Checking controllability numerically

In the following an algorithm will be derived which only uses unitary transformations to transform an arbitrary pencil $\lambda F + G \in \mathbb{C}[\lambda]_1^{p,q}$ into a canonical form. This canonical form reveals the controllability of the system $\mathfrak{B}(\lambda F + G)$.

To do so we repeatedly have to compress the rows (and then also the columns) of an arbitrary matrix $F \in \mathbb{C}^{p,q}$, i.e., we repeatedly have to compute a unitary matrix $X \in \mathbb{C}^{p,p}$ such that

$$XF = \begin{bmatrix} F_1 \\ 0 \end{bmatrix},$$

where F_1 has full row rank.

Lemma 1 (Rank revealing QR-decomposition)**.** *Let $F \in \mathbb{C}^{p,q}$. Then there exists a finite number of Householder transformations*

$$H_i := I - 2\frac{v_i v_i^*}{v_i^* v_i} \in \mathbb{C}^{p,p}, \qquad v_i \in \mathbb{C}^p, \qquad \textit{for } i = 1, \ldots, r$$

such that

$$H_r \cdots H_1 F = \begin{bmatrix} L \\ 0 \end{bmatrix},$$

where $L \in \mathbb{C}^{r,q}$ has full row rank r.

Proof. The proof is an algorithm. Initialize $r := 0$. If $F = 0$ there is nothing to do. Otherwise, increase $r := r + 1 = 1$ and let $P_1 \in \mathbb{C}^{q,q}$ be a permutation matrix, that moves a column of F which has the greatest norm $\|\cdot\|_2$ to the front. Let H_1 be the Householder transformation which makes this first column of FP a multiple of the first unit vector. Then we have the form

$$H_1 F P_1 = \left[\begin{array}{c|ccc} \square & \times & \cdots & \times \\ \hline 0 & \times & \cdots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \cdots & \times \end{array}\right] =: \begin{bmatrix} \square & \times \\ 0 & F_2 \end{bmatrix},$$

where $\times$ denotes arbitrary entries and $\square$ denotes non-singular scalars (and later also non-singular matrices). The element in the (1,1) position is nonzero, since here $F \neq 0$.

If $F_2 = 0$ we are done. Otherwise, increase $r := r + 1 = 2$ let $P_2 \in \mathbb{C}^{q-1,q-1}$ be a permutation matrix, that moves a column of F_2 which has the greatest norm $\|\cdot\|_2$ to the front. Let H_2 be the Householder transformation which makes this first column of $F_2 P_2$ a multiple of the first unit vector. Then we have the form

$$\underbrace{\begin{bmatrix} 1 & \\ & \tilde{H}_2 \end{bmatrix}}_{=:H_2} H_1 F P_1 \begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} = \left[\begin{array}{cc|ccc} \square & \times & \times & \cdots & \times \\ 0 & \square & \times & \cdots & \times \\ \hline 0 & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \times & \cdots & \times \end{array}\right] =: \begin{bmatrix} \square & \times \\ 0 & F_3 \end{bmatrix},$$

where the element in the (2,2) position is nonzero, since here $F_2 \neq 0$.

If $F_3 = 0$ we are done. Otherwise, we continue inductively as before until we computed

$H_1, \dots, H_r$ and a permutation matrix $P \in \mathbb{C}^{q,q}$ such that

$$H_r \cdots H_1 F P = \begin{bmatrix} \Box & \times & \cdots & \times & \cdots & \times \\ 0 & \ddots & \ddots & \vdots & & \vdots \\ \vdots & \ddots & \Box & \times & \cdots & \times \\ \vdots & & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{bmatrix} =: \tilde{F}.$$

Since P is invertible we conclude that the first r rows of $\tilde{F}P^{-1}$ still have full row rank. Thus, $H_r \cdots H_1 F$ is in the desired form. □

Remark 2. a) Asymptotically the rank revealing QR-decomposition takes as much time as the usual QR-decomposition $\mathcal{O}(pq^2)$.

b) The Householder transformations H_i can be stored by only storing the Householder vectors v_i in the subdiagonal entries which become zero. This is also more efficient to compute the product of a matrix with a Householder transformation (which will be necessary in the following).

c) On a computer the test $F_i = 0$ has to be replaced by a test of the form

$$\|F_i\| < \text{tol} \tag{1}$$

(for some suitable norm $\| \cdot \|$) due to roundoff errors. We refer to the tests (1) as numerical rank decisions, because whenever $\|F_i\| \geq \text{tol}$, the algorithm increases r by one.

Lemma 3. *For $\lambda F + G \in \mathbb{C}[\lambda]_1^{p,q}$ there exist unitary matrices $X \in \mathbb{C}^{p,p}$ and $Y \in \mathbb{C}^{q,q}$ such that*

$$X\,(\lambda F + G)\,Y = \lambda \begin{bmatrix} F_1 & \tilde{F}_1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G_1 & \tilde{G}_1 \\ 0 & R_1 \end{bmatrix},$$

where $R_1 \in \mathbb{C}^{r,a}$ has full column rank a and $\begin{bmatrix} F_1 & \tilde{F}_1 \end{bmatrix} \in \mathbb{C}^{d,q}$ has full row rank d.

