# Analysis of the rational Krylov subspace method for large-scale algebraic Riccati equations


V. Simoncini

Dipartimento di Matematica, Università di Bologna

valeria.simoncini@unibo.it


*Partly joint works with M. Monsalve, Y.Lin, D. Szyld*

## The problem

Find $X \in \mathbb{R}^{n\times n}$ such that

$$AX + XA^\top - XBB^\top X + C^\top C = 0$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{s\times n}$, $p, s = \mathcal{O}(1)$

**Rich literature on analysis, applications and numerics:**

Lancaster-Rodman 1995, Bini-Iannazzo-Meini 2012, Mehrmann etal 2003 ...

# The problem

Find $X \in \mathbb{R}^{n \times n}$ such that

$$AX + XA^\top - XBB^\top X + C^\top C = 0$$

with $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{s \times n}$, $p, s = \mathcal{O}(1)$

**Rich literature on analysis, applications and numerics:**

Lancaster-Rodman 1995, Bini-Iannazzo-Meini 2012, Mehrmann etal 2003 ...

We focus on the large scale case: $n \gg 1000$

Different strategies

- (Inexact) Kleinman iteration (Newton-type method)
- Projection methods
- Invariant subspace iteration
- (Sparse) multilevel methods
- ....

## Galerkin projection method for the Riccati equation

Given the basis $V_k$ for an approximation space, determine approximation

$$X_k = V_k Y_k V_k^\top$$

to the Riccati solution matrix by orthogonal projection:

$$V_k^\top (AX_k + X_k A^\top - X_k BB^\top X_k + C^\top C) V_k = 0 \qquad \text{(Galerkin condition)}$$

giving

$$(V_k^\top A V_k) Y + Y (V_k^\top A^\top V_k) - Y (V_k^\top BB^\top V_k) Y + (V_k^\top C^\top)(CV_k) = 0$$

(Heyouni-Jbilou 2009)

## Galerkin projection method for the Riccati equation

Given the basis $V_k$ for an approximation space, determine approximation

$$X_k = V_k Y_k V_k^\top$$

to the Riccati solution matrix by orthogonal projection:

$$V_k^\top (AX_k + X_k A^\top - X_k BB^\top X_k + C^\top C) V_k = 0 \qquad \text{(Galerkin condition)}$$

giving

$$(V_k^\top A V_k) Y + Y (V_k^\top A^\top V_k) - Y_k (V_k^\top BB^\top V_k) Y_k + (V_k^\top C^\top)(C V_k) = 0$$

(Heyouni-Jbilou 2009)

Key questions:

- Which approximation space?
- Is this meaningful from a Control Theory perspective?

## On the choice of approximation space

Approximate solution $X_k = V_k Y_k V_k^\top$, with

$$(V_k^\top A V_k)Y + Y(V_k^\top A^\top V_k) - Y_k(V_k^\top BB^\top V_k)Y_k + (V_k^\top C^\top)(CV_k) = 0$$

Krylov-type subspaces: (from Lyapunov case)

- $\mathcal{K}_k(A, C^\top) := \mathrm{Range}([C^\top, AC^\top, \ldots, A^{k-1}C^\top])$ (Polynomial)
- $\mathcal{EK}_k(A, C^\top) := \mathcal{K}_k(A, C^\top) \cup \mathcal{K}_k(A^{-1}, A^{-1}C^\top)$ (EKSM, Rational)
- $\mathcal{RK}_k(A, C^\top, \mathbf{s}) :=$

  $$\mathrm{Range}([C^\top, (A - s_2 I)^{-1}C^\top, \ldots, \prod_{j=1}^{k-1}(A - s_{j+1}I)^{-1}C^\top])$$

  (RKSM, Rational)

## On the choice of approximation space

Approximate solution $X_k = V_k Y_k V_k^\top$, with

$$(V_k^\top A V_k) Y + Y (V_k^\top A^\top V_k) - Y_k (V_k^\top B B^\top V_k) Y_k + (V_k^\top C^\top)(C V_k) = 0$$

Krylov-type subspaces: (from Lyapunov case)

