UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS FÍSICAS E MATEMÁTICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA PURA E APLICADA

Felipe Wisniewski

Métodos de Resolução da Equação de Lyapunov e Aplicações em Redução de Modelo

Florianópolis
2019

Felipe Wisniewski

Métodos de Resolução da Equação de Lyapunov e Aplicações em Redução de Modelo

Tese submetida ao Programa de Pós-Graduação em Matemática Pura e Aplicada da Universidade Federal de Santa Catarina para a obtenção do título de doutor em Matemática com área de concentração em Matemática Aplicada.
Orientador: Prof. Licio Hernanes Bezerra, Dr.

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Wisniewski, Felipe
Métodos de Resolução da Equação de Lyapunov e Aplicações em Redução de Modelo / Felipe Wisniewski ; orientador, Licio H. Bezerra, 2019.
151 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências Físicas e Matemáticas, Programa de Pós-Graduação em Matemática Pura e Aplicada, Florianópolis, 2019.

Inclui referências.

1. Matemática Pura e Aplicada. 2. Matemática Aplicada. 3. Análise Numérica. 4. Equações de Lyapunov. 5. Sistemas Dinâmicos. I. H. Bezerra, Licio. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Matemática Pura e Aplicada. III. Título.

Felipe Wisniewski

Métodos de Resolução da Equação de Lyapunov e Aplicações em Redução de Modelo

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Licio Hernanes Bezerra, Dr.
Presidente e Orientador
Universidade Federal de Santa Catarina

Prof. Douglas Gonçalves Soares, Dr.
Universidade Federal de Santa Catarina

Prof. Francisco Damasceno Freitas, Dr.
Universidade de Brasília

Prof. Nelson Martins, Dr.
Centro de Pesquisas em Energia Elétrica (CEPEL-Eletrobrás)

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Matemática com área de concentração em Análise.

Assinado de forma digital por Marcelo Sobottka:88286096053
Dados: 2019.12.12 17:12:55 -03'00'

Prof. Marcelo Sobottka, Dr.
Coordenador do Programa

Documento assinado digitalmente
Licio Hernanes Bezerra
Data: 06/12/2019 15:35:01-0300
CPF: 431.380.047-68

Prof. Licio Hernanes Bezerra, Dr.
Orientador

Florianópolis, 2019

Este trabalho é dedicado
à minha família.

"The good thing about science is that it's true whether or not you believe in it."
(Neil deGrasse Tyson)

Resumo

Neste trabalho nós tratamos de métodos de resolução da equação de Lyapunov $AP + PA^T = -BB^T$, com $A \in \mathbb{R}^{n\times n}$ e $B \in \mathbb{R}^{n\times m}$. Num primeiro momento abordamos métodos já conhecidos na literatura como o método ADI (*Alternating Direction Implicit*) e um método baseado em subespaços de Krylov racionais (RKSM). Ambos os métodos são testados em sistemas dinâmicos descritores esparsos e, para isso, desenvolvemos uma implementação do RKSM específica para esse tipo de sistema, inspirando-se em implementações já feitas com o método ADI. Nós também fazemos uma análise da solução explícita da equação de Lyapunov para identificar em P um autoespaço de A que seja dominante num certo sentido que será definido no trabalho. A partir disso, propomos uma escolha de parâmetros para o método ADI . Essa escolha mostrou-se promissora em testes numéricos que fizemos, principalmente em situações em que o método ADI é utilizado para redução de modelo em sistemas descritores. Essa noção de dominância também é utilizada para determinar a região que contém os parâmetros utilizados no método baseado em subespaços de Krylov racionais (RKSM). Ao realizar testes numéricos notamos uma melhora significativa do método RKSM ao restringir essa região de busca de parâmetros. Nesse trabalho nós também introduzimos um critério de parada auxiliar para redução de modelo via polos dominantes, da qual surge uma nova definição de polos dominantes para sistemas dinâmicos. Por fim, nós introduzimos um método novo para resolução da equação de Lyapunov, baseado em métodos iterativos do tipo *splitting* para sistemas lineares. O método é construído a partir da formulação de Kronecker da equação de Lyapunov. Apresentamos uma breve análise de convergência do método e ilustramos com algumas aplicações numéricas. Ao final do trabalho fazemos um comparativo entre os métodos para a equação de Lyapunov. A comparação é feita com base na performance dos métodos em exemplos numéricos de redução de modelo. Esses testes evidenciam, dentre outras coisas, as melhorias significativas nos métodos já existentes para resolução da equação de Lyapunov, bem como destaca o potencial do novo método proposto neste trabalho.

Palavras-chave: Equaçao de Lyapunov; Redução de modelo; Autovalores Dominantes; Método Splitting.

Abstract

This doctoral dissertation addressed methods of solving the Lyapunov equation $AP + PA^T = {}^{-}BB^T$, with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. At the outset, we discussed methods already known in the literature, such as the ADI (Alternating Direction Implicit) method and the rational Krylov subspace method (RKSM). Those methods were tested on sparse dynamical descriptor systems and, to that end, we developed a way to implement the RKSM method specifically to that type of systems, resorting to implementations already tested with the ADI method. Furthermore, an analysis of the explicit solution of the Lyapunov equation was carried out to identify in P an eigenspace of A which is dominant in a certain direction, which was later defined. Based on that, we proposed a choice of parameters for the ADI method. The choice proved promising in the numerical tests done, especially in situations where the ADI method was used for model reduction of descriptor systems. This notion of dominance likewise served to determine the region containing the parameters used in the RKSM method. In the numerical tests, we noticed a significant improvement of the RKSM method when restricting this parameter search region. In this study, we also introduced an additional stopping criterion for model reduction via the dominant poles, from which comes a new definition of dominant poles for dynamical systems. Lastly, we introduced a new method for solving the Lyapunov equation based on iterative splitting methods for linear systems. The method was created from the Kronecker formulation of the Lyapunov equation. Subsequently, we presented a brief convergence analysis of the method, illustrated with some numerical applications. At the end of the study, we made a comparison between the methods for the Lyapunov equation. The comparison drew on how well the numerical methods performed under model reduction. The tests evinced, among other things, the significant improvements on existing methods, which we proposed to solve the Lyapunov equation, and pointed as well to the potential of the new method herein proposed.

Keywords: Lyapunov Equation; Model Reduction; Dominant Eigenspaces; Splitting Method.

Lista de Tabelas

Lista de Figuras

Sumário

Lista de Notações

$\mathbb{R}$ - Conjunto dos números reais
$\mathbb{R}_+$ - Conjunto dos números reais não negativos
$\mathbb{R}_-$ - Conjunto dos números reais não positivos
$\mathbb{R}^n$ - Espaço vetorial de todas as n-uplas de números reais
$\mathbb{C}$ - Espaço vetorial dos números complexos
$\overline{z}$ - Complexo conjugado de z (salvo notação específica)
$Re(z)$ - Parte real do número complexo z
$\mathbb{C}_+$ - Conjunto de números complexos z tais que $Re(z) > 0$
$\mathbb{C}_-$ - Conjunto de números complexos z tais que $Re(z) < 0$
j - Quando não for sub-índice, denota o número complexo $\sqrt{-1}$
$\langle \cdot, \cdot \rangle$ = Produto interno euclidiano
I - Identidade matricial de ordem n
$A > 0$ - A matriz A é positiva definida
$A \geq 0$ - A matriz A é positiva semi-definida
A^T - Transposta da matriz A
A^H - Transposta conjugada da matriz A
$rank(A)$ - Posto de A
$Im(A)$ - Imagem de A
$ker(A)$ - Núcleo de A
$span\{V\}$ - Espaço vetorial gerado pelo conjunto V
$diag(v)$ - Matriz diagonal cujas entradas da diagonal são as entradas do vetor v
e_i - i-esimo vetor canônico.
$\lambda_i(A)$ - i-ésimo autovalor da matriz A (pressupõe-se que eles estejam ordenados conforme grandeza do módulo, do maior para o menor)
$\sigma_i(A)$ - i-ésimo valor principal de A(pressupõe-se que eles estejam ordenados, do maior para o menor)
$\rho(A)$ - Raio espectral de A.
$L^2(\Omega)$ - Espaço das funções 2-integráveis, ou seja, funções f tais que $\left(\int_\Omega |f(x)|^2 \, dx\right)^{1/2} \leq \infty$, com a integral no sentido de Lebesgue e Ω sendo um subconjunto do domínio de f
$\| \cdot \|_{L^2(\Omega)}$ - Norma induzida pelo produto interno em $L^2(\Omega)$
$\| \cdot \|_2$ - Norma euclidiana
$\| \cdot \|_F$ - Norma de Frobenius
$\approx$ - Aproximadamente
C^1 - Classe de funções contínuas, com todas derivadas parciais contínuas num aberto
$supp(f)$ - Suporte da função f.

Introdução

Vamos considerar a equação de Lyapunov dada por

$$AP + PA^T = -BB^T, \tag{1}$$

com $A \in \mathbb{R}^{n\times n}$ e $B \in \mathbb{R}^{n\times m}$. Sob certas condições, há garantia de unicidade da solução P e, nessas circunstâncias, a matriz P será semi-definida positiva. A equação (1) está presente em inúmeras situações da matemática e da engenharia, conforme pode ser visto em [40]. Neste trabalho, a principal aplicação dessa equação refere-se ao cálculo das matrizes de Gram P e Q de um sistema dinâmico dado por um sistema de equações diferenciais da forma:

$$\Sigma \equiv \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \quad t \geq t_0, x(t_0) = x_0, \end{cases}$$

com $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, $x(t) \in \mathbb{R}^n$ e $u(t) \in \mathbb{R}^m$. Tais sistemas, bem como as matrizes de Gram associadas, são muito bem explicados no livro de Antoulas [1], que é nossa principal referência nesse assunto.

Dentre outras informações pertinentes ao sistema Σ, as colunas da matriz de Gram P são geradoras de todo o subespaço invariante por A (à direita) que atua de maneira efetiva no sistema Σ. O mesmo ocorre com a matriz Q que é relacionada ao subespaço de invariante por A à esquerda. Esse elo de ligação motivou inúmeras pesquisas que visam a redução de tamanho do modelo Σ em aplicações da engenharia. Dentre esses estudos podemos destacar o método de *redução por balanceamento*, que teve contribuições significativas feitas por [12], [44] e [9].

Outro método de redução de modelo é baseado em polos dominantes do sistema Σ, que nada mais são que autovalores da matriz A que exercem alguma dominância na relação entre as variáveis de entrada saída do sistema dinâmico. Nesse trabalho fazemos algumas contribuições que relacionam as matrizes P e Q a esse tipo de redução, que é chamada de *modal*. Uma das principais referências que adotamos para tratar desse assunto é [32].

Em muitas situações, como ocorre no caso de sistemas dinâmicos oriundos de modelos elétricos [24], o sistema Σ é esparso e de ordem elevada. Por isso, a abordagem dos métodos para equações de Lyapunov que fazemos nesse trabalho é direcionada para esse tipo de problema.

Simoncini, em um trabalho publicado em 2016 [40], faz um apanhado geral dos principais métodos para resolução de equações de Lyapunov dos últimos tempos. Para problemas de pequeno porte, existem métodos baseados na decomposição de Schur da matriz A, como o método Bartels-Stewart e o método de Golub-Nash-Van Lohan. Dentre os métodos utilizados em sistemas de médio e grande porte, destacam-se os métodos baseados em subespaços de Krylov, introduzidos pela própria autora, que têm demonstrado robustez e eficiência em aplicações numéricas. Esses métodos consistem em resolver a projeção da equação de Lyapunov (1) em algum subespaço de dimensão menor do que o tamanho do sistema original. Em [39] é sugerido o subespaço de Krylov

por blocos dado por

$$\mathbf{K}_l(A,B) = span\{B, A^{-1}B, AB, A^{-2}B, A^2B, A^{-3}B, ..., A^{l-2}B, A^{-(l-1)}B\}, \tag{2}$$

A vantagem de utilizar potências inversas de A, além de potências de A, é que a construção da base do espaço agrega informações tanto dos maiores autovalores de A como dos maiores autovalores de A^{-1}. Outra abordagem é feita em [8], considerando o subespaço de Krylov racional

$$K_l(A,B,\mu) := \text{span}\left\{B, (A-\mu_1 I)^{-1}B, ..., \left(\prod_{j=1}^{l}(A-\mu_j I)^{-1}\right)B\right\}, \tag{3}$$

para algum $\mu = (\mu_1, \mu_2, ..., \mu_l) \in \mathbb{C}^l$. A construção do subespaço poderia ser feita a partir de qualquer outra matriz, de ordem compatível, no lugar de B. Porém, a utilização da matriz B se mostra vantajosa por incorporar mais informações da equação original ao subespaço de projeção. Além disso, através da escolha do parâmetro μ, podem ser incorporadas informações do espectro de A que fazem com que o subespaço (3) contenha boas propriedades de aproximação para a equações de Lyapunov.

Outro método bastante utilizado em equações de Lyapunov e, inclusive, bem mais antigo que os métodos baseados em subespaços de Krylov, é o método conhecido como ADI (Alternating Direction Implicit). Uma das primeiras aparições desse método foi na versão apresentada por Smith em 1968 [42]. A partir disso, em 1988, [50] introduziu uma técnica que consiste em variar um determinado parâmetro durante a execução do método (essa técnica é o que motiva o nome de ADI). Durante décadas esse método recebeu vários aperfeiçoamentos. Dentre eles, destaca-se o trabalho de Penzl, em (1999) [30], que introduz um método que utiliza um conjunto de parâmetros fixos, de maneira cíclica, utilizando fatores de Cholesky para baratear significativamente os custos computacionais da execução do método em sistemas dinâmicos de grande porte definidos por matrizes de posto relativamente pequeno. Recentemente, uma atualização proposta por Freitas, Rommes e Martins [11], permite aplicar o método ADI de maneira eficiente em sistemas dinâmicos descritores. Um fato bastante interessante mostrado por Simoncini em [6] é que, sob certas condições, o método ADI é equivalente ao método de Krylov com a base racional dada em (3).

As nossas contribuições presentes nesse trabalho estão voltadas para os métodos de resolução de equações de Lyapunov e podem ser divididas em dois tópicos principais. O primeiro deles resgata uma abordagem elementar na qual a solução da equação de Lyapunov é dada analiticamente em função da decomposição espectral da matriz A. É obvio que, em problemas de grande porte, é computacionalmente inviável calcular todos os autovalores de A apenas para calcular a solução P. No entanto, essa representação analítica permite desenvolver uma análise a fim de evidenciar um subespaço invariante por A, de dimensão significativamente menor do que a ordem de A, que exerce uma certa dominância na solução P da equação (3). Isso dá origem à expressão *autovalores de A dominantes na solução P*, bastante usada nesse trabalho. A principal motivação para essa linha de raciocínio vem da teoria de polos dominantes em redução de modelo. Em trabalhos como [32], [47], [26], [34] e [46], é evidenciado que, em certos modelos, uma redução modal feita a partir de alguns poucos polos dominantes é capaz de gerar um modelo reduzido muito fiel ao original. Além disso, a atuação que um sistema dinâmico promove entre suas variáveis de entrada e saída está intimamente relacionado ao produto PQ das matrizes de Gram do sistema. Isso sugere que deve existir também uma dominância exercida por esses polos em cada uma das matrizes P e Q, de maneira isolada. Além disso, conforme mencionado por Penzl [30], em problemas esparsos de grande porte, é comum o chamado "fenômeno de baixo posto"(em inglês, *low-rank phenomenon*),

que faz com que o posto da solução P da equação (1) seja extremamente pequeno se comparado à ordem n do sistema. Como bem comenta Simoncini [40], ainda há muito o que se descobrir sobre a relação entre o decaimento dos autovalores de P e o espectro da matriz A da equação (1). Nosso ponto de partida nesse assunto é a análise de decaimento da solução P feita por Antoulas, Sorensen e Zhou (2002) [2].

No que se refere às aplicações dos conceitos do parágrafo anterior, nós propomos uma escolha de parâmetros para o método ADI baseada em um conjunto de autovalores de A dominantes em P. De maneira muito similar, nós também propomos uma região para a escolha dos parâmetros do método baseado em subespaços de Krylov racionais. Em ambos os casos nós atestamos a eficiência das escolhas através de testes numéricos. Além disso, nós também fazemos uso da expressão de P dada em função do espectro de A para definir um novo critério de dominância para polos de um sistema dinâmico. Grande parte das definições de polos dominantes presentes na literatura caracterizam um polo dominante de maneira isolada. A definição proposta visa caracterizar um polo dominante de maneira relativa a um conjunto de polos dominantes previamente determinados.

Note que o termo "dominante"é utilizado em duas situações diferentes. Para que não haja confusão, sempre que utilizamos a expressão *polos dominantes*, estamos nos referindo aos polos da redução modal. Quando nos dizemos *autovalores dominantes*, estamos nos referindo aos autovalores de A dominantes na solução P da equação de Lyapunov.

A nossa segunda contribuição significativa nesse trabalho é a criação de um método novo para resolução da equações de Lyapunov baseado em métodos iterativos para sistemas lineares construídos a partir de uma cisão da matriz de coeficientes do sistema. Esses métodos, comumente chamados de *splitting*, são devidamente explicados, por exemplo, em [13]. O método que propomos é construído a partir da representação da equação de Lyapunov (1) dada por um sistema linear do tipo $\tilde{A}\tilde{x} = \tilde{b}$. Essa representação é construída utilizando-se o produto de Kronecker. Como dito por Simoncini em [40], os métodos inspirados nesse tipo de representação não tem sido muito explorados na literatura pois o sistema linear $\tilde{A}\tilde{x} = \tilde{b}$ é de ordem n^2, o que torna caro qualquer método que utilize tais matrizes. No entanto, as iterações do método proposto operam utilizando apenas as matrizes A e B da equação original, ou seja, utiliza matrizes de tamanho n apenas, o que torna o método viável computacionalmente. Nesse trabalho nós também discutimos alguns aspectos relacionados a convergência do método apresentado e exibimos alguns testes numéricos.

Este trabalho está subdividido em cinco partes principais. A primeira delas contém um resumo dos conceitos relacionados aos sistemas dinâmicos que constituem o cenário das aplicações dos métodos para equações de Lyapunov. No segundo capítulo tratamos de dois métodos de redução de modelo: a redução modal, feita a partir de polos dominantes, e a redução por balanceamento, que faz uso das matrizes de Gram do sistema. O terceiro capítulo é destinado aos métodos de resolução da equações de Lyapunov, já conhecidos na literatura. Primeiramente, fazemos uma explicação resumida sobre o método ADI e sobre os métodos baseados em subespaços de Krylov. Em seguida tratamos da solução das relações entre a decomposição espectral de A e a solução P. No penúltimo capítulo, introduzimos um método novo para a resolução da equação de Lyapunov. Esse método é inspirado em algoritmos *splitting* para equações de Lyapunov. Por fim, no capítulo 5, expomos um comparativo, através de testes numéricos, entre todas as estratégias de redução de modelo abordadas nesse trabalho.

Capítulo 1

Sistemas Dinâmicos

Este capítulo tem por objetivo apresentar os conceitos mais elementares da teoria de controle em sistemas dinâmicos lineares. Esses conceitos são importantes para que se possa entender com clareza os métodos de redução de modelo, discutidos no capítulo seguinte, que constituem um dos principais cenários para as aplicações dos métodos discutidos e aprimorados nesse trabalho.

Grosso modo, um sistema dinâmico pode ser visto como um modelo matemático que recebe uma variável de entrada u e retorna uma variável de saída y, ambas dependentes do tempo t.

Figura 1.1: Planta básica de um sistema $\mathbf{\Sigma}$

Vamos considerar um sistema dinâmico linear $\mathbf{\Sigma}$, representado por um conjunto finito de equações diferenciais (dimensão finita), dado por:

$$\Sigma \equiv \begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t), & t > t_0, \quad x(t_0) = x_0, \\ y(t) = C(t)x(t) + D(t)u(t), & t \geq t_0, \end{cases} , \tag{1.1}$$

com $A(t) \in \mathbb{R}^{n\times n}$, $B(t) \in \mathbb{R}^{n\times m}$, $C(t) \in \mathbb{R}^{p\times n}$, $D(t) \in \mathbb{R}^{p\times m}$ e $t_0 \in \mathbb{R}$ fixado. Na maioria dos casos vamos considerar $t_0 = 0$. Quando a variável t é contínua, dizemos que o sistema Σ é contínuo. Caso contrário, diz-se que o sistema Σ é discreto.

Uma definição axiomática mais geral de sistemas dinâmicos pode ser vista em [22]. A representação (1.1) é chamada de *descrição interna* de um sistema dinâmico e é apresentada com mais detalhes em [1]. O Teorema a seguir estabelece condições para a unicidade da solução $x(t)$ do sistema dinâmico (1.1).

Teorema 1.1 *Supondo que haja uma condição inicial $x_0 = x(t_0)$ para o sistema (1.1) e assumindo que as entradas de $A(t)$, de $B(t)$ e de $u(t)$ são funções contínuas com relação ao tempo, então o sistema (1.1) admite uma única solução. A saber, tal solução é dada por:*

$$x(t) = \phi(x(t_0), t_0, u, t),$$

com

$$\phi(x(t_0), t_0, u, t) := \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)\ d\tau, \quad t \geq t_0 \tag{1.2}$$

e $\Phi(t, t_0) \in \mathbb{R}^{n\times n}$ dita uma matriz de transição.

Demonstração. A demonstração da unicidade pode ser vista em [5]. Já a expressão para a solução pode ser encontrada em [22]. □

Daqui em diante, vamos considerar apenas sistemas dinâmicos lineares invariantes no tempo do tipo a seguir:

$$\Sigma \equiv \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), & t > t_0, x(t_0) = x_0, \\ y(t) = Cx(t) + Du(t), & t \geq t_0, \end{cases} \tag{1.3}$$

com $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ e $D \in \mathbb{R}^{p\times m}$, matrizes cujas entradas são constantes. O vetor x é chamado de variável de *estado*, enquanto u e y são chamadas de variáveis de *entrada* e de *saída*, respectivamente. Daqui em diante, sempre que mencionarmos um sistema dinâmico (A, B, C, D), estamos nos referindo a um sistema dinâmico linear contínuo e invariante no tempo, como definido em (1.3). Nas situações em que a matriz D é nula, denotaremos o sistema dinâmico apenas por (A, B, C). Considerando a representação (1.3), a planta básica apresentada na Figura 1.1 pode ser substituída pelo diagrama de blocos da Figura 1.2, a seguir, que proporciona uma visualização mais detalhada do sistema dinâmico que tratamos aqui.

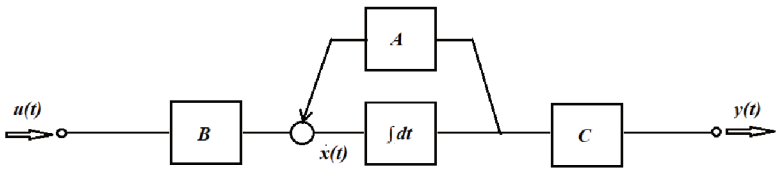


Figura 1.2: Representação do sistema (A, B, C) por um diagrama de blocos.

No caso particular em que $m = p = 1$ dizemos que o sistema dinâmico (1.3) é de *entrada única e saída única* ou, em inglês, *single input and single output* (SISO). Quando $m > 1$ e $p > 1$, diz-se que o sistema é de *entrada múltipla e saída múltipla* ou, em inglês, *multiple input and multiple output* (MIMO). Neste trabalho, exceto quando houver menção específica, estaremos tratando de sistema do tipo SISO.

Para o sistema (A, B, C, D), a solução $x(t)$ dada em (1.2) passa a ter uma representação mais simples dada por:

$$x(t) = \Phi(t - t_0)x_0 + \int_{t_0}^{t} \Phi(t - \tau)Bu(\tau)\ d\tau, \quad t \geq t_0. \tag{1.4}$$

Mais especificamente, $\Phi(t-t_0) = e^{A(t-t_0)}$, ou seja

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\ d\tau, \quad t \geq t_0. \tag{1.5}$$

Com efeito, se derivarmos (1.5) com relação a t, temos

$$\begin{aligned}
\dot{x}(t) &= \frac{d}{dt}\left[e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\ d\tau\right] && (1.6)\\
&= \frac{d}{dt}\left[e^{A(t-t_0)}x_0 + e^{At}\int_{t_0}^{t} e^{A(-\tau)}Bu(\tau)\ d\tau\right] && (1.7)\\
&= Ae^{A(t-t_0)}x_0 + Ae^{At}\int_{t_0}^{t} e^{A(-\tau)}Bu(\tau)\ d\tau + e^{At}e^{A(-t)}Bu(\tau) && (1.8)\\
&= A\left[e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\ d\tau\right] + Bu(t) && (1.9)\\
&= Ax(t) + Bu(t). && (1.10)
\end{aligned}$$

A unicidade dessa solução segue do Teorema 1.1, pois as funções envolvidas no sistema (1.3) são contínuas.