Especially, this implies that $\operatorname{rank}(F_1) \geq d - a$ *(since the number of columns of $\tilde{F}_1$ is equal to a) and that $r \geq a$ (since R_1 has full column rank).*

Proof. Use Lemma 1 to compute a unitary X which compresses the rows of F, i.e., such that

$$X(\lambda F + G) =: \lambda \begin{bmatrix} L \\ 0 \end{bmatrix} + \begin{bmatrix} \hat{G}_1 \\ \hat{G}_2 \end{bmatrix}.$$

Then use Lemma 1 again to compute a unitary Y which compresses the *columns* of $\hat{G}_2$, i.e., apply Lemma 1 to $\hat{G}_2^*$ to obtain a unitary $\tilde{Y}^*$ such that

$$\tilde{Y}^* \hat{G}_2^* = \begin{bmatrix} R_1^* \\ 0 \end{bmatrix}, \qquad \Rightarrow \underbrace{\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \tilde{Y}^*}_{=:Y^*} \hat{G}_2^* = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} R_1^* \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ R_1^* \end{bmatrix}, \qquad \Rightarrow \hat{G}_2 Y = \begin{bmatrix} 0 & R_1 \end{bmatrix},$$

where R_1^* has full row rank and thus R_1 has full column rank. We conclude that

$$X(\lambda F + G)Y \quad = \quad \left(\lambda \begin{bmatrix} L \\ 0 \end{bmatrix} + \begin{bmatrix} \hat{G}_1 \\ \hat{G}_2 \end{bmatrix} \right) Y =: \lambda \begin{bmatrix} F_1 & \tilde{F}_1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G_1 & \tilde{G}_1 \\ 0 & R_1 \end{bmatrix},$$

and thus the claim is shown. □

Lemma 4. *For $\lambda F + G \in \mathbb{C}[\lambda]_1^{p,q}$ there exist unitary matrices $X \in \mathbb{C}^{p,p}$ and $Y \in \mathbb{C}^{q,q}$ and $d, a, r \in \mathbb{N}_0$ such that*

$$X\,(\lambda F + G)\,Y = \lambda \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix} + \begin{bmatrix} G_{11} & G_{12} \\ 0 & G_{22} \end{bmatrix},$$

where $F_{11} \in \mathbb{C}^{d,q-a}$ has full row rank and $\lambda F_{22} + G_{22} \in \mathbb{C}[\lambda]^{r,a}$ is right prime.

Proof. In this proof, × denotes arbitrary blocks of matching dimension which shall not be further specified.

Introduce the notation $d_0 := p$, $a_0 := 0$, and $r_0 := 0$. If F already has full row rank then we set $d := d_0$, $a := a_0$, and $r := r_0$ and the proof is finished. Otherwise, if F does not have full row rank, using Lemma 3, we find that there exist unitary matrices X_1, Y_1 and $d_1, a_1, r_1 \in \mathbb{N}_0$ such that $r_1 > 0$ and

$$X_1(\lambda F + G)Y_1 = \lambda \begin{bmatrix} F_1 & \tilde{F}_1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G_1 & \times \\ 0 & R_1 \end{bmatrix},$$

where $F_1 \in \mathbb{C}^{d_1, q-a_1}$ and $R_1 \in \mathbb{C}^{r_1, a_1}$ with $\operatorname{rank}\left(\begin{bmatrix} F_1 & \tilde{F}_1 \end{bmatrix}\right) = d_1$, $\operatorname{rank}(R_1) = a_1 \leq r_1$, and $\operatorname{rank}(F_1) \geq d_1 - a_1$.

If F_1 has full row rank we set $d := d_1$, $a := a_1$, $r := r_1$, $X := X_1$, and $Y := Y_1$ to obtain

$$X\,(\lambda F + G)\,Y = \lambda \begin{bmatrix} F_1 & \tilde{F}_1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G_1 & \times \\ 0 & R_1 \end{bmatrix} =: \lambda \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix} + \begin{bmatrix} G_{11} & G_{12} \\ 0 & G_{22} \end{bmatrix},$$

which implies that for all $\lambda_0 \in \mathbb{C}$ we have

$$\operatorname{rank}(\lambda_0 F_{22} + G_{22}) = \operatorname{rank}(\lambda_0 0 + R_1) = \operatorname{rank}(R_1) = a$$

i.e., that $\lambda F_{22} + G_{22} \in \mathbb{C}[\lambda]^{r,a}$ is right prime. Thus in this case we are also finished. Otherwise, if F_1 does not have full row rank, using Lemma 3 again, we find that there exist unitary matrices X_2, Y_2 and $d_2, a_2, r_2 \in \mathbb{N}$ such that $r_2 > 0$ and

$$X_2(\lambda F_1 + G_1)Y_2 = \lambda \begin{bmatrix} F_2 & \tilde{F}_2 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G_2 & \times \\ 0 & R_2 \end{bmatrix},$$

where $F_2 \in \mathbb{C}^{d_2, q-a_1-a_2}$ and $R_2 \in \mathbb{C}^{r_2, a_2}$ with $\operatorname{rank}\left(\begin{bmatrix} F_2 & \tilde{F}_2 \end{bmatrix}\right) = d_2$, $\operatorname{rank}(R_2) = a_2 \leq r_2$, and $\operatorname{rank}(F_2) \geq d_2 - a_2$. In this case we have

$$\begin{array}{rcl} d_2 & = & \operatorname{rank}\left(\begin{bmatrix} F_2 & \tilde{F}_2 \end{bmatrix}\right) = \operatorname{rank}(F_1) < \operatorname{rank}\left(\begin{bmatrix} F_1 & \tilde{F}_1 \end{bmatrix}\right) = d_1, \\ r_2 & = & d_1 - d_2 = d_1 - \operatorname{rank}(F_1) \leq d_1 - (d_1 - a_1) = a_1 \end{array} \tag{2}$$

