squares problems and there exist computational experience and satisfactory backward error bounds which justify the need for row interchanges; see e.g. [3], [43], [8]. In what follows, we try to contribute to further understanding of the nature of the backward error and its implications to the forward perturbation of the orthogonal QR factor. We first recall the following result of Powell and Reid.

Proposition 2.3 *Let X, δX, $\tilde{Q}$, $\tilde{R}$ be as in Proposition 2.1, and let the QR factorization be computed by a sequence of p Householder reflections. Let $\tilde{X}^{(k)}$, $k \in \{1, \ldots, p\}$, denote the floating-point matrix computed in the kth step of the algorithm, let $X^{(k)}$, $k \in \{1, \ldots, p\}$, denote the matrix in the kth step of exact computation, and let $\tilde{X} = X + \delta X$ and*

$$\rho_i(\tilde{X}) = \max_{j,k} |(\tilde{X}^{(k)})_{ij}|, \quad \rho_i(X) = \max_{j,k} |(X^{(k)})_{ij}|, \quad i = 1, \ldots, m. \tag{16}$$

Then there exists a modest polynomial $h(p)$ such that $|\delta X_{ij}| \leq h(p)\epsilon\rho_i(\tilde{X})$. Furthermore, if the columns of X are permuted following the pivoting of Golub [23], and if, in addition, the rows of the matrices $X^{(k)}$ are permuted so that, for all k, $|(X^{(k)})_{kk}| = \max_{i \geq k} |(X^{(k)})_{ik}|$, then

$$\max_j |(X^{(k+1)})_{k,j}| \leq \sqrt{m}|(X^{(k)})_{kk}|, \quad \textit{and} \quad \rho_i(X) \leq (1+\sqrt{2})^{i-1}\sqrt{m}\max_j |X_{ij}|.$$

Powell and Reid report that the pivot growth factors

$$\mu_i(X) = \frac{\rho_i(X)}{\max_j |X_{ij}|}, \quad 1 \leq i \leq m, \tag{17}$$

are usually moderate and that the exponential growth is attained only in pathological cases.

Remark 2.3 Barlow [3] has shown that the G–algorithm of Bareiss [2] has similar backward error bound as in Proposition 2.3. Hence, the forthcoming analysis can be applied to the G–algorithm as well.

Let us continue with two observations.

(i) In the computation of an orthonormal basis for span(X), column scaling $X := XD$ (D diagonal, $\det(D) \neq 0$) is admissible transformation, while the row scaling $X := D'X$ (D' diagonal, $\det(D') \neq 0$) is not. In other words, we may scale the columns of X to achieve better numerical stability or tighter perturbation bounds, but we must not scale the rows of X.

(ii) Let $X = X_c'D$, where D is a diagonal matrix of powers of the base of the floating–point arithmetic. Then, in the absence of underflow and overflow, the QR factorizations of X and X_c' are numerically equivalent in the sense that both compute the same floating–point approximation $\tilde{Q}_x \approx Q_x$. (Q_x denotes the exact orthogonal factor of X.)

To simplify the notation, we assume that the initial matrix is permuted so that no column or row interchanges are necessary in the Powell–Reid QR factorization with pivoting. We also assume that we can write $X = D'X_cD = X_c'D$, where D, D' are nonsingular diagonal scalings, the diagonals of D are powers of the base of the floating–point arithmetic, and that no column interchanges are necessary to compute the QR factorization with column pivoting of X_c'. (In that case, neither the row interchanges are necessary in the Powell–Reid row pivoting.) We let $\tilde{Q}_x$ denote the computed approximation of the orthonormal basis Q_x. Using Proposition 2.3, we conclude that there exist an exactly orthonormal matrix Q_x' and a backward error δX such that the computed triangular factor $\tilde{R}_x$ satisfies

$$X + \delta X = Q_x'\tilde{R}_x, \quad |\delta X_{ij}| \leq h(p)\epsilon\mu_i(\tilde{X}) \max_j |X_{ij}|. \tag{18}$$

By observation *(ii)*, we can also write

$$\tilde{X}_c' \equiv X_c' + \delta(X_c') = Q_x'(\tilde{R}_x D^{-1}), \quad |(\delta(X_c'))_{ij}| \leq h(p)\epsilon\mu_i(\tilde{X}_c') \max_j |(X_c')_{ij}| \tag{19}$$

We can rewrite relation (19) to

$$D'(X_c + D'^{-1}\delta(X_c'))D = Q_x'\tilde{R}_x, \quad \left|\frac{(\delta(X_c'))_{ij}}{D_{ii}'}\right| \leq h(p)\epsilon\mu_i(\tilde{X}_c') \max_j |(X_c)_{ij}|. \tag{20}$$

Since the computed matrix $\tilde{Q}_x$ is nearly orthonormal and since $||\tilde{Q}_x - Q_x'||_2$ is (up to a factor of the dimensions m, p) of order ϵ, the main issue in perturbation of $\mathcal{X} = \mathrm{span}(X)$ is how the matrix Q_x changes in the presence of the following perturbation:

$$X \equiv D'X_cD \longmapsto X + \delta X = D'(X_c + \delta X_c)D, \quad |(\delta X_c)_{ij}| \leq h(p)\epsilon\mu_i(\tilde{X}_c') \max_j |(X_c)_{ij}|. \tag{21}$$

The existing perturbation results for the QR factorization can be roughly divided into two groups. In the first group, we have error bounds in terms of $||\delta X||_F/||X||_2$ and typical estimate is of the form

$$||\delta Q_x||_F \leq \sqrt{2}\kappa_2(X)\frac{||\delta X||_F}{||X||_2}, \tag{22}$$

as in [40] (derived using fixed–point and operator theory; see also [33], [38]) or

$$||\delta Q_x||_F \leq \sqrt{2} \max_{0\leq t\leq 1} ||(X + t\cdot(\delta X))^{-1}||_2 ||\delta X||_F, \tag{23}$$

as in [5] (derived for $m \times m$ nonsingular matrices using calculus on the manifold $\mathbf{GL}(m)$). In the second group are the results of Sun [39] and Zha [46]. These results are best represented by the following theorem due to Zha: If $|\delta X| \leq \epsilon G_X|X|$, with $G_X \geq 0$, $||G_X||_\infty = 1$, and with sufficiently small ϵ, then

$$||\delta Q_x||_\infty \leq \mathbf{z}(m,p)\epsilon|| \, |R_x| \cdot |R_x^{-1}| \, ||_\infty, \tag{24}$$

where $\mathbf{z}(m,p)$ is modestly growing function. Zha has shown that the bound (24) is sharp. (Here the matrix absolute values and the inequalities involving matrices are understood element–wise; $||\cdot||_\infty$ is the matrix norm induced by the ℓ_∞ vector norm.) An important feature of the bound (24) is that it is invariant under replacing $X + \delta X$ with $(X + \delta X)D_x$, where D_x is arbitrary diagonal nonsingular matrix. Hence, the size of the error in the case of column–wise perturbations (such as in Proposition 2.1) is essentially determined by $\mathrm{cond}(X) = \min_{D_x=\mathrm{diag}} \kappa_2(XD_x)$. However, $\mathrm{cond}(X)$ may be large if X has heavily weighted rows, for example if X is composed as $X = D'X_cD$, where D and D' are ill–conditioned diagonal scalings, and X_c has moderate (say) $\kappa_2(X_c)$. Thus, the bound (24) is not sharp in the case of perturbation (21).

It does not seem simple to deal with row scaling in the perturbation analysis of the QR factorization. In the case of column scaling, we use the fact that both $X = Q_xR_x$ and $XD_x = Q_x(R_xD_x)$ are the essentially unique QR factorizations, and we can take advantage of the fact that $\kappa_2(XD_x)$

might be much smaller that $\kappa_2(X)$. In other words, if the ill–conditioning can be "filtered out" by column scaling, it is artificial and it does not affect the accuracy of the computation. On the other hand, the relation between the orthonormal QR factors of the matrices X, X_c, $X+\delta X$, $X_c+\delta X_c$ in relation (21) is not obvious. (For an asymptotic analysis of the QR factorization see [34].) We discuss the solution to this problem in the next section, where we describe a new algorithm that is based on another fundamental matrix factorization, namely the LU factorization.