- $\mathcal{K}_k(A, C^\top) := \mathrm{Range}([C^\top, AC^\top, \ldots, A^{k-1}C^\top])$ (Polynomial)
- $\mathcal{EK}_k(A, C^\top) := \mathcal{K}_k(A, C^\top) \cup \mathcal{K}_k(A^{-1}, A^{-1}C^\top)$ (EKSM, Rational)
- $\mathcal{RK}_k(A, C^\top, \mathbf{s}) :=$

  $$\mathrm{Range}([C^\top, (A - s_2 I)^{-1} C^\top, \ldots, \prod_{j=1}^{k-1} (A - s_{j+1} I)^{-1} C^\top])$$

  (RKSM, Rational)

$\star$ Matrix $BB^\top$ **not** involved

## On the choice of approximation space

Approximate solution $X_k = V_k Y_k V_k^\top$, with

$$(V_k^\top A V_k)Y + Y(V_k^\top A^\top V_k) - Y_k(V_k^\top BB^\top V_k)Y_k + (V_k^\top C^\top)(CV_k) = 0$$

Krylov-type subspaces: (from Lyapunov case)

- $\mathcal{K}_k(A, C^\top) := \mathrm{Range}([C^\top, AC^\top, \ldots, A^{k-1}C^\top])$ (Polynomial)
- $\mathcal{EK}_k(A, C^\top) := \mathcal{K}_k(A, C^\top) \cup \mathcal{K}_k(A^{-1}, A^{-1}C^\top)$ (EKSM, Rational)
- $\mathcal{RK}_k(A, C^\top, \mathbf{s}) :=$

  $$\mathrm{Range}([C^\top, (A - s_2 I)^{-1}C^\top, \ldots, \prod_{j=1}^{k-1}(A - s_{j+1}I)^{-1}C^\top])$$

  (RKSM, Rational)

⋆ Matrix $BB^\top$ **not** involved

⋆ Parameters $s_j$ (adaptively) chosen in field of values of $-A$

## Performance of solvers

Problem: $A$: 3D Laplace operator, $B$, $C$ `randn` matrices, tol$=10^{-8}$

$(n,p,s) = (125000, 5, 5)$

| | its | inner its | time | space dim | rank($X_f$) |
|---|---|---|---|---|---|
| Newton $X_0 = 0$ | 15 | 5, . . . , 5 | 808 | 100 | 95 |
| GP-EKSM | 20 | | 531 | 200 | 105 |
| GP-RKSM | 25 | | 524 | 125 | 105 |

$(n,p,s) = (125000, 20, 20)$

| | its | inner its | time | space dim | rank($X_f$) |
|---|---|---|---|---|---|
| Newton $X_0 = 0$ | 19 | 5, . . . , 5 | 2332 | 400 | 346 |
| GP-EKSM | 15 | | 622 | 600 | 364 |
| GP-RKSM | 20 | | 720 | 400 | 358 |

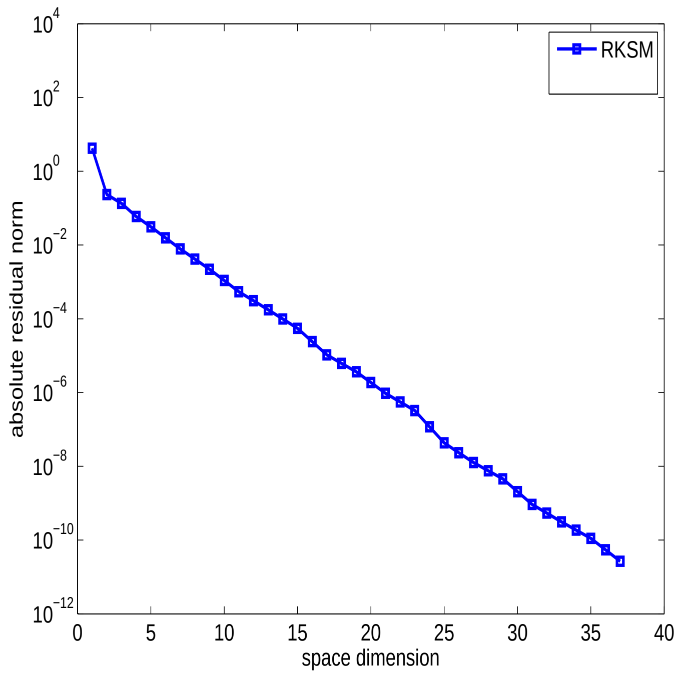
GP=Galerkin projection

(V.Simoncini, D.Szyld, M.Monsalve, 2014)