and also

$$\begin{array}{rl} & \begin{bmatrix} X_2 & \\ & I_{r_1} \end{bmatrix} X_1\,(\lambda F + G)\,Y_1 \begin{bmatrix} Y_2 & \\ & I_{a_1} \end{bmatrix} \\ = & \begin{bmatrix} X_1 & \\ & I_{r_1} \end{bmatrix} \left(\lambda \left[\begin{array}{c|c} F_1 & \tilde{F}_1 \\ \hline 0 & 0 \end{array}\right] + \left[\begin{array}{c|c} G_1 & \times \\ \hline 0 & R_1 \end{array}\right] \right) \begin{bmatrix} Y_1 & \\ & I_{a_1} \end{bmatrix} \\ = & \lambda \left[\begin{array}{cc|c} F_2 & \tilde{F}_2 & \times \\ 0 & 0 & \times \\ \hline 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{cc|c} G_2 & \times & \times \\ 0 & R_2 & \times \\ \hline 0 & 0 & R_1 \end{array}\right]. \end{array}$$

If F_2 has full row rank d_2 we set $d := d_2$, $a := a_1 + a_2$, $r := r_1 + r_2$, $X := \begin{bmatrix} X_2 & \\ & I_{r_1} \end{bmatrix} X_1$, and $Y := Y_1 \begin{bmatrix} Y_2 & \\ & I_{a_1} \end{bmatrix}$, to obtain

$$\begin{array}{rcl} X(\lambda F + G)Y & = & \lambda \left[\begin{array}{c|cc} F_2 & \tilde{F}_2 & \times \\ \hline 0 & 0 & \times \\ 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{c|cc} G_2 & \times & \times \\ \hline 0 & R_2 & \times \\ 0 & 0 & R_1 \end{array}\right] \\ & =: & \lambda \left[\begin{array}{c|c} F_{11} & F_{12} \\ \hline 0 & F_{22} \end{array}\right] + \left[\begin{array}{c|c} G_{11} & G_{12} \\ \hline 0 & G_{22} \end{array}\right], \end{array}$$

which implies that for all $\lambda_0 \in \mathbb{C}$ we have

$$\operatorname{rank}(\lambda_0 F_{22} + G_{22}) = \operatorname{rank}\left(\lambda_0 \begin{bmatrix} 0 & \times \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} R_2 & \times \\ 0 & R_1 \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} R_2 & \times \\ 0 & R_1 \end{bmatrix}\right) = a,$$

since R_2 and R_1 have full column rank. This means that $\lambda F_{22} + G_{22} \in \mathbb{C}[\lambda]^{a+r,a}$ is right prime. Thus in this case we are also finished. Otherwise, if F_2 does not have full row rank, we can repeat the process over and over. In this way we obtain a decreasing sequence $d_1 > d_2 > d_3 > \ldots > d_s$ with $s \in \mathbb{N}$ until at some point $d_s = 0$ or the corresponding F_s has full row rank d_s. Since one can think of F_s as a matrix with zero rows we see that after a finite number of steps (namely s) we have inductively constructed unitary X, Y such that

$$\begin{aligned} X\,(\lambda F + G)\,Y &= \lambda \left[\begin{array}{c|cccc} F_s & \tilde{F}_s & \times & \cdots & \times \\ \hline & 0 & \times & \cdots & \times \\ & & 0 & \ddots & \vdots \\ & & & \ddots & \times \\ & & & & 0 \end{array}\right] + \left[\begin{array}{c|cccc} G_s & \times & \times & \cdots & \times \\ \hline & R_s & \times & \cdots & \times \\ & & R_{s-1} & \ddots & \vdots \\ & & & \ddots & \times \\ & & & & R_1 \end{array}\right] \\ &=: \lambda \left[\begin{array}{c|c} F_{11} & F_{12} \\ \hline 0 & F_{22} \end{array}\right] + \left[\begin{array}{c|c} G_{11} & G_{12} \\ \hline 0 & G_{22} \end{array}\right]. \end{aligned}$$

Setting $a := a_1 + \ldots + a_s$ and $r := r_1 + \ldots + r_s$ we see that $\lambda F_{22} + G_{22} \in \mathbb{C}[\lambda]^{r,a}$ is right prime since for all $\lambda_0 \in \mathbb{C}$ we have that

$$\operatorname{rank}\,(\lambda_0 F_{22} + G_{22}) \quad = \quad \operatorname{rank}\left(\begin{bmatrix} R_s & \times & \cdots & \times \\ & R_{s-1} & \ddots & \vdots \\ & & \ddots & \times \\ & & & R_1 \end{bmatrix}\right) = a,$$

since $R_1, \ldots, R_s$ all have full column rank. Thus the claim is shown. □

Remark 5. In the proof of the previous Lemma 4 the equation (2) was not needed. However, by the same construction one can show inductively that for the sequences $r_1, \ldots, r_s$ and $a_1, \ldots, a_s$ it holds

$$r_1 \geq a_1 \geq r_2 \geq a_2 \geq \ldots \geq r_s \geq a_s.$$

Theorem 6. *For $\lambda F + G \in \mathbb{C}[\lambda]_1^{p,q}$ there exist unitary matrices $X \in \mathbb{C}^{p,p}$ and $Y \in \mathbb{C}^{q,q}$ and $p_1, p_2, p_3, q_1, q_3 \in \mathbb{N}_0$ such that*

$$X\,(\lambda F + G)\,Y \quad = \quad \lambda \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ 0 & F_{22} & F_{23} \\ 0 & 0 & F_{33} \end{bmatrix} + \begin{array}{cc} \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ 0 & G_{22} & G_{23} \\ 0 & 0 & G_{33} \end{bmatrix} & \begin{matrix} p_1 \\ p_2 \\ p_3 \end{matrix} \\ \begin{matrix} q_1 & \quad p_2 & \quad q_3 \end{matrix} & \end{array} \tag{3}$$

where

1. *$\lambda F_{11} + G_{11} \in \mathbb{C}[\lambda]^{p_1,q_1}$ is left prime and $F_{11} \in \mathbb{C}^{p_1,q_1}$ has full row rank,*
2. *$F_{22} \in \mathbb{C}^{p_2,p_2}$ is square and invertible, and*
3. *$\lambda F_{33} + G_{33} \in \mathbb{C}[\lambda]^{p_3,q_3}$ is right prime.*