3 The new algorithm

The main difference between our new algorithm and the algorithm of Björck and Golub is in the computation of the orthonormal bases of $\mathcal{X} = \mathrm{span}(X)$ and $\mathcal{Y} = \mathrm{span}(Y)$. Instead of the QR factorization applied directly to the matrices X and Y, we first compute the LU factorizations of X and Y using Gaussian elimination with complete (or partial) pivoting. Then we use the computed unit lower trapezoidal LU factors as new bases for $\mathcal{X}$, $\mathcal{Y}$. (Note that the numbers of parameters in the unit lower trapezoidal LU factors of X and Y are equal to the dimensions of the corresponding Stiefel manifolds of $m \times p$ and $m \times q$ orthonormal matrices.)

Algorithm 3.1 CC(X, Y)

Input $X \in \mathbf{R}^{m\times p}$, $Y \in \mathbf{R}^{m\times q}$ full column rank matrices with $p \geq q$.

Step 1 Compute the LU factorizations with pivoting, $P_1 X P_2 = L_x U_x$, $P_3 Y P_4 = L_y U_y$. (For partial pivoting, $P_2 = I_p$, $P_4 = I_q$.)

Step 2 Compute the QR factorizations $L_x = Q_x R_x$, $L_y = Q_y R_y$, using the modified Gram–Schmidt algorithm.

Step 3 Compute the matrix $S = Q_x^T((P_1 P_3^T) Q_y)$ and the SVD of S, $S = W \Sigma V^T$.

Output Return the matrices Σ, $P_1^T Q_x W$, $P_3^T Q_y V$.

Consider the computational complexity of Algorithm 3.1. We estimate the number of elementary floating point operations (*flops*), where we consider only the highest order terms in polynomial expressions for the actual *flop* count. The cost of the LU factorizations in Step 1 is $m(p^2+q^2) - (p^3+q^3)/3$ *flops* and, in the case of complete pivoting, $m(p^2+q^2)/2 - (p^3+q^3)/6$ floating–point absolute value comparisons. Since L_x and L_y are lower trapezoidal, the cost of the modified Gram–Schmidt orthogonalization can be reduced using the following algorithm:

Algorithm 3.2 MGS_LT(L)

```
for j = p, p-1, ..., 1
    L_x(j:m, j) := (1/||L_x(j:m, j)||_2) L_x(j:m, j)
    for i = j-1, j-2, ..., 1
        L_x(j:m, i) := L_x(j:m, i) - ((L_x(j:m, j))^T L_x(j:m, i)) L_x(j:m, j)
    end for
end for
```

Note that this algorithm overwrites L_x with a lower trapezoidal orthonormal basis of $\mathrm{span}(L_x)$. The computation completes in $2mp^2 - 4p^3/3$ *flops*. Our current implementation of Algorithm 3.2 uses BLAS 1 procedures `SAXPY()`, `SDOT()`, `SNRM2()` and `SSCAL()`. To compute the matrix S, we first use the LAPACK's procedure `SLASWP()` to apply the permutation $P_1 P_3^T$ to the computed matrix $\tilde{Q}_y$. Then we use the triangularity of the leading $p \times p$ submatrix of $\tilde{Q}_x$ to compute S by application of BLAS 3 procedures `STRMM()` and `SGEMM()`. The cost of this computation is $2mpq - p^2 q$ *flops*. Hence, starting with X and Y, we compute the matrix S in a total of $3m(p^2+q^2) + 2mpq - (5/3)(p^3+q^3) - p^2 q$ *flops* and $m(p^2+q^2)/2 - (p^3+q^3)/6$ floating–point absolute value comparisons. For comparison, the Björck–Golub algorithm computes S in a total of $2m(p^2+2q^2) + 4mpq - (2/3)(p^3+2q^3) - 2p^2 q$ *flops*.

3.1 Error analysis

First and the most important fact used in the analysis is that span(X) = span(L_x), span(Y) = span(L_y). The central idea of Algorithm 3.2 is then to concentrate all perturbations into perturbations of L_x and L_y. The motivation is well–conditioning of L_x and L_y – they are lower trapezoidal matrices with unit diagonal and with off–diagonal elements less than one in modulus. This fact also ensures that the QR factorizations of L_x and L_y can be accurately computed using the modified Gram–Schmidt algorithm. Furthermore, since L_x and L_y are well–conditioned bases for $\mathcal{X}$ and $\mathcal{Y}$ we expect the backward error to be small in the angle metric.

We begin detailed error analysis by pointing out an important difference between the LU and the QR factorization. Consider the floating–point LU factorization of X (cf. [24], [27]):

Proposition 3.1 *Let X be a real $m \times p$ matrix, and let its LU factorization be computed by the Gaussian elimination. If no zero pivots are encountered during the elimination process, the computed factors $\tilde{L}_x$ and $\tilde{U}_x$ satisfy*

$$\tilde{L}_x\tilde{U}_x = X + \delta X, \quad |\delta X| \leq \epsilon_{LU}(p)|\tilde{L}_x| \cdot |\tilde{U}_x|, \quad \epsilon_{LU}(p) \leq \frac{p\epsilon}{1-p\epsilon}. \tag{25}$$

Let us now assume that in Step 1 of Algorithm 3.2 the rows and the columns of X are so permuted that

$$X + \delta X = \tilde{L}_x\tilde{U}_x, \quad |\delta X| \leq \epsilon_{LU}(p)|\tilde{L}_x| \cdot |\tilde{U}_x|, \tag{26}$$

is the LU factorization with complete pivoting. (That is, we simplify the notation by identifying $X \equiv P_1XP_2 = L_xU_x$, where P_1 and P_2 are permutation matrices determined by the complete pivoting.) Let $X = D_1ZD_2$, where D_1 and D_2 are diagonal scalings, and let δZ be defined by the relation

$$X + \delta X = D_1(Z + \delta Z)D_2, \tag{27}$$

that is, $\delta Z = D_1^{-1}\delta X D_2^{-1}$. If $Z = L_zU_z$ and $Z + \delta Z = \tilde{L}_z\tilde{U}_z$ are the LU factorizations, then

$$Z = (D_1^{-1}L_xD_1)(D_1^{-1}U_xD_2^{-1}), \quad Z + \delta Z = (D_1^{-1}\tilde{L}_xD_1)(D_1^{-1}\tilde{U}_xD_2^{-1}),$$

and, by the uniqueness of the LU factorization,

$$L_z = D_1^{-1}L_xD_1, \quad U_z = D_1^{-1}U_xD_2^{-1}, \quad \tilde{L}_z = D_1^{-1}\tilde{L}_xD_1, \quad \tilde{U}_z = D_1^{-1}\tilde{U}_xD_2{}^{-1}.$$

Furthermore, from relations (26) and (27) it follows that

$$\tilde{L}_z\tilde{U}_z = Z + \delta Z, \quad |\delta Z| \leq \epsilon_{LU}(p)|\tilde{L}_z| \cdot |\tilde{U}_z|. \tag{28}$$

Note that $L_x - \tilde{L}_x = D_1(L_z - \tilde{L}_z)D_1^{-1}$ and that

$$\frac{\|(L_x - \tilde{L}_x)e_i\|_2}{\|L_xe_i\|_2} \leq \max_{j>i}\left|\frac{(D_1)_{jj}}{(D_1)_{ii}}\right| \frac{\|(L_z - \tilde{L}_z)e_i\|_2}{\|L_ze_i\|_2}\frac{\|L_ze_i\|_2}{\|L_xe_i\|_2} \tag{29}$$

$$\leq \max_{j>i}\left|\frac{(D_1)_{jj}}{(D_1)_{ii}}\right| \sqrt{m-i+1}\frac{\|(L_z - \tilde{L}_z)e_i\|_2}{\|L_ze_i\|_2}. \tag{30}$$

Similarly, writing $U_x - \tilde{U}_x = (D_1D_2)D_2^{-1}(U_z - \tilde{U}_z)D_2$, we obtain

$$\frac{\|(U_x - \tilde{U}_x)^Te_i\|_2}{\|U_x^Te_i\|_2} \leq \max_{j\geq i}\left|\frac{(D_2)_{jj}}{(D_2)_{ii}}\right| \sqrt{p-i+1}\frac{\|(U_z - \tilde{U}_z)^Te_i\|_2}{\|U_z^Te_i\|_2}. \tag{31}$$

The important relations (30) and (31) are interpreted as follows: If $X \equiv P_1XP_2$ can be written as $X = D_1ZD_2$, where the diagonal entries of the diagonal matrices $|D_1|$ and $|D_2|$ are graded from large to small and Z in (28) admits an accurate LU factorization, then the floating–point

LU factorization of X is accurate as well. Furthermore, if for some diagonal matrices D_L, D_U the scaled matrices $L_z D_L$ and $D_U U_z$ are well conditioned, then the matrices $L_x D_L$ and $D_U U_x$ are well conditioned as well. (Cf. [13], [17], [20], [35], [12].)