## A numerical example

Consider the $500 \times 500$ Toeplitz matrix

$$A = \mathrm{toeplitz}(-1, \underline{2.5}, 1, 1, 1), \quad C = [1, -2, 1, -2, \ldots], B = \mathbf{1}$$



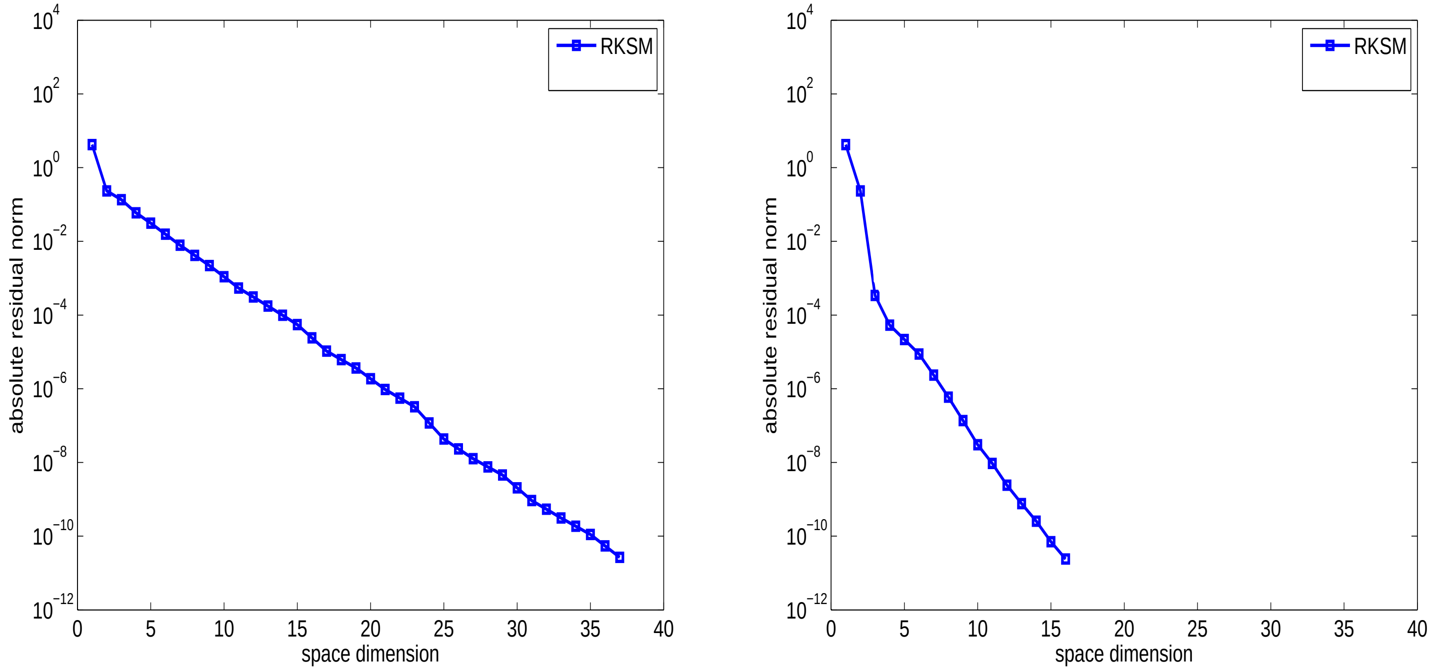

Parameter computation:

**Left**: adaptive RKSM on $A$

# A numerical example

Consider the $500 \times 500$ Toeplitz matrix

$$A = \text{toeplitz}(-1, \underline{2.5}, 1, 1, 1), \quad C = [1, -2, 1, -2, \ldots], B = \mathbf{1}$$




Parameter computation:
**Left**: adaptive RKSM on $A$   **Right**: adaptive RKSM on $A - BB^\top X_k$

(Lin, Simoncini 2015)

## Connection to dynamical systems

$$A\mathbf{X} + \mathbf{X}A^\top - \mathbf{X}BB^\top\mathbf{X} + C^\top C = 0$$

Time-invariant linear system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0 \\ y(t) = Cx(t), \end{cases}$$