The system $\mathfrak{B}(\lambda F + G)$ is controllable if and only if $p_2 = 0$.

Proof. Use Lemma 4 to obtain unitary $\tilde{X}_1, \tilde{Y}_1$ such that

$$\tilde{X}_1(\lambda F + G)\tilde{Y}_1 = \lambda \begin{bmatrix} \tilde{F}_{11} & \tilde{F}_{12} \\ 0 & F_{33} \end{bmatrix} + \begin{bmatrix} \tilde{G}_{11} & \tilde{G}_{12} \\ 0 & G_{33} \end{bmatrix},$$

where $\tilde{F}_{11}$ has full row rank and $\lambda F_{33} + G_{33}$ is right prime. Then use Lemma 4 again to obtain $\tilde{X}_2, \tilde{Y}_2$ such that

$$\tilde{X}_2 \left(\lambda \tilde{F}_{11}^* + \tilde{G}_{11}^*\right) \tilde{Y}_2 = \lambda \begin{bmatrix} F_{22}^* & F_{12}^* \\ 0 & F_{11}^* \end{bmatrix} + \begin{bmatrix} G_{22}^* & G_{12}^* \\ 0 & G_{11}^* \end{bmatrix}, \tag{4}$$

where F_{22}^* has full row rank and $\lambda F_{11}^* + G_{11}^*$ is right prime. This implies that F_{22} has full column rank. Furthermore, since $\tilde{F}_{11}$ has full row rank we see that

$$\tilde{F}_{11}^* = \begin{bmatrix} F_{22}^* & F_{12}^* \\ 0 & F_{11}^* \end{bmatrix}$$

has full column rank. This implies that F_{22}^* has full column rank and thus F_{22} has full row rank. We deduced that F_{22} has full row rank and full column rank; in other words F_{22} is invertible. Taking the conjugate-transposed of (4) this implies that

$$\tilde{Y}_2^* \left(\lambda \tilde{F}_{11} + \tilde{G}_{11}\right) \tilde{X}_2^* = \lambda \begin{bmatrix} F_{22} & 0 \\ F_{12} & F_{11} \end{bmatrix} + \begin{bmatrix} G_{22} & 0 \\ G_{12} & G_{11} \end{bmatrix}$$

and that $\lambda F_{11} + G_{11}$ is left prime (compare Series 5, Task 6). Exchanging the block rows and block columns we see that there exist unitary $\tilde{X}_3, \tilde{Y}_3$ such that

$$\tilde{X}_3 \left(\lambda \tilde{F}_{22} + \tilde{G}_{22}\right) \tilde{Y}_3 = \lambda \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix} + \begin{bmatrix} G_{11} & G_{12} \\ 0 & G_{22} \end{bmatrix}.$$

Partitioning the transformation of the matrix $\lambda \tilde{F}_{12} + \tilde{G}_{12}$ accordingly into

$$\tilde{X}_3 \left(\lambda \tilde{F}_{12} + \tilde{G}_{12}\right) =: \lambda \begin{bmatrix} F_{13} \\ F_{23} \end{bmatrix} + \begin{bmatrix} G_{13} \\ G_{23} \end{bmatrix},$$

we obtain that

$$\begin{aligned}
\underbrace{\begin{bmatrix} \tilde{X}_3 & \\ & I \end{bmatrix} \tilde{X}_1}_{=:X} (\lambda F + G) \underbrace{\tilde{Y}_1 \begin{bmatrix} \tilde{Y}_3 & \\ & I \end{bmatrix}}_{=:Y} &= \begin{bmatrix} \tilde{X}_3 & \\ & I \end{bmatrix} \left(\lambda \begin{bmatrix} \tilde{F}_{11} & \tilde{F}_{12} \\ 0 & F_{33} \end{bmatrix} + \begin{bmatrix} \tilde{G}_{11} & \tilde{G}_{12} \\ 0 & G_{33} \end{bmatrix} \right) \begin{bmatrix} \tilde{Y}_3 & \\ & I \end{bmatrix} \\
&= \lambda \begin{bmatrix} \tilde{X}_3 \tilde{F}_{11} \tilde{Y}_3 & \tilde{X}_3 \tilde{F}_{12} \\ 0 & F_{33} \end{bmatrix} + \begin{bmatrix} \tilde{X}_3 \tilde{G}_{11} \tilde{Y}_3 & \tilde{X}_3 \tilde{G}_{12} \\ 0 & G_{33} \end{bmatrix} \\
&= \lambda \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ 0 & F_{22} & F_{23} \\ 0 & 0 & F_{33} \end{bmatrix} + \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ 0 & G_{22} & G_{23} \\ 0 & 0 & G_{33} \end{bmatrix},
\end{aligned}$$

which proves the claim. □

The proof of Theorem 6 is constructive. However, to avoid copying one would in practice not simply apply (the algorithm behind) Lemma 4 to the conjugated transposed of the remaining part as done in (4). It is faster to implement the corresponding algorithm such that it works directly on the original system.