1.1 Estabilidade

Seguindo os passos de [1], para discutir sobre a estabilidade de sistemas dinâmicos, vamos desconsiderar as influências externas ao sistema dinâmico (A, B, C, D) e estudar apenas o sistema de equações diferenciais autônomo dado por

$$\dot{x}(t) = Ax(t), \quad A \in \mathbb{R}^{n\times n}. \tag{1.11}$$

O sistema (1.11) é dito *estável* se todas as soluções $x(t)$ são limitadas para $t > 0$. O sistema (1.11) é dito *assintoticamente estável* se todas as soluções $x(t)$ tendem a zero quando t tende a infinito.

1.1.1 Estabilidade e Autovalores

O teorema a seguir, retirado de [27], estabelece uma relação entre a estabilidade do sistema (1.11) e os autovalores da matriz A.

Teorema 1.2 *O sistema (1.11) é:*

- **assintoticamente estável** *se, e somente se, todos os autovalores de A tiverem parte real negativa, isto é, se A for Hurwitz.*
- **estável** *se, e somente se, todos os autovalores de A tiverem parte real não positiva e, além disso, para cada um dos autovalores imaginários puros, as multiplicidades algébrica e geométrica, coincidirem.*

Demonstração. A demonstração desse Teorema está no Apêndice A. □

1.1.2 Estabilidade no Sentido de Lyapunov

A descrição dos conceitos apresentados nesta subseção é inspirada, principalmente, em [5] e [22].

Considere o sistema autônomo

$$\dot{x}(t) = f(x(t)), \quad x(t) \in \mathbb{R}^n, \quad t \in \mathbb{R}, \tag{1.12}$$

com $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ de classe C^1. Seja x_0 um ponto de equilíbrio de (1.12). Vamos assumir, por questão de simplicidade, que $x_0 = 0$. Nessas condições, $x(t) = 0$ é solução de (1.12).

Definição 1.3 *Seja Ω um conjunto que contém a bola fechada com centro na origem e raio $r > 0$. Uma função $\Theta : \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, de classe C^1, é dita uma função de Lyapunov para o sistema (1.12) se, para toda solução $x : \mathbb{R} \longrightarrow \mathbb{R}^n$ de (1.12), tivermos*

$$\begin{aligned} &\tfrac{d}{dt}\Theta(x(t)) = \nabla\Theta(x(t))^H \cdot f(x(t)) \leq 0, \\ &\Theta(x(t)) > 0 \quad se \quad x \neq 0 \quad e \quad \Theta(0) = 0. \end{aligned}$$

Além disso, dizemos que Θ é uma função de Lyapunov estrita, quando $\frac{d}{dt}\Theta(x(t)) < 0$ para todo $x(t)$ que não é identicamente nulo.

Teorema 1.4 *Se existe uma função de Lyapunov para o sistema (1.12), então o sistema é estável. Se a função de Lyapunov for estrita, então o sistema é assintoticamente estável.*

Demonstração. A demonstração desse Teorema está no Apêndice A. □

Resta agora saber como encontrar a função Θ que satisfaz a Definição 1.3. Conforme foi feito na subseção anterior, vamos restringir a busca dessa função para o sistema linear (1.11).

Consideremos uma matriz simétrica positiva definida P. Desta forma, uma candidata é a função $\Theta(x(t)) = x(t)^T P x(t)$ [22]. Note que

$$\begin{aligned} \frac{d}{dt}(x(t)^T P x(t)) &= \dot{x}(t)^T P x(t) + x(t)^T P \dot{x}(t) \\ &= (Ax(t))^T P x(t) + x(t)^T P (Ax(t)) \\ &= x(t)^T (A^T P + PA) x(t) \\ &= x(t)^T E x(t), \end{aligned}$$

com $E = A^T P + PA$. Portanto, para determinar uma função de Lyapunov para o sistema (1.12), é suficiente escolher uma matriz $E \leq 0$ e encontrar uma solução P positiva definida para a equação

$$A^T P + PA = E. \tag{1.13}$$

O Teorema a seguir apresenta um resumo das construções relacionadas à estabilidade no sentido de Lyapunov, descritas até então.

Teorema 1.5 *Se existe uma solução $P > 0$ para a equação (1.13), com $E < 0$, então o sistema (1.12) é assintoticamente estável. Além disso, se existir uma solução $P > 0$ para (1.13), com $E \leq 0$, então o sistema (1.13) é estável.*

O Teorema 1.5 estabelece uma das relações entre equações de Lyapunov e sistemas dinâmicos apresentadas nesse trabalho. Porém, na maioria dos casos, resolver a equação (1.13) para verificar

a estabilidade do sistema é um procedimento caro, pois, de acordo com o Teorema 1.5, a matriz E precisa ter posto cheio para que essa verificação seja possível. Sendo assim, optamos por não nos aprofundar em métodos aplicados em tais situações. Ao invés disso, trabalharemos com as equações de Lyapunov cujas soluções são as matrizes de Gram definidas na subseção a seguir. Essas matrizes, por sua vez, são úteis para redução de modelo apresentada no próximo capítulo.

1.2 Matrizes de Gram e Estabilidade via Inércia

Nesta seção, trataremos de duas importantes matrizes relacionadas aos conceitos de observabilidade e controlabilidade de sistemas lineares dinâmicos. Essas matrizes tem a forma de matrizes de Gram [17] e, por isso, levam esse nome. Os conceitos abordados nesta seção podem ser vistos com mais detalhes em [1] e [22].

1.2.1 Acessibilidade e Controlabilidade

Definição 1.6 *Dado um sistema (A, B, C, D), um estado $\overline{x}(t) \in X$, tal que $\overline{x}(t_0) = 0$, é dito acessível a partir do estado zero se existe uma função de entrada $\overline{u}(t)$ tal que $\|u\|_{L^2(0,\infty)} < \infty$, e um tempo $\overline{t_0} < \infty$, tais que*

$$\overline{x}(t) = \phi(0, \overline{t_0}, \overline{u}(\tau), t), \quad \forall t > t_0.$$

O subespaço acessível $X^{acess} \subset X$ de (A, B, C, D) é o conjunto que contém todos os estados acessíveis do sistema. A quádrupla (A, B, C, D) é dita completamente acessível se $X^{acess} = X$.

A fim de melhorar o entendimento sobre a última definição, lembremos que, pela expressão (1.5), um estado $x(t)$, definido a partir de uma função de entrada $u(t)$, no sistema (A, B, C, D) pode ser representado por

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\ d\tau, \quad t \geq t_0.$$

Na Definição 1.6, temos $\overline{x}(t_0) = 0$. Então, para que o estado $\overline{x}(t)$ seja acessível, precisamos garantir a existência de uma função $\overline{u}(t)$ tal que

$$\overline{x}(t) = \int_{t_0}^{t} e^{A(t-\tau)}B\overline{u}(\tau)\ d\tau, \quad t \geq t_0,$$

Utilizando o Teorema de Cayley-Hamilton [3, p. 79], podemos escrever

$$e^{A(t-\tau)} = \sum_{i=0}^{n-1} \alpha_i(t, \tau)A^i,$$

com funções escalares $\alpha_i(t, \tau)$, $i = 1, 2, ..., n-1$, que dependem do tempo t e da variável de integração τ. Desta forma temos

$$\overline{x}(t) = \sum_{i=0}^{n-1} A^i B \int_{t_0}^{t} \alpha_i(t, \tau)\overline{u}(\tau)\ d\tau, \quad t \geq t_0.$$

ou seja,

$$\overline{x}(t) = \left[\begin{array}{cccc} B & AB & ... & A^{n-1}B \end{array} \right] \left[\begin{array}{c} \int_{t_0}^{t} \alpha_0(t,\tau)\overline{u}(\tau)\ d\tau \\ \int_{t_0}^{t} \alpha_2(t,\tau)\overline{u}(\tau)\ d\tau \\ \vdots \\ \int_{t_0}^{t} \alpha_{n-1}(t,\tau)\overline{u}(\tau)\ d\tau \end{array} \right].$$

A matriz

$$\mathcal{R}_n(A,B) = [B\ AB\ A^2B\ ...\ A^{n-1}B]. \tag{1.14}$$

é chamada de *matriz de acessibilidade* de (A,B,C,D). Com a discussão feita até o momento, foi provado o seguinte resultado:

Teorema 1.7 *Seja (A,B,C,D) um sistema dinâmico. Então, tanto no caso de tempo contínuo como no caso de tempo discreto, X^{acess} é um subespaço linear de X, definido por*

$$X^{acess} = Im(\mathcal{R}_n(A,B)).$$

Corolário 1.8 *Um sistema (A,B,C,D) é completamente acessível se, e somente se, rank$(\mathcal{R}_n(A,B)) = n$.*

O conceito de controlabilidade de um sistema definido a seguir está intimamente ligado à acessibilidade do sistema, conforme definido a seguir.

Definição 1.9 *Dado um tempo inicial t_0, um estado $\overline{x}(t) \in X$, o sistema (A,B,C,D) é dito controlável para o estado zero se existe uma função de entrada $\overline{u}(t)$ e um tempo $\overline{t}$ tais que*

$$\phi(\overline{x}(t_0), t_0, \overline{u}(\tau), \overline{t}) = 0.$$

O subespaço controlável $X^{contr} \in X$ de (A,B,C,D) é o conjunto de todos os estados controláveis. Um sistema (A,B,C,D) é dito controlável quando $X^{contr} = X$

O Teorema 1.10 estabelece uma equivalência entre *acessibilidade* e *controlabilidade* de sistemas para tempo contínuo.

Teorema 1.10 *Para um sistema (A,B,C,D) tem-se que*

$$X^{contr} = X^{acess}.$$

Demonstração. Por definição, $x \in X^{contr}$ significa que, dado um instante inicial t_0, existe uma função de entrada u e um tempo t tais que $\phi(x(t_0), t_0, u(\tau), t) = 0$. Disto e de (1.5) segue que

$$e^{At}x = -\int_{t_0}^{\tau} e^{A(t-\tau)}Bu(\tau)\ d\tau = -\phi(0, t_0, u(\tau), t). \tag{1.15}$$

Sendo assim $e^{At}x \in\ X^{acess}$ e portanto

$$x \in e^{-At}X^{acess}. \tag{1.16}$$

Pelo Teorema 1.7 tem-se que

$$Ax \in A \cdot Im(\mathcal{R}(A,B)) = Im(A \cdot \mathcal{R}(A,B)) = Im([AB\ A^2B\ ...\ A^nB\ ...]) \subset$$
$$Im([B\ AB\ ...\ A^{n-1}B\ ...]) = X^{acess}.$$

Desta forma, X^{acess} é invariante por A. Utilizando a expansão em série de potências de e^{-At} pode-se perceber que X^{acess} é também invariante por e^{-At}, logo, por (1.16), $X^{contr} \subset X^{acess}$. Por outro lado, supondo que $x \subset X^{acess}$, verifica-se de forma análoga à feita anteriormente que $-e^{At}x(t) \in X^{acess}$. Isto significa que existe um instante t_0 e uma função de entrada u tais que $-e^{At}x(t) = \phi(0, t_0, u(\tau), t) = \int_{t_0}^{\tau} e^{A(t-\tau)}Bu(\tau)\ d\tau$. Segue disto que $\phi(x, t_0, u(\tau), t) = 0$, mostrando assim que $x \in X^{contr}$. Isso nos permite concluir que $X^{acess} \subset X^{contr}$. □

A seguir, é definido um conceito que será bastante utilizado no decorrer deste trabalho.

Definição 1.11 *A matriz de Gram de acessibilidade (ou de controlabilidade) de um sistema (A, B, C, D), para um tempo $t < \infty$, é definida como*

$$P(t) = \int_0^t e^{A\tau}BB^T e^{A^T\tau}\ d\tau, \quad t \in \mathbb{R}_+. \tag{1.17}$$

De acordo com o Teorema 1.10, a matriz de Gram de *acessibilidade* pode também ser chamada de matriz de Gram de *controlabilidade*. Por questões de objetividade, as definições e teoremas que seguem mencionarão apenas o termo controlabilidade.

Proposição 1.12 *A matriz de Gram de controlabilidade possui as seguintes propriedades:*

(i) $P(t) = P^T(t) \geq 0$.

(ii) Suas colunas geram o subespaço de controlabilidade, ou seja,

$$ImP(t) = Im\mathcal{R}(A, B) \quad \forall t > 0.$$

Demonstração. (i) O fato da matriz ser simétrica e positiva semi-definida decorre diretamente da definição. (ii) A fim de provar a segunda afirmação, vamos verificar que um determinado vetor v é ortogonal a $P(t)$ se, e somente se, é ortogonal também a $\mathcal{R}(A, B)$. Uma vez que $P(t)$ é simétrica e positiva semi-definida, v é ortogonal às colunas de $P(t)$ se, e somente se, $v^H P(t)v = 0$. Além disso,

$$0 = v^* P(t)v = \int_0^t v^* e^{A\tau}BB^T e^{A^T\tau}v\ d\tau = \int_0^t \|B^T e^{A^T\tau}v\|\ d\tau.$$

Isso é equivalente a dizer que $B^T e^{A^T\tau}v = 0$. Como a função exponencial é analítica e positiva em $t = 0$, então a função $B^T e^{A^T\tau}v$ é analítica e, além disso, ela e suas derivadas são zero em $t = 0$. Portanto, $B^T(A^T)^{k-1}v = 0$ para todo k natural. Pela expressão (1.14) isto é equivalente a dizer que $v \perp \mathcal{R}(A, B)$.

□

Corolário 1.13 *Um sistema (A, B, C, D) é controlável se, e somente se, $P(t)$ é positiva definida para algum $t > 0$.*

Demonstração. A demonstração segue diretamente da Proposição 1.12. □

Corolário 1.14 *Se o sistema (A, B, C, D) é controlável, então não existe algum autovetor v à esquerda de A que pertença ao núcleo à esquerda de B, ou seja,*

$$v^T A = \lambda v^T \Rightarrow v^T B \neq 0.$$

Demonstração. Se existisse um autovetor v tal que $v^T A = \lambda v^T$ e $v^T B = 0$, claramente teríamos $v^T\mathcal{R}(A, B) = 0$. Isso implicaria que o sistema não é controlável. □

1.2.2 Observabilidade

Uma pergunta natural a se fazer é a seguinte: dada uma variável de saída $y(\tau)$ gerada por um sistema (A, B, C, D), com τ pertencente a algum intervalo apropriado $[t_0, t]$, seria possível, a partir disto, reconstruir o estado inicial $x(t_0)$? Essa pergunta costuma receber o nome de *Problema da Observabilidade* [1]. Sem perda de generalidade, assumamos de agora em diante $t_0 = 0$.

Definição 1.15 *Um estado $\overline{x} \in X$ de um sistema (A, B, C, D) é dito inobservável se*

$$C\phi(\overline{x}(0), \overline{t_0}, 0, t) = 0 \quad \forall t \geq 0.$$

O conjunto de todos os estados inobserváveis de (A, B, C, D) é chamado de subespaço inobservável de X denotado por X^{inobs}. Um Sistema é dito completamente observável se $X^{inobs} = \{0\}$.

Na Definição 1.15 a função de entrada é considerada como sendo nula, sendo assim, no sistema (1.3), o estado $\overline{x}$ é representado por

$$\overline{x}(t) = e^{A(t-t_0)}\overline{x}_{t_0}, \quad \forall t > 0.$$

Portanto, a variável de saída do sistema será a função

$$y(t) = C\phi(\overline{x}(0), \overline{t_0}, 0, t) = Ce^{A(t-t_0)}\overline{x}_{t_0}, \quad \forall t > 0,$$

que, claramente, é analítica. Pensando em escrever essa função em termos de combinações de suas derivadas no ponto t_0, faz sentido notar que

$$\begin{aligned}
&y(t_0) = C\overline{x}(t_0)\\
&\dot{y}(t_0) = C\dot{\overline{x}}(t_0) = CA\overline{x}(t_0)\\
&\ddot{y}(t_0) = C\ddot{\overline{x}}(t_0) = CA^2\overline{x}(t_0)\\
&\vdots\\
&y^{(n-1)}(t_0) = C\overline{x}^{(n-1)}(t_0) = CA^{(n-1)}\overline{x}(t_0)\\
&\vdots
\end{aligned}$$

Pelo Teorema de Cayley-Hamilton [3, p. 79], a função y pode ser inteiramente determinada num ponto t_0 fazendo uso de derivadas até a ordem $n-1$. Sendo assim, a última expressão pode ser representada por

$$\begin{bmatrix} y(t_0) \\ \dot{y}(t_0) \\ \ddot{y}(t_0) \\ \vdots \\ y^{(n-1)}(t_0) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{(n-1)} \end{bmatrix} \overline{x}(t_0).$$

A matriz

$$\mathcal{O}_n(C, A) := (C^T \ A^TC^T \ ... \ (A^T)^{n-1}C^T)^T \tag{1.18}$$

é chamada de matriz de *observabilidade* do sistema (A, B, C, D). O Teorema a seguir resume as discussões feitas acerca do conceito de observabilidade até aqui.

Teorema 1.16 *Dado um sistema (A, B, C, D), tanto para $t \in \mathbb{Z}$ como para $t \in \mathbb{R}$, X^{inobs} é um subespaço linear de X dado por*

$$X^{inobs} = ker\mathcal{O}(C, A) = \{x \in X : CA^{i-1}x = 0, i > 0\} \tag{1.19}$$

Corolário 1.17 *Um sistema (A, B, C, D) é observável se, e somente se,*

$$rank(\mathcal{O}_n(C, A)) = n$$

Definição 1.18 *Para um sistema (A, B, C, D), definimos a matriz de Gram de observabilidade para um tempo $t < \infty$ como sendo*

$$Q(t) = \int_0^t e^{A^T\tau} C^T C e^{A\tau}\ d\tau \quad t \in \mathbb{R}_+. \tag{1.20}$$

A seguir é apresentado um resultado que relaciona os conceitos de controlabilidade e observabilidade. Para tal resultado, é preciso o conhecimento da definição de dualidade de um sistema (A, B, C, D).

Definição 1.19 *O sistema dual $(A, B, C, D)^*$ de (A, B, C, D) é definido por*

$$\Sigma \equiv \begin{cases} \dot{x}(t) = -A^T x(t) - C^T u(t), & t > t_0, \quad x(t_0) = x_0, \\ y(t) = B^T x(t) + D^T u(t), & t \geq t_0, \end{cases}, \tag{1.21}$$

Teorema 1.20 *Dado um sistema (A, B, C, D) tem-se que*

$$\left(X^{contr}_{(A,B,C,D)}\right)^{\perp} = X^{inobs}_{(A,B,C,D)^*},$$

em que $\left(X^{contr}_{(A,B,C,D)}\right)^{\perp}$ é o complemento ortogonal do subespaço controlável de (A, B, C, D) e $X^{inobs}_{(A,B,C,D)^*}$ é o subespaço inobservável do sistema dual $(A, B, C, D)^*$

Demonstração. A prova segue imediatamente das expressões (1.14) e (1.18) e do fato que, dada uma matriz M, $Im(M)^{\perp} = ker(M^T)$. □

Corolário 1.21 *Um sistema (A, B, C, D) é controlável se, e somente se, seu dual $(A, B, C, D)^*$ é observável.*

Este último Corolário garante a veracidade do teorema a seguir, por meio da dualidade e a utilização dos Corolários 1.8 e 1.13.

Teorema 1.22 (Condições de Observabilidade) *As seguintes afirmações são equivalentes:*

(i) O sistema (A, B, C, D) é observável.

(ii) A matriz de observabilidade tem posto completo.

(iii) A matriz de Gram de observabilidade é positiva definida, isto é, $Q(t) > 0$ para todo $t > 0$.

1.2.3 Equações de Lyapunov e Matrizes de Gram

No caso em que $t = \infty$ pode-se redefinir as matrizes de Gram (1.17) e (1.20), respectivamente, da seguinte maneira:

$$P = \int_0^\infty e^{A\tau} B B^T e^{A^T\tau}\, d\tau, \tag{1.22}$$

$$Q = \int_0^\infty e^{A^T\tau} C^T C e^{A\tau}\, d\tau. \tag{1.23}$$

As matrizes de Gram infinitas P e Q são também soluções de duas equações de Lyapunov específicas, relacionadas ao sistema dinâmico (A, B, C), conforme vemos no teorema a seguir.

Teorema 1.23 *Dado um sistema (A, B, C, D) contínuo no tempo e estável, a matriz de Gram de controlabilidade infinita P associada satisfaz a equação de Lyapunov de tempo contínuo*

$$AP + PA^T = -BB^T, \tag{1.24}$$

enquanto a matriz de Gram de observabilidade infinita Q satisfaz

$$A^T Q + QA = -C^T C. \tag{1.25}$$

Demonstração. Para provar (1.24) perceba que

$$\frac{d}{d\tau}\left(e^{A\tau} B B^T e^{A^T\tau}\right) = A e^{A\tau} B B^T e^{A^T\tau} + e^{A\tau} B B^T e^{A^T\tau} A^T.$$

Assim, por (1.22),

$$\begin{aligned} AP + PA^T &= \int_0^\infty A e^{A\tau} B B^T e^{A^T\tau} + e^{A\tau} B B^T e^{A^T\tau} A^T\, d\tau \\ &= \int_0^\infty \frac{d}{d\tau}\left(e^{A\tau} B B^T e^{A^T\tau}\right)\, d\tau = -BB^T, \end{aligned}$$

pois, como o sistema é estável, então $\lim_{t\longrightarrow\infty}\left(e^{A\tau}\right) = \lim_{t\longrightarrow\infty}\left(e^{A^T\tau}\right) = 0$. Isto prova (1.24). A prova para (1.25) é inteiramente análoga. □

No que se refere à equação de Lyapunov em sistemas dinâmicos, o foco desse nosso trabalho é resolver as equações (1.24) e (1.25), para obter a matriz de controlabilidade (1.22) e a matriz de observabilidade (1.23), respectivamente. Ao tratar de métodos para tais equações, faremos referência apenas à equação (1.24), já que a equação (1.25) é dual da primeira. Vamos estudar agora as condições para que tais equações admitam solução. Para isso lançaremos mão do conceito de produto de Kronecker que permite representar a equação de Lyapunov de maneira mais simples.

Definição 1.24 *O produto de Kronecker entre duas matrizes $A = (a_{ij}) \in \mathbb{R}^{p\times q}$ e $B \in \mathbb{R}^{r\times s}$ dadas, é definido por*

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1q}B \\ \vdots & \ddots & \vdots \\ a_{p1}B & \cdots & a_{pq}B \end{pmatrix}$$

Dado um número real α e matrizes A, B, C e D, o *produto de Kronecker* goza das seguintes propriedades [16]:

$$\begin{aligned}
&(a)\ A \otimes (\alpha B) = \alpha(A \otimes B) = (\alpha A) \otimes B \\
&(b)\ (A + B) \otimes C = A \otimes C + B \otimes C \\
&(c)\ A \otimes (B + C) = A \otimes B + A \otimes C \\
&(d)\ (A \otimes B)(C \otimes D) = AC \otimes BD
\end{aligned} \tag{1.26}$$

Lema 1.25 *Se λ_i e λ_j são autovalores das matrizes $A \in \mathbb{R}^{n\times n}$ e $B \in \mathbb{R}^{m\times m}$ respectivamente, então $\lambda_i + \lambda_j$ são autovalores de $(I_n \otimes A) + (B^T \otimes I_m)$.*

Demonstração. Sejam z_i e w_j autovetores de A e B^T associados aos autovalores λ_i e μ_j, respectivamente. Utilizando as propriedades vistas em (1.26), verificamos que

$$\begin{aligned}
\Big[(I_n \otimes A) + (B^T \otimes I_m)\Big](w_j \otimes z_i) &= (I_n \otimes A)(w_j \otimes z_i) + (B^T \otimes I_m)(w_j \otimes z_i) \\
&= (I_n w_j \otimes A z_i) + (B^T w_j \otimes I_m z_i) \\
&= (w_j \otimes \lambda_i z_i) + (\mu_j w_j \otimes z_i) \\
&= (\lambda_i + \mu_j)(w_j \otimes z_i),
\end{aligned}$$

mostrando assim que $\lambda_i + \lambda_j$ é autovalor de $(I_n \otimes A) + (B^T \otimes I_m)$ com autovetor associado $(w_j \otimes z_i)$. □

Teorema 1.26 *Dada uma matriz $E \neq 0$, a equação $AP + PA^T = E$ admite única solução se, e somente se, $\lambda_i + \overline{\lambda}_j \neq 0$ para quaisquer autovalores λ_i e λ_j da matriz A.*

Demonstração. A equação $AP + PA^T = E$, com $E \in \mathbb{R}^{n\times n}$ é equivalente ao sistema vetorizado

$$\tilde{A}\tilde{P} = \tilde{E}, \tag{1.27}$$

com $\tilde{P}$ e $\tilde{E}$ sendo as vetorizações de P e E, respectivamente, e

$$\tilde{A} = I \otimes A + A \otimes I.$$

O sistema (1.27) admite única solução se, e somente se, a matriz $\tilde{A}$ não possui autovalores nulos. De acordo com o Lema 1.25, os autovalores de $\tilde{A}$ são da forma $\lambda_i + \lambda_j$, sendo λ_i e λ_j autovalores de A. Nessas condições, o sistema (1.27) admite única solução se, e somente se, $\lambda_i + \lambda_j \neq 0$. □

O Corolário 1.27, a seguir, é uma consequência direta do Teorema 1.26.