On the other hand, relations (4) and (5) from Proposition 2.1 show that the error estimates of the floating–point QR factorization are only invariant under column scalings. This limits the accuracy of the QR factorization without complete (row and column) pivoting in the least squares computation with heavy row weighting (cf. [32], [43], [3], [4], [8, § 4.4.2]), in the SVD computation (cf. [13]) and in the principal angle computation (cf. Examples 2.2, 2.3, 2.4). In the case of the QR factorization with complete pivoting, we can use the results for the LU perturbations to understand the higher accuracy. Recall relation (21),

$$X \equiv D' X_c D \longmapsto X + \delta X = D'(X_c + \delta X_c)D, \quad |(\delta X_c)_{ij}| \leq h(p)\epsilon \mu_i(\tilde{X}'_c) \max_j |(X_c)_{ij}|,$$

and assume that the diagonals of D, D' are are graded ($|D_{ii}| \geq |D_{i+1,i+1}|$, $|D'_{ii}| \geq |D'_{i+1,i+1}|$) and that X_c admits accurate LU factorization in the presence of the perturbation δX_c. (Note that pivoting ensures that D, D' nearly meet the ordering assumption.) In that case, the LU factorization of $X = L_x U_x$ is accurate as well and $\max_i \|\delta L_x e_i\|_2 / \|L_x e_i\|_2 \ll 1$. Now note that from $X = L_x U_x = Q_x R_x$ it follows that $L_x = Q_x (R_x U_x^{-1})$ is the QR factorization of L_x. In other words, the orthonormal QR factors of X and L_x are essentially the same (up to orientation of the columns of Q_x, depending on the signs of the pivots). Similarly, if $X + \delta X = (L_x + \delta L_x)(U_x + \delta U_x) = (Q_x + \delta Q_x)(R_x + \delta R_x)$, then $Q_x + \delta Q_x$ is orthonormal QR factor of $L_x + \delta L_x$. This means that we can develop perturbation theory for δQ_x as function of L_x and δL_x. The good news is that δL_x is from the column–wise class of perturbations and the relevant condition number is $\min_{D_L = \mathrm{diag}} \kappa_2(L_x D_L)$. This condition number is moderate if the unit lower trapezoidal LU factor of X_c is well–conditioned. In that case we can derive sharp perturbation estimates for the QR factorization of the perturbed matrix X from relation (21). For example, we can prove the following proposition.

Proposition 3.2 *Let $X = Q_x R_x$, $X + \delta X = (Q_x + \delta Q_x)(R_x + \delta R_x)$ be the QR factorizations of X and $X + \delta X$, respectively. Let L_x and $L_x + \delta L_x$ be the unit lower triangular factors of X and $X + \delta X$, and let $(L_x)_c = L_x \mathrm{diag}(\|L_x e_i\|_2)^{-1}$, $(\delta L_x)_c = \delta L_x \mathrm{diag}(\|L_x e_i\|_2)^{-1}$, $\|(\delta L_x) L_x^\dagger\|_2 < 1/2$. There exist an upper triangular matrix E such that*

$$\left(I + (\delta L_x) L_x^\dagger\right) Q_x = (Q_x + \delta Q_x)(I + E)$$

and

$$\begin{aligned} \|E\|_F &\leq \sqrt{2} \| \left((\delta L_x) L_x^\dagger\right)^\tau + (\delta L_x) L_x^\dagger + \left((\delta L_x) L_x^\dagger\right)^\tau \left((\delta L_x) L_x^\dagger\right) \|_F & (32) \\ &\leq \sqrt{2} \|(L_x)_c^\dagger\|_2 \left(2\|(\delta L_x)_c\|_F + \|(L_x)_c^\dagger\|_2 \|(\delta L_x)_c\|_F^2\right), & (33) \\ \|\delta Q_x\|_F &\leq \frac{\|(\delta L_x) L_x^\dagger\|_F + \|E\|_F}{1 - \|E\|_2}. & (34) \end{aligned}$$

Furthermore,

$$\sin \angle(\mathrm{span}(Q_x + \delta Q_x), \mathrm{span}(X)) \leq \|(L_x)_c^\dagger\|_2 \|(\delta L_x)_c\|_2. \quad (35)$$

We now return to the analysis of Algorithm 3.1. Since L_x and L_y are unit lower trapezoidal matrices computed by Gaussian elimination with pivoting, the spectral condition numbers of L_x and L_y are bounded by a function of the dimensions. Although the theoretical bound of the condition numbers is exponential function of the dimension, the values of $\kappa_2(L_x)$ and $\kappa_2(L_y)$ are almost always moderate (cf. e.g. [41], [36], [44]). Hence, we can safely use the modified Gram–Schmidt algorithm to compute nearly orthogonal bases for $\mathrm{span}(L_x)$ and $\mathrm{span}(L_y)$. The numerical

properties of the modified Gram–Schmidt algorithm are well understood; see [6], [11], [7]. The two most important facts are summarized in the following theorem due to Higham [27, § 18.7, Theorem 8.12].

Theorem 3.1 *Suppose the modified Gram–Schmidt algorithm is applied to $A \in \mathbf{R}^{m\times p}$ of rank p. If $\tilde{Q}$ and $\tilde{R}$ are the computed matrices, then there exist backward perturbation δA and moderate polynomials $\wp_{MGS}(m,p)$, $\wp'_{MGS}(m,p)$ such that*

$$A+\delta A = \tilde{Q}\tilde{R}, \quad \|\delta A e_i\|_2 \leq \epsilon\wp_{MGS}(m,p)\|Ae_i\|_2, \tag{36}$$

$$\|\tilde{Q}^\tau\tilde{Q}-I\|_2 \leq \epsilon\wp'_{MGS}(m,p)\kappa_2(A_c)+O((\epsilon\wp'_{MGS}(m,p)\kappa_2(A_c))^2), \tag{37}$$

where $A_c = A\mathrm{diag}(\|Ae_i\|_2)^{-1}$.

Theorem 3.2 *Let $\tilde{L}_x = L_x+\delta L_x$, $\tilde{L}_y = L_y+\delta L_y$ be the computed lower triangular factors in Step 1 of Algorithm 3.1, let $\mathrm{rank}(L_x+\delta L_x) = \mathrm{rank}(L_x)$, $\mathrm{rank}(L_y+\delta L_y) = \mathrm{rank}(L_y)$, and let*

$$\eta_x \equiv \max_{1\leq i\leq p}\frac{\|\delta L_x e_i\|_2}{\|L_x e_i\|_2} < 1, \quad \eta_y \equiv \max_{1\leq i\leq q}\frac{\|\delta L_y e_i\|_2}{\|L_y e_i\|_2} < 1.$$

Further, let in Step 2 the computed approximations $\tilde{Q}_x$, $\tilde{Q}_y$ of Q_x and Q_y, respectively, satisfy

$$\omega \equiv \max\{\|\tilde{Q}_x^\tau\tilde{Q}_x - I_p\|_F, \|\tilde{Q}_y^\tau\tilde{Q}_y - I_q\|_F\} < 1, \tag{38}$$

where ω is derived from Theorem 3.1. Then there exist subspaces $\hat{\mathcal{X}}$, $\hat{\mathcal{Y}}$ and moderate function $f(m,p,q)$ such that

(i) The subspaces $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$ are close approximations of $\mathcal{X}$ and $\mathcal{Y}$, respectively. More precisely, it holds that

$$\sin\angle(\mathcal{X},\hat{\mathcal{X}}) \leq \sqrt{p}(\eta_x+\epsilon\wp_{MGS}(m,p)(1+\eta_x))\|(L_x)_c^\dagger\|_2, \tag{39}$$

$$\sin\angle(\mathcal{Y},\hat{\mathcal{Y}}) \leq \sqrt{q}\left(\eta'_y+\frac{\epsilon f(m,p,q)(1+\eta'_y)}{(1-\omega)^2}\right)\|(L_y)_c^\dagger\|_2, \tag{40}$$

where $\eta'_y = \eta_y+\epsilon\wp_{MGS}(m,q)(1+\eta_y)$.