The single steps that have to be done practically are summarized by the following graphic. Here R denotes (constant) right prime matrices (i.e., matrices with full column rank), L denotes (constant) left prime matrices (i.e., matrices with full row rank), S denotes a (constant) invertible matrix, and "x" arbitrary other blocks.

$$\begin{aligned}
& ([\, x \,], [\, x \,]) \to \left(\left[\begin{array}{c} L \\ \hline 0 \end{array}\right], \left[\begin{array}{c} x \\ \hline x \end{array}\right] \right) \to \left(\left[\begin{array}{cc} x & x \\ \hline 0 & 0 \end{array}\right], \left[\begin{array}{cc} x & x \\ \hline 0 & R \end{array}\right] \right) \\
= \; & \left(\left[\begin{array}{c|c} \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 \end{array}\right], \left[\begin{array}{c|c} x & x \\ \hline 0 & R \end{array}\right] \right) \to \left(\left[\begin{array}{c|c} \mathbf{L} & \mathbf{x} \\ \mathbf{0} & \mathbf{x} \\ \hline 0 & 0 \end{array}\right], \left[\begin{array}{c|c} x & x \\ x & x \\ \hline 0 & R \end{array}\right] \right) \\
= \; & \left(\left[\begin{array}{c|c} L & x \\ \hline 0 & x \\ \hline 0 & 0 \end{array}\right], \left[\begin{array}{c|c} \mathbf{x} & x \\ \hline \mathbf{x} & x \\ \hline \mathbf{0} & R \end{array}\right] \right) \to \left(\left[\begin{array}{cc|c} x & x & x \\ \hline 0 & 0 & x \\ \hline 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|c} \mathbf{x} & \mathbf{x} & x \\ \hline \mathbf{0} & \mathbf{R} & x \\ \hline \mathbf{0} & \mathbf{0} & R \end{array}\right] \right) \\
= \; & \left(\left[\begin{array}{c|cc} \mathbf{x} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & x \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} x & x & x \\ \hline 0 & R & x \\ 0 & 0 & R \end{array}\right] \right) \to \left(\left[\begin{array}{c|cc} \mathbf{L} & \mathbf{x} & \mathbf{x} \\ \mathbf{0} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & x \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} x & x & x \\ x & x & x \\ \hline 0 & R & x \\ 0 & 0 & R \end{array}\right] \right) \\
= \; & \left(\left[\begin{array}{c|cc} L & x & x \\ \hline 0 & x & x \\ \hline 0 & 0 & x \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} \mathbf{x} & x & x \\ \hline \mathbf{x} & x & x \\ \hline \mathbf{0} & R & x \\ \mathbf{0} & 0 & R \end{array}\right] \right) \to \left(\left[\begin{array}{cc|cc} x & x & x & x \\ \hline 0 & 0 & x & x \\ \hline 0 & 0 & 0 & x \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} \mathbf{x} & \mathbf{x} & x & x \\ \hline \mathbf{0} & \mathbf{R} & x & x \\ \hline \mathbf{0} & \mathbf{0} & R & x \\ \mathbf{0} & \mathbf{0} & 0 & R \end{array}\right] \right) \\
& \to \ldots \to \left(\left[\begin{array}{c|ccc} L & x & \cdots & x \\ \hline 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & x \\ 0 & \cdots & 0 & 0 \end{array}\right], \left[\begin{array}{c|ccc} x & x & \cdots & x \\ \hline 0 & R & \ddots & \vdots \\ \vdots & \ddots & \ddots & x \\ 0 & \cdots & 0 & R \end{array}\right] \right) \qquad \text{(“right prime reduction” finished)} \\
=: \; & \left(\left[\begin{array}{c|c} \mathbf{L} & x \\ \hline \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{c|c} x & x \\ \hline 0 & G_{33} \end{array}\right] \right) \to \left(\left[\begin{array}{cc|c} \mathbf{0} & \mathbf{S} & x \\ \hline \mathbf{0} & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{cc|c} x & x & x \\ \hline 0 & 0 & G_{33} \end{array}\right] \right)
\end{aligned}$$

$$= \left(\left[\begin{array}{c|cc} 0 & S & x \\ \hline 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{c|cc} \mathbf{x} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & G_{33} \end{array}\right] \right) \rightarrow \left(\left[\begin{array}{c|cc} 0 & L & x \\ 0 & L & x \\ \hline 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{c|cc} \mathbf{L} & \mathbf{x} & \mathbf{x} \\ \mathbf{0} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & G_{33} \end{array}\right] \right)$$

$$= \left(\left[\begin{array}{c|c|c} 0 & \mathbf{L} & x \\ \hline 0 & \mathbf{L} & x \\ \hline 0 & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{c|c|c} L & x & x \\ \hline 0 & x & x \\ \hline 0 & 0 & G_{33} \end{array}\right] \right) \rightarrow \left(\left[\begin{array}{c|cc|c} 0 & \mathbf{S} & \mathbf{x} & x \\ \hline 0 & \mathbf{0} & \mathbf{S} & x \\ \hline 0 & \mathbf{0} & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{c|cc|c} L & x & x & x \\ \hline 0 & x & x & x \\ \hline 0 & 0 & 0 & G_{33} \end{array}\right] \right)$$