Corolário 1.27 *Se $Re(\lambda) < 0$ para todo autovalor λ de A, então a equação (1.24) (bem como a equação* (1.27)*) admite única solução.*

É possível também verificar a estabilidade do sistema utilizando uma equação de Lyapunov via conceito de inércia, conforme descrevemos na subseção a seguir. No entanto, acrescentamos esse trecho apenas como curiosidade. Daqui em diante vamos assumir sempre que todos os autovalores da matriz A do sistema (A, B, C, D) possuem parte real negativa, ou seja, o sistema é assintoticamente estável. Isso ficará subentendido quando dissermos simplesmente que o sistema dinâmico é estável, ou então que a matriz A é estável. Com essas assunções, além de garantir a existência e unicidade da solução P de (1.24), podemos afirmar que P é simétrica semi-definida positiva, pois possui expressão analítica dada em (1.22). Essas são informações importantes para a teoria de redução por balanceamento, que será apresentada no capítulo 2, bem como para os métodos de resolução da equações de Lyapunov apresentados no capítulo 3.

1.2.4 Estabilidade via Inércia

Nesta seção veremos como verificar algumas propriedades de um sistema dinâmico (A, B, C, D) a partir de informações extraídas da equação

$$AP + PA^T = E$$

para os casos em que a matriz E é semi-definida, isto é, quando $E \leq 0$ ou $E \geq 0$. Tais propriedades decorrem da relação entre autovalores da matriz A com os autovalores da solução da equação. Para isso, se faz necessária a definição seguinte.

Definição 1.28 *Dada uma matriz quadrada $A \in \mathbb{R}^{n\times n}$, definimos a inércia de A como a tripla*

$$in(A) = (\nu(A), \delta(A), \pi(A)),$$

em que $\nu(A)$, $\delta(A)$ e $\pi(A)$ representam o número de autovalores no semiplano esquerdo, no eixo imaginário e no semiplano direito do plano complexo, respectivamente.

Matrizes cuja inércia é $(n, 0, 0)$ são chamadas de *matrizes estáveis*. Pois, uma vez que todos os autovalores de A tem parte real negativa, o sistema (1.12) definido a partir de A será assintoticamente estável. De maneira análoga, se a inércia for $(0, 0, n)$, então A é chamada de matriz *anti-estável*. A sequência de resultados enunciados a seguir estabelece relações entre a inércia de A e a inércia de P na equação de Lyapunov (1.13).

Lema 1.29 *Dada uma matriz $E < 0$, a equação de Lyapunov (1.13) admite uma única solução $P > 0$ se, e somente se, A é uma matriz estável, isto é, $in(A) = (n, 0, 0)$.*

De maneira análoga, se $E > 0$, a equação de Lyapunov (1.13) admite uma única solução $P > 0$ se, e somente se, A é uma matriz anti-estável, isto é, $in(A) = (0, 0, n)$

Demonstração. Para ver a demonstração desse teorema, consulte o Apêndice A.

Teorema 1.30 *Sejam A, P e E satisfazendo (1.13) com $E > 0$. Então $in(A) = in(P)$.*

Demonstração. Para ver a demonstração desse teorema, consulte o Apêndice A.

Uma versão análoga ao Teorema 1.30 pode ser enunciada para o caso em que $E > 0$. Nessas condições, verifica-se também que $in(A) = (n, 00)$ e $in(P) = (0, 0, n)$.

Por fim, o Teorema a seguir aborda o caso em que a matriz E é semidefinida e, para tal, invoca o conceito de controlabilidade definido a pouco.

Teorema 1.31 *Sejam A, P e E satisfazendo (1.13) com $E \leq 0$ (ou $E \geq 0$). Se o par (A, E) é controlável, então*

$$in(A) = in(P).$$

Demonstração. Para ver a demonstração desse teorema, consulte o Apêndice A. Em resumo, dado que existe uma solução $P > 0$ para

$$AP + PA^T = -BB^T,$$

uma vez que $BB^T > 0$, o Teorema 1.4 garante a estabilidade assintótica do sistema. No entanto, se o sistema for muito grande, o custo computacional para encontrar a matriz P pode ser alto nessas condições. Se BB^T tem posto relativamente pequeno, os cálculos podem ser reduzidos. Porém, nesse caso, precisamos da garantia que o sistema seja controlável (conforme Teorema 1.31). Do contrário, o fato de termos $P > 0$ garante apenas que o sistema é estável (conforme Teorema 1.4).

1.3 Sistemas Descritores

Em muitas aplicações, como ocorre, por exemplo, em sistemas elétricos de potência, os modelos de sistemas dinâmicos lineares e invariantes no tempo são construídos utilizando-se uma matriz Jacobiana esparsa, de ordem N e separada em blocos

$$J = \begin{bmatrix} J_1 & J_2 \\ J_3 & J_4 \end{bmatrix}, \tag{1.28}$$

com $J_4 \in \mathbb{R}^{N-n \times N-n}$ e $J_1 \in \mathbb{R}^{(n)\times(n)}$. A matriz J é calculada a fim de gerar uma aproximação linear local do sistema original, avaliada sobre um ponto de equilíbrio do sistema. Essa aproximação linear para o sistema dinâmico toma a seguinte forma [24]:

$$\begin{cases} \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{w}(t) \end{bmatrix} = \begin{bmatrix} J_1 & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} x(t) \\ w(t) \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t) \\ y(t) = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x(t) \\ w(t) \end{bmatrix} + D_J u(t), \end{cases} \tag{1.29}$$

sendo B_1, B_2, C_1, C_2 e D_J matrizes esparsas dadas e $w(t) \in \mathbb{R}^{N-n}$ um vetor de variáveis algébricas. O sistema (1.29) pode ainda ser representado de maneira mais geral por

$$\begin{cases} E\dot{x_J}(t) = Jx_J(t) + B_J u(t) \\ y(t) = C_J x_J(t) + D_J u(t) \end{cases} \tag{1.30}$$

com $x_J(t) = [x(t)\ w(t)]^T \in \mathbb{R}^N$ $B_J = [B_1^T\ B_2^T]^T$ e $C_J = [C_1^T\ C_2^T]^T$ e

$$E = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}.$$

As variáveis $x_J(t)$ e $w(t)$ são chamadas de *estado* e *variável algébrica*, respectivamente. Se $x(t) \in \mathbb{R}^n$ dizemos que o sistema dinâmico possui n estados e $N - n$ variáveis algébricas.

Para este trabalho vamos considerar que J_4 é sempre não singular. Para que o sistema (1.29) seja escrito nos moldes do sistema (1.3) é preciso relacionar as matrizes da seguinte forma:

$$\begin{aligned} A &= J_1 - J_2 J_4^{-1} J_3 \\ B &= B_1 - J_2 J_4^{-1} B_2 \\ C &= C_1 - C_2 J_4^{-1} J_3 \\ D &= D_J - C_2 J_4^{-1} B_2. \end{aligned} \tag{1.31}$$

Deste modo, podemos relacionar a equação de Lyapunov

$$AP + PA^T = -BB^T$$

ao sistema (1.29) de maneira análoga ao que foi feito com o sistema (1.1), desde que as relações (1.31) sejam satisfeitas.

Veremos posteriormente que, em muitos casos, não é conveniente obter explicitamente a matriz A do sistema dinâmico, pois as operações explícitas envolvendo a matriz A em (1.31) demandam um custo computacional elevado em sistemas de grande porte. Por isso vamos estudar agora maneiras práticas de realizar algumas operações matriciais nesses casos.

Para avaliar o produto da matriz $(A + \mu I)^{-1}$ pela matriz B, é apresentada uma técnica em

[11] que consiste em resolver o sistema

$$\begin{bmatrix} J_1 + \mu I & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} \Gamma \\ \Upsilon \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \tag{1.32}$$

pois, pela expressão (1.32), podemos aferir que

$$\begin{aligned} &(J_1 + \mu I)\Gamma + J_2\Upsilon = B_1, \quad \text{e} \\ &J_3\Gamma + J_4\Upsilon = B_2, \end{aligned} \tag{1.33}$$

então, $\Upsilon = J_4^{-1}B_2 - J_4^{-1}J_3\Gamma$. Substituindo Υ em (1.33), temos

$$(J_1 + \mu I)\Gamma + J_2(J_4^{-1}B_2 - J_4^{-1}J_3\Gamma) = B_1.$$

Disto segue que

$$\Gamma = (J_1 - J_2J_4^{-1}J_3 + \mu I)^{-1}(B_1 - J_2J_4^{-1}B_2),$$

ou seja, pelas relações (1.31),

$$\Gamma = (A + \mu I)^{-1}B.$$

Vamos considerar agora uma outra matriz S, sem ser necessariamente B, mas de mesma ordem. Então, o produto da matriz $(A + \mu I)^{-1}$ pela matriz S pode ser obtido resolvendo-se o sistema

$$\begin{bmatrix} J_1 + \mu I & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} \Gamma \\ \Upsilon \end{bmatrix} = \begin{bmatrix} S \\ 0 \end{bmatrix}. \tag{1.34}$$

Com uma verificação análoga à descrita no caso anterior, conclui-se que, para o sistema (1.34), é válido que

$$\Gamma = (A + \mu I)^{-1}S.$$

Por fim, vamos considerar a situação em que seja preciso efetuar o produto de $(A + \mu I)$ por uma matriz S, de ordem compatível, que não é necessariamente B. Como $A = J_1 - J_2J_4^{-1}J_3$, o processo pode ser dividido nas seguintes etapas:

$$\begin{aligned} &(1)\ \textit{calcula-se os produto } J_3S \text{ e } J_1S; \\ &(2)\ \textit{resolve-se o sistema linear } J_4X = J_3S; \\ &(3)\ \textit{por fim, calcula-se } \Gamma = J_1S - J_2X + S. \end{aligned} \tag{1.35}$$

A matriz Γ do passo (3) satisfaz

$$\Gamma = (A + \mu I)S.$$

1.3.1 Principais Modelos Estudados Nesse Trabalho

Nesta subseção destacamos os principais modelos para sistemas descritores que utilizamos para os testes numéricos.

O primeiro é o `brasilsemtcsc`, um modelo do tipo SISO, oriundo do Sistema Elétrico Interligado do Brasil e que já foi considerado em outros trabalhos como [25] e [33]. Nesse modelo, a matriz jacobiana J é uma matriz esparsa de ordem $N = 13251$, com densidade de $0,028\%$. A matriz E é diagonal dada por $E = diag(1, 1, ..., 1, 0, ..., 0)$ e possui 1664 elementos não nulos. Portanto, a matriz $A = J_1 - J_2J_4^{-1}J_3$ é de ordem $n = 1664$. O vetor B possui apenas 4 elementos

não nulos: $B(524) = B(1442) = 1$ e $B(1884) = B(1918) = -1$. O vetor C, também tem 4 elementos não nulos dadas por $C(11558) = 26,5721$, $C(11559) = -13,1127$, $C(12502) = -29,2954$, e $C(12503) = 3,7609$. A matriz D, nesse caso, é nula. A esparsidade da matriz J é ilustrada na Figura 3.11.

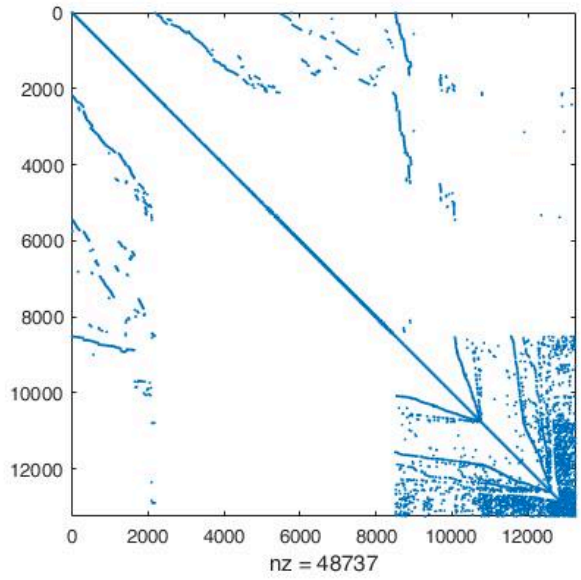


Figura 1.3: Elementos não nulos da matriz jacobiana J.

Para fins didáticos, utilizamos o Matlab para calcular a matriz A explicitamente, bem como todos os seus autovalores (algo que não é recomendado para sistemas de grande porte). Isso permitiu ter conhecimento da distribuição do espectro de A que é exibida na Figura 1.4.

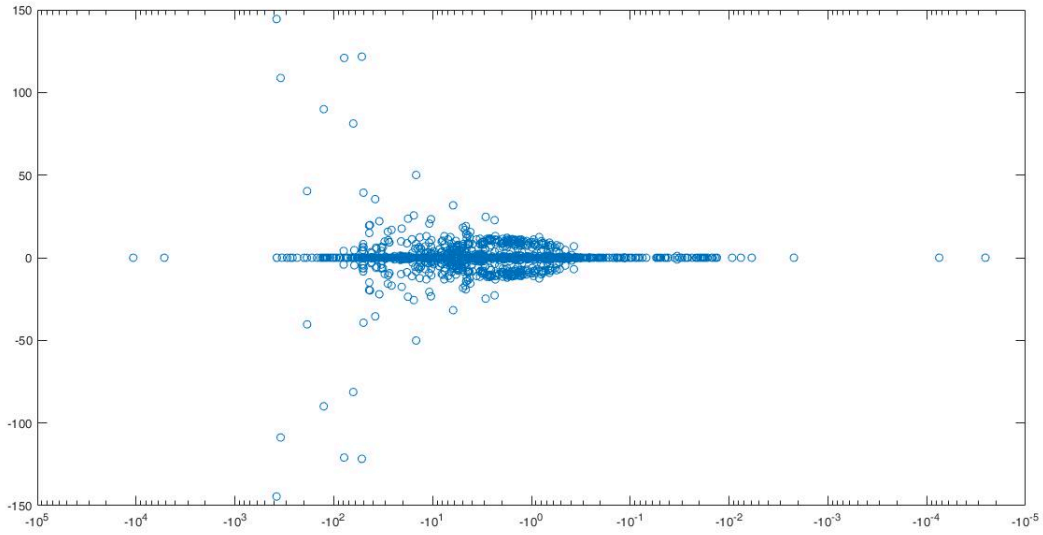


Figura 1.4: Autovalores da matriz A obtida a partir do sistema descritor.

Consideramos também o modelo `ww_vref_6405`, disponível em *https://sites.google.com/site/rommes/so*

que utiliza a mesma matriz A do sistema `brasilsemtcsc`, diferindo apenas na formação dos vetores B e C, cujas entradas são $B(2037) = C(2028) = 1$ e as restantes todas nulas.

Capítulo 2

Função de Transferência e Redução de Modelo

2.1 Função de Transferência e Operador de Hankel

Esta seção destina-se a um estudo mais aprofundado da relação entre entrada e saída de um sistema dinâmico estável (A, B, C, D). Esse estudo é feito, basicamente, sobre uma função de transferência associada ao sistema. A definição dessa função faz uso da Transformada de Laplace, que é o primeiro conceito que vamos definir aqui.

Definição 2.1 *[15] [29] Dada uma função $f : \mathbb{R}_+ \longrightarrow \mathbb{R}^{p\times q}$, a transformada de Laplace $Z = \mathcal{L}(f)$, com $Z : \mathbb{D} \longrightarrow \mathbb{C}^{p\times q}$, é definida por*

$$Z(s) = \int_0^\infty e^{-st} f(t)\, dt.$$

O conjunto $\mathbb{D}$ é o domínio, ou seja, uma região de convergência de Z.

Pelo Teorema de Lerch [4], se uma função F é tal que $F(s) = \int_0^\infty e^{-st} f(t)\, dt$, para alguma função contínua f, então F é unicamente determinada por f. Mais do que isso, o Teorema de Paley-Winer [29] garante que a transformada de Laplace $\mathcal{L}$ leva, unitariamente, funções de $L^2(\mathbb{R}_+)$ em funções pertencentes ao espaço de Hardy $H^2(\mathbb{C}_+)$, constituído de funções f analíticas sobre o conjunto $\{s \in \mathbb{C}_+ \ : \ |s| < 1\}$ (consulte [29] para mais detalhes) tais que

$$\sup_{0<r<1} \left(\int_{\{s\in\mathbb{C}_+ : |s|=1\}} |f(rs)|^2 d\mathbf{m}(s) \right)^{1/2} < \infty,$$

com $\mathbf{m}$ representando a medida de Lebesgue normalizada em $\{s \in \mathbb{C}_+ : |s| = 1\}$.

Nessas condições, podemos definir a transformada de Laplace inversa $\mathcal{L}^{-1}$ sobre funções do espaço de Hardy $H^2(\mathbb{C}_+)$.

Vamos supor que, no sistema (A, B, C, D) de (1.3), exista uma espécie de aplicação G, à qual damos o nome de *planta*, que leve a variável de entrada u até a variável de saída y, ou seja, $y = Gu$. Consideremos o operador

$$\mathcal{G} = \mathcal{L}G\mathcal{L}^{-1}, \tag{2.1}$$

definido sobre o espaço de Hardy $H^2(\mathbb{C}_+)$. Dessa forma, pode-se verificar que [29]

$$\mathcal{G}(e^{-az}\varphi) = e^{-az}\mathcal{G}\varphi, \quad \forall \varphi \in H^2(\mathbb{C}^+), \quad a > 0.$$

Neste caso, a notação z está representando uma função identidade de $\mathbb{C}$ em $\mathbb{C}$. Disto segue que, se φ e ψ são combinações lineares de funções e^{-az}, $a > 0$, então

$$\psi(\mathcal{G}\varphi) = (\mathcal{G}\psi)\varphi.$$

Sendo assim, se φ é uma combinação linear das funções e^{-az}, $a > 0$, então

$$e^{-z}\mathcal{G}\varphi = (\mathcal{G}e^{-z})\varphi \Longrightarrow \mathcal{G}\varphi = \frac{(\mathcal{G}e^{-z})\varphi}{e^{-z}}.$$

Portanto

$$\mathcal{G}\varphi = H\varphi, \tag{2.2}$$

com

$$H(s) = \frac{(\mathcal{G}e^{-z})}{e^{-s}},$$

para s com módulo suficientemente grande.

A aplicação $H(s)$ é chamada de *função de transferência* do sistema (1.3). Uma vez que o espaço de combinações lineares de e^{-az}, $a > 0$, é denso em $H^2(\mathbb{C}^+)$ [29], então a expressão (2.2) pode ser estendida para uma função arbitrária de $H^2(\mathbb{C}^+)$.

Pelo que acabamos de ver, para calcular a função H, basta encontrar a transformada de Laplace avaliada na variável de saída do sistema:

$$\mathcal{L}(y(t))(s).$$

Supondo uma condição inicial $x_0 = 0$ para o sistema $\dot{x}(t) = Ax(t) + Bu(t)$ de (1.3) e assumindo que o sistema é estável, tem-se que

$$\begin{aligned} \mathcal{L}(\dot{x}(t))(s) &= \int_0^\infty e^{-st}\dot{x}(t)\,dt \\ &= e^{-st}x(t)|_{t=0}^{t\to\infty} + s\int_0^\infty e^{-st}x(t)\,dt \\ &= x_0 + s\mathcal{L}(x(t))(s) = s\mathcal{L}(x(t))(s). \end{aligned} \tag{2.3}$$

Além disso, como $\mathcal{L}$ é um operador linear,

$$\mathcal{L}(\dot{x}(t))(s) = A\mathcal{L}(x(t))(s) + B\mathcal{L}(u(t))(s). \tag{2.4}$$

Substituindo (2.4) em (2.3), temos

$$\mathcal{L}(x(t))(s) = (sI - A)^{-1}B\mathcal{L}(u(t))(s).$$

Como $y(t) = Cx(t) + Du(t)$ em (1.3), então

$$\mathcal{L}(y(t))(s) = \left[C(sI - A)^{-1}B + D\right]\mathcal{L}(u(t))(s).$$

Portanto

$$H(s) = C(sI - A)^{-1}B + D, \tag{2.5}$$

para s com módulo suficientemente grande.

Para o caso do sistema descritor definido na seção 1.3, um raciocínio muito parecido com o que

fizemos para chegar até a expressão (2.5) permite verificar que a função de transferência, nesse caso, é dada por

$$H(s) = C_J(sE - A_J)^{-1}B_J + D_J.$$

O sistema (A, B, C, D) é comumente chamado de uma *realização* da função de transferência $H(s)$.

A fim de definir uma norma para a função de transferência (2.5), consideramos um domínio de frequências formado por elementos do tipo $j\omega$, com ω sendo uma variável escrita em unidades radiadas e j representando a unidade imaginária. Assim,

$$\|H(j\omega)\|_\infty = \sup_\omega \|H(j\omega)\|. \tag{2.6}$$

Da teoria de Álgebra Linear sabemos que [13], na norma euclidiana, para um $j\omega$ fixado,

$$\|H(j\omega)\| = \max_{1\leq i\leq n}\{\sigma_i^2(H(j\omega))\} = \max_{1\leq i\leq n}\{\lambda_i(H^H(j\omega)H(j\omega))\},$$

em que $\sigma_i(M)$ indica o i-ésimo valor principal de M e $\lambda_i(M)$ única o i-ésimo autovalor de M. Note que a função (2.5) representa um sistema de equações diferenciais com um sistema puramente algébrico.

Seguindo o raciocínio apresentado em [29], vamos agora utilizar o conceito de função de transferência para construir uma representação direta para a planta G, que receberá o nome de *operador de Hankel*. Para isso, será preciso conhecer o conceito de convolução, que possui propriedades muito úteis envolvendo a transformada de Laplace.

Definição 2.2 *Dadas duas funções f e g, integráveis em $\mathbb{R}$, a convolução $f * g$ entre as funções f e g é definida por*

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)\,d\tau.$$

Teorema 2.3 (Teorema da convolução) *Se $\overline{f}$ é a transformada de Laplace de uma função f e $\overline{g}$ é a transformada de Laplace de uma função g, então*

$$\mathcal{L}^{-1}(\overline{f}\,\overline{g})(t) = (f * g)(t) = \int_0^t f(\tau)g(t-\tau)\,d\tau. \tag{2.7}$$

Demonstração. A prova desse resultado pode ser encontrada em [4]. □

O intervalo da integral da convolução em (2.7) está definido a partir de zero em virtude das funções do domínio da transformada de Laplace estarem definidas em $[0, \infty)$.

Vamos supor que a função de entrada $u(t)$ é suficientemente suave e seja tal que $supp(u) \subset (-\infty, 0)$. Assim, podemos considerar $D \equiv 0$ em (1.3). Nessas condições, a função de transferência torna-se

$$H(s) = C(sI - A)^{-1}B. \tag{2.8}$$

Vamos supor ainda que exista $M > 0$ tal que $\text{supp}(u) \subset (-M, 0)$ e consideremos as translações:

$$u_M(t) = u(t - M), \quad y_M(t) = y(t - M), \quad t \in \mathbb{R}.$$

De maneira análoga ao que foi feito anteriormente, pode-se verificar que

$$\mathcal{L}(y_M(t))(s) = \left[C(sI - A)^{-1}B\right]\mathcal{L}(u_M(t))(s)$$

Consideremos o operador h definido em $\mathbb{R}$ como sendo $h(t) = Ce^{tA}B$ e apliquemos, sobre essa função, a transformada de Laplace. Nas passagens a seguir supomos que a função exponencial é escrita na forma de série de potências $e^{At} = I + At + \frac{A^2t^2}{2!} + ... + \frac{A^nt^n}{n!} + ...$ Então

$$\begin{aligned}(\mathcal{L}h)(s) = \int_0^\infty e^{-st}Ce^{tA}B\,dt &= \int_0^\infty e^{-(sI-A)t}B\,dt \\ &= \lim_{T\longrightarrow\infty}\left[-(sI-A)Ce^{-(sI-A)t}B\right]_0^T \\ &= C(sI-A)^{-1}B,\end{aligned}$$

para $s \in \mathbb{C}$ tal que $|s|$ é suficientemente grande.

Portanto, utilizando o teorema 2.3, podemos concluir que y_M é a convolução entre as funções h e u_M, isto é,

$$y_M(t) = (h * u_M)(t) \int_0^\infty Ce^{A(t-\tau)}Bu_M(\tau)\,d\tau.$$

Por fim, nota-se que

$$y(t) = \int_0^\infty Ce^{A(t+\tau)}Bu(-\tau)\,d\tau = \int_0^\infty h(t+\tau)Bv(\tau)\,d\tau =: (\mathbf{H}v)(t), \quad \tau > 0, \tag{2.9}$$

com $v(t) := u(-t)$ para $t \in \mathbb{R}$.

O operador $\mathbf{H}$, definido em (2.9), é chamado de *operador de Hankel* [29] e sua utilidade nesse trabalho está concentrada em poder mensurar a ação do sistema (1.3), como pode ser visto na análise de valores principais resumida a seguir, mas que pode ser vista com mais detalhes em [12].