(ii) If $\sigma'_1\geq\cdots\geq\sigma'_q$ are the exact cosines of the principal angles between $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$, then, for all i, either $\tilde{\sigma}_i = \sigma'_i = 0$ or $|\tilde{\sigma}_i-\sigma'_i|/\sigma'_i$ is less than ω plus ϵ times a moderate polynomial of the space dimension.

Proof: The floating–point QR factorization of $\tilde{L}_x$ can be represented as $\tilde{Q}_x\tilde{R}_x = \tilde{L}_x+\delta\tilde{L}_x$, where (cf. Theorem 3.1) $\|\delta\tilde{L}_x e_i\|_2 \leq \epsilon\wp_{MGS}(m,p)\|\tilde{L}_x e_i\|_2$, $1\leq i\leq p$. Let $\Delta L_x = \delta L_x+\delta\tilde{L}_x$, and let, as in Theorem 2.2, $\tilde{Q}_x = Q'_x(I+T'_x)$ be the QR factorization of $\tilde{Q}_x$. (Q'_x is exactly orthonormal and T'_x is upper triangular with $\|T'_x\|_2\leq\omega$.) Then $L_x+\Delta L_x = Q'_x(I+T'_x)\tilde{R}_x$. Note that $\mathrm{rank}(L_x+\Delta L_x) = \mathrm{rank}(L_x)$. Define $\hat{\mathcal{X}} = \mathrm{span}(L_x+\Delta L_x)$ and note that the sine of the angle between $\mathcal{X}$ and $\hat{\mathcal{X}}$ equals $\sin\angle(\mathcal{X},\hat{\mathcal{X}}) = \|((Q'_x)^\perp)^\tau Q_x\|_2$, where $(Q'_x)^\perp$ is orthonormal basis of the orthogonal complement of $\hat{\mathcal{X}}$. An easy calculation shows that

$$((Q'_x)^\perp)^\tau Q_x = -((Q'_x)^\perp)^\tau(\Delta L_x)R_x^{-1}, \quad \sin\angle(\mathcal{X},\hat{\mathcal{X}}) \leq \|\Delta L_x R_x^{-1}\|_2.$$

Similarly, we can write $L_y+\Delta' L_y = Q'_y(I+T'_y)\tilde{R}_y$, $\Delta' L_y = \delta L_y+\delta\tilde{L}_y$. As in the proof of Theorem 2.4, we write $\tilde{S}\equiv \boldsymbol{fl}(\tilde{Q}_x^\tau\tilde{Q}_y) = \tilde{Q}_x^\tau\tilde{Q}_y+E_S$ and we represent the computed singular values $\tilde{\sigma}_1\geq\cdots\geq\tilde{\sigma}_q$ as exact singular values of $\tilde{S}+\delta\tilde{S}$, where $\delta\tilde{S}$ as well as E_S is small in the spectral norm ($\|E_S\|_2\ll 1$, $\|\delta\tilde{S}\|_2\ll 1$). Then we write

$$\begin{aligned}\tilde{S}+\delta\tilde{S} &= \tilde{Q}_x^\tau\tilde{Q}_y+E_S+\delta\tilde{S} = \tilde{Q}_x^\tau(\tilde{Q}_y+(\tilde{Q}_x^\tau)^\dagger(E_S+\delta\tilde{S}))\\ &= (I+T'_x)^\tau\left((Q'_x)^\tau Q''_y\right)(I+T''_y),\end{aligned}$$

where $\tilde{Q}_y + (\tilde{Q}_x^T)^\dagger(E_S + \delta\tilde{S}) = Q_y''(I + T_y'')$ is the QR factorization of an almost orthonormal matrix with $\|T_y''\|_2 \ll 1$. Define

$$L_y + \Delta L_y \equiv L_y + \Delta' L_y + (\tilde{Q}_x^T)^\dagger(E_S + \delta\tilde{S})\tilde{R}_y = Q_y''(I + T_y'')\tilde{R}_y,$$

and $\hat{\mathcal{Y}} = \mathrm{span}(L_y + \Delta L_y)$. The proof completes by elementary calculations of the upper bounds for $\|\Delta L_x e_i\|_2/\|L_x e_i\|_2$, $\|\Delta L_y e_j\|_2/\|L_y e_j\|_2$, $1 \le i \le p$, $1 \le j \le q$, and by invoking Theorem 2.1. Q.E.D.

Remark 3.1 The backward error bounds (39) and (40) can be improved as follows. Note that it also holds that $\sin\angle(\mathcal{X}, \hat{\mathcal{X}}) = \|(Q_x^\perp)^T Q_x'\|_2$, where $Q_x^\perp$ is orthonormal basis of the orthogonal complement of $\mathcal{X}$. An easy calculation shows that

$$(Q_x^\perp)^T Q_x' = (Q_x^\perp)^T(\Delta L_x)\tilde{R}_x^{-1}(I + T_x')^{-1}.$$

Let now $\Delta L_x = Q_{\Delta_x} R_{\Delta_x}$ be the QR factorization of ΔL_x and let $\mathcal{L}_{\Delta_x} = \mathrm{span}(\Delta L_x)$. Then

$$(Q_x^\perp)^T Q_x' = \left((Q_x^\perp)^T Q_{\Delta_x}\right) R_{\Delta_x}\tilde{R}_x^{-1}(I + T_x')^{-1}, \quad \sin\angle(\mathcal{X}, \hat{\mathcal{X}}) \le \sin\angle(\mathcal{X}, \mathcal{L}_{\Delta_x})\frac{\|\Delta L_x \tilde{R}_x^{-1}\|_2}{1 - \|T_x'\|_2}.$$

Hence, if $\mathrm{span}(\Delta L_x) \subset \mathcal{X}$, then $\sin\angle(\mathcal{X}, \hat{\mathcal{X}}) = 0$.

Remark 3.2 In this paper, we consider only the classical partial and complete pivoting in the Gaussian elimination. Other choices include, for example, the pivoting for stability and sparsity due to Björck and Duff [9], the maximal transversal pivoting due to Olschowka and Neumaier [30], and pivoting for forward stable Gaussian elimination due to Demmel et al [13].

In Figure 4, we summarize the difference between the Björck–Golub algorithm and Algorithm 3.1.