$$= \left(\left[\begin{array}{c|c|cc} 0 & S & x & x \\ \hline 0 & 0 & S & x \\ \hline 0 & 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{c|c|cc} L & x & x & x \\ \hline \mathbf{0} & \mathbf{x} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & 0 & G_{33} \end{array}\right] \right) \rightarrow \left(\left[\begin{array}{c|c|cc} 0 & S & x & x \\ \hline 0 & 0 & L & x \\ 0 & 0 & L & x \\ \hline 0 & 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{c|c|cc} L & x & x & x \\ \hline \mathbf{0} & \mathbf{L} & \mathbf{x} & \mathbf{x} \\ \mathbf{0} & \mathbf{0} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & 0 & G_{33} \end{array}\right] \right)$$

$$= \left(\left[\begin{array}{cc|c|c} 0 & S & \mathbf{x} & x \\ 0 & 0 & \mathbf{L} & x \\ \hline 0 & 0 & \mathbf{L} & x \\ \hline 0 & 0 & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{cc|c|c} L & x & x & x \\ 0 & L & x & x \\ \hline 0 & 0 & x & x \\ \hline 0 & 0 & 0 & G_{33} \end{array}\right] \right) \rightarrow \left(\left[\begin{array}{cc|cc|c} 0 & S & \mathbf{x} & \mathbf{x} & x \\ 0 & 0 & \mathbf{S} & \mathbf{x} & x \\ \hline 0 & 0 & \mathbf{0} & \mathbf{S} & x \\ \hline 0 & 0 & \mathbf{0} & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{cc|cc|c} L & x & x & x & x \\ 0 & L & x & x & x \\ \hline 0 & 0 & x & x & x \\ \hline 0 & 0 & 0 & 0 & G_{33} \end{array}\right] \right)$$

$$= \left(\left[\begin{array}{cc|c|cc} 0 & S & x & x & x \\ 0 & 0 & S & x & x \\ \hline 0 & 0 & 0 & S & x \\ \hline 0 & 0 & 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{cc|c|cc} L & x & x & x & x \\ 0 & L & x & x & x \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{x} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & 0 & 0 & G_{33} \end{array}\right] \right) \rightarrow \left(\left[\begin{array}{cc|c|cc} 0 & S & x & x & x \\ 0 & 0 & S & x & x \\ \hline 0 & 0 & 0 & L & x \\ 0 & 0 & 0 & L & x \\ \hline 0 & 0 & 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{cc|c|cc} L & x & x & x & x \\ 0 & L & x & x & x \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{L} & \mathbf{x} & \mathbf{x} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{x} & \mathbf{x} \\ \hline 0 & 0 & 0 & 0 & G_{33} \end{array}\right] \right)$$

$$= \left(\left[\begin{array}{ccc|c|c} 0 & S & x & \mathbf{x} & x \\ 0 & 0 & S & \mathbf{x} & x \\ 0 & 0 & 0 & \mathbf{L} & x \\ \hline 0 & 0 & 0 & \mathbf{L} & x \\ \hline 0 & 0 & 0 & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{ccc|c|c} L & x & x & x & x \\ 0 & L & x & x & x \\ 0 & 0 & L & x & x \\ \hline 0 & 0 & 0 & x & x \\ \hline 0 & 0 & 0 & 0 & G_{33} \end{array}\right] \right) \rightarrow \left(\left[\begin{array}{ccc|cc|c} 0 & S & x & \mathbf{x} & \mathbf{x} & x \\ 0 & 0 & S & \mathbf{x} & \mathbf{x} & x \\ 0 & 0 & 0 & \mathbf{S} & \mathbf{x} & x \\ \hline 0 & 0 & 0 & \mathbf{0} & \mathbf{S} & x \\ \hline 0 & 0 & 0 & \mathbf{0} & \mathbf{0} & F_{33} \end{array}\right], \left[\begin{array}{ccc|cc|c} L & x & x & x & x & x \\ 0 & L & x & x & x & x \\ 0 & 0 & L & x & x & x \\ \hline 0 & 0 & 0 & 0 & x & x \\ \hline 0 & 0 & 0 & 0 & 0 & G_{33} \end{array}\right] \right)$$

$$\rightarrow \ldots \rightarrow \left(\left[\begin{array}{cccc|c|c} 0 & S & x & \cdots & x & x \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ddots & 0 & S & x & x \\ \hline 0 & \cdots & 0 & 0 & S & x \\ \hline 0 & \cdots & 0 & 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{cccc|c|c} L & x & \cdots & x & x & x \\ 0 & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \ddots & L & x & x & x \\ \hline 0 & \cdots & 0 & 0 & x & x \\ \hline 0 & \cdots & 0 & 0 & 0 & G_{33} \end{array}\right] \right) \quad \text{(“left prime reduction” finished)}$$

$$=: \left(\left[\begin{array}{c|c|c} F_{11} & x & x \\ \hline 0 & F_{22} & x \\ \hline 0 & 0 & F_{33} \end{array}\right], \left[\begin{array}{c|c|c} G_{11} & x & x \\ \hline 0 & G_{22} & x \\ \hline 0 & 0 & G_{33} \end{array}\right] \right)$$

Remark 7. As described here the algorithm needs $(\max(q,p))^4$ operations but it can be modified to take pq^2 [BD88] (this modified algorithm is very complicated). An algorithm which does this is implemented and called GUPTRI (from Generlized UPper TRIangular), cf. [DK93a, DK93b].

For the special case $\lambda F + G = \lambda \begin{bmatrix} 0 & I \end{bmatrix} + \begin{bmatrix} -B & -A \end{bmatrix} \in \mathbb{C}[\lambda]^{n,m+n}$ there exists a simpler (and faster) version of the algorithm which keeps the structure of the matrix $F = \begin{bmatrix} 0 & I \end{bmatrix}$. It will be discussed in the following.