$u(t)$ : control (input) vector; $\quad y(t)$ : output vector

$x(t)$ : state vector; $\quad x_0$ : initial state

## Connection to dynamical systems

$$A\mathbf{X} + \mathbf{X}A^\top - \mathbf{X}BB^\top\mathbf{X} + C^\top C = 0$$

Time-invariant linear system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), & x(0) = x_0 \\ y(t) = Cx(t), & \end{cases}$$

$u(t)$ : control (input) vector; $y(t)$ : output vector

$x(t)$ : state vector; $x_0$ : initial state

Minimization problem for a Cost functional: (simplified form)

$$\inf_u \mathcal{J}(u, x_0) \qquad \mathcal{J}(u, x_0) := \int_0^\infty (x(t)^\top C^\top Cx(t) + u(t)^\top u(t))dt$$

## Connection to dynamical systems

$$A\mathbf{X} + \mathbf{X}A^\top - \mathbf{X}BB^\top\mathbf{X} + C^\top C = 0$$

$$\inf_u \mathcal{J}(u, x_0) \qquad \mathcal{J}(u, x_0) := \int_0^\infty (x(t)^\top C^\top C x(t) + u(t)^\top u(t))dt$$

THEOREM. Let the pair $(A, B)$ be stabilizable and $(C, A)$ observable. Then there is a unique solution $\mathbf{X} \geq 0$ of the Riccati equation. Moreover,

i) For each $x_0$ there is a unique optimal control, and it is given by $u_*(t) = -B^\top\mathbf{X}\exp((A - BB^\top\mathbf{X})t)x_0$ for $t \geq 0$;

ii) $\mathcal{J}(u_*, x_0) = x_0^\top\mathbf{X}x_0$ for all $x_0 \in \mathbb{R}^n$

## Order reduction of dynamical systems by projection

Let $V_k \in \mathbb{R}^{n \times d_k}$ have orthonormal columns, $d_k \ll n$

Let $\quad T_k = V_k^\top A V_k, \quad B_k = V_k^\top B, \quad C_k^\top = V_k^\top C^\top$

Reduced order dynamical system:

$$\begin{cases} \dot{\widehat{x}}(t) = T_k \widehat{x}(t) + B_k \widehat{u}(t), & \widehat{x}(0) = \widehat{x}_0 := V_k^\top x_0 \\ \widehat{y}(t) = C_k \widehat{x}(t) \end{cases}$$

$x_k(t) = V_k \widehat{x}(t) \approx x(t)$

Typical frameworks:

- Transfer function approximation
- Model reduction

## The role of the projected Riccati equation

$$(V_k^\top A V_k)\mathbf{Y} + \mathbf{Y}(V_k^\top A^\top V_k) - \mathbf{Y}(V_k^\top B B^\top V_k)\mathbf{Y} + (V_k^\top C^\top)(C V_k) = 0$$

that is

$$T_k \mathbf{Y} + \mathbf{Y} T_k^\top - \mathbf{Y} B_k B_k^\top \mathbf{Y} + C_k^\top C_k = 0 \qquad (*)$$

Theorem. Let the pair $(T_k, B_k)$ be stabilizable and $(C_k, T_k)$ observable. Then there is a unique solution $\mathbf{Y}_k \geq 0$ of (*) that for each $\widehat{x}_0$ gives the feedback optimal control

$$\widehat{u}_*(t) = -B_k^* \mathbf{Y}_k \exp((T_k - B_k B_k^* \mathbf{Y}_k)t)\widehat{x}_0, \quad t \geq 0$$

for the reduced system.

## The role of the projected Riccati equation

$$(V_k^\top A V_k)\mathbf{Y} + \mathbf{Y}(V_k^\top A^\top V_k) - \mathbf{Y}(V_k^\top B B^\top V_k)\mathbf{Y} + (V_k^\top C^\top)(C V_k) = 0$$

that is

$$T_k \mathbf{Y} + \mathbf{Y} T_k^\top - \mathbf{Y} B_k B_k^\top \mathbf{Y} + C_k^\top C_k = 0 \qquad (*)$$

Theorem. Let the pair $(T_k, B_k)$ be stabilizable and $(C_k, T_k)$ observable. Then there is a unique solution $\mathbf{Y}_k \geq 0$ of (*) that for each $\widehat{x}_0$ gives the feedback optimal control

$$\widehat{u}_*(t) = -B_k^* \mathbf{Y}_k \exp((T_k - B_k B_k^* \mathbf{Y}_k)t)\widehat{x}_0, \quad t \geq 0$$

for the reduced system.