O operador transposto conjugado de $\mathbf{H}$ é dado por

$$(\mathbf{H}^*y)(t) = \int_0^\infty B^*e^{A^*(t+\tau)}Cy(\tau)\,dt.$$

A fim de encontrar os valores principais de $\mathbf{H}$, suponha que σ_i é um valor singular associado ao autovetor v de $\mathbf{H}^*\mathbf{H}$, isto é, $(\mathbf{H}^*\mathbf{H})v = \sigma_i^2 v$. Seja

$$\begin{aligned}y(t) := (\mathbf{H}v)(t) &= \int_0^\infty Ce^{A(t+\tau)}Bu(\tau)\,d\tau \\ &= Ce^{A(t)}\int_0^\infty e^{A(\tau)}Bu(\tau)\,d\tau \\ &= Ce^{A(t)}\overline{x},\end{aligned}$$

com

$$\overline{x} = \int_0^\infty e^{A(\tau)}Bu(\tau)\,d\tau. \tag{2.10}$$

Então

$$\begin{aligned}[(\mathbf{H}^*\mathbf{H})v](t) &= [\mathbf{H}^*y](t) \\ &= B^*e^{A^*t}\int_0^\infty e^{A^*\tau}C^*Ce^{A\tau}\overline{x}\,d\tau \\ &= B^*e^{A^*t}Q\overline{x} \\ &= \sigma^2 v(t),\end{aligned}$$

com Q definida em (1.23). Assim

$$v(t) = B^* e^{A^* t} Q \overline{x} {\sigma_i}^{-2} \tag{2.11}$$

e, substituindo (2.11) em (2.10), obtemos

$$PQ\overline{x} = \sigma_i^2 \overline{x}.$$

Portanto, se σ_i é valor principal do operador $\mathbf{H}$, então σ_i^2 é autovalor da matriz PQ, ou seja,

$$\sigma_i^2(\mathbf{H}) = \lambda_i(PQ). \tag{2.12}$$

Vamos agora relacionar os conceitos de *planta*, *função de transferência* e *operador de Hankel* do sistema (1.3). Pelo que acabamos de ver, dada uma variável de entrada u do sistema (1.3), $\mathbf{H}v = Gu$, com $v(t) = u(-t)$, para $t \in \mathbb{R}_-$, com G sendo a *planta* do sistema. Além disso, por (2.1) e (2.2), $H(s)u = (\mathcal{L}G\mathcal{L}^{-1})u$.

Tudo isso motiva a definição de uma norma induzida para o operador

$$\mathbf{H} : L^2(0,\infty) \longrightarrow L^2(0,\infty),$$

baseada na análise de valores próprios, e que serve também como uma norma para a função $H(s)$.

Definição 2.4 *[12] Seja $H(s) = C(sI - A)^{-1}B$ a função de transferência do sistema (1.3), com s suficientemente grande e $Re(\lambda_i(A)) < 0$ para $i = 1, 2, ..., n$. Então a norma Hankel de $H(s)$ é definida como*

$$\|H(s)\|_H := \tilde{\sigma}(\mathbf{H}) = \max_{j=1,...n} \{\lambda_j^{1/2}(PQ)\}, \tag{2.13}$$

com $\mathbf{H}$, P e Q definidos por (2.9), (1.22) e (1.23), respectivamente.

A Definição 2.13 está atrelada ao fato de que a variável de entrada satisfaz $u(t) = 0$ para $t \geq 0$. Isso faz com que a *norma Hankel* não dependa da matriz D do sistema (1.3). Nessas condições, é válido também que [12]

$$\|H(s)\|_H = \sup_{u \in L^2(-\infty,0)} \frac{\|y\|_{L^2(0,\infty)}}{\|u\|_{L^2(-\infty,0)}}.$$

Em geral, as normas definidas em (2.13) e (2.6) satisfazem a seguinte desigualdade [1]:

$$\|H(s)\|_H \leq \|H(j\omega)\|_\infty. \tag{2.14}$$

Para finalizar essa seção, vamos definir um sistema de ordem “mínima"e apresentar condições necessárias e suficientes para que a dimensão do sistema seja a menor possível.

Definição 2.5 *Uma realização (A, B, C, D) de uma função de transferência $H(s)$ é chamada de “realização mínima"ou “sistema de ordem mínima"de $H(s)$ se a matriz A possui a menor dimensão possível para manter a relação entrada-saída do sistema.*

Teorema 2.6 *Uma realização (A, B, C, D) é minimal se, e somente se, é controlável e observável.*

Demonstração. [51]

□

2.2 Método de redução modal

Vamos supor que a função de transferência $H(s)$ do sistema linear (A, B, C) é uma função racional cujos polos são todos simples. Nesse caso, podemos escrever $H(s)$ na forma de frações parciais, conforme expressão a seguir.

$$H(s) = \sum_{i=1}^{n} \frac{R_i}{s - \lambda_i}. \tag{2.15}$$

Os numeradores dos termos da soma (2.15) são constantes dadas por

$$R_i := \lim_{s \to \lambda_i} H(s)(s - \lambda_i),$$

que chamaremos de *resíduos* e os números $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ são os polos de $H(s)$. Perceba que os polos de $H(s)$ nada mais são do que singularidades de uma função pois $\lim_{s \to \lambda_i} H(s) = \infty$ para $i = 1, 2, ..., n$. Propositalmente, denotamos os polos da função $H(s)$ com a mesma simbologia dos autovalores da matriz A. A seguir, explicamos o motivo disso.

Vamos supor que o espectro de A é simples (autovalores todos distintos). Vamos considerar a decomposição espectral $A = V\Lambda W^*$, com $V = [v_1|v_2|...|v_n]$ sendo a matriz cujas colunas são autovetores à direita de A, $W = [w_1^*|w_2^*|...|w_n^*]^*$ sendo a matriz cujas linhas são os autovetores à esquerda de A e $\Lambda = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_n\}$. Nesse caso é possível verificar que

$$H(s) = \sum_{i=1}^{n} \frac{C^T v_i w_i^* B}{s - \lambda_i},$$

ou seja, os polos são justamente os autovalores de A e os resíduos são dados por

$$R_i = C^T v_i w_i^* B, \quad \text{para} \quad i = 1, 2, ..., n.$$

Na redução de modelo, o objetivo é diminuir a ordem das matrizes A, B e C de modo que a relação entre entrada e saída do sistema (A, B, C) seja preservada. Para mensurar a diferença entre o modelo original e o modelo reduzido, passamos agora a um breve estudo sobre algumas normas que podem ser aplicadas de maneira direta ou indireta para aferir a magnitude da variável de saída $y(t)$ do sistema (A, B, C). Vamos considerar primeiro a norma

$$\|y(t)\|_{L_2(0,\infty)} := \int_0^\infty |y(t)|^2 \, dt \quad \forall y(t) \in L_2(0, \infty).$$

O Teorema de Parseval nos fornece a seguinte identidade envolvendo o domínio das frequências:

$$\int_0^\infty |y(t)|^2 \, dt = \int_0^\infty |\mathcal{L}\{y\}(\omega j)|^2 \, d\omega, \tag{2.16}$$

para toda função $y(t)$ no domínio da transformada de Laplace $\mathcal{L}$, com $j = \sqrt{-1}$ e $\mathcal{L}$ repressentando a transformada de Laplace. Sendo assim, se u é a variável de entrada do sistema (A, B, C), temos

$$\begin{aligned}\int_0^\infty |y(t)|^2\, dt &= \int_0^\infty |C^T(\omega j I - A)^{-1}B|^2 |\mathcal{L}\{u\}(\omega j)|^2\, d\omega, \\ &\leq \int_0^\infty |H(\omega j)|^2\, d\omega \int_0^\infty |\mathcal{L}\{u\}(\omega j)|^2\, d\omega.\end{aligned}$$

Isso motiva o uso da norma em um domínio de frequência $\mathcal{L}_2(0,\infty)$, conhecida por representar a *energia do sistema*, [1] dada por

$$\|H(s)\|_{\mathcal{L}_2(0,\infty)} := \frac{1}{\pi}\int_0^\infty |H(\omega j)|^2\, d\omega.$$

Seja $H(s)$ a função de transferência do modelo (A, B, C) original e $\widehat{H}(s)$ a função de transferência associada a um modelo reduzido. Podemos estimar o erro de aproximação calculando

$$\|H(s) - \widehat{H}(s)\|_{\mathcal{L}_2(0,\infty)}.$$

Sendo assim, faz muito sentido analisar o comportamento da função de transferência H no domínio das frequências para buscar boas alternativas de redução de modelo. Para tratar desse assunto, vamos seguir os passos de [32] daqui em diante nesta seção.

Consideramos um polo $\lambda_m = \alpha_m + j\beta_m$ associado ao resíduo R_m e aplicamos a função $H(s)$ em $s = j\beta_m$:

$$H(j\beta_m) = -\frac{R_m^H}{\alpha_m} + \sum_{\substack{i=1,\\ i\neq m}}^{n} \frac{R_i^H}{j\beta_m - \lambda_i} \tag{2.17}$$

Note que, pela expressão (2.17), um polo λ_m associado a um valor de $|R_m/\alpha_m|$ relativamente elevado tende a dominar a resposta de $H(s)$ causando um pico de amplitude em $s = j\beta_m$. Isso motiva a seguinte definição:

Definição 2.7 *[34] Um polo $\lambda_m = \alpha_m + \beta_m j$ de $H(s)$ com os autovetores associados v_m e w_m à direita e à esquerda, respectivamente, ($w_m^H v_m = 1$) é chamado dominante se $|R_m/\alpha_m| > |R_n/\alpha_n|$ sempre que $m \neq n$.*

Embora a definição 2.7 refira-se a um único polo dominante apenas, ela pode ser interpretada a fim de se obter um conjunto $\Omega_k = \{\lambda_1, \lambda_2, ..., \lambda_k\}$ de polos dominantes em $H(s)$, escolhidos de modo a satisfazer $|R_m/\alpha_m| > |R_n/\alpha_n|$, sempre que $1 \leq m \leq k$ e $n > k$. Isso motiva a construção de um modelo reduzido utilizando apenas esse conjunto de $k << n$ polos dominantes. A função de transferência do modelo reduzido terá, então, a seguinte forma:

$$H(s) \backsimeq H_k(s) = \sum_{\substack{l=1,\\ \lambda_l \in \Omega_k}}^{k} \frac{R_l^H}{s - \lambda_l}, \tag{2.18}$$

O processo que consiste em obter a expressão (2.18) é chamado de *redução modal*. O sistema reduzido nos moldes do sistema (A, B, C) passa a ser

$$\begin{aligned} \dot{x}(t) &= A_k x_k(t) + B_k u(t) \\ y(t) &= C_k^T x_k(t), \end{aligned} \tag{2.19}$$

em que $A \in \mathbb{C}^{k\times k}$ é uma matriz diagonal de polos dominantes, $B_k \in \mathbb{C}^k$ é um vetor contendo os resíduos e $C_k \in \mathbb{R}^k$ é um vetor de "uns".

Além da forte noção intuitiva de que os polos dominantes definidos em 2.7 tendem a ser mais relevantes na formação da função H, existem várias situações numericamente testadas em que um conjunto de poucos polos dominantes é suficiente para obter-se um modelo reduzido fiel ao original, como pode ser visto em [46], [26], [34], [47], [46] e [14]. No entanto, não se pode afirmar que o erro $\|H(s) - H_k(s)\|$ decresce seguindo a mesma proporção dos valores $|R_k/\alpha_k|$. Talvez por isso, na literatura, existem outras definições de polos dominantes em funções de transferência, como pode ser visto em [32], por exemplo. Nesse trabalho, propomos uma nova definição de polos dominantes que será apresentada na subseção 3.3.6 e visa caracterizar a dominância de um polo de maneira relativa a um conjunto de polos dominantes previamente escolhidos.

Na literatura existem vários métodos capazes de encontrar um conjunto de polos dominantes para gerar a aproximação (2.18) sem ter que calcular, necessariamente, todo o espectro de A. Dentre os mais utilizados podemos citar o SADPA (*Subspace Accelerate Dominant Pole Algorithm*)[34] e o DPSE [25].

Uma das principais dificuldades nesse tipo de redução de modelo é estimar o quão próxima uma função $H_k(s)$ do modelo reduzido está da função $H(s)$ original, pois o cálculo do erro $\|H(s) - H_k(s)\|_{\mathcal{L}_2(0,\infty)}$ requer que conheçamos todos os polos de $H(s)$. A subseção a seguir trata de estimativas para o erro de aproximação da redução modal.

2.2.1 Critério de parada

Visando estabelecer um critério de parada para a redução modal que depende da quantidade de polos dominantes do sistema a serem considerados, vamos descrever agora uma estratégia apresentada em [46] utilizando estimativas para a função

$$E(H) := \|H(s)\|_{\mathcal{L}(\omega_0,\omega_f)} := \frac{1}{\pi}\int_{\omega_0}^{\omega_f} |H(j\omega)|^2\, d\omega. \tag{2.20}$$

Vamos considerar os polos de $H(s)$ escritos na forma retangular: $\lambda_j = \alpha_i + \beta_i j$. Em [46] utiliza-se a decomposição em frações parciais (2.15) para aferir que a expressão (2.20) pode ser reescrita da seguinte forma:

$$E(H) = \frac{1}{\pi}\sum_{i=1}^{n/2} e_i, \tag{2.21}$$

com

$$e_i = a_i g_i + b_i h_i.$$

Os coeficientes a_i e b_i são dados por

$$a_i = Re\{R_i^H H(-\lambda_i)\}, \tag{2.22}$$

$$b_i = Im\{R_i^H H(-\lambda_i)\} \tag{2.23}$$

Já os números g_i e h_i podem ser calculados como segue:

$$g_i = 2\left[atan\left(\frac{\omega_0 - \beta_i}{\alpha_i}\right) - atan\left(\frac{\omega_f - \beta_i}{\alpha_i}\right) + atan\left(\frac{\omega_0 + \beta_i}{\alpha_i}\right) - atan\left(\frac{\omega_f + \beta_i}{\alpha_i}\right)\right],$$

$$h_i = ln\left[(\omega_f - \beta_i)^2 + \alpha_i^2\right] - ln\left[(\omega_0 - \beta_i)^2 + \alpha_i^2\right] + ln\left[(\omega_0 + \beta_i)^2 + \alpha_i^2\right] - ln\left[(\omega_f - \beta_i)^2 + \alpha_i^2\right].$$

A partir disto, são consideradas duas aproximações para (2.20). A primeira é construída utilizando um conjunto de k polos dominantes:

$$E(H) \cong E(H_k) = \frac{1}{\pi}\sum_{i=1}^{k/2} e_i.$$

Os coeficientes a_i e b_i dependem da função H, portanto, com base em (2.22) e (2.23), pode-se dizer que

$$a_i \cong \tilde{a}_i := Re\{R_i^H H_k(-\lambda_i)\} \tag{2.24}$$

$$b_i \cong \tilde{b}_i := Im\{R_i^H H_k(-\lambda_i)\}. \tag{2.25}$$

Isso sugere a segunda aproximação para (2.20):

$$E(H) \cong \tilde{E}(H_k) = \frac{1}{\pi}\sum_{i=1}^{k/2} \tilde{e},$$

com $\tilde{e}_i = \tilde{a}_i g_i + \tilde{b}_i h_i$ para $i = 1, , 2, ..., k/2$.

Uma vez que $H_k(s) \overset{k\to n}{\to} H(s)$, então $\tilde{E}(H_k) \overset{k\to n}{\to} E(H_k) \overset{k\to n}{\to} E(H)$. Os estudos de [46] se baseiam nessas informações para definir o seguinte erro em cada polo considerado na aproximação:

$$\varepsilon_i = \frac{e_i - \tilde{e}_i}{E(H_k)}. \tag{2.26}$$

Disto surge a estimativa para o erro global utilizando uma técnica de quadrados mínimos (RMS) aplicada nos polos considerados na aproximação:

$$\varepsilon_{RMS}(k) = \sqrt{\frac{2\sum_{i=1}^{k}\varepsilon_i^2}{k}}.$$

Dada uma tolerância ε_{max}, o número de polos dominantes k da redução modal pode ser considerado adequado para determinadas situações quando

$$\varepsilon_{RMS}(k) < \varepsilon_{max}. \tag{2.27}$$

Vários exemplos numéricos apresentados em [46] atestam a eficácia do critério de parada (2.27) para determinadas classes de problemas. Além disso, sua viabilidade em termos de custo computacional é notável, pois a imagem da função $\tilde{E}(H_k)$ depende apenas dos k pólos considerados na aproximação modal.

Embora se verifique a efetividade do critério de parada (2.27) para muitos casos, apontamos o fato de que a diminuição da diferença entre $\tilde{E}(H_k)$ e $E(H_k)$, à medida que k aumenta, não implica, necessariamente que $E(H_k)$ se aproxima de $E(H)$ numa mesma proporção. A influência de um polo (que também pode ser visto como uma singularidade da função H) na magnitude da imagem da função H tende a ser mais intensa numa região próxima do polo quando comparado com pontos mais distantes no domínio de $H(s)$. A figura 2.1 ilustra esse fato por meio de curvas de nível. Note que, uma vez que a matriz A do sistema (A, B, C) é real, os polos de $H(s)$ são dispostos em pares complexos conjugados. No entanto, apenas para facilitar a visualização, na função do exemplo a seguir consideramos apenas um representante de cada par complexo conjugado.

Acreditamos que uma maneira de melhorar a estimativa (2.27) seria considerar uma norma ponderada definida por

$$\|H(j\omega)\|_{W\mathcal{L}(\omega_0,\omega_f)} := \int_{\omega_0}^{\omega_f} |H(j\omega)W(j\omega)|\, d\omega, \tag{2.28}$$

com $E, F \in \mathbb{R}^{p\times 1}$ e $W \in \mathbb{R}^{p\times p}$ sendo matrizes dadas. A matriz W pode ser construída para ter autovalores $\{\lambda_1^W, \lambda_2^W, ..., \lambda_p^W\}$ diferentes dos autovalores de A.

A vantagem do uso da norma ponderada (2.28) é que, no caso de avaliar o erro entre função original e a aproximação feita por redução modal, o usuário poderia escolher uma função $W(s)$ cujos polos sejam distribuídos de maneira homogênea num subconjunto de interesse no domínio da função $H(s)$. Além disso, todo o desenvolvimento apresentado em [46] poderia ser aproveitado para chegar a uma expressão análoga a (2.27) com essa nova norma.

A desvantagem é que essa técnica encareceria o processo por necessitar da avaliação da função $H(s)$ em mais pontos. Além disso, poderia haver incerteza de precisão, pois os polos da ponderação $W(s)$, dependendo da sua localização no plano complexo e de seus respectivos resíduos, poderiam fazer com que o conjunto de polos dominantes do produto $H(s)W(s)$ seja completamente diferente do conjunto de polos dominantes de $H(s)$. Diante dessas conjecturas, optamos por não seguir essa linha de raciocínio. Ao invés disso, no final da seção 3.3, apresentamos uma nova estimativa de erro para redução modal baseada em soluções de equações de Lyapunov associadas ao sistema (A, B, C).

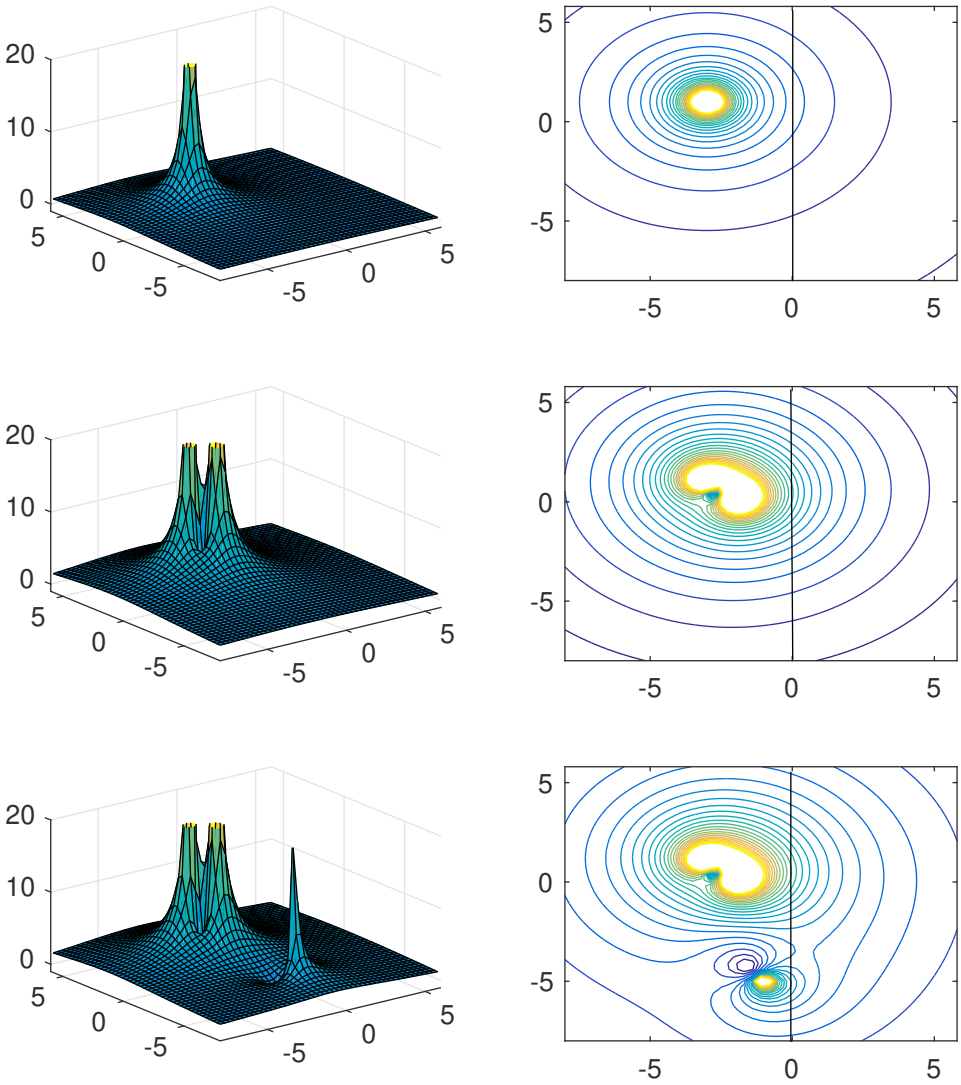


Figura 2.1: Gráficos e curvas de nível da função $|H(s)| = \left|\frac{5-3i}{s-\lambda_1} + \frac{7+2i}{s-\lambda_2} + \frac{2+i}{s-\lambda_3}\right|$, com $\lambda_1 = -3+i$, $\lambda_2 = -2+0,2i$ e $\lambda_3 = -1-5i$. A primeira linha de gráficos apresenta a função apenas com o polo λ_1. A segunda linha exibe os gráficos para a função com os polos λ_1 e λ_2 e, por último, é apresentada a função com os três polos.

2.3 Redução por Balanceamento (mudança de base)

Para os estudos desta seção, vamos reforçar a hipótese de que todos os autovalores da matriz A do sistema (1.3) possuem parte real negativa. Isso implica, pelo Corolário 1.27, que as equações de Lyapunov (1.24) e (1.25) admitem solução única dada por (1.22) e (1.23), respectivamente.

Seja T uma transformação linear inversível aplicada ao estado x do sistema , de modo que tenhamos a mudança de base $\widehat{x} = Tx$. Desta forma, para que seja mantida a relação entre a

entrada e a saída do sistema, é preciso substituir as matrizes A, B, C e D do sistema por matrizes $\widehat{A}$, $\widehat{B}$, $\widehat{C}$ e $\widehat{D}$, respectivamente, definidas como segue:

$$\widehat{A} = TAT^{-1},\ \widehat{B} = TB,\ \widehat{C} = CT^{-1} \text{ e } \widehat{D} = D.$$

Existem maneiras de definir uma transformação T que torne mais viável a redução da ordem do sistema (A, B, C, D) [12] [44]. Apresentamos aqui algumas dessas estratégias, focando na utilização das matrizes de gram P e Q, definidas em (1.22) e (1.23), respectivamente. Essas estratégias são motivadas pelas explanações feitas na seção 2.1.

Supomos, inicialmente, que a matriz P (ou Q) seja positiva definida, com decomposições de Cholesky $P = UU^T$ e $Q = LL^T$. Seja K uma matriz ortogonal tal que

$$U^TQU = K\Sigma^2K^T,$$

com $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_n)$ sendo a matriz dos valores principais de U^TQU.

Assim, se definirmos

$$T = \Sigma^{1/2}K^TU^{-1} \quad \text{e} \quad T^{-1} = UK\Sigma^{-1/2},$$

por (1.22) e (1.23), as novas matrizes de Gram de controlabilidade e observabilidade serão

$$\widehat{P} = TPT^T = \Sigma \quad \text{e} \quad \widehat{Q} = T^{-T}QT^{-1} = \Sigma,$$

ou seja, as matrizes de Gram $\widehat{P}$ e $\widehat{Q}$ são matrizes diagonais idênticas entre si. Nesse caso dizemos que o sistema (1.3) está *balanceado* [1].