If one only applies unitary transformations V in the form

$$\begin{aligned} V \left(\lambda \begin{bmatrix} 0 & I \end{bmatrix} + \begin{bmatrix} B & A \end{bmatrix} \right) \begin{bmatrix} I & \\ & V^* \end{bmatrix} &= \lambda \begin{bmatrix} 0 & VV^* \end{bmatrix} + \begin{bmatrix} VB & VAV^* \end{bmatrix} \\ &= \lambda \begin{bmatrix} 0 & I \end{bmatrix} + \begin{bmatrix} VB & VAV^* \end{bmatrix}, \end{aligned}$$

it has the advantage, that we keep the structure of the problem, i.e., in a computer it is sufficient to save the matrices A and B (or the transformed quantities VAV^* and VB) but one does not have to store $F := \begin{bmatrix} 0 & I \end{bmatrix}$ in the memory. Note that this is the same transformation which is used for the Kalman decomposition.

Since $F = \begin{bmatrix} 0 & I \end{bmatrix}$ has full row rank, it is clear that the *“right prime reduction”* is not necessary since the pencil is already in the wanted form (i.e., in Lemma 3 we have $p_2 = q_2 = 0$). Thus we can directly start with the *“left prime reduction”* . The result of this process is then given by the following Theorem.

Theorem 8. *Let $A \in \mathbb{C}^{n,n}$ and $B \in \mathbb{C}^{n,m}$. Then there exists a unitary matrix $V \in \mathbb{C}^{n,n}$, $s \in \mathbb{N}$, $s > 0$, and a finite sequence $n_1 \geq n_2 \geq \ldots \geq n_{s-1} \geq n_s \geq 0$ with $n_{s-1} > 0$ such that*

$$(VB, VAV^*) = \left(\begin{bmatrix} B_1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \left[\begin{array}{cccc|c} A_{1,1} & \cdots & \cdots & A_{1,s-1} & A_{1,s} \\ A_{2,1} & \ddots & & \vdots & \vdots \\ & \ddots & \ddots & \vdots & \vdots \\ & & A_{s-1,s-2} & A_{s-1,s-1} & A_{s-1,s} \\ \hline & & & & A_{s,s} \end{array}\right] \right), \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ n_{s-1} \\ \hline n_s \end{array} \tag{5}$$

$$\begin{array}{ccccc|c} m & n_1 & \cdots & \cdots & n_{s-1} & n_s \end{array}$$

where B_1, $A_{2,1}$, ..., $A_{s-1,s-2}$ have full row rank and $A_{s,s}$ is invertible. This implies that

$$\lambda F_{11} + G_{11} := \lambda \begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B_1 & A_{1,1} & \cdots & \cdots & A_{1,s-1} \\ 0 & A_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & A_{s-1,s-2} & A_{s-1,s-1} \end{bmatrix}$$

is left prime and thus $\mathfrak{B}(\lambda F_{11} + G_{11})$ is controllable. Especially, this means that (5) *gives a Kalman decomposition of (A, B). Furthermore, setting $\lambda F_{22} + G_{22} := \lambda I - A_{s,s}$ and $\lambda F_{21} + G_{21} := \lambda \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix} - \begin{bmatrix} A_{1,s}^T & \cdots & A_{s-1,s}^T \end{bmatrix}^T$, we see that we arrived at the form* (3), *since then*

$$V \left(\lambda \begin{bmatrix} 0 & I \end{bmatrix} - \begin{bmatrix} B & A \end{bmatrix} \right) \begin{bmatrix} I & \\ & V^* \end{bmatrix} = \lambda \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix} + \begin{bmatrix} G_{11} & G_{12} \\ 0 & G_{22} \end{bmatrix},$$

and thus the system is controllable if and only if $n_s = 0$.

Proof. The proof is an algorithm. If $B = 0$ set $s := 1$, $V := I$, and $n_1 := n$. Then we are done. Otherwise, set $n_1 := \operatorname{rank}(B)$. If then B has full row rank set $s := 2$, $V := I$, and $n_2 := 0$. Thus, in this case we are also done.

Otherwise, there exists a unitary $V_0 \in \mathbb{C}^{n,n}$ that $V_0 B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$ where $B \in \mathbb{C}^{n_1,m}$ has full row rank n_1 with $n > n_1 > 0$. We then have

$$V_0 \left(\lambda \begin{bmatrix} 0 & I \end{bmatrix} - \begin{bmatrix} B & A \end{bmatrix} \right) \begin{bmatrix} I & \\ & V_0^* \end{bmatrix} =: \left(\lambda \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B_1 & A_{1,1} & \tilde{A}_{1,2} \\ 0 & \tilde{A}_{2,1} & \tilde{A}_{2,2} \end{bmatrix} \right). \tag{6}$$

If $\tilde{A}_{2,1} = 0$ set $s := 2$, $V := V_0$, and $n_2 := n - n_1$. Then we are done. Otherwise, set $n_2 := \operatorname{rank}\left(\tilde{A}_{2,1}\right)$. If then $\tilde{A}_{2,1}$ has full row rank set $s := 3$, $V := V_0$, and $n_3 := 0$. Then we are also done.