If there exists a matrix $K$ such that $A - BK$ is passive, then the pair $(T_k, B_k)$ is stabilizable.

## Projected optimal vs approximate optimal control functions

⋆ Our projected optimal control function:

$$\widehat{u}_*(t) = -B_k^\top Y_k \exp((T_k - B_k B_k^\top Y_k)t)\widehat{x}_0, \quad t \geq 0$$

with $X_k = V_k Y_k V_k^\top$

⋆ Typically used approximate control function:

$$\widetilde{u}(t) := -B^\top \widetilde{X} x(t)$$

where $\widetilde{x}(t) := \exp((A - BB^\top \widetilde{X})t)x_0$ for some $\widetilde{X} \approx X$

$$\widehat{u}_* \neq \widetilde{u}$$

$\boxed{\text{They induce different actions on the cost functional } \mathcal{J}\text{, even for } \widetilde{X} = X_k}$

## Projected optimal vs approximate optimal control functions

Residual matrix:

$$R_k := AX_k + X_k A - X_k BB^\top X_k + C^\top C$$

THEOREM. Assume that $A - BB^\top X_k$ is stable and that $\widetilde{u}(t) := -B^\top X_k x(t)$. Then

$$|\mathcal{J}(\widetilde{u}, x_0) - \widehat{\mathcal{J}}_k(\widehat{u}_*, \widehat{x}_0)| \leq \frac{\|R_k\|}{2\alpha} x_0^\top x_0,$$

where $\alpha > 0$ is such that $\|e^{(A-BB^\top X_k)t}\| \leq e^{-\alpha t}$ for all $t \geq 0$.

Note: $|\mathcal{J}(\widetilde{u}, x_0) - \widehat{\mathcal{J}}_k(\widehat{u}_*, \widehat{x}_0)|$ is nonzero for $R_k \neq 0$

## On the residual matrix and adaptive RKSM

$$R_k := AX_k + X_kA - X_kBB^\top X_k + C^\top C$$

THEOREM. Let $\mathcal{T}_k = T_k - B_kB_k^\top Y_k$. Then

$$R_k = \widehat{R}_kV_k^\top + V_k\widehat{R}_k^\top, \qquad \text{with} \quad \widehat{R}_k = AV_kY_k + V_kY_k\mathcal{T}_k^\top + C^\top(CV_k)$$

so that $\|R_k\|_F = \sqrt{2}\|\widehat{R}_k\|_F$

At least formally:

$\Rightarrow V_kY_kV_k^\top$ is a solution to the Riccati equation ($R_k = 0$) if and only if $Z_k = V_kY_k$ is the solution to the Sylvester equation ($\widehat{R}_k = 0$)

## On the residual matrix and adaptive RKSM

$$R_k = \widehat{R}_k V_k^\top + V_k \widehat{R}_k^\top$$

Expression for the semi-residual $\widehat{R}_k$:

THEOREM. Assume $C^\top \in \mathbb{R}^n$, Range($V_k$)= $\mathcal{RK}_k(A, C^\top, \mathbf{s})$. Assume that $\mathcal{T}_k = T_k - B_k B_k^\top Y_k$ is diagonalizable. Then

$$\widehat{R}_k = \psi_{k,T_k}(A) C^\top C V_k (\psi_{k,T_k}(-\mathcal{T}_k^\top))^{-1}.$$

where

$$\psi_{k,T_k}(z) = \frac{\det(zI - T_k)}{\prod_{j=1}^{k}(z - s_j)}$$

Here $\mathcal{T}_k = V_k^\top (A - BB^\top X_k) V_k = T_k - B_k B_k^\top Y_k$

(see also Beckermann 2011)

## On the choice of the next parameters $s_{k+1}$

$$\widehat{R}_k = \psi_{k,T_k}(A)C^\top C V_k(\psi_{k,T_k}(-\mathcal{T}_k^\top))^{-1}.$$

with $\psi_{k,T_k}(z) = \frac{\det(zI - T_k)}{\prod_{j=1}^{k}(z - s_j)}$

$\star$ Greedy strategy: Next shift should make $(\psi_{k,T_k}(-\mathcal{T}_k^\top))^{-1}$ smaller

$$\Downarrow$$

Determine for which $s$ in the spectral region of $\mathcal{T}_k$ the quantity $(\psi_{k,T_k}(-s))^{-1}$ is large, and add a root there