Feito isso, uma ideia intuitiva de reduzir o sistema seria truncar a matriz T de modo que os menores valores da diagonal de Σ^2 sejam "jogados fora", pois, por (2.12), eles representam os menores valores singulares do operador de Hankel $\mathbf{H}$. Porém, esse método, da forma como está apresentado, não faz sentido se a matriz P (ou Q) não é inversível. Pensando nisso, há de se observar que

$$\sigma_i(G(s)) = \lambda_i^{1/2}(PQ) = \lambda_i^{1/2}(UU^TLL^T) = \lambda_i^{1/2}(U^TLL^TU) = \sigma_i(U^TL).$$

Assim, se tomarmos a decomposição de SVD $U^TL = W\Sigma V^T$, podemos definir

$$T = \Sigma^{-1/2}V^TL^T \quad \text{e} \quad T^\dagger = UW\Sigma^{-1/2}. \tag{2.29}$$

O símbolo † é empregado para representar a pseudo-inversa de uma matriz. Essa notação se faz necessária para os casos em que U^TL não tem posto cheio.

Analogamente ao caso anterior, utilizando as matrizes expressas em (2.29), conseguimos tornar o sistema (1.3) balanceado, ou seja, com $\widehat{P} = \widehat{Q} = \Sigma$, com Σ sendo uma matriz diagonal. Esse

fato está mostrado com mais detalhes no Teorema 2.8.

Vamos considerar a seguinte partição das matrizes W, Σ e V:

$$W = \left[\begin{array}{cc} W_1 & W_2 \end{array}\right], \quad \Sigma = \left[\begin{array}{cc} \Sigma_1 & \\ & \Sigma_2 \end{array}\right] \quad \text{e} \quad V = \left[\begin{array}{cc} V_1 & V_2 \end{array}\right], \tag{2.30}$$

com $\Sigma_1 > 0$. Isso permite definir as seguintes transformações, originárias de um "truncamento"das matrizes (2.29).

$$T_L := \Sigma_1^{-1/2} V_1^T L^T \quad \text{e} \quad T_R := U W_1 \Sigma_1^{-1/2}, \tag{2.31}$$

O novo sistema, definido pelas matrizes

$$\widehat{A} = T_L A T_R, \quad \widehat{B} = T_L B, \quad \widehat{C} = C T_R \quad \text{e} \quad \widehat{D} = D, \tag{2.32}$$

possui propriedades muito úteis para este trabalho. Tais propriedades são exibidas nos próximos dois teoremas.

A fim de facilitar o entendimento da redução de modelo, consideremos a seguinte partição do sistema $(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$, baseada nas partições de (2.30).

$$\widehat{A} = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right], \quad \widehat{B} = \left[\begin{array}{c} B_1 \\ B_2 \end{array}\right] \quad \mathbf{e} \quad \widehat{C}\left[\begin{array}{cc} C_1 & C_2 \end{array}\right], \tag{2.33}$$

de modo que a dimensão de A_{11} é $k \times k$.

Teorema 2.8 *O sistema $(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$ definido conforme (2.32), é uma realização balanceada, truncada em k, do sistema (A, B, C, D).*

Demonstração. Ver Apêndice B □

Teorema 2.9 *[44] Seja A uma matriz Hurvitz, ou seja, $Re(\lambda_i(A)) < 0$ para $i = 1, ..., n$. Se a partição (2.30) for feita de modo que $\Sigma_2 \equiv 0$ e $\Sigma_1 > 0$, então o sistema de ordem reduzida (A_{11}, B_1, C_1, D_1) é completamente controlável e observável e a função de transferência $\mathbf{G_k(s)}$ do modelo reduzido é idêntica à função $\mathbf{G(s)}$ do sistema original.*

Demonstração. Ver Apêndice B. □

Pelos teoremas 2.9 e 2.6, a redução por balanceamento é uma maneira prática de transformar o sistema (A, B, C, D) numa realização minimal.

Teorema 2.10 *[9] Vamos supor que a matriz Σ de (2.30) seja*

$$\Sigma = \left[\begin{array}{cccc} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_n \end{array}\right]$$

com $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k \geq \sigma_{k+1} \geq ... \geq \sigma_n$. *Então, a função de transferência* $\mathbf{G}_k(s)$ *do sistema reduzido satisfaz*

$$\|\mathbf{G}(s) - \mathbf{G}_k(s)\|_\infty \leq 2 \sum_{i=k+1}^{n} \sigma_i. \tag{2.34}$$

Demonstração. Ver Apêndice B. □

Por fim, o Teorema 1.31, juntamente ao Teorema 1.13, garante que, se $\Sigma_1 > 0$ em (2.30), então o sistema reduzido $(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$ é assintoticamente estável.

Existe também, um método de redução de modelo que consiste na combinação da redução modal, apresentada na seção 2.2, que é aplicada por primeiro, seguida da redução por balanceamento. Essa técnica de redução híbrida aparece muito bem escrita e devidamente testada em [46].

O método de redução por balanceamento permite reduzir a ordem do modelo de maneira significativa preservando fielmente a relação entrada-saída de um sistema dinâmico linear. No entanto, há o inconveniente de que o cálculo das soluções P e Q das equações de Lyapunov associadas ao modelo pode ter um custo computacional muito alto em sistemas de ordem elevada. Sendo assim, efetuar uma redução modal antes de aplicar a redução por balanceamento pode reduzir significativamente esse custo.

Capítulo 3

Alguns Métodos para Equação de Lyapunov

Vamos considerar a equação de Lyapunov

$$AP + PA^T = -BB^T, \tag{3.1}$$

com $B \in \mathbb{R}^{n\times m}$ e $A \in \mathbb{R}^{n\times n}$ sendo uma matriz de Hurwitz (todos autovalores possuem parte real negativa), diagonalizável e de espectro simples (autovalores todos distintos).

Este capítulo é destinado ao estudo de métodos para a equação (3.1), bem como a verificação de desempenho de alguns deles em redução de modelo. Nosso foco está em problemas cujas matrizes são esparsas e de ordem elevada. Iniciamos tratando de métodos conhecidos e bastante utilizados na literatura, como o método ADI e métodos baseados em subespaços de Krylov. Em seguida, apresentamos uma abordagem que envolve a solução explícita para (3.1) a partir da decomposição espectral de A. Por fim, na última seção, introduzimos um método para equações de Lyapunov baseado em métodos iterativos para sistemas lineares a partir de uma cisão da matriz A, popularmente conhecidos como métodos do tipo *splitting*.

3.1 Método ADI

Nesta seção discutiremos sobre um método iterativo bastante utilizado para aproximar soluções de equações de Lyapunov. Fazemos uma dedução do algoritmo ADI a partir do método de Smith [42], para depois falar de aprimoramentos feitos por [49], [23], e [30].

Inicialmente vamos descrever a dedução do método Smith [42] para a equação de Lyapunov (3.1). Escolhendo um escalar $\mu \in \mathbb{R}$ cuja parte real é negativa, a equação (3.1) pode ser escrita como a expressão (3.2), a seguir:

$$(A - \mu I)P(A^T - \mu I) - (A + \mu I)P(A^T + \mu I) = 2\mu BB^T. \tag{3.2}$$

A equação (3.2) pode ser multiplicada pela esquerda por $(A-\mu I)^{-1}$ e pela direita por $(A^T-\mu I)^{-1}$ a fim de obter-se

$$P - UPV = W, \tag{3.3}$$

com $U = (A-\mu I)^{-1}(A^T+\mu I)$, $V = (A^T+\mu I)(A^T-\mu I)^{-1}$ e $W = 2\mu(A-\mu I)^{-1}BB^T(A^T-\mu I)^{-1}$. Como A é uma matriz de Hurwitz e $\mu < 0$, então o raio espectral de U, assim como o de V, é menor do que 1. Isso ocorre em virtude da transformação $\xi(z) = (z-\mu)^{-1}(\mu+z)$ levar o semi-plano $\{z \in \mathbb{C} : Re(z) < 0\}$ no disco $\{z \in \mathbb{C} : |z| < 1\}$. Portanto, fazendo uso do Teorema de Banach para Álgebra Linear [38], é possível verificar que

$$P = \sum_{k=1}^{\infty} U^{k-1}WV^{k-1} \tag{3.4}$$

é solução de (3.3). Definindo $P_\nu = \sum_{k=1}^{2^\nu} U^{k-1}WV^{k-1}$, é possível provar que existe um número real r, entre zero e um, tal que a sequência $\{P_\nu\}_{\nu\in\mathbb{N}}$ converge a uma taxa $(1-r)^{-1}r^{2^\nu}$ para solução da equação (3.1) [42].

Em vez de fazer uso de um parâmetro μ que se mantém constante nos termos da série (3.4), pode-se utilizar um conjunto de parâmetros $\{\mu_1, \mu_2, ...\}$ a fim de acelerar a convergência do método [49], [23]. Nesses moldes, uma expressão similar a (3.4) seria

$$P \approx \sum_{k=1}^{\infty} \tilde{U}_i^{k-1}\tilde{W}_i\tilde{V}_i^{k-1}. \tag{3.5}$$

em que

$$\tilde{U}_i = (A-\mu_i I)^{-1}(A+\mu_i I), \quad \tilde{V}_i = (A^T+\mu_i I)(A^T-\mu_i I)^{-1}$$

$$e \quad \tilde{W}_i = 2Re(\mu_i)(A+\mu_i I)^{-1}BB^T(A^T+\mu_i I)^{-1}.$$

Definindo $P_0 = 0_{n\times n}$, cada soma parcial da série (3.5) pode ser representada por meio das iterações

$$(A+\mu_i I)P_{(i-1/2)} = -P_{(i-1)}(A^T-\mu_i I) + BB^T \tag{3.6}$$

$$(A+\mu_i I)P_i^H = -P_{(i-1/2)}^H(A^T-\mu_i I) + BB^T \tag{3.7}$$

com $i = 2, 3, ...$ Com efeito, a partir de (3.6) obtém-se que

$$P_{(i-1/2)} = -(A+\mu_i I)^{-1}P_{i-1}(A^T-\mu_i I) + (A+\mu_i I)^{-1}BB^T,$$

então

$$\begin{aligned} P_i^H =& (A+\mu_i I)^{-1}(A-\mu_i I)P_{i-1}^T(A^T+\mu_i I)^{-1}(A^T-\mu_i I) \\ & - (A+\mu_i I)^{-1}BB^T(A^T+\mu_i I)^{-1}(A^T-\mu_i I) + (A+\mu_i I)^{-1}BB^T. \end{aligned}$$

Portanto,

$$\begin{aligned}
P_i =&(A-\mu_i I)(A+\mu_i I)^{-1}P_{i-1}(A^T-\mu_i I)(A^T+\mu_i I)^{-1}\\
&-(A-\mu_i I)(A+\mu_i I)^{-1}BB^T(A^T+\mu_i I)^{-1}+BB^T(A^T+\mu_i I)^{-1}\\
=&\tilde{U}P_{i-1}\tilde{V}-(A-\mu_i I)(A+\mu_i I)^{-1}BB^T(A^T+\mu_i I)^{-1}\\
&+(A+\mu_i I)(A+\mu_i I)^{-1}BB^T(A^T+\mu_i I)^{-1}\\
=&\tilde{U}_i P_{i-1}\tilde{V}_i+\tilde{W}_i,
\end{aligned}$$

mostrando assim a equivalência entre (3.6) e as somas parciais de (3.5).

O método regido pelas iterações mostradas em (3.6) é comumente chamado de ADI [48] (*Alternating Direction Implicit*). Os números $\{\mu_1, \mu_2, ...\}$ são chamados de parâmetros ADI. Embora essa dedução do método seja feita considerando-se parâmetros μ_i reais, é mais comum na literatura a versão do método que faz uso de parâmetros complexos. Daqui em diante, consideraremos $\mu_i \in \mathbb{C}$.

Vamos estudar agora uma variação do método ADI que melhora sua performance quando as matrizes envolvidas são de ordem elevada. A estratégia consiste em utilizar a decomposição de Cholesky de baixo posto para matrizes esparsas. Pais detalhes podem ser vistos em [30]. Seja

$$P_i = Y_i Y_i^H \tag{3.8}$$

a fatoração de Cholesky de baixo posto (em inglês: *Low-Rank Cholesky Factorization* do iterado P_i do método ADI. Então

$$Y_1 = \sqrt{-2re(\mu_1)}(A+\mu_1 I)^{-1}B \quad \text{e}$$

$$Y_i = \left[\begin{array}{cc} \sqrt{-2Re(\mu_i)}(A+\mu_i I)^{-1}B & \tilde{U}_i Y_{i-1} \end{array}\right], \quad i = 2, 3, ...$$

Perceba que

$$Y_i = \left[\sqrt{-2Re(\mu_i)}A_{+\mu_i}^{-1}B,\ A_{+\mu_i}^{-1}A_{-\mu_i}A_{+\mu_i}^{-1}\sqrt{-2Re(\mu_{i-1})}B,\ ...,\ A_{+\mu_2}^{-1}A_{-\mu_2}A_{+\mu_1}^{-1}\sqrt{-2Re(\mu_1)}B\right],$$

com

$$A_{+\mu_i}^{-1} = (A+\mu_i I)^{-1}, \quad \text{e} \quad A_{-\mu_i} = (A-\mu_i I).$$

Como as matrizes $A_{+\mu_i}^{-1}$ e $A_{-\mu_i}$ comutam, então

$$Y_i = \left[\sqrt{-2Re(\mu_i)}A_{+\mu_i}^{-1}B,\ K_{i-1}B,\ K_{i-2}K_{i-1}B\ ...,\ K_1K_2...K_{i-1}\sqrt{-2Re(\mu_1)}A_{+\mu_i}^{-1}B\right], \tag{3.9}$$

com

$$K_i := \sqrt{\frac{-Re(\mu_i)}{-Re(\mu_{i+1})}} A^{-1}_{+\mu_i} A_{-\mu_{i+1}} \qquad = \sqrt{\frac{-Re(\mu_i)}{-Re(\mu_{i+1})}} \left[I - (\mu_{i+1} + \mu_i)(A + \mu_i I)^{-1} \right]. \tag{3.10}$$

Como a ordem dos parâmetros $\mu_1, \mu_2, ...$ é irrelevante para o método, podemos inverter a ordem dos índices $1, 2, ..., i$ em (3.9). Podemos então reescrever as iterações do método conforme segue:

$$\begin{aligned} &S_1 := Y_1 = \sqrt{-2Re(\mu_1)}(A + \mu_1 I)^{-1} B \\ &S_i = \sqrt{\tfrac{-Re(\mu_i)}{-Re(\mu_{i-1})}} \left[S_{i-1} - \gamma_i (A + \mu_i I)^{-1} S_{i-1} \right] \\ &Y_i = \begin{bmatrix} Y_{i-1} & S_i \end{bmatrix}, \quad i = 2, 3, ..., \end{aligned} \tag{3.11}$$

com $\gamma_i = 2Re(\mu_i)$.

Seguindo a linha de raciocínio contida em [51], vamos buscar uma estimativa para o decaimento dos autovalores da solução P da equação de Lyapunov (3.1). Essa estimativa fornecerá também estimativa para o erro da sequência gerada pelo método ADI. Perceba que, como

$$P = \tilde{U}_i P \tilde{V}_i + \tilde{W}_i,$$

então

$$P - P_i = \tilde{U}_i (P - P_{i-1}) \tilde{V}_i.$$

Repetindo esse processo por sucessivas vezes, assumindo que $P_0 = 0_{n \times n}$, concluímos que

$$P - P_k = \prod_{i=1}^{k} \tilde{U}_i P \tilde{V}_i = \prod_{i=1}^{k} (A - \overline{\mu}_i I)^{-1} (A + \mu_i I) P (A^T + \overline{\mu}_i I)(A^T - \mu_i I)^{-1}. \tag{3.12}$$

A partir de (3.8) e (3.9), tem-se que $rank(P_{i-1}) \leq rank(P_i) \leq rank(P_{i-1}) + m$, para algum m natural. Deste modo, $rank(P_k) \leq km$ para todo k natural.

Suponhamos que os autovalores de P estejam dispostos em ordem decrescente, ou seja, $\lambda_1(P) \leq \lambda_2(P) \leq ... \leq \lambda_n(P)$. Então, pelo Teorema de Schmidt-Mirsky, o decaimento dos autovalores de P é dado por

$$\frac{\lambda_i(P)}{\lambda_1(P)} = \min_{\overline{P} \in \mathbb{R}^{n \times n}, rank(\overline{P}) \leq i-1} \frac{\|P - \overline{P}\|_2}{\|P\|_2}. \tag{3.13}$$

Combinando (3.12) com (3.13) obtemos

$$\frac{\lambda_{mk+1}(P)}{\lambda_1(P)} \leq \frac{\|P - P_k\|_2}{\|P\|_2} \leq \left\| \prod_{i=1}^{k} \tilde{U}_i \right\|_2 \left\| \prod_{i=1}^{k} \tilde{V}_i \right\|_2 \tag{3.14}$$

$$= \left\| \prod_{i=1}^{k} \tilde{U}_i \right\|_2^2 \tag{3.15}$$

pois $\tilde{U}_i = \tilde{V}_i^T$.

Como estamos supondo A diagonalizável, existe uma matriz invertível T, cujas colunas são autovetores de A, e uma matriz $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_n)$, cujos elementos da diagonal são autovalores de A, satisfazendo

$$AT = T\Lambda.$$

Então

$$\begin{aligned}
\| \prod_{i=1}^{k} \tilde{U}_i \|_2^2 &= \left\| \prod_{i=1}^{k} (A - \overline{\mu}_i I)^{-1}(A + \mu_i I) \prod_{i=1}^{k} (A^T + \overline{\mu}_i I)(A^T - \mu_i I)^{-1} \right\|_2 \\
&= \left\| T \left\{ \prod_{i=1}^{k} (\Lambda - \overline{\mu}_i I)^{-1}(\Lambda + \mu_i I) \right\} T^{-1} T^{-T} \left\{ \prod_{i=1}^{k} (\Lambda^T + \overline{\mu}_i I)(\Lambda^T - \mu_i I)^{-1} \right\} T^T \right\|_2 \\
&\leq \kappa(T)\kappa(T^T) \left\{ \max_{1 \leq l \leq n} \prod_{i=1}^{k} \left| \frac{\overline{\mu}_i - \lambda_l}{\mu_i + \lambda_l} \right| \right\} \left\{ \max_{1 \leq p \leq n} \Pi_{i=1}^{k} \left| \frac{\mu_i - \overline{\lambda}_p}{\overline{\mu}_i + \overline{\lambda}_p} \right| \right\} \\
&= \kappa^2(T) \left\{ \max_{1 \leq l \leq n} \prod_{i=1}^{k} \left| \frac{\overline{\mu}_i - \lambda_l}{\mu_i + \lambda_l} \right| \right\}^2 .
\end{aligned}$$

Essa estimativa é válida para qualquer escolha de $\{\mu_1, \mu_2, ..., \mu_k\} \subset \mathbb{C}_-$. Portanto, a busca por parâmetros que tornem a última expressão menor possível caracteriza o problema *min-max*:

$$\underset{\{\tilde{\mu}_1, ..., \tilde{\mu}_k \in \mathbb{C}_-\}}{\arg\min} \left\{ \max_{1 \leq l \leq n} \Pi_{i=1}^{k} \left| \frac{\overline{\tilde{\mu}}_i - \lambda_l}{\tilde{\mu}_i + \lambda_l} \right| \right\}. \tag{3.16}$$

O problema de otimização (3.16) pode ser difícil, principalmente nos casos em que não se conhece os autovalores de A. Por isso, costuma-se definir o problema "min-max"conforme segue:

$$\underset{\{\tilde{\mu}_1, ..., \tilde{\mu}_k \in \mathbb{C}_-\}}{\arg\min} \left\{ \max_{x \in \mathcal{R}} \left| \prod_{i=1}^{k} \frac{\overline{\tilde{\mu}}_i - x}{\tilde{\mu}_i + x} \right| \right\}, \tag{3.17}$$

em que $\lambda_1(A), \lambda_2(A), ..., \lambda_n(A) \in \mathcal{R} \subset \mathbb{C}^-$. Mais detalhes sobre como obter essa região R e sobre como resolver o problema (3.16) de maneira *quase ótima* são dados na seção 3.3.4.

Os parâmetros $\{\mu_1, \mu_2, ...\}$ podem ser utilizados de maneira cíclica, ou seja, calcula-se uma quantidade de parâmetros $\{\mu_1, \mu_2, ..., \mu_J\}$, com $J \in \mathbb{N}$, e repete-se o uso desses parâmetros a partir da J+1-ésima iteração do método. Obtém-se assim, o método introduzido por Penzl e conhecido como *ADI* cíclico para equações de Lyapunov de ordem elevada e posto baixo (*Cyclic Low-Rank ADI Method for Large Sparse Lyapunov Equations*)[30].

O critério de parada para o método (3.11) pode ser estabelecido checando-se a diferença relativa entre os iterados Y_i e Y_{i-1} [30]:

$$\frac{\|S_i\|_F^2}{\|Y_i\|_F^2} \leq \varepsilon, \tag{3.18}$$

com uma tolerância $\varepsilon \ll 1$.

Vejamos agora como adaptar o algoritmo ADI para sistemas descritores. Na primeira linha do processo (3.11), é preciso avaliar o produto da matriz $(A + \mu_1 I)^{-1}$ matriz B. Na sequência, é preciso efetuar $(A + \mu_i I)^{-1} S_{i-1}$. Na seção 1.3 já vimos como efetuar essas operações de maneira prática no caso des matrizes serem oriundas do sistema definido em (1.29). Estas técnicas, somadas às construções esboçadas até aqui, compõem o método ADI sistemas esparsos de ordem elevada (*Sparce Low-Rank Cholesky Factor ADI-method* - SLRCF-ADI) [11], descrito no algoritmo a seguir

Algoritmo 1: SLRCF-ADI

Entrada: Matrizes do sistema Jacobiano $(J_1, J_2, J_3, J_4, B_1, B_2, C_1, C_2 D_a)$ e os parâmetros ADI μ_i, $i = 1, 2, ..., J$

Saída: Fator de Cholesky de baixo posto Y de $P = YY^T$

1 **início**

2 Calcule Γ de

$$\begin{bmatrix} J_1 + \mu_1 I & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} \Gamma \\ \Upsilon \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}.$$

3 $Y_1 \leftarrow S_1 \leftarrow \sqrt{-2Re(\mu_1)}\Gamma$.

4 **para** *i=2,3,..., até convergir (critério* (3.18)*,* **faça**

5 $\mu_i \leftarrow \mu_i^{i(\text{mod } J)}$

6 Calcule Γ de

$$\begin{bmatrix} J_1 + \mu_i I & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} \Gamma \\ \Upsilon \end{bmatrix} = \begin{bmatrix} S_{i-1} \\ 0 \end{bmatrix}.$$

7 $S_i \leftarrow \sqrt{\frac{-Re(\mu_i)}{-Re(\mu_{i-1})}} \left(S_{i-1} - \gamma_i \Gamma\right)$

8 $Y_i \leftarrow \left[Y_{i-1}\ S_i\right]$.

9 **fim**

10 **fim**

11 $Y \leftarrow Y_i$.

Para o algoritmo 1 consideramos o mesmo critério de parada que definimos em (3.18). Neste trabalho, o sistema da linha 6 é resolvido via decomposição LU. Acreditamos que outras estratégias mais eficientes podem ser utilizadas para realizar esse passo. Isso pode ser objeto de estudo em projetos futuros.

3.2 Método baseados em subespaços de Krylov

Nesta seção descrevemos métodos que geram aproximações para a solução da equação de Lyapunov (3.1) por meio de iterações que consistem em projetá-la num subespaço de ordem reduzida. Tais técnicas foram muito bem difundidas em trabalhos como [39], [8], [43], [40], [21], [6] e [45].

Seguindo os passos descritos em [39], consideremos um subespaço $\mathcal{V} \subset \mathbb{R}^n$, e seja V uma matriz cujas colunas formam uma base ortonormal para $\mathcal{V}$, que satisfaz $B = VN$, para alguma matriz N e B sendo a matriz que aparece em (3.1). Em outras palavras, as colunas de B devem pertencer ao subespaço $\mathcal{V}$. Assim, pode-se projetar a equação de Lyapunov (3.1) no subespaço $\mathcal{V}$ e encontrar uma solução da forma $\tilde{P} = VYV^T$, para alguma matriz $Y \in \mathbb{R}^{n\times n}$. A equação projetada é obtida pela imposição de que o erro $R = A\tilde{X} + \tilde{X}A^T + BB^T$ seja ortogonal ao espaço de aproximação $\mathcal{V}$. Essa propriedade é chamada de condição de Galerkin. Em termos matriciais, isso significa que $\langle R, VMV_T\rangle = 0$ para todo $M \in \mathbb{R}^{n\times n}$, ou seja, $\langle V^TRV, M\rangle = 0$ para todo $M \in \mathbb{R}^{n\times n}$. Isto ocorre se, e somente se, $V^TRV = 0$. Disto segue que

$$V^TAVY + YV^TA^TV = -V^TBB^TV, \tag{3.19}$$

cuja solução Y define a aproximação $\tilde{X}$ para a solução da equação de Lyapunov (3.1).