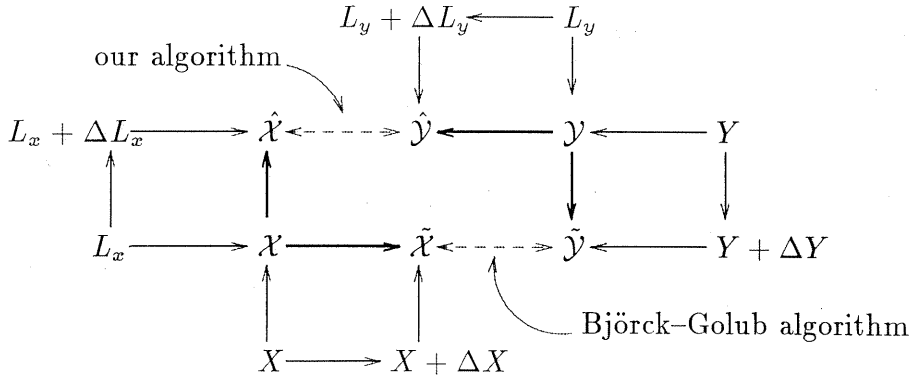


Figure 4: Commutative diagram for principal angle computation between $\mathcal{X} = \mathrm{span}(X)$ and $\mathcal{Y} = \mathrm{span}(Y)$. The Björck–Golub algorithm computes the principal angles between $\tilde{\mathcal{X}} = \mathrm{span}(X + \Delta X)$ and $\tilde{\mathcal{Y}} = \mathrm{span}(Y + \Delta Y)$. On the other hand, our algorithm computes the principal angles between $\hat{\mathcal{X}} = \mathrm{span}(L_x + \Delta L_x)$ and $\hat{\mathcal{Y}} = \mathrm{span}(L_y + \Delta L_y)$.

3.2 Comments on the SVD computation in principal angle algorithms

Next, we analyze more closely the SVD computation of the matrix $\tilde{S} = \boldsymbol{fl}(\tilde{Q}_x^T\tilde{Q}_y) = \tilde{Q}_x^T\tilde{Q}_y + E_S$. Recall that the computed singular values are denoted by $\tilde{\sigma}_1 \ge \cdots \ge \tilde{\sigma}_q$, and that in the proofs of Theorem 2.2 and Theorem 3.2 we do not explicitly mention which SVD algorithm is used.

Implicitly, however, we use the fact that the algorithm is backward stable where the backward error is small in the matrix norm. Thereby, we indicate that the QR or the divide and conquer algorithm are good choices, and, at the same time, we raise the following question: *If the goal is maximal accuracy, why don't we use the more accurate Jacobi SVD algorithm?*. The question is reasonable, for the Jacobi SVD algorithm is more accurate than any other algorithm that first bidiagonalizes the matrix (cf. [15]). We argue that SVD computation with higher accuracy (using the Jacobi SVD algorithm) generally does not improve the accuracy of the overall computation. First indication for that is fairly sharp estimate of Theorem 3.2, which is derived using norm–wise backward stability (hence, only high absolute accuracy) of the SVD computation. Secondly, we can show that in the cases when the Jacobi SVD algorithm can compute much more accurate singular values of $\tilde{S}$, the error in $\tilde{S}$ is such that the initial uncertainty of the singular values cannot be corrected. We can show this using backward error analysis. If $S' = \tilde{Q}_x^T \tilde{Q}_y$, the issue is how well the singular values of S' are determined by $\tilde{S} = S' + E_S$.

Backward analysis: If we use the Jacobi SVD algorithm, then we can estimate $\delta\tilde{S}$ by (cf. [18]) $\|\delta\tilde{S}e_i\|_2 \leq g(p,q)\epsilon\|\tilde{S}e_i\|_2$, $1 \leq i \leq q$, where $g(p,q)$ is modest polynomial. This means that the backward error in small columns of $\tilde{S}$ is correspondingly small. However, if for some j the column $\tilde{S}e_j$ is small (of order ϵ, say), it means that $\tilde{S}e_j$ is generally accurate only to an absolute uncertainty of order $m\epsilon$ ($\|E_S e_j\|_2 \leq O(m\epsilon)$) and that the relative error might be large ($\|E_S e_j\|_2/\|S'e_j\|_2 \gg \epsilon$). Hence, $\|(E_S + \delta\tilde{S})e_j\|_2$ might be large and we cannot expect high relative accuracy of the corresponding small singular value. (This holds even if the backward error $\delta\tilde{S}$ is zero.)

It is instructive to show the same fact in the forward mode of the analysis. The QR or the divide and conquer algorithms are less accurate than the Jacobi algorithm if in the factorization $S' = S'_c D_s$, $D_s = \mathrm{diag}(\|S'e_i\|_2)$, the condition number $\kappa_2(S')$ is much larger than $\kappa_2(S'_c)$, $\kappa_2(S') \gg \kappa_2(S'_c)$. In that case, $\kappa_2(D_s) \gg 1$ and $\min_{1\leq j\leq q}\|S'e_j\|_2 \ll 1$. To estimate the relative difference between the singular values of S' and $\tilde{S} + \delta\tilde{S}$, we assume full column rank S' and we use the factorization

$$\tilde{S} + \delta\tilde{S} = \left(I + (E_S + \delta\tilde{S})(S')^\dagger\right) S',$$

and Theorem 2.1 to obtain relative error bound $\eta = \|(E_S + \delta\tilde{S})D_s^{-1}\|_2 \cdot \|(S'_c)^\dagger\|_2$. If j_0 is such that the column norm $(D_s)_{j_0 j_0}$ is minimal, and if we cannot guarantee that $\|(E_S + \delta\tilde{S})e_{j_0}\|_2 \ll (D_s)_{j_0 j_0}$, then the bound

$$\|(S'_c)^\dagger\|_2 \frac{\|(E_S + \delta\tilde{S})e_{j_0}\|_2}{(D_s)_{j_0 j_0}} \leq \eta \leq \frac{\|E_S + \delta\tilde{S}\|_2}{\min_{1\leq j\leq q}\|S'e_j\|_2}\|(S')^\dagger\|_2.$$

for the parameter η shows that the relative accuracy might depend on $\kappa_2(S')$. Moreover, if one of the subspace is close to the orthogonal complement of the other one, the matrix $\tilde{S}$ might be merely the roundoff noise and computation with high relative accuracy is not feasible.

Example 3.1 We illustrate the above discussion in the case $p = q = 1$. Let $X = [x]$, $Y = [y]$, where $x, y \in \mathbf{R}^m$ are unit vectors, and let δx, δy be small perturbations ($\|\delta x\|_2 \ll 1$, $\|\delta y\|_2 \ll 1$). Then $\angle(\mathcal{X}, \mathcal{Y}) = \arccos(x^T y)$ and

$$|(x + \delta x)^T(y + \delta y) - x^T y| \leq \|\delta x\|_2 + \|\delta y\|_2 + \|\delta x\|_2\|\delta y\|_2.$$

Hence, the relative accuracy of the singular value $\sigma_1 = x^T y$ is in the presence of errors δx and δy determined by $(\|\delta x\|_2 + \|\delta y\|_2)/(x^T y)$. If $\|\delta x\|_2$ and $\|\delta y\|_2$ are of the order of the machine precision ϵ, then we see that floating–point computation of σ_1 is feasible to only (roughly) $-\lfloor \log_{10}(\epsilon/(x^T y))\rfloor$ decimal places. In other words, small singular values are poorly determined if the corresponding subspaces are nearly orthogonal (cf. [25]).

Remark 3.3 If X and Y are normally scaled ($p = q$, $X^T Y = I_p$) then the canonical correlations of X and Y are the singular values of XY^T. In that case the canonical correlations are determined

to high relative accuracy if the condition numbers $\kappa_2(X_c)$, $\kappa_2(Y_c)$ of column scaled matrices X, Y are moderate (cf. [25]). For accurate SVD computation of XY^T see [17], [19].

3.3 Cross–product implementation

In this subsection, we briefly discuss an implementation of Algorithm 3.1 that may be an efficient alternative in the case $m \gg \max\{p, q\}$.

Algorithm 3.3 X–CC(X, Y)

Input $X \in \mathbf{R}^{m\times p}$, $Y \in \mathbf{R}^{m\times q}$ full column rank matrices with $p \geq q$.

Step 1 Compute the LU factorizations with pivoting, $P_1 X P_2 = L_x U_x$, $P_3 Y P_4 = L_y U_y$. (For partial pivoting, $P_2 = I_p$, $P_4 = I_q$.)

Step 2 Compute the matrices $H_{xx} = L_x^T L_x$, $H_{yy} = L_y^T L_y$, $H_{xy} = L_x^T((P_1 P_3^T) L_y)$, and the Cholesky factorizations $H_{xx} = R_x^T R_x$, $H_{yy} = R_y^T R_y$. Exploit symmetry as much as possible.