Otherwise, there exists a unitary $V_1 \in \mathbb{C}^{n-n_1,n-n_1}$ such that $V_1 \tilde{A}_{2,1} = \begin{bmatrix} A_{2,1} \\ 0 \end{bmatrix}$ where $A_{2,1} \in \mathbb{C}^{n_2,n_1}$ has full row rank n_2 with $n - n_1 > n_2 > 0$. We then have

$$\begin{aligned} &\begin{bmatrix} I & \\ & V_1 \end{bmatrix} \left(\lambda \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B_1 & A_{1,1} & \tilde{A}_{1,2} \\ 0 & \tilde{A}_{2,1} & \tilde{A}_{2,2} \end{bmatrix} \right) \begin{bmatrix} I & & \\ & I & \\ & & V_1^* \end{bmatrix} \\ =: &\left(\lambda \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B_1 & A_{1,1} & A_{1,2} & \tilde{A}_{1,3} \\ 0 & A_{2,1} & A_{2,2} & \tilde{A}_{2,3} \\ 0 & 0 & \tilde{A}_{3,2} & \tilde{A}_{3,3} \end{bmatrix} \right). \end{aligned}$$

If $\tilde{A}_{3,2} = 0$ set $s := 3$, $V := \operatorname{diag}(I, V_1) V_0$, and $n_3 := n - n_1 - n_2$. Then we are done. Otherwise, set $n_3 := \operatorname{rank}\left(\tilde{A}_{3,2}\right)$. If then $\tilde{A}_{3,2}$ has full row rank set $s := 4$, $V := \operatorname{diag}(I, V_1) V_0$, and $n_4 := 0$. Then we are also done.

Otherwise, let $V_2 \in \mathbb{C}^{n-n_1-n_2,n-n_1-n_2}$ be unitary such that $V_2\tilde{A}_{3,2} = \begin{bmatrix} A_{3,2} \\ 0 \end{bmatrix}$ where $A_{3,2} \in \mathbb{C}^{3,2}$ has full row rank n_3 with $n - n_1 - n_2 > n_3 > 0$. We then have

$$\begin{bmatrix} I & & \\ & I & \\ & & V_2 \end{bmatrix} \left(\lambda \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B_1 & A_{1,1} & A_{1,2} & \tilde{A}_{1,3} \\ 0 & A_{2,1} & A_{2,2} & \tilde{A}_{2,3} \\ 0 & 0 & \tilde{A}_{3,2} & \tilde{A}_{3,3} \end{bmatrix} \right) \begin{bmatrix} I & & & \\ & I & & \\ & & I & \\ & & & V_2^* \end{bmatrix}$$

$$=: \left(\lambda \begin{bmatrix} 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B_1 & A_{1,1} & A_{1,2} & A_{1,3} & \tilde{A}_{1,4} \\ 0 & A_{2,1} & A_{2,2} & A_{2,3} & \tilde{A}_{2,4} \\ 0 & 0 & A_{3,2} & A_{3,3} & \tilde{A}_{3,4} \\ 0 & 0 & 0 & \tilde{A}_{4,3} & \tilde{A}_{4,4} \end{bmatrix} \right)$$

If $\tilde{A}_{4,3} = 0$ set $s := 4$, $V := \operatorname{diag}(I, I, V_2) \operatorname{diag}(I, V_1) V_0$, and $n_4 := n - n_1 - n_2 - n_3$. Then we are done. Otherwise, set $n_4 := \operatorname{rank}\left(\tilde{A}_{4,3}\right)$. If then $\tilde{A}_{4,3}$ has full row rank set $s := 5$, $V := \operatorname{diag}(I, I, V_2) \operatorname{diag}(I, V_1) V_0$, and $n_5 := 0$. Then we are also done.

Otherwise, continue in the very same way until at some point $\tilde{A}_{s,s-1}$ becomes zeros (in this case $n_s > 0$) or until $\tilde{A}_{s,s-1}$ has full row rank (in this case $n_s = 0$). Since the size of $\tilde{A}_{s,s-1}$ decreases in every step by at least one it is assured that the iteration stops after a finite number of steps. □

Remark 9. If the number of inputs m is considered to be constant the algorithm of the proof of Theorem 8 can be implemented so that it has an asymptotic runtime of $\mathcal{O}(n^3)$. Therefore, it is necessary that one does not compute the unitary matrices V_i explicitly but to store the Householder transformations H_i from Lemma 1 in form of the Householder vectors v_i (as indicated in Remark 2 b)) since one can then compute the product H_iX with asymptotic costs of $\mathcal{O}(n^2)$. If m is constant (and thus small, since we only analyze the case where n grows but not m) this implies that also the necessary matrix product in (6)

$$V_0AV_0^* = H_1 \cdots H_rAH_r^* \cdots H_1^*,$$

can be computed with asymptotic costs of $\mathcal{O}(n^2)$. Since the algorithm of the proof of Theorem 8 takes at most n iterations the overall complexity is then $\mathcal{O}(n^3)$.

Remark 10. Although unitary matrices have nice numerical properties they have the disadvantage that they do not well conserve the sparsity of matrices. Thus, for large sparse matrices it might be better to use non-unitary transformations. Note that all the constructions/algorithms work as well, when X and Y are not unitary, but only invertible.

References

[BD88] Th. Beelen and P. Van Dooren. An improved algorithm for the computation of Kronecker's canonical form of a singular pencil. *Linear Algebra and its Applications*, 105(0):9 – 65, 1988.

[DK93a] J. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil A - zB: robust software with error bounds and applications. Part I: theory and algorithms. *ACM Trans. Math. Software*, 19(2):160–174, 1993.

[DK93b] J. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil A - zB: robust software with error bounds and applications. Part II: software and applications. *ACM Trans. Math. Software*, 19(2):175–201, 1993.