O subespaço de projeção $\mathcal{V}$ pode ser, por exemplo, um espaço da forma

$$K_q(A, B) = span\{B, AB, A^2B, ..., A^qB\}, \tag{3.20}$$

sendo q algum número natural menor ou igual a n. Vale lembrar que cada elemento da base de (3.20) é uma matriz de mesma ordem que a matriz B. Essa estrutura dá origem ao que chamamos de base de Krylov por blocos [37].

Supondo que a matriz A é estável, a equação de Lyapunov (3.1) admite única solução que é dada pela expressão (1.22), cuja imagem é igual a imagem da matriz (1.14), como foi visto na seção 1.2. Esse fato é usado em [39] para justificar o uso da base (3.20) para construir aproximações para a solução da equação de Lyapunov (3.1). Nossa contribuição, nesse sentido, consiste em reforçar essa justificativa enunciando o Teorema 3.1, a seguir, que independe do conceito de controlabilidade de um sistema dinâmico.

Teorema 3.1 *Supondo que exista uma solução simétrica semi-definida positiva P para a equação de Lyapunov (3.1). Então, a imagem de P é um subespaço invariante por A.*

Demonstração. Suponhamos inicialmente que $ker(P) = \{0\}$. Nesse caso não há nada a provar pois, pelo Teorema do Núcleo e da Imagem, $Im(P) = \mathbb{R}^n$. Seja $x \neq 0$ um vetor que está no núcleo de P. Como P é simétrica semi-definida positiva, podemos decompor $P = LL^T$, para alguma matriz $L \in \mathbb{R}^{n\times n}$. Então,

$$x \in ker(P) \Leftrightarrow x \in ker(L^T).$$

Além disso, por (3.1)

$$\underbrace{ALL^Tx}_{=0} + LL^TA^Tx = -BB^Tx \implies LL^TA^Tx = -BB^Tx. \tag{3.21}$$

Disto segue que

$$0 = x^T L L^T A^T x = -x^T B B^T x,$$

ou seja, $B^T x = 0$. Assim, por (3.21), $LL^T A^T x = 0$. Então, $A^T x \in \ker(P)$, em outras palavras, $\ker(P)$ é invariante por A^T. Como $Im(P) \perp \ker(P)$, pode-se dizer que $Im(P)$ é invariante por A.

□

Note que, se A é estável, o corolário 1.27 garante que a solução P é única e, nesse caso, é dada pela expressão (1.22). A matriz (1.22) é, claramente, semi-definida positiva. Isso garante as hipóteses do Teorema 3.1.

3.2.1 Base de Krylov estendida

A fim de enriquecer a base (3.20) com mais informações, é sugerido em [39] iniciar a construção da base a partir do par $\{B, A^{-1}B\}$, gerando um subespaço que contenha informações de A e A^{-1}, adicionando vetores por meio de multiplicações por A e A^{-1}, da forma:

$$\mathbf{K}_l(A, B) = span\{B, A^{-1}B, AB, A^{-2}B, A^2B, A^{-3}B, ..., A^{l-2}B, A^{-(l-1)}B\}. \tag{3.22}$$

Perceba que, se considerarmos $\tilde{B} = A^{-(l-1)}B$ e reorganizarmos os termos em (3.22), teremos

$$\mathbf{K}_l(A, B) = \{\tilde{B}, A\tilde{B}, A^2\tilde{B}, ..., A^{l-1}\tilde{B}\}.$$

Portanto, o subespaço (3.22) tem a mesma estrutura do subespaço de Krylov convencional (3.20), diferindo apenas na escolha do bloco inicial. Além disso, a maneira como a base é construída proporciona a seguinte propriedade:

$$\mathbf{K}_l(A, B) \subset \mathbf{K}_{l+1}(A, B), \quad \forall l \in \mathbb{N}.$$

Uma vez escolhido o espaço de aproximação $\mathcal{V} = \mathbf{K}_l(A, B)$, para algum número natural l, constrói-se uma base ortogonal V_l para $\mathcal{V}$ por meio do método de Arnoldi por blocos [18]. Esse método aparece embutido no algoritmo 2, desenvolvido por Simoncini [39] e chamado de *Krylov-*

Plus-Inverted-Krylov (K-PIK).

Algoritmo 2: MÉTODO K-PIK

Entrada: Matrizes B, A, e $V_0 = 0$.

Saída: Aproximação X para a equação $AX + XA^T = -BB^T$

1 **início**

2 $V_1 \leftarrow$ ortogonalização (Gram-Schmidt modificado) [13]) de $[B, A^{-1}B]$

3 **para** *l=2,3,...* **faça**

4 $V_l \leftarrow [V_{l-1},\ V_l]$

5 $H_l \leftarrow V_l^T A V_l$, e $E \leftarrow V_l^T B$

6 Resolva $H_l Y + Y H_l^T + EE^T = 0$ e considere $Y_l = Y$

7 Se convergir, faça $P = V_l Y_l V_l^T$ e pare.

8 Do contrário, sejam $V_l^{(1)}$: primeiras s colunas de V_l e $V_l^{(2)}$: colunas restantes de$V_l^{(2)}$.

9 $V_{l+1} \leftarrow [AV_l^{(1)}, A^{-1}V_l^{(2)}]$

10 $\widehat{V}_{l+1} \leftarrow$ ortogonalização de V'_{l+1}com respeito a V_l

11 $V_{l+1} =$ ortogonalização (Gram-Shmidt modificado) de $\widehat{V}_{l+1}$.

12 **fim**

13 **fim**

14 **retorna** V_{l+1}

O critério de parada utilizado na linha de comando 6 do algoritmo 2 é definido por

$$\frac{\|AP_l + P_l A^T + BB^T\|_F}{2\|A\|_F\|Y_l\|_F + \|B\|_F^2} \leq tol, \tag{3.23}$$

com $P_l = V_l Y_l V_l^T$, $tol > 0$ sendo alguma tolerância dada dada. A motivação para essa escolha vem do fato que

$$\|AP + PA^T + BB^T\| \leq \|AP\| + \|PA^T\| + \|BB^T\| \leq 2\|A\|\|P\| + \|BB^T\|.$$

Mais detalhes sobre os motivos dessa escolha e também sobre como calcular a estimativa (3.23) de maneira prática durante a execução do método podem ser vistos em [39].

A equação de Lyapunov que aparece como um subproblema na linha **6** do algoritmo 2 pode ser resolvido com auxílio do pacote computacional LYAPACK, disponível em *https://www.tu-chemnitz.de/sfb393/lyapack/*.

Na linha 5 do algoritmo 2, em virtude da ortogonalidade da matriz V_l, a matriz H_l é "quase"Hessenberg superior. Além disso, é possível gerar essa matriz sem realizar os produtos $V_l^T A V_l$, utilizando de uma recorrência às entradas das matrizes da iteração anterior [39].

Assim como ocorre em métodos para sistemas lineares do tipo $Ax = b$, o algoritmo 2 tem terminação finita na aritmética exata. Contudo, como são adicionados dois blocos a cada iteração,

pode ocorrer de de um bloco ser múltiplo do outro. No entanto, a seguinte proposição mostra que, caso isso ocorra, implica na convergência do método.

Proposição 3.2 *[39] Assumamos que A é estável e que $l-1$ iterações do método K-PIK sejam realizadas, com $B \in \mathbb{R}^n$. Na l-ésima iteração, assumamos que $\widehat{V}_{l+1}$, na linha 10, possua posto menor do que dois. Então $V_l \cup \widehat{V}_{l+1}$ é um subespaço invariante por A com respeito a B. Portanto, $Im(V_l \cup \widehat{V}_{l+1})$ contém a solução exata.*

Demonstração. Para ver a demonstração dessa proposição, consulte [39]. □

A Proposição 3.2 considera B como sendo um vetor. Porém a demonstração é análoga para o caso em que B é uma matriz.

3.2.2 Método de Arnoldi para o Modelo Descritor

Nesta subseção estudamos meios de utilizar o algoritmo 2 em situações que envolvem o sistema (1.30).

A proposta inicial de Simoncini para adequar o método K-PIK a sistemas descritores depende da matriz E, dada em (1.30), ser não singular. Vamos buscar uma alternativa que não dependa dessa hipótese.

Nossa ideia consiste em considerar a equação de Lyapunov usual (3.1), obedecendo as relações (1.31). Resta então saber como calcular os blocos $\{AB, A^{-1}B, A^2B, A^{-2}B...\}$. Note que as potências inversas $\{A^{-1}B, A^{-2}B, A^{-3}B...\}$ podem ser calculadas pelas fórmulas dadas em (1.32) e (1.34) considerando o caso particular em que $\mu = 0$. Já as potências $\{AB, A^2B, ...\}$ podem ser obtidas através da sequência de passos descrita em (1.35).

3.2.3 Subespaços de Krylov Racionais

Esta seção tem por objetivo apresentar métodos para resolução da equação de Lyapunov (3.1) por meio de projeções em subespaços de Krylov racionais, originalmente proposto por [36], da forma

$$K_l(A, B, \mu) := \text{span}\left\{B, (A-\mu_1 I)^{-1}B, ..., \left(\prod_{j=1}^{l}(A-\mu_j I)^{-1}\right)B\right\}, \tag{3.24}$$

com $\mu = [\mu_1, \mu_2, ..., \mu_l]$ sendo um vetor de entradas complexas.

Os subespaços da forma (3.24) foram utilizados, inicialmente, em problemas de autovalores, com eficácia comprovada [36]. No entanto, estudos mais recentes mostram que tais subespaços podem ser muito eficientes para aproximar soluções de equações de Lyapunov [8].

Vamos agora descrever os passos da construção da base (3.24) e também dos "shifts"utilizados, seguindo os passos de [8]. A priori, vamos supor que o vetor $\mu = [\mu_1, \mu_2, ..., \mu_l]$ seja previamente fornecido, e seja $D_l = diag(\mu_2, ..., \mu_l)$. Uma base ortonormal pode ser gerada utilizando o método de Arnoldi padrão com Gram-Schmidt modificado e reortogonalização, partindo do bloco $(A -$

$\mu_1 I)^{-1}B$, e continuando com um novo "shift"a cada iteração. Denotemos por V_l a matriz cujas colunas são os vetores da base ortonormal calculada.

Para construir a matriz $V_l^T A V_l$ sem ter que realizar dois produtos de matrizes a cada iteração, pode ser utilizada uma relação de recorrência que utiliza apenas um produto matriz vetor a cada iteração. Mais detalhes podem ser vistos em [35] [8].

Os procedimentos para gerar a base e a matriz de representação de A no subespaço de Krylov racional é sumarizado no algoritmo 3. Esse método é conhecido como *Rational Krylov Subspace Method* (RKSM) e foi introduzido por Simoncini [8]. A priori, o método considera um vetor $B = b$, mas pode ser generalizado para estrutura de blocos de forma semelhante ao que foi feito no Algoritmo 2. Além disso, omitimos a parte do algoritmo que serve para aproximar a solução da equação de Lyapunov (3.1), pois ela é inteiramente análoga à etapa correspondente no algoritmo 2. O que muda aqui é o método de calcular a base V_l apenas.

Algoritmo 3: RKSM-Método baseado em Subespaços de Krylov Racionais para resolução da Equação de Lyapunov

Entrada: Matriz A, vetor b, escalares $\mu_0^{(1)}$ e $\mu_0^{(2)}$ e um número máximo de iterações l_{max}

Saída: Base de Krylov ortonormal $V_{l_{max}}$ e a matriz projetada H

início

- $\mu_1 = \mu_0^{(1)}$
- $\tilde{v}_1 \leftarrow (A - \mu_1 I)^{-1} b, \quad v_1 = \tilde{v}_1 / \|\tilde{v}_1\|$
- $\alpha \leftarrow v_1^* A v_1$
- $\mu_2 \leftarrow \texttt{newpole}(\alpha, \mu_0^{(1)}, \{\mu_0^{(1)}, \mu_0^{(2)}, \alpha\})$
- **para** *l=2,3,...,l_{max}* **faça**
  - $v_{l+1} \leftarrow (A - \mu_{l+1} I)^{-1} v_l$
  - Ortogonalize o vetor v_{l+1} com respeito a $v_1, v_2, ..., v_l$ e armazene a matriz H_l formada pelos coeficientes da ortogonalização (Gram-Schmidt Modificado)
  - $v_{l+1} \leftarrow \tilde{v}_l / \|\tilde{v}_l\|, \quad V_{l+1} \leftarrow [V_l, v_{l+1}]$
  - $g \leftarrow V_l^* A v_{l+1}$ e $D_l \leftarrow diag(\mu_1, \mu_2, ..., \mu_{l+1})$
  - $T_l \leftarrow (I_l + H_l D_l - g h_{l+1,l} e_l^T) H_l^{-1}$
  - Calcule os autovalores $\{\lambda_1(T_l), \lambda_2(T_l), ..., \lambda_l(T_l)\}$
  - $\mu_{l+2} \leftarrow \texttt{newpole}(\{\lambda_1(T_l), ..., \lambda_l(T_l)\}, \{\mu_1, ..., \mu_l\}, \{\mu_0^{(1)}, \mu_0^{(2)}, \mu_1, ..., \mu_{l+1}\})$
- **fim**

fim

retorna V_{l+1}

A função `newpole` é responsável por calcular um novo parâmetro μ_{l+1} a cada iteração. Sua construção, que pode ser vista em [8] é inspirada na estimativa de erro para sistemas do tipo $(A - \mu I)x = b$, apresentada em [7]. Os parâmetros iniciais $\mu_0^{(1)}$ e $\mu_0^{(2)}$ devem ser estimativas, em

valor absoluto, para o menor e para o maior autovalor de A, respectivamente, em módulo, ou seja

$$\mu_0^{(1)} \approx \min_{i=1,\ldots,n} |\lambda_i| \text{ e}$$
$$\mu_0^{(2)} \approx \max_{i=1,\ldots,n} |\lambda_i|.$$

Algoritmo 4: NEWPOLE-FUNÇÃO QUE CALCULA NOVO PARÂMETRO A PARTIR DOS DADOS DA ITERAÇÃO ANTERIOR

Entrada: $\{\lambda_1(T_l), ..., \lambda_l(T_l)\}$, $\{\mu_1, ..., \mu_l\}$ e um conjunto de valores $\{\eta_1, \eta_2, ..., \eta_l\}$

início
  para *j=1,2,...,l − 1* **faça**
    $\beta_j \leftarrow arg\max_{\beta\in[\eta_j,\eta_{j+1}]} \frac{1}{|r_l(\beta)|}$ $r_l(\beta) = \prod_{k=1}^{l} \frac{\beta-\lambda_k}{\beta-\mu_k}$
  fim
  $\mu_{l+2} \leftarrow arg\max_{j=1,2,\ldots,l-1} \frac{1}{|r_l(\beta_j)|}$
fim

Os algoritmos 3 e 4, da forma como estão dispostos, estão aptos a receber e calcular apenas parâmetros reais. No entanto, quando a matriz A não é simétrica, acaba sendo inevitável trabalhar com autovalores e parâmetros complexos. Para estes casos, a ideia apresentada em [8] consiste em duas possibilidades: considerar os parâmetros sempre reais, tomados sobre o intervalo da parte real dos autovalores espelhados; ou considerar parâmetros complexos, escolhidos sobre o envoltório convexo que contém os autovalores e os autovalores espelhados. Ambas as possibilidades tem seus prós e contras pois, por um lado, o uso de parâmetros reais facilita a resolução dos sistemas envolvidos em cada iteração, enquanto o uso de parâmetros complexos podem promover uma melhora significativa na convergência do método para alguns problemas [8].

Recentemente, foi demonstrado o Teorema 3.3 que fornece uma estimativa para a taxa de convergência do método RKSM aplicado a equações de Lyapunov [6].

Teorema 3.3 *Seja P a solução para (3.1), com A estável. Seja P_l a sequência de aproximações gerada pelo método RKSM aplicado à equações de Lyapunov (nos moldes do algoritmo 2). Então, existe uma constante α tal que*

$$\|P - P_l\|_F \leq \alpha \max_{z\in W(A)} \prod_{j=1}^{l} \frac{|z - \overline{\mu}_j|^2}{|z + \mu_j|^2} \|b\|_2^2, \tag{3.25}$$

em que $\mu_1, \mu_2, ...\mu_l$ são os parâmetros do método RKSM e $W(A)$ é o alcance numérico de A definido por:

$$W(A) = \left\{ \frac{x^*Ax}{x^*x} \mid x \in \mathbb{C}^n, x \neq 0 \right\}.$$

O Teorema 3.3 reforça a importância de computar bons parâmetros para o método RKSM e acreditamos que haja muito a se explorar nesse sentido. Nesse trabalho propomos uma modificação

na escolha dos parâmetros iniciais $\mu_0^{(1)}$ e $\mu_0^{(2)}$, restringindo o intervalo a apenas uma região do espectro que exerce uma certa dominância na solução P. Essa proposta será apresentada na subseção 3.3.5.

Um fato curioso é que, em [6], é verificada uma equivalência entre o método RKSM e o método ADI, desde que os parâmetros $\mu_1, \mu_2, ...$ de ambos os métodos sejam ótimos no sentido da expressão (3.17). Esse fato sugere que, uma vez encontrada uma boa estratégia para a busca de parâmetros do método RKSM, ela pode ser aplicada também para o método ADI, com a diferença que, no método ADI, as aproximações dos autovalores de A precisariam ser feitas num algoritmo a parte.

Vejamos agora como o método RKSM pode ser implementado em modelos descritores. Consideremos, novamente, o sistema (1.29), regido pelas relações (1.31). Durante a execução do método RKSM, é preciso resolver sistemas da forma $(A - \mu_j I)X = B$. Neste caso, basta recorrer ao sistemas (1.32) e (1.34). Vale lembrar que, durante a confecção desse trabalho, notamos que Hossain e Uddin já apresentaram uma ideia bastante similar em [45]. Nesse mesmo trabalho, são feitas comparações, através de testes numéricos, entre o método RKSM e o método SLRCF-ADI. Nessas comparações, o método RKSM parece levar uma vantagem significativa. Já nos resultados apresentados ao final do presente trabalho, o método baseado em iterações ADI calcula soluções mais precisas. Isso somado ao fato do método SLRCF-ADI possuir uma complexidade numérica menor, coloca o método baseado em iterações ADI em destaque nessa disputa.

O critério de parada para o método 3 pode ser estabelecido de maneira análoga ao que foi feito para o método de Krylov estendido para problemas descritores.

3.3 Solução Explícita em Função da Decomposição Espectral da Matriz A

Nosso objetivo nesta seção é estudar as relações entre a solução P de (3.1) e o espectro de A. Mais especificamente, queremos classificar os autovalores de A de acordo com o nível de significância do subespaço invariante associado a esses autovalores da solução P, donde surge o termo *dominância*. A inspiração para esta análise é proveniente do estudo de redução modal, apresentado no capítulo anterior, combinado com o método de redução de modelo por balanceamento, apresentado na seção 2.3.

Na seção 2.2 foi definido que um autovalor λ_i de A (ou polo da função $H(s)$) é dominante quando a razão $|R_i/Re(\lambda_i)|$ é relativamente elevada. Foi visto também que o resíduo R_i é dado por $R_i = C^T v_i w_i^H B$, com v_i e w_i sendo autovetores à direita e à esquerda de A, respectivamente, associados ao autovalor λ_i. Por outro lado, a redução por balanceamento apresentada no capítulo 2.3 leva em consideração os valores principais de maior magnitude do produto de matrizes PQ. Do ponto de vista da função de transferência $H(s) = C^T(A - Is)^{-1}B$, pode-se dizer também que a norma da matriz P representa a energia adquirida na entrada do sistema dinâmico (A, B, C), ou

seja, está relacionada à transformação $(A - Is)^{-1}B$, enquanto a norma de Q representa a energia obtida na saída do sistema e está relacionada a $C^T(A - Is)^{-1}$. Portanto, é natural pensar que a influência de um autopar (λ_i, w_i^H) de A na solução P de (3.1) está associado, de alguma forma, à magnitude de $Re(\lambda_i)$ e de $w_i^H B$.

Nesta seção, vamos assumir que a matriz $A = V\Lambda W^H$, com $\Lambda = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_n\}$. As colunas das matrizes $V = [v_1|v_2|...|v_n]$ e $W = [w_1|w_2|...|w_n]$ são autovetores à direita e à esquerda de A, respectivamente, tais que $WV^H = W^H V = I$.

3.3.1 Solução explícita dada por similaridade

Em [40], é relatado histórico sobre métodos computacionais para equações de Lyapunov. Um deles consiste em construir a solução P por meio de similaridade, considerando-se $\tilde{B} = -W^H BB^T V$ a fim de se obter:

$$P = V\tilde{P}V^H, \quad \text{com} \quad \tilde{P}_{ij} = \frac{\tilde{B}_{ij}}{\lambda_i + \lambda_j^*} \tag{3.26}$$

O Teorema a seguir fornece uma formulação equivalente a (3.26), apenas com uma terminologia diferente que permite analisar com mais clareza algumas características da solução P.

Teorema 3.4 *A solução P da equação (3.1), com $B \in \mathbb{R}^{n\times 1}$, é dada por:*

$$P = -\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^H BB^T w_j^H}{\lambda_i + \overline{\lambda_j}} v_i v_j^H. \tag{3.27}$$

ou, na forma matricial,

$$P = -VXC^{\Lambda}X^H V^H, \tag{3.28}$$

com $X = diag(w_1^H B, w_i^H B, ..., w_n^H B)$ e C^{Λ} sendo uma matriz de Cauchy tal que $C_{ij}^{\Lambda} := \frac{1}{\lambda_i + \overline{\lambda_j}}$.

Demonstração. Substituindo a expressão (3.27) no lado esquerdo da equação (3.1), obtemos

$$\begin{aligned}
&-A\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^H BB^T w_j}{\lambda_i + \overline{\lambda_j}} v_i v_j^H - \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^H BB^T w_j}{\lambda_i + \overline{\lambda_j}} v_i v_j^H \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\lambda_i w_i^H BB^T w_j}{\lambda_i + \overline{\lambda_j}} v_i v_j^H - \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^H BB^T w_j \overline{\lambda_j}}{\lambda_i + \overline{\lambda_j}} v_i v_j^H \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} (w_i^H BB^T w_j) v_i v_j^H = -VW^H BB^T WV^H = -BB^T.
\end{aligned}$$

A verificação de (3.28) consiste apenas em desenvolver o produto de matrizes e por isso é deixada a cargo do leitor. □

Perceba que, se desenvolvermos o produto $XC^{\Lambda}X^H$, deixaremos a expressão (3.28) idêntica a (3.26).

O corolário a seguir contém uma versão da solução (3.27) para o caso em que B possui mais de uma coluna.

Corolário 3.5 *Nas hipóteses do teorema 3.4, considere $B \in \mathbb{R}^{n\times m}$, com $m \leq n$, e $\mathbf{X} := W^H B$. Então, a solução da equação (3.1) é dada por*

$$P = -\sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\mathbf{x}_{il}\mathbf{x}_{jl}^H}{\lambda_i + \overline{\lambda_j}} v_i v_j^H, \tag{3.29}$$

em que $\mathbf{x}_{jl}$ indica as entradas de $\mathbf{X}$. Ou, de maneira análoga,

$$P = \sum_{l=1}^{m} -V\mathbf{X}_l C^{\Lambda}\mathbf{X}_l^H V^H, \tag{3.30}$$

com $\mathbf{X}_l$ sendo uma matriz diagonal, cujas entradas da diagonal são os elementos da l-esima coluna de $\mathbf{X}$.

Demonstração. Supondo que $\{b_1, b_2, ..., b_m\}$ sejam as colunas da matriz B, então

$$BB^T = b_1 b_1^T + b_2 b_2^T + ... + b_m b_m^T.$$

Sendo assim, fazendo uso da linearidade da aplicação $AP + PA^T$ com relação a P, basta aplicar o teorema 3.4 nos termos $b_1 b_1^T$, $b_1 b_1^T$,... e $b_m b_m^T$ separadamente para obter o resultado. □

Um fato curioso é que a construção da solução (3.29) fornece uma demonstração alternativa para um fato conhecido sobre matrizes de Cauchy, que apresentamos no Teorema 3.6 a seguir. Uma das demonstrações já existentes pode ser vista em [10].

Teorema 3.6 *Sejam $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ e $\{x_1, x_2, ..., x_n\}$ dois conjuntos de números complexos e C uma matriz de Cauchy generalizada simétrica, cujas entradas são da forma*

$$C_{ij} = \frac{x_i\overline{x_j}}{\lambda_i + \overline{\lambda_j}}.$$

Se $Re(\lambda_i) < 0$, para $i = 1, 2, ..., n$, então C é negativa definida.