Step 3 Compute the matrix $S = R_x^{-T} H_{xy} R_y^{-1}$ and the SVD of S, $S = W\Sigma V^T$.

Output Return the matrix Σ.

The use of the Cholesky factors of the cross–product matrices is similar to the Peters–Wilkinson [31] algorithm for least squares solution using normal equation. (Recall that the principal angle problem in the case $q = 1$ is closely related to the classical least squares problem, cf. [10]. Also note that in the case of sparse X and Y we may use complete pivoting of Björck and Duff [9] which is designed to preserve as much of the original sparsity as possible. For related results see also Barlow [3] and Barlow and Handy [4].) Perturbation analysis of Algorithm 3.3 can be done as in [19]. We omit the details for the sake of brevity.

4 Numerical examples

We conclude this work with several numerical examples.

Example 4.1 In this example, we generate test pairs (X, Y) as in Example 2.1, and we measure the errors in the computed orthonormal bases $\tilde{Q}_x$, $\tilde{Q}_y$ of $\mathcal{X} = \mathrm{span}(X)$, $\mathcal{Y} = \mathrm{span}(Y)$, respectively. We record for each generated matrix X the following values:

$$\begin{aligned}
&\mathbf{e}_x = \sin\angle(\mathrm{span}(\tilde{Q}_x), \mathcal{X}),\\
&\mathbf{g}_x = \|\tilde{Q}_x^T \tilde{Q}_x - I_p\|_2,\\
&\kappa_2(X_c), \quad X_c = X\mathrm{diag}(\|Xe_i\|_2)^{-1},\\
&\kappa_2((\tilde{L}_x)_c), \quad (\tilde{L}_x)_c = \tilde{L}_x\mathrm{diag}(\|\tilde{L}_x e_i\|_2)^{-1}.
\end{aligned}$$

(Similarly for Y.) The results for all 900 values of X are given in Figure 5. Recall that the test matrices $\{X \equiv D'X_s D\}$ are divided into five classes (180 examples each) with $\kappa_2(X_s) = 10^2, 10^3, 10^4, 10^5, 10^6$. These classes are clearly recognizable in Figure 5 if one follows the growth of $\mathbf{e}_x$. Also note that the deviation from orthonormality of $\tilde{Q}_x$ is of the order of $m\epsilon$. We also observe similar accuracy in a variant of Algorithm 3.1 with LU factorizations with partial pivoting, see Figure 6.

Example 4.2 In this example, we generate a set of rather ill–conditioned bases. We first generate an X as in Example 2.1, and then we partition X as $X = [X_1, X_2]$ and we introduce heavy weighting into the rows of X_2. Both Algorithm 3.1 and the Björck–Golub algorithm are sensitive

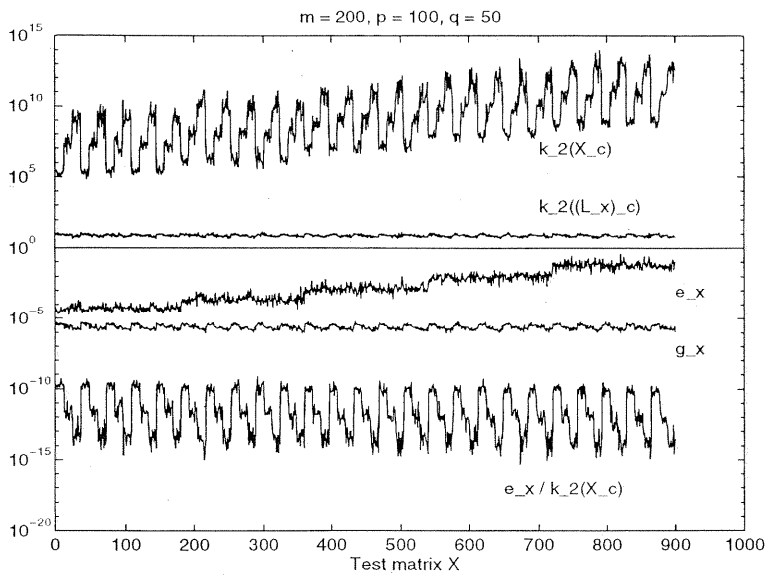


Figure 5: The values of $\mathbf{e}_x$, $\mathbf{g}_x$, $\kappa_2(X_c)$, $\kappa_2((\tilde{L}_x)_c)$ for 900 test in Example 4.1. The LU factorizations are computed with complete pivoting.

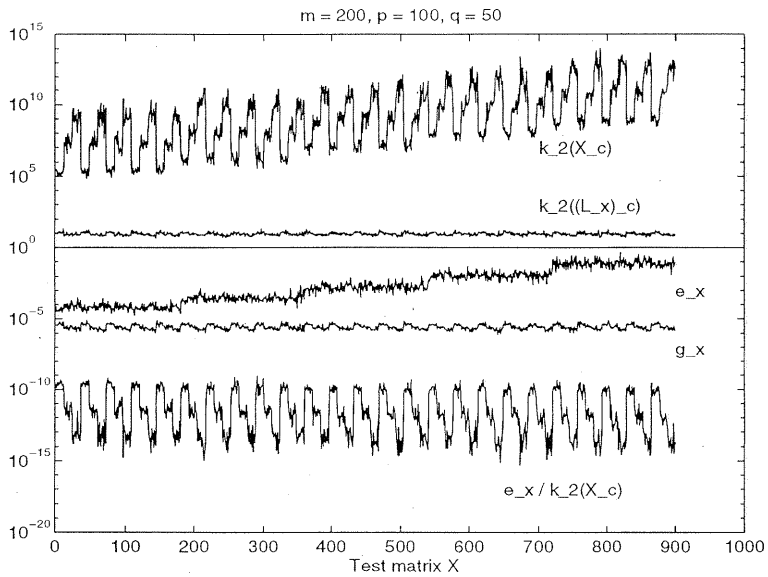


Figure 6: The values of $\mathbf{e}_x$, $\mathbf{g}_x$, $\kappa_2(X_c)$, $\kappa_2((\tilde{L}_x)_c)$ for 900 tests in Example 4.1. The LU factorizations are computed with partial (row) pivoting.

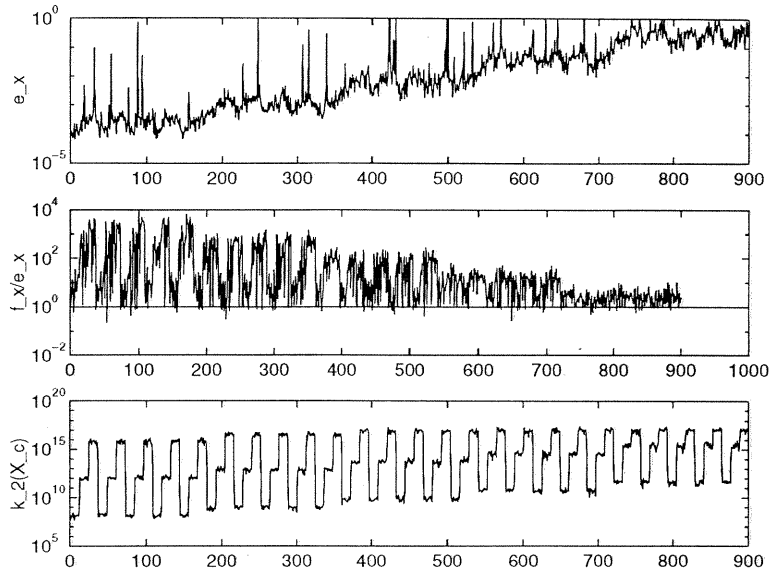


Figure 7: The values of $\mathbf{e}_x$, $\mathbf{f}_x/\mathbf{e}_x$, $\kappa_2(X_c)$ for all 900 matrices $\{X\}$ in Example 4.2. The LU and the QR factorizations are computed with complete pivoting.

to ill–conditioning introduced in this way. However, Algorithm 3.1 retains its accuracy properties in most of the cases, while the QR based approach computes with much larger errors. In Figure 7, $\mathbf{e}_x$ is defined as in Example 4.1 ($\tilde{Q}_x$ computed by Algorithm 3.1) and $\mathbf{f}_x = \sin\angle(\mathrm{span}(\tilde{\tilde{Q}}_x), \mathcal{X})$, where $\tilde{\tilde{Q}}_x$ is computed by the QR factorization with complete pivoting. The variant of Algorithm 3.1 with partial pivoting is also less accurate; see Figure 8.