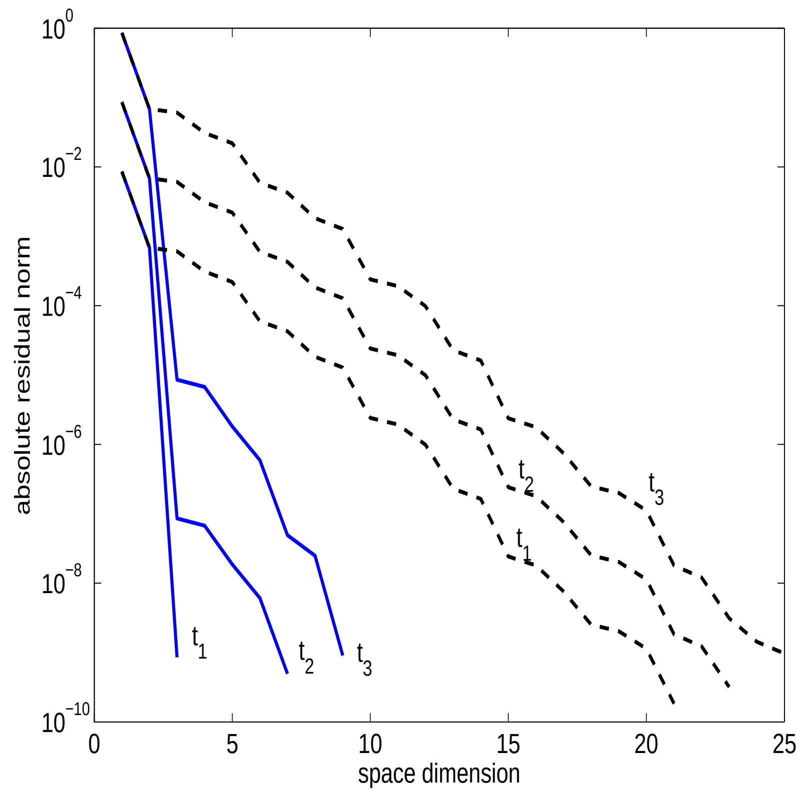
$$s_{k+1} = \arg\max_{s \in \partial\mathbb{S}_k} \left| \frac{1}{\psi_{k,T_k}(s)} \right|$$

$\mathbb{S}_k$ region enclosing the eigenvalues of $-\mathcal{T}_k$

(This argument is new also for Sylvester/Lyapunov equations)

## Selection of $s_{k+1}$ in RKSM. An example

$A$: $900 \times 900$ 2D Laplacian, $B = t\mathbf{1}$ with $t_j = 5 \cdot 10^{-j}$, $C = [1, -2, 1, -2, 1, -2, ...]$




RKSM convergence with and without modified shift selection as $t$ varies

**Solid curves:** use of $\mathcal{T}_k$ **Dashed curves:** use of $T_k$

## Further results and conclusions

**Not presented but relevant:**

- Stabilization properties of approximate solution $X_k$
- Approximation tracking as subspace grows
- Invariant subspace approximation

**Wrap-up:**

- Projection-type methods fill the gap between MOR and Riccati equation
- Clearer role of the non-linear term during the projection

# Further results and conclusions

## Not presented but relevant:

- Stabilization properties of approximate solution $X_k$
- Approximation tracking as subspace grows
- Invariant subspace approximation

## Wrap-up:

- Projection-type methods fill the gap between MOR and Riccati equation
- Clearer role of the non-linear term during the projection

References


- V. Simoncini, Daniel B. Szyld and Marlliny Monsalve,
  *On two numerical methods for the solution of large-scale algebraic Riccati equations* IMA Journal of Numerical Analysis, v.34, n.3, (2014)
- Yiding Lin and V. Simoncini,
  *A new subspace iteration method for the algebraic Riccati equation*
  Numerical Linear Algebra w/Appl., v.22, n.1, (2015)
- V.Simoncini, *Analysis of the rational Krylov subspace projection method for large-scale algebraic Riccati equations* arXiv: 1602.00649. To appear in SIMAX