Demonstração. Como $Re(\lambda_i) < 0$, para $i = 1, 2, ..., n$, pelo Corolário 1.27, a equação de Lyapunov $\Lambda P + P\Lambda^H = -BB^T$, com $B = diag(1, 1, ..., 1) \in \mathbb{R}^{n\times n}$, possui única solução. Nesse caso, por 3.30 essa solução é dada por $P = \sum_{l=1}^{n}\sum_{m=1}^{n} -e_l C^{\Lambda} e_m^T = -C^{\Lambda}$, com $C_{ij}^{\Lambda} = \frac{1}{\lambda_i + \overline{\lambda_j}}$ e e_m indicando o m-ésimo vetor canônico. Por outro lado, pelo Teorema 1.23, a solução P também pode ser definida por

$$P = \int_0^{\infty} e^{\Lambda\tau} BB^T e^{\Lambda^H\tau}\, d\tau,$$

que é uma matriz positiva definida. Portanto, C^{Λ} é negativa definida. Note que $C = YC^{\Lambda}Y$, com $Y = diag(x_1, x_2, ..., x_n)^H$. Então, C também é negativa definida. □

Obs.: Note que a demonstração do Teorema 3.6 referencia o Corolário 1.27 e o Teorema 1.23. Embora os resultados estejam enunciados apenas para o caso real, as mesmas demonstrações podem ser facilmente estendidas para o caso complexo.

3.3.2 Autovalores de A dominantes em P

Nesta seção, o objetivo é aproximar a solução P utilizando informações apenas do subespaço invariante por A que possui maior influência na soma (3.27). Inicialmente, vamos considerar $B \in \mathbb{R}^{n\times 1}$ e definir $X := \text{diag}(w_1^H B, w_i^H B, ..., w_n^H B)$. Uma aproximação para P pode ser definida da forma

$$P_k := -\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{w_i^H B B^T w_j}{\lambda_i + \overline{\lambda_j}} v_i v_j^H, \tag{3.31}$$

ou, na forma matricial,

$$P_k = -V_k X_k C_k^{\Lambda} X_k^H V^H, \tag{3.32}$$

com $k << n$, $X_k = \text{diag}(w_1^H B, w_i^H B, ..., w_k^H B)$, $V = [v_1|v_2|...|v_k]$ e C_k^{Λ} sendo a matriz de Cauchy construída a partir de um subconjunto $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ do espectro de A. Note que P_k é uma projeção da solução P no espaço invariante por A gerado por $\left\{v_i v_j^H\right\}_{i,j=1}^{k}$.

Na literatura existem tentativas de relacionar o decaimento dos autovalores de P com características dos autovalores de A. Em [2], por exemplo, considera-se a decomposição $C^{\Lambda} = LDL^H$, em que $D = \text{diag}\{\delta_1, \delta_2, ..., \delta_n\}$, com $\delta_1 \geq \delta_2 \geq ... \geq \delta_n$, e L é uma matriz triangular inferior que possui somente "uns"na diagonal. Cada coluna Le_j de L satisfaz $\|Le_j\|_\infty = 1$. A partir disso, pode ser calculada uma aproximação de (3.26) conforme é dada no teorema a seguir, com $z_i = XVLe_i$, para $i = 1, 2, ..., n$.

Teorema 3.7 *Sejam $\delta_1, \delta_2, ..., \delta_n$ tais que*

$$\delta_k = \max_{k\leq i,j\leq n} \left\{ \frac{-1}{2Real(\lambda_k)} \prod_{i=1}^{k-1} \left| \frac{\lambda_k - \lambda_i}{\overline{\lambda_k} + \lambda_i} \right|^2 \right\}, \quad k = 1, 2, ..., n. \tag{3.33}$$

Se

$$P_k = \sum_{i=1}^{k} \delta_i z_i z_i^H,$$

então

$$\|P - P_k\| \leq (n-k)^2 \delta_{k+1} (\kappa_2(XV)\|b\|_2)^2, \tag{3.34}$$

com κ_2 sendo o número de condição perante a norma euclidiana e $X := diag(w_1^H B, w_i^H B, ..., w_n^H B)$.

Demonstração. A demonstração desse teorema pode ser vista em [2]. □

A estimativa (3.34) sugere que, caso o produto XV seja bem condicionado, é possível obter uma boa aproximação para a solução P considerando apenas a estrutura matricial de (3.26) correspondente aos maiores valores de δ_i, com $i = 1, 2, .., n$.

A desvantagem dessa abordagem é que, em geral, não é possível garantir que o produto XV seja bem condicionado. Note que a estrutura da matriz XV está intimamente relacionada aos efeitos que B proporciona a P.

Penzl mostra em [31] que é impossível prever o decaimento da solução de (3.1) com estimativas que não dependem de B e que levam em consideração apenas os autovalores de A.

Teorema 3.8 *Seja $A \in \mathbb{R}^{n\times n}$ uma matriz cuja parte real de cada um de seus autovalores é negativa e $B \in \mathbb{R}^n$. Suponha que o par (A, B) seja observável e considere os valores reais $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n > 0$. Então existe alguma matriz $\widehat{A}$, similar a A, e um vetor $\widehat{B} \in \mathbb{R}^{\kappa}$ tais que a solução $\widehat{P}$ da equação de Lyapunov*

$$\widehat{A}\widehat{P} + \widehat{P}\widehat{A}^T = \widehat{B}\widehat{B}^T$$

possui os valores principais $\sigma_1, \sigma_2, ..., \sigma_n$.

Demonstração. Para ver a demonstração desse teorema, consulte [31]. □

Exemplo 3.9 *Considere*

$$A = \begin{bmatrix} -2 & 0 \\ 0 & -10 \end{bmatrix} \quad e \quad B = \begin{bmatrix} 1 \\ 100 \end{bmatrix}.$$

A solução para a equação $AP + PA^T = -BB^T$ é

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{20} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix} = \frac{1}{4}e_1e_1^T + \frac{100}{12}e_2e_1^T + \frac{100}{12}e_1e_2^T + 500e_2e_2^T,$$

que possui valores singulares $\sigma_1 \approx 500,14$ e $\sigma_2 \approx 0,11$. De acordo com a estimativa (3.33), a melhor aproximação de posto um para P, dada pela expressão (3.31), seria

$$P_1 = \frac{1}{4}e_1e_1^T.$$

No entanto, P_1 possui valor principal $\sigma = \frac{1}{4}$, enquanto o valor principal de $P_4 = 100$ é $\sigma = 500$.

Ao invés de utilizar decomposições de matrizes, decidimos partir para uma análise direta dos termos que aparecem na soma (3.27). Vamos supor que $\|v_i\|_2 = 1$, para $i = 1, 2, ..., n$, em (3.27). Então,

$$\|P\|_2 = \left\| \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^H BB^T w_j^H}{\lambda_i + \overline{\lambda_j}} v_i v_j^H \right\|_2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n} \left| \frac{w_i^H BB^T w_j^H}{\lambda_i + \overline{\lambda_j}} \right|$$

A partir daí, queremos determinar quais autopares de A são responsáveis pelos maiores coeficientes $\left|\frac{w_i^H BB^T w_j}{\lambda_i + \overline{\lambda_j}}\right|$.

Intuitivamente, é natural pensar que os autovalores de A com parte real mais próxima de zero tem mais potencial de influenciar na solução P pois estão próximos de causar singularidade nos termos da diagonal da matriz de Cauchy C^{Λ}. Para buscar indícios mais concretos dessa afirmação, vamos nos ater, inicialmente, ao caso particular em que os autovalores de A são todos reais. Vamos denotar as entradas de C^{Λ} por $c^{\Lambda}_{i,j}$, para $i, j = 1, 2, ..., n$.

Lema 3.10 *Sejam* $\{\lambda_1, \lambda_2, ...\lambda_n\}$ *autovalores de A, todos reais negativos e satisfazendo* $|\lambda_1| \leq |\lambda_2| \leq ... \leq |\lambda_n|$. *Seja ainda* $\epsilon_k = \max\{\{|c^{\Lambda}_{k,j}|\}_{j=1}^{k}, \{|c^{\Lambda}_{i,k}|\}_{i=1}^{k}\}$. *Então, para cada* $k = 1, 2, ..., n$,

$$|c^{\Lambda}_{i,j}| \leq \epsilon_k,$$

sempre que $i \geq k$ *ou* $j \geq k$.

Demonstração. Uma vez que $|c^{\Lambda}_{i,i}| = \frac{1}{2|\lambda_i|}$, então os valores absolutos das entradas da diagonal da matriz C^{Λ} estão dispostos em ordem decrescente, ou seja, $|c^{\Lambda}_{i,i}| \geq |c^{\Lambda}_{i+1,i+1}|$ para todo $i = 1, 2, ..., n$. Perceba ainda que, dado um $k \in \mathbb{N}$, para as colunas e linhas de C^{Λ}, respectivamente, são válidas as seguintes relações:

$$|c^{\Lambda}_{i,j}| = \left|\frac{1}{\lambda_i + \lambda_j}\right| \leq \left|\frac{1}{\lambda_k + \lambda_j}\right| = |c^{\Lambda}_{k,j}|, \quad \forall\ i > k, ,$$

$$|c^{\Lambda}_{i,j}| = \left|\frac{1}{\Lambda_i + \lambda_j}\right| \leq \left|\frac{1}{\lambda_i + \lambda_k}\right| = |c^{\Lambda}_{i,k}|, \quad \forall\ j > k,$$

Como $\epsilon_k = \max\{\{|c^{\Lambda}_{k,j}|\}_{j=1}^{k}, \{|c^{\Lambda}_{i,k}|\}_{i=1}^{k}\}$, segue que $|c^{\Lambda}_{i,j}| \leq \epsilon_k$, sempre que $i \geq k$ ou $j \geq k$. □

A figura 3.1 ilustra o lema 3.10. O tamanho do símbolo "+"simboliza a magnitude da entrada c^{Λ}_{ij}.

Figura 3.1: Magnitude das dentradas de C^{Λ}

$$\begin{bmatrix} + & + & + & + & + & + \\ + & + & + & + & + & + \\ + & + & + & + & + & + \\ + & + & + & + & + & + \\ + & + & + & + & + & + \\ + & + & + & + & + & + \end{bmatrix}$$

No caso mais geral, em que os autovalores de A são complexos, vamos supor que exista uma constante $\beta \geq 1$ que limita o amortecimento dos autovalores, ou seja, se cada autovalor é da forma $\lambda_i = a_i + b_i j$, com $a, b \in \mathbb{R}$ e $j = \sqrt{-1}$, então

$$|b_i| \leq \beta |a_i| \quad \forall\ i = 1, 2, ..., n. \tag{3.35}$$

Chamaremos β de *constante de amortecimento* dos autovalores de A. Para essa situação, enunciamos o seguinte resultado:

Lema 3.11 *Sejam $\{\lambda_1, \lambda_2, ...\lambda_n\}$ autovalores de A, cujas partes reais são todas negativas e satisfazem $Re(\lambda_1) \geq Re(\lambda_2) \geq ... \geq Re(\lambda_n)$. Seja ainda*

$$\epsilon_k = max\{\{|Re(c^{\Lambda}_{k,j})|\}^k_{j=1}, \{|Re(c^{\Lambda}_{i,k})|\}^k_{i=1}\},$$

para algum k natural, e β a constante de amortecimentoo, conforme (3.35). Então, para cada $k = 1, 2, ..., n$, sempre que $i \geq k$ ou $j \geq k$, é válido que

$$|c^{\Lambda}_{i,j}| \leq (1 + \beta)\epsilon_k.$$

Demonstração. Perceba que, para cada $i = 1, 2, ..., n$, $|c^{\Lambda}_{i,i}| = \frac{1}{2|\mathrm{Re}(\lambda_i)|}$. Portanto, analogamente ao início da demonstração do Lema 3.10, verifica-se a mesma ordem decrescente dos elementos da diagonal de C^{Λ}. Agora, para algum k fixado, vamos considerar as entradas $c^{\Lambda}_{k,j}$ e $c^{\Lambda}_{i,j}$, com $j \leq k$ e $i \geq k$. Vamos considerar as representações retangulares:

$$\begin{aligned} \lambda_k &= a_k + b_k j, \\ \lambda_i &= a_i + b_i j, \\ \lambda_j &= a_j + b_j j. \end{aligned}$$

É importante ressaltar que, nesse caso, são atribuídos dois significados para a notação j, pois quando ele não aparece como sub-índice, denota o número imaginário $\sqrt{-1}$. Por hipótese temos que $(a_k + a_j)^2 \leq (a_i + a_j)^2$. Além disso, como $(b_k - b_j)^2 \leq \beta^2(a_k + a_j)^2$, então $(b_k - b_j)^2 \leq \beta^2(a_k + a_j)^2 + (b_i - b_j)^2$. Destas observações segue que

$$\frac{1}{(a_i + a_j)^2 + (b_i - b_j)^2 + \beta^2(a_k + a_j)^2} \leq \frac{1}{(a_k + a_j)^2 + (b_k - b_j)^2} = |c^{\Lambda}_{k,j}|. \tag{3.36}$$

Uma vez que $(a_k + a_j)^2 \leq (b_i - b_j)^2 + (a_i + a_j)^2$, então

$$\begin{aligned} \frac{\frac{1}{(a_i + a_j)^2 + (b_i - b_j)^2 + \beta^2(a_k + a_j)^2}}{|c^{\Lambda}_{ij}|} &= \frac{(a_i + a_j)^2 + (b_i - b_j)^2}{(a_i + a_j)^2 + (b_i - b_j)^2 + \beta^2(a_k + a_j)^2} \\ &\geq \frac{(a_k + a_j)^2}{(a_k + a_j)^2 + \beta^2(a_k + a_j)^2} = \frac{1}{1 + \beta} \end{aligned} \tag{3.37}$$

. De (3.36) e (3.37) segue que

$$|c^{\Lambda}_{ij}| \leq (1 + \beta)|c^{\Lambda}_{kj}| \leq (1 + \beta)|\mathrm{Re}(c^{\Lambda}_{kj})| \leq (1 + \beta)\epsilon_k.$$

Um raciocínio inteiramente análogo pode ser utilizado para verificar a desigualdade para as linhas. Com isto concluímos que

$$|c^{\Lambda}_{i,j}| \leq (1+\beta)\epsilon_k,$$

sempre que $i \geq k$ ou $j \geq k$. □

Deste resultado, segue diretamente o corolário a seguir:

Corolário 3.12 *Nas hipóteses do Lema 3.11, se*

$$P_k = -\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{w_i^H B B^T w_j}{\lambda_i + \overline{\lambda_j}} u_i u_j^H,$$

com $k < n$, então

$$\|P - P_k\|_2 \leq (n^2 - k^2)(1+\beta)\epsilon_k \xi_k, \tag{3.38}$$

com

$$\xi_k = \max_{i=1,\ldots,n,\ j=k,\ldots,n} |w_i^H B B w_j| \tag{3.39}$$

.

Baseando-se nas explanações feitas até o momento, vamos enunciar a seguinte definição:

Definição 3.13 *Seja λ_i um autovalor de A associado aos autovetores w_i^H e v_i à direita e à esquerda respectivamente, com $w_i^H v_i = 1$ e $\|v_i\| = 1$. Dizemos que λ_i é dominante na solução P da equação* (3.1) *se*

$$\delta_i := \left|\frac{(w_i^H B)^2}{2Re(\lambda_i)}\right| \geq \left|\frac{(w_j^H B)^2}{2Re(\lambda_j)}\right|, \tag{3.40}$$

para todo $j \neq i$.

A escolha da expressão (3.40) visa incorporar a grandeza dos valores que compõem (3.39) à análise feita, a priori, sobre a parte real dos autovalores de A. Vale lembrar também que os termos da forma (3.13) são exatamente os elementos da diagonal da matriz de Cauchy generalizada $XC^{\lambda}X^H$. Note que, se o objetivo for obter uma projeção de P num subespaço invariante por A de dimensão 1 (ou 2, no caso complexo conjugado), a definição 3.13 fornece o autovalor ótimo para essa situação.

Se escolhermos $\lambda_1, \lambda_2, ..., \lambda_k$, com $2 < k < n$ tais que

$$\left|\frac{(w_i^H B)^2}{2Re(\lambda_i)}\right| \geq \left|\frac{(w_j^H B)^2}{2Re(\lambda_j)}\right|,$$

para todo $j > k$, não há garantia de que a projeção P_k no subespaço associado a esses autovalores seja ótima. No entanto, alguns resultados numéricos apresentados na subseção a seguir mostram que esse subespaço invariante tem uma forte predominância na solução P.

Obs. 1: Se considerarmos $k << n$ em problemas de grande porte, o número $(n^2 - k^2)$ torna-se relativamente elevado, fazendo com que a estimativa (3.38), por si só, não forneça muita precisão. No entanto, há um outro fenômeno envolvido nessa situação que devemos considerar. Nos casos em que o valor de β é suficientemente pequeno, suponha que dois valores consecutivos $\left|\frac{(w_i^H B)^2}{2Re(\lambda_i)}\right|$ e $\left|\frac{(w_{i+1}^H B)^2}{2Re(\lambda_{i+1})}\right|$ sejam significativamente grandes. Se os autovalores λ_i e λ_{i+1} estiverem próximos um do outro, as duas linhas (e colunas) correspondentes a esses dois parâmetros na matriz $XC^{\Lambda}X$ estão próximas de serem múltiplas entre si. Portanto, para selecionar um conjunto de autovalores dominantes, além da magnitude dos valores (3.40), faz sentido considerar a diferença entre os autovalores. Isso significa que, se os autovalores estiverem dispostos em "clusters", será necessário um número pequeno de autovalores dominantes para uma boa aproximação de P.

Obs. 2: Com base na expressão (3.29) definição 3.13 pode ser facilmente estendida para o caso em que $B \in \mathbb{R}^{n\times m}$, com $m > 1$, substituindo δ_i por

$$\Delta_i = \frac{\|w_i^H B\|_2^2}{|2Re(\lambda_i)|}.$$

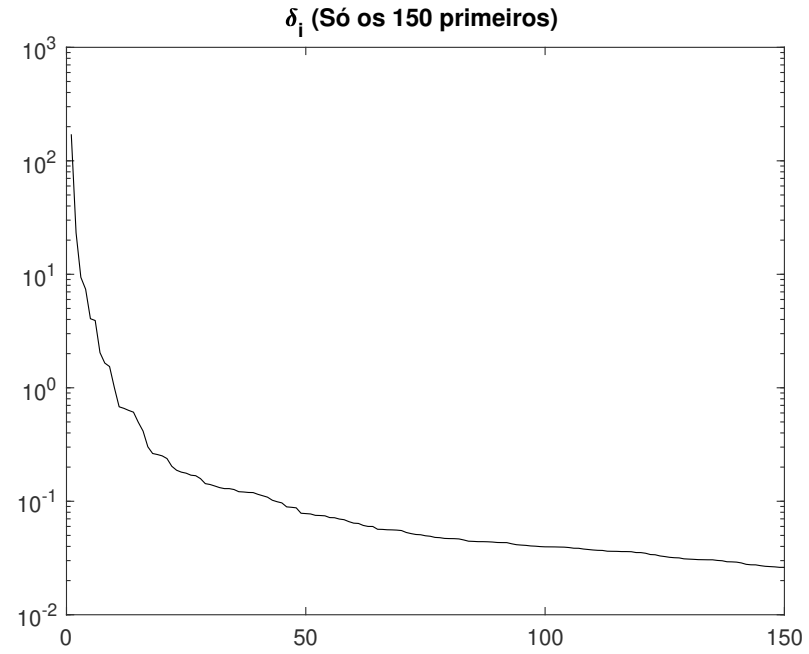
3.3.3 Exemplos numéricos

Num primeiro exemplo, vamos considerar a equação de Lyapunov 3.1 com $A = T \otimes I + I \otimes T$ de ordem n^2 e T sendo uma matriz tridiagonal de ordem n dada por

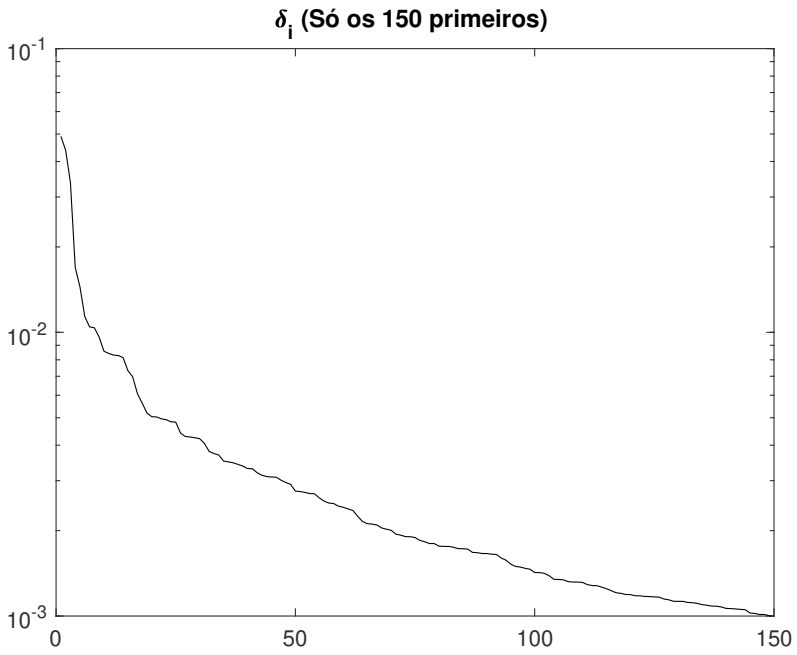
$$T = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \tag{3.41}$$

Nessas condições, a matriz A possui autovalores todos reais, distribuídos de maneira quase uniforme no intervalo $[0, 8]$. Para $n^2 = 900$ tem-se $\min_{i=1,..n}\{|\lambda_i|\} \approx -0.02$ e $\max_{i=1,2,...,n}\{|\lambda_i|\} \approx -7.98$.

Os gráficos da figura 3.2 apresentam os valores δ_i, da definição 3.13, em ordem decrescente para o $B = \texttt{rand}(n^2, 1)$ (o termo "`rand`"refere-se ao comando do Matlab que gera matrizes randômicas) e também para o caso em que B é um vetor esparso com apenas três entradas não nulas, iguais a 1.



(a) Com B randômico

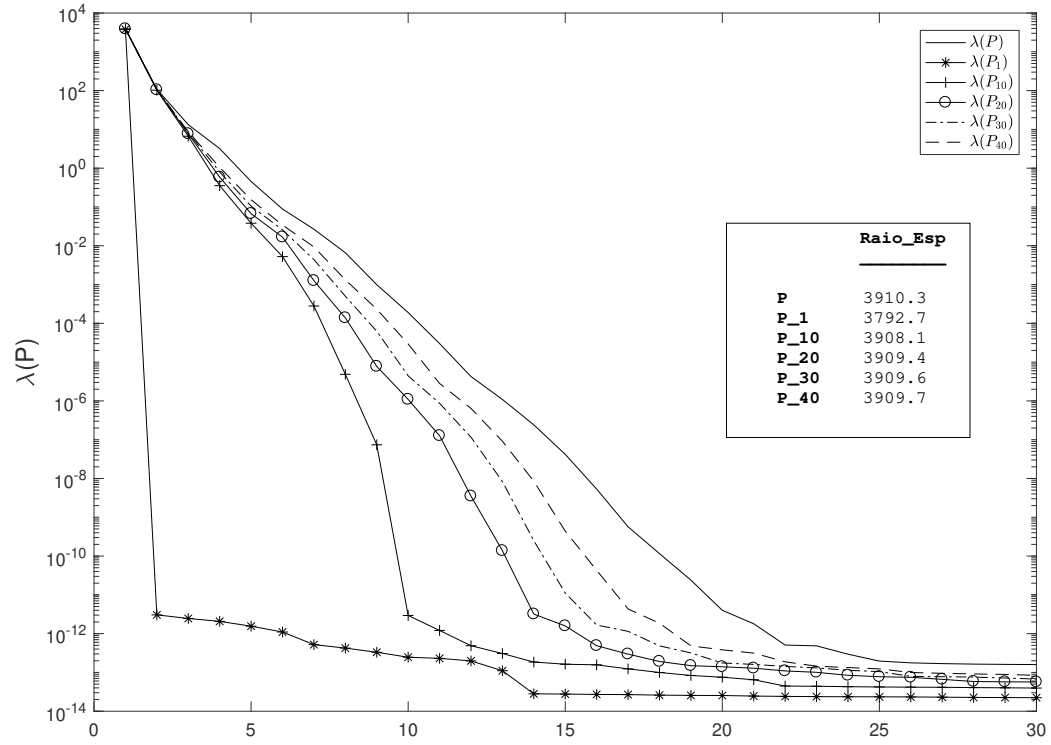


(b) Com B esparso

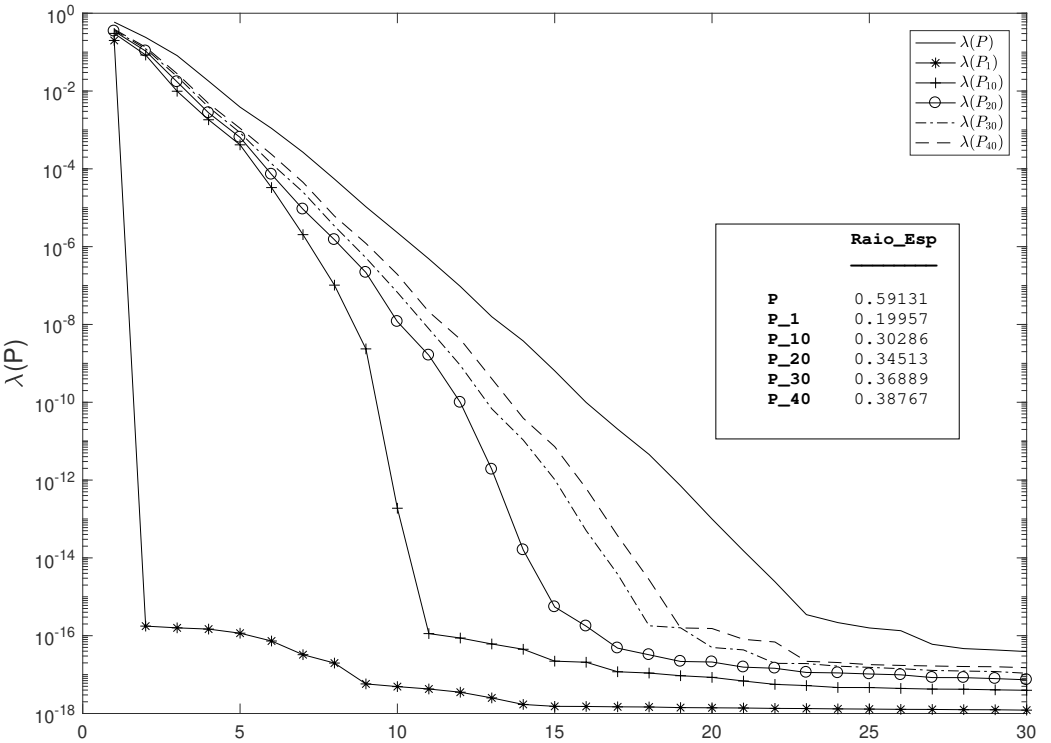
Figura 3.2: Valores δ do critério de dominância definido em (3.13).