Example 4.3 Examples where Algorithm 3.1 is guaranteed to achieve high accuracy include structured matrices where various combinatorial and algebraic conditions (sparsity, sign pattern) ensure forward stable Gaussian elimination with pivoting, cf. [13]. (For further references on highly accurate Gaussian elimination see [27].) In such cases, Algorithm 3.1 has an advantage over the modified Björck–Golub algorithm.

Example 4.4 In this example, we measure the forward error in the computed canonical correlations. As reference values we use the approximate canonical correlations $\sigma_1^{(D)} \geq \cdots \geq \sigma_q^{(D)}$ computed by the double precision Algorithm 3.1. The test problems are generated as in Example 4.1. We test the accuracy of the Björck–Golub algorithm with complete pivoting, Algorithm 3.1 with complete and partial pivoting, and Algorithm 3.3 with complete pivoting. For single precision approximations $\sigma_1^{(S)} \geq \cdots \geq \sigma_q^{(S)}$ computed by each of the four algorithms, we compute

$$\epsilon_{CC} = \frac{\max_{1\leq i\leq q} |\sigma_i^{(D)} - \sigma_i^{(S)}|}{\max\{\kappa_2(X_s), \kappa_2(Y_s)\}}.$$

The expected values of ϵ_{CC} are of order of the machine precision ϵ. The computed values are shown in Figure 9.

Remark 4.1 Our software, written in FORTRAN 77, is based on the BLAS and the LAPACK libraries. The procedure for the QR factorization with complete pivoting is a simple modification of the LAPACK procedure `SGEQPF()`, the LU factorization is computed using the LAPACK procedure `SGETRF()`, and the LU factorization with complete pivoting is computed by a modification of the

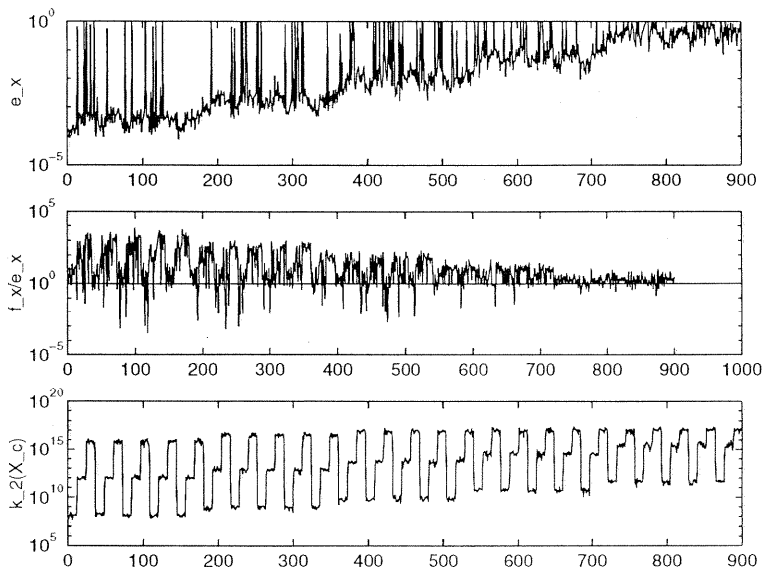


Figure 8: The values of $\mathbf{e}_x$, $\mathbf{f}_x/\mathbf{e}_x$, $\kappa_2(X_c)$ for all matrices $\{X\}$ in Example 4.2. The LU factorizations are computed with partial pivoting and the QR factorizations are computed with complete pivoting.

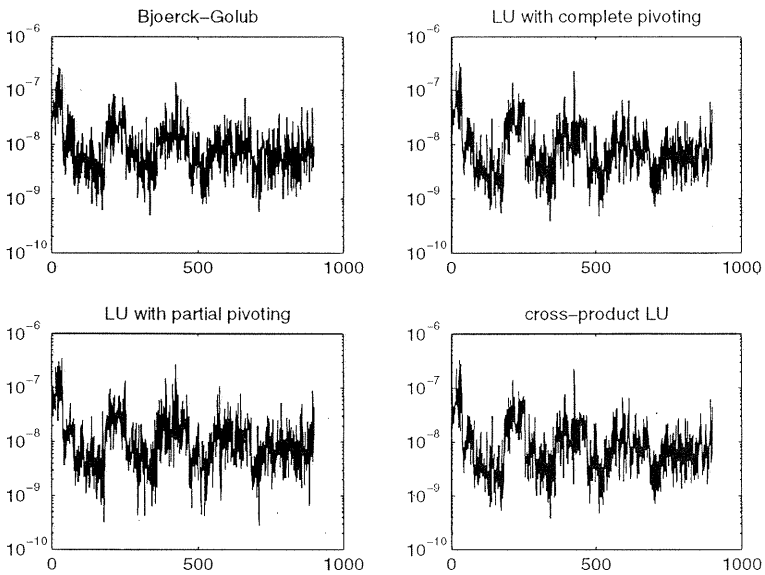


Figure 9: The values of ϵ_{CC} for all 900 pairs $\{(X, Y)\}$ in Example 4.4. Note that all four algorithm have nearly the same accuracy.

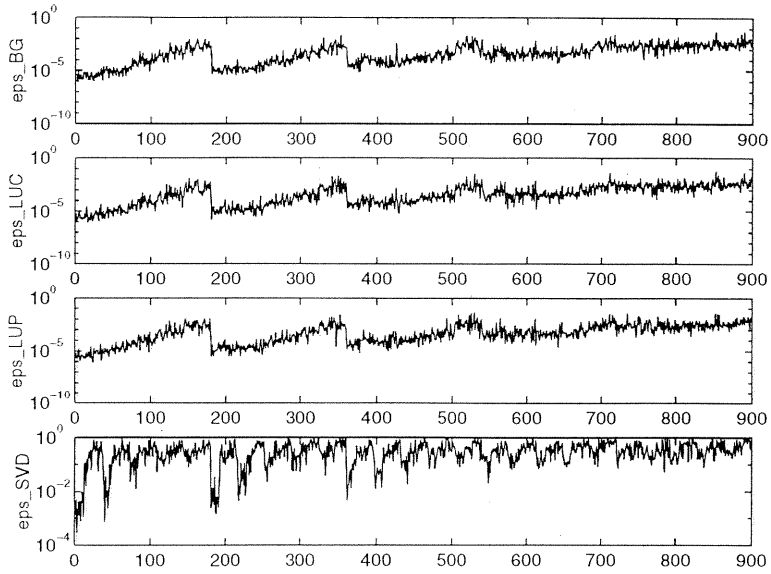


Figure 10: The computed forward errors in Example 4.5.

LAPACK procedure `SGETF2()`. The modified Gram–Schmidt Algorithm 3.2 is implemented on top of BLAS 1. Matrix multiplications are performed using the BLAS 3 procedures `SGEMM()` and `STRMM()`. All experiments were done on a DEC alpha workstation. We have observed that Algorithm 3.1 with partial pivoting is fastest in reducing the pair (X, Y) to the single matrix S. For instance, if $m = 400$, $p = 100$, $q = 50$, it requires 0.67 of the time needed in the Björck–Golub algorithm with complete pivoting, or 0.63 of the time needed in Algorithm 3.1 with complete pivoting. Algorithm 3.1 with complete pivoting requires in this case 1.06 of the time of the Björck–Golub algorithm with complete pivoting. For a thorough analysis of the efficiency of these algorithms, more software engineering has to be done.

Example 4.5 In our last example, we compare the algorithms based on the QR and the LU factorizations with pivoting with the algorithm based on the use of the SVD in the computation of the orthonormal bases for span(X), span(Y). (The use of the SVD in the principal angle computation is discussed in [10] in connection with ill–conditioned and rank deficient cases.) In this example, we compute the SVD using the LAPACK procedure `SGESVD()`. The test is performed as in Example 4.4 and with the dimensions $m = 100$, $p = q = 50$. For each of 900 examples, we compute the maximal forward errors ϵ_{BG} (for the Björck–Golub algorithm with complete pivoting), ϵ_{LUC} (for Algorithm 3.1 with complete pivoting), ϵ_{LUP} (for Algorithm 3.1 with partial pivoting) and ϵ_{SVD} for the computation based on the SVD. The results shown in Figure 10 show that the SVD approach is less accurate than the QR and LU based algorithms with complete pivoting. (We conjecture that similar situation occurs in the weighted least squares computation if we compare the Peters–Wilkinson algorithm, the QR approach with complete pivoting and the algorithm based on the SVD.)

References

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenny, S. Ostrouchov, and D. Sorensen. *LAPACK users' guide, second edition.* SIAM, Philadelphia, PA, 1992.

[2] E. H. Bareiss. Numerical solution of the weighted linear least squares problem by G–transformations. Technical Report 82–03–NAM–03, Department of Electrical Engineering and Computer Science, Northwestern University, Evanstons, Illinois, April 1982.

[3] J. Barlow. Stability analysis of the G–algorithm and a note on its application to sparse least squares problems. *BIT*, 25:507–520, 1985.

[4] J. Barlow and S. Handy. The direct solution of weighted and equality constrained least–squares problems. *SIAM J. Sci. Stat. Comp.*, 9(4):704–716, 1988.

[5] R. Bhatia and K. Mukherjea. Variation of the unitary part of a matrix. *SIAM J. Matrix Anal. Appl.*, 15:1007–1014, 1994.

[6] Å. Björck. Solving linear least squares problems by Gram–Schmidt orthogonalization. *BIT*, 7:1–21, 1967.

[7] Å. Björck. Numerics of Gram–Schmidt orthogonalization. *Linear Algebra Appl.*, 197/198:297–316, 1994.

[8] Å. Björck. *Numerical Methods for Least Squares Problems.* SIAM, 1996.

[9] Å. Björck and I. S. Duff. A direct method for the solution of sparse linear least squares. *Linear Algebra Appl.*, 34:43–67, 1980.

[10] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Math. Comp.*, 27:579–594, 1973.

[11] Å. Björck and C. C. Paige. Loss and recapture of orthogonalty in the modified Gram–Schmidt algorithm. *SIAM J. Matrix Anal. Appl.*, 13:176–190, 1992.

[12] X.-W. Chang, C. C. Paige, and G. W. Stewart. New perturbation analyses for the Cholesky factorization. *IMA Journal of Numerical Analysis*, 16:457–484, 1996.

[13] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač. Computing the singular value decomposition with high relative accuracy. Submitted to Lin. Alg. Appl., 1997.

[14] J. Demmel and A. McKenney. A test matrix generation suite. LAPACK Working Note 9, Courant Institute, New York, March 1989.

[15] J. Demmel and K. Veselić. Jacobi's method is more accurate than QR. *SIAM J. Matrix Anal. Appl.*, 13(4):1204–1245, 1992.

[16] Z. Drmač. *Computing the Singular and the Generalized Singular Values.* PhD thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, 1994.

[17] Z. Drmač. Accurate computation of the product induced singular value decomposition with applications. Department of Computer Science, University of Colorado at Boulder, Technical report CU-CS-816-96, submitted to SIAM J. Numer. Anal., July 1995.

[18] Z. Drmač. A tangent algorithm for computing the generalized singular value decomposition. Department of Computer Science, University of Colorado at Boulder, Technical report CU-CS-815-96, submitted to SIAM J. Numer. Anal., July 1995.

[19] Z. Drmač. Fast and accurate algorithms for canonical correlations, weighted least squares and related generalized eigenvalue and singular value decompositions. Department of Computer Science, University of Colorado at Boulder, Technical report CU-CS-833-97, March 1997.

[20] Z. Drmač and E. R. Jessup. On accurate generalized singular value computation in floating–point arithmetic. Department of Computer Science, University of Colorado at Boulder, Technical Report CU-CS-811-96, submitted to SIAM J. Matrix Anal. Appl., October 1996.

[21] S. Eisenstat and I. Ipsen. Relative perturbation techniques for singular value problems. *SIAM J. Num. Anal.*, 32(6):1972–1988, 1995.

[22] S. K. Godunov, A. G. Antonov, O. P. Kirilyuk, and V. I. Kostin. *Garantirovannaya tochnost resheniya sistem lineĭnykh uravneniĭ v evklidovykh prostranstvakh.* Novosibirsk Nauka, Sibirskoe Otdelenie, 1988.

[23] G. H. Golub. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7:206–216, 1965.

[24] G. H. Golub and C. F. Van Loan. *Matrix Computations, second edition.* The Johns Hopkins University Press, 1989.

[25] G. H. Golub and H. Zha. Perturbation analysis of the canonical correlations of matrix pairs. *Linear Algebra Appl.*, 210:3–28, 1994.

[26] G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. In *Linear Algebra for Signal Processing, The IMA volumes in mathematics and its applications (A. Bojanczyk and G. Cybenko, eds.)*, volume 69, pages 27–49, 1995.

[27] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, 1996.

[28] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[29] Ren-Cang Li. Relative perturbation theory: (I) Eigenvalue and singular value variations. Technical report, Mathematical Science Section, Oak Ridge National Laboratory, Oak Ridge, TN 37831–6367, January 1996.

[30] M. Olschowka and A. Neumaier. A new pivoting strategy for Gaussian eleimination. 1993.

[31] G. Peters and J. H. Wilkinson. The least squares problem and pseudoinverses. *Comput. J.*, 13:309–316, 1970.

[32] M. J. D. Powell and J. K. Reid. On applying Householder transformations to linear least squares problems. In *Information Processing 68, Proc. International Federation of Information Processing Congress, Edinburgh, 1968*, pages 122–126. North Holland, Amsterdam, 1969.

[33] G. W. Stewart. Perturbation bounds for the QR decomposition of a matrix. *SIAM J. Numer. Anal.*, 14:509–518, 1977.

[34] G. W. Stewart. On the asymptotic behavior of scaled singular value and QR decompositions. *Math. Comp*, 43:483–489, 1984.

[35] G. W. Stewart. On the perturbation of LU and Cholesky factors. Technical Report TR–3535, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, 1995.

[36] G. W. Stewart. The triangular matrices of Gaussian elimination and related decompositions. Technical Report TR–3533, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, 1995.

[37] G. W. Stewart and Ji-Guang. Sun. *Matrix Perturbation Theory.* Academic Press, 1990.

[38] Ji-Guang Sun. Perturbation bounds for the Cholesky and QR factorizations. *BIT*, 31:341–352, 1991.

[39] Ji-Guang Sun. Componentwise perturbation bounds for some matrix decompositions. *BIT*, 32:702–714, 1992.

[40] Ji-Guang Sun. On perturbation bounds for the QR factorization. *Linear Algebra Appl.*, 215:95–111, 1995.

[41] N. L. Trefethen and R.S. Schreiber. Average–case stability of gaussian elimination. *SIAM J. Matrix Anal. Appl.*, 11:335–360, 1990.

[42] A. van der Sluis. Condition numbers and equilibration of matrices. *Numer. Math.*, 14:14–23, 1969.

[43] C. F. Van Loan. A generalized SVD analysis of some weighting methods for equality constrained least squares. In *Matrix Pencils Proceedings of a Conference.* Springer Verlag, 1982.

[44] D. Viswanath and L. N. Trefethen. Condition numbers of random triangular matrices. Preprint at `http://simon.cs.cornell.edu/home/lnt/`, 1996.

[45] P. Å. Wedin. On angles between subspaces of a finite dimensional inner product space. In *Matrix Pencils Proceedings of a Conference.* Springer Verlag, 1982.

[46] H. Zha. A componentwise perturbation analysis of the QR decomposition. *SIAM J. Matrix Anal. Appl.*, 4:1124–131, 1993.