Nos dois exemplos escolhidos existem algumas dezenas de autovalores apenas que se destacam por ter dominância δ_i bem maior que os demais autovalores. Os gráficos da figura 3.3 apresentam um comparativo entre o espectro de cada aproximação P_k e o espectro de P. O número k determina

a quantidade de autopares de A considerados na aproximação da solução dada por (3.31) seguindo o critério da definição 3.13. Como as linhas dos gráficos não deixam bem visível a diferença entre os maiores autovalores, optamos por exibir uma tabela na própria figura, contendo o maior autovalor de cada P_k.



(a) Com B randômico

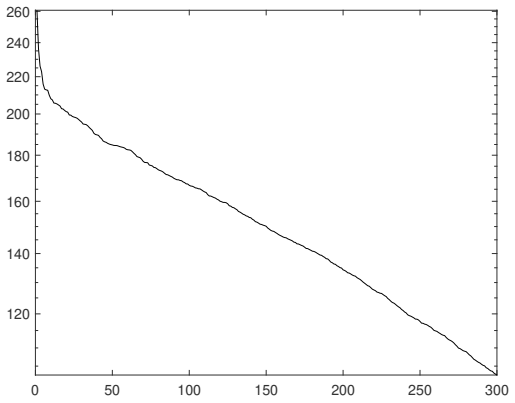


(b) Com B esparso

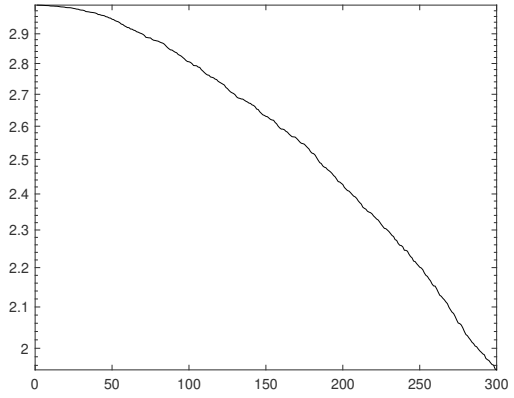
Figura 3.3: Autovalores das aproximações P_k em comparação com os autovalores de P

Observando os gráficos da figura 3.3, percebemos que, nos exemplos escolhidos, o decaimento do espectro das aproximações P_k se aproximam rapidamente do decaimento dos autovalores de P, como era previsto pela dominância exibida na figura 3.2.

Nos gráficos da figura 3.4 é possível visualizar o comportamento do erro em termos da equação de Lyapunov, bem como o decaimento do erro relativo entre as aproximações P_k.

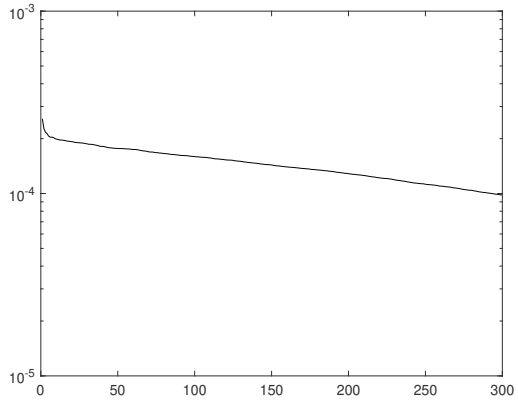


(a) Norma absoluta $\|AP_k + P_kA^T + BB^T\|_F$

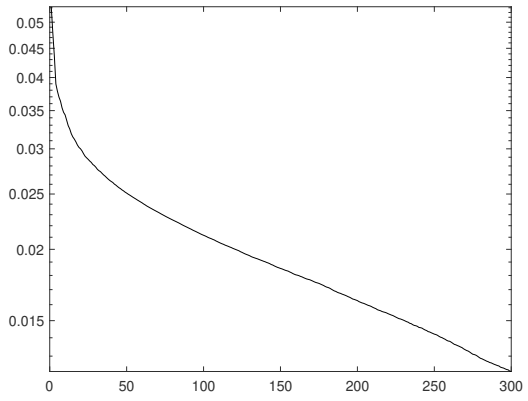


(b) Norma absoluta $\|AP_k + P_kA^T + BB^T\|_F$, com B esparso.

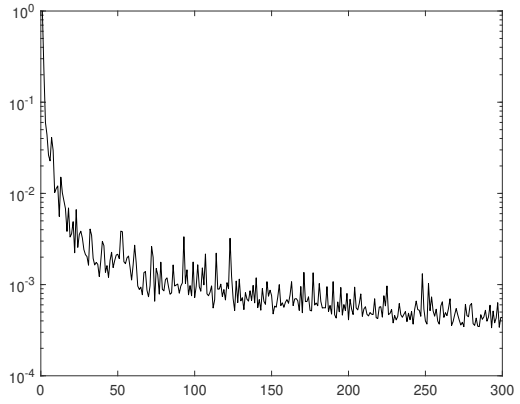
Figura 3.4: Erro perante a equação de Lyapunov e decaimento da diferença relativa entre as aproximações P_k



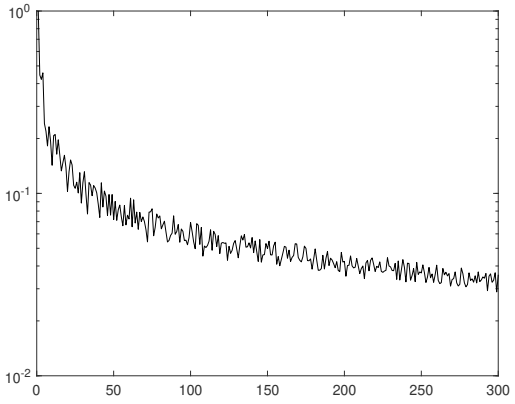
(a) Norma relativa $\frac{\|AP_k+P_kA^T+BB^T\|_F}{\|A\|_F\|P_k\|_F+\|BB^T\|_F}$.



(b) Norma relativa $\frac{\|AP_k+P_kA^T+BB^T\|_F}{\|A\|_F\|P_k\|_F+\|BB^T\|_F}$, com B esparso.



(a) Norma relativa $\frac{\|P-P_k\|_F}{\|P_{k-1}\|_F}$.



(b) Norma relativa $\frac{\|P-P_k\|_F}{\|P_{k-1}\|_F}$, com B esparso.

Note que o decaimento do erro ocorre de maneira brusca à medida que os primeiros polos dominantes são utilizados na aproximação P_k, tendendo a estabilizar-se na sequência.

Vamos considerar agora o modelo `brasilsemtcsc`, exibido de maneira detalhada na subseção 1.3.1. Para esse caso específico, o gráfico da figura 3.7 contém os valores δ_i, definidos em (3.13), dispostos em ordem decrescente.

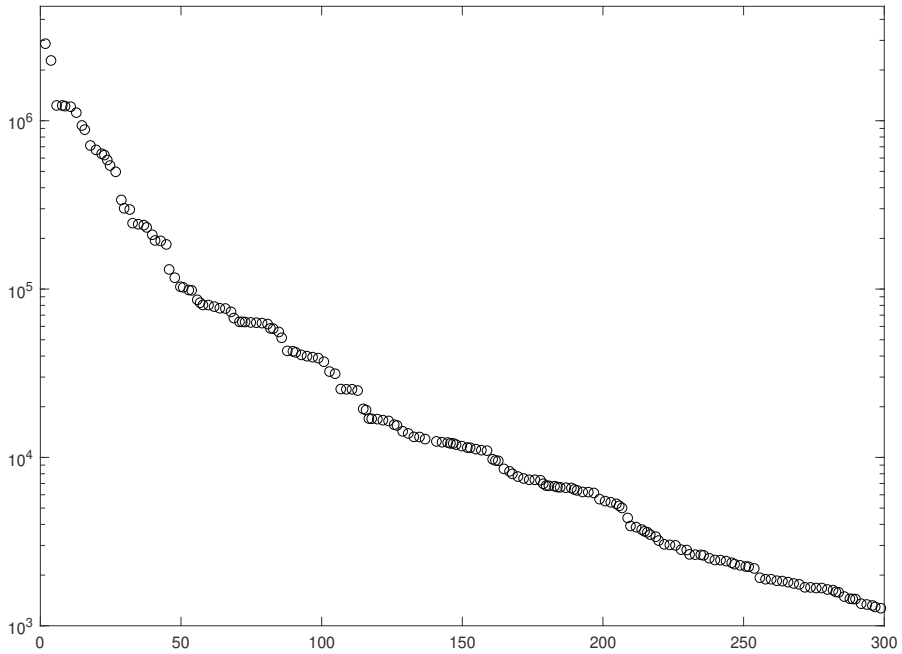


Figura 3.7: Valores δ_i do critério de dominância 3.13 dos autovalores de A na solução P.

Com base no gráfico da figura 3.7 podemos pressupor que as primeiras dezenas de autopares dominantes de A são muito mais significantes na solução P que os demais. Esse fato pode ser verificado na figura 3.8, na qual são plotados os autovalores da solução P da equação de Lyapunov

$AP + PA^T = -BB^T$, construída a partir de todos os autopares da matriz A conforme a expressão (3.27), juntamente com os autovalores das aproximações P_k definidas pela expressão (3.31). A título de curiosidade, exibimos na tabela 3.1 os 10 autovalores que apresentam maior dominância δ_i.

Tabela 3.1: Autovalores com maior dominância δ_i (Def. 3.13)

Real	Imag.	Dominância δ_i
-34,2775	0,0971	$2{,}845 \cdot 10^6$
-6,9172	3,2292	$2{,}262 \cdot 10^6$
-30,5425	9,8183	$1{,}222 \cdot 10^6$
-61,9642	4,5734	$1{,}221 \cdot 10^6$
-75,5448	–	$1{,}210 \cdot 10^6$
-1,4790	8,2551	$1{,}201 \cdot 10^6$
-5,8148	4,8704	$1{,}108 \cdot 10^6$
-11,9414	0,3932	$0{,}929 \cdot 10^6$
-12,2546	–	$0{,}878 \cdot 10^6$
-0,0335	1,0787	$0{,}707 \cdot 10^6$

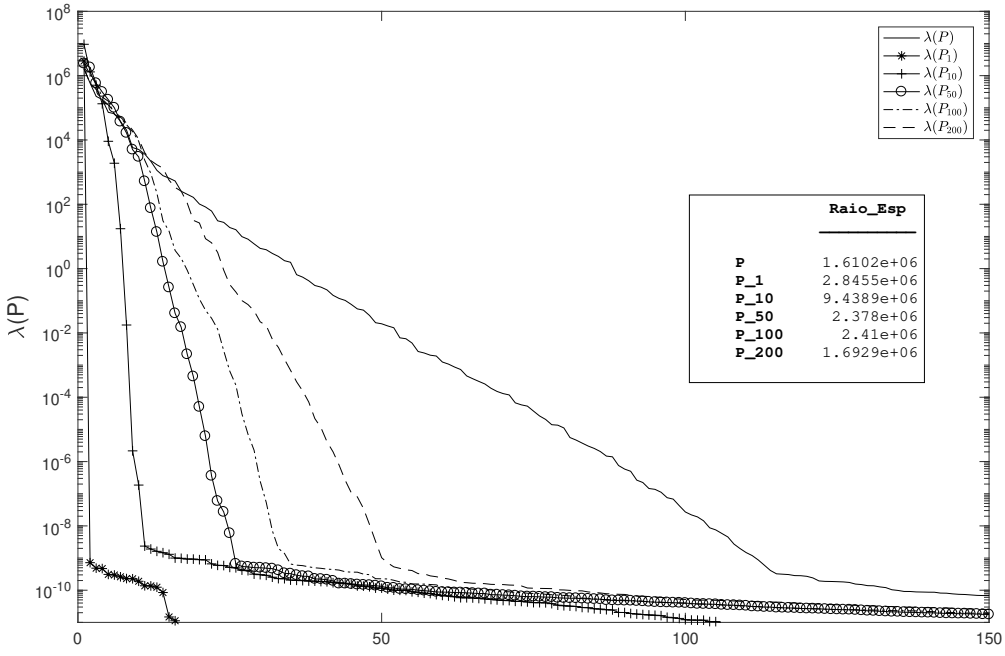
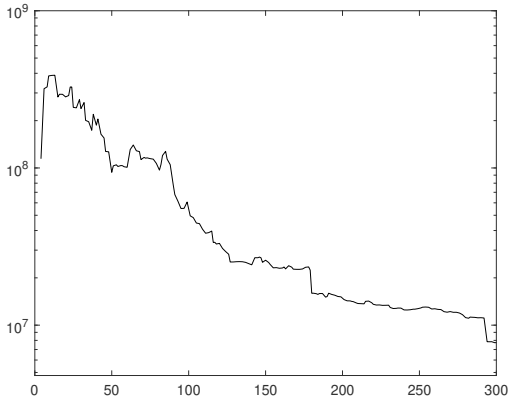


Figura 3.8: Comparativo entre o decaimento dos autovalores da solução P e dos autovalores das aproximações P_k.

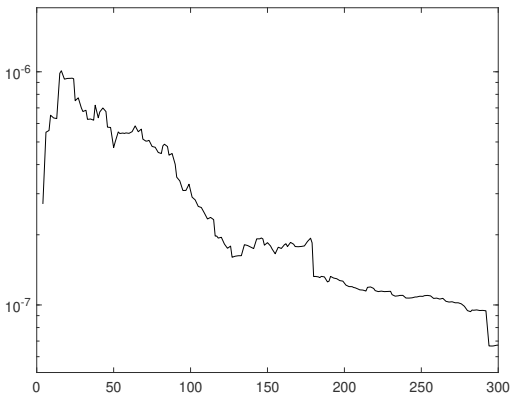
É possível notar que o posto numérico das aproximações P_k é mais sensível aos autovalores

dominantes. Os gráficos da figura 3.9 exibem o erro perante equação de Lyapunov, bem como o decaimento da diferença relativa entre as aproximações P_k. Para os autovalores que, por ventura, aparecem em pares complexos conjugados, cada par é considerado como um polo pelo índice k.

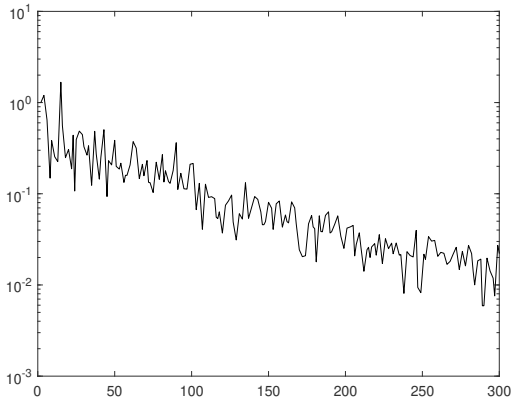
Figura 3.9: Erro na equação de Lyapunov e diferença relativa entre aproximações.



(a) Norma relativa absoluta .



(b) Norma relativa



(c) Norma relativa $\frac{\|P_{k-1}-P_k\|_F}{\|P_k\|_F}(Frobenius)$.

Nos gráficos da Figura 3.9 percebemos que, por mais que a diferença relativa entre as aproximações P_k decaia significativamente, o erro perante a equação de Lyapunov permanece elevado. Isso significa que, na prática, calcular uma boa aproximação para a solução P utilizando a parte da decomposição espectral de A correspondente aos polos dominantes pode ser um processo computacionalmente caro em situações como a do modelo `brasilsemtcsc`. No entanto, um conjunto de polos dominantes pode ser útil como parâmetros que auxiliam na convergência de outros métodos, como é visto na subseção a seguir.

3.3.4 Parâmetros ADI calculados a partir de autovalores dominantes

Na seção 3.1 vimos que os parâmetros ótimos para o método ADI são os que satisfazem o desafiador problema de minimização (3.17). Na mesma seção, foi descrita uma estratégia que consiste em utilizar alguns poucos parâmetros ADI de maneira cíclica para problemas esparsos, dando origem ao método LRCF-ADI. No entanto, a pergunta sobre como calcular de maneira prática um conjunto de parâmetros $\{\mu_1, \mu_2, ..., \mu_k\}$ a fim de tornar o método o mais eficiente possível ainda não tem uma resposta precisa.

Em [30] é proposta uma estratégia heurística para o cálculo desses parâmetros. Para explicar esse processo, vamos dividi-lo em duas etapas. Na primeira etapa é construído um conjunto discreto para ocupar o lugar de $\mathcal{R}$ em (3.17). Para isso é calculada uma matriz V_{k+}, cujas colunas formam uma base ortonormal para o subespaço de Krylov estendido

$$K_{k_+}(A, B) = span\{B, AB, A^2B, ..., A^{k_+-1}B\},$$

e também uma matriz V_{k-}, cujas colunas formam uma base para o subespaço de Krylov

$$K_{k_-}(A, B) = span\{B, A^{-1}B, A^{-2}B, ..., A^{-(k_--1)}B\}.$$

Essas bases podem ser facilmente obtidas pelo método de Arnoldi. A partir disso são calculados os conjuntos $\mathcal{R}_+$ e $\mathcal{R}_-$ de *valores de Ritz* de A que são autovalores das matrizes de Hessenberg $H_{k_+} = V_{k_+}^T A V_{k_+}$ e $H_{k_-} = V_{k_-}^T A V_{k_-}$, respectivamente. O conjunto $\mathcal{R}$ é dado então por

$$\mathcal{R} := \mathcal{R}_+ \cup 1/\mathcal{R}_-. \tag{3.42}$$

A justificativa para tal escolha é que os valores de $\mathcal{R}_+$ costumam ser aproximações dos autovalores de maior magnitude de A enquanto os elementos do conjunto $1/\mathcal{R}_-$ geralmente são aproximações para os autovalores de A que estão mais próximos da origem. Consequentemente o conjunto (3.42) pode ser considerado uma provável aproximação do espectro de A. A segunda etapa do método heurístico de Penzl consiste em calcular um conjunto $\mu = \{\mu_1, \mu_2, ..., \mu_l\} \subset \mathcal{R}$, de elementos chamados de *parâmetros quase ótimos* que resolvem

$$\min_{\mu \subset (\mathcal{R})} \max_{t \in \mathcal{R}} s_\mu(t), \tag{3.43}$$

com

$$s_\mu(t) = \frac{\prod_{i=1}^{l}(t - \mu_i)}{\prod_{i=1}^{l}(t + \mu_i)}. \tag{3.44}$$

Esse método está implementado no pacote LYAPAC, disponível em
https://www.netlib.org/lyapack/.

Nesse trabalho propomos duas estratégias diferentes das de Penzl. A primeira consiste em

substituir o conjunto $\mathcal{R}_+ \cup 1/\mathcal{R}_-$ pelos valores de Ritz obtidos a partir da base de Krylov estendida

$$K_l(A, B) = span\{A^{-(k-1)}B, ..., A^{-1}B, B, AB, A^2B, ..., A^{k-1}B\}. \tag{3.45}$$

O processo para calcular uma base ortonormal V_l para o espaço (3.45) pode ser feito utilizando o método de Arnoldi de maneira idêntica a descrita na seção 3.2.1. Uma análise de convergência para aproximar funções matriciais utilizando esse tipo de subespaço para uma grande classe de funções pode ser encontrada em [20].

A segunda estratégia é considerar $\mathcal{R}$ como sendo um conjunto de autovalores dominantes descritos na subseção 3.3.2 pois, conforme vimos anteriormente, autovalores dominantes (Def. 3.13) estão associados a uma base invariante por A que é predominante na solução P. Portanto, faz sentido dar prioridade a esses polos ao tentar resolver o problema *minmax* (3.17).

Testamos a eficiência dos parâmetros ADI escolhidos de três maneiras diferentes, incluindo o método heurístico original (de Penzl), no caso particular da equação de Lyapunov $AP + PA^T = -BB^T$ associada ao modelo `brasilsemtcsc`. Por se tratar de um modelo descritor com matrizes esparsas, empregamos o método SRCF-ADI dado pelo algoritmo 1.

Tabela 3.2: Erro relativo do método SLRCF-ADI com parâmetros calculados pelo método heurístico de Penzl.

	ADI-Penzl (10 parâm.)		ADI-Base estend. (10 parâm.)	
	k_+=k_-=30	k_+=k_-=50	k=30	k=50
Erro Rel.	$3,1 \cdot 10^{-10}$	$2,0 \cdot 10^{-8}$	$2,0 \cdot 10^{-9}$	$4,0 \cdot 10^{-8}$
Nº de It.	100	100	100	100

Tabela 3.3: Erro relativo do método SLRCF-ADI com parâmetros calculados a partir de polos dominantes (Def. 3.13).

	ADI-Polos dominantes (10 parâm.)		
	10 polos	30 polos	50 polos
Erro Rel.	$3,1 \cdot 10^{-9}$	$4,2 \cdot 10^{-12}$	$4,2 \cdot 10^{-12}$
Nº de It.	100	100	100

Tabela 3.4: Erro relativo do método SLRCF-ADI com apenas um parâmetro

	ADI-Penzl (1 parâm.)			ADI-Polos dom. (1 parâm.)	
	k_+=k_-=10	k_+=k_-=20	k_+=k_-=30	1 polo	10 polos
Erro Rel.	$1,7 \cdot 10^{-7}$	$1,2 \cdot 10^{-7}$	$9,4 \cdot 10^{-8}$	$7,6 \cdot 10^{-10}$	$3,2 \cdot 10^{-9}$
Nº de It.	130	130	130	130	130

Em todos os testes das tabelas 3.2, 3.3 e 3.4 foi estabelecida uma tolerância de 10^{-6} para o critério de parada definido em (3.18). Essa tolerância não foi atingida, fazendo com que todos os testes fossem até o número máximo de iterações pré-determinada. O erro relativo é definido por

$$\frac{\|AP + PA^T + BB^T\|_F}{\|A\|_F\|P\|_F + \|BB^T\|_F}.$$

Nos casos em que um conjunto de l parâmetros são calculados a partir de p polos ou valores de Ritz, com $l < k$, utilizamos a rotina `lp_mnmx` disponível no pacote LYAPAC para minimizar a função (3.43). Nas tabelas 3.2 e 3.3, cada polo do pares complexos conjugados está sendo contabilizado. Na tabela 3.4, o parâmetro ADI utilizado é sempre a parte real do número que minimiza (3.43). Note que, quando o número de parâmetros buscados é igual ao número de polos dominantes considerados, os parâmetros que satisfazem (3.43) são os polos em si.

Observando as tabelas percebemos que, para o exemplo escolhido, não existem diferenças significativas entre o uso da estratégia de Penzl da maneira original e a utilização da base estendida para o cálculo dos parâmetros. Já o uso de parâmetros calculados a partir dos polos dominantes da definição 3.13 apresentou resultados iguais ou melhores em em quase todos os cenários testados. Chama a atenção, inclusive, o fato de conseguirmos uma convergência razoável utilizando apenas a parte real do polo dominante. Resumindo, de acordo com o que foi visto na subseção 3.3.2, se não pudéssemos lançar mão da decomposição espectral completa de A, bastaria um método que calculasse o autovalor que maximiza

$$\delta_i = \left|\frac{(w_i^H B)^2}{2Re(\lambda_i)}\right|,$$

com w_i^H sendo o autovetor à esquerda de A associado a λ_i, satisfazendo $w_i^H v = 1$ para o autovetor unitário à direita de A associado a λ_i.

No caso em que utilizamos apenas um parâmetro, a motivação para a escolha da parte real do polo dominate como parâmetro ADI é baseada no fato que, para minimizar a função (3.44), o parâmetro escolhido deve estar igualmente próximo aos pares complexos conjugados de autovalores de região $\mathcal{R}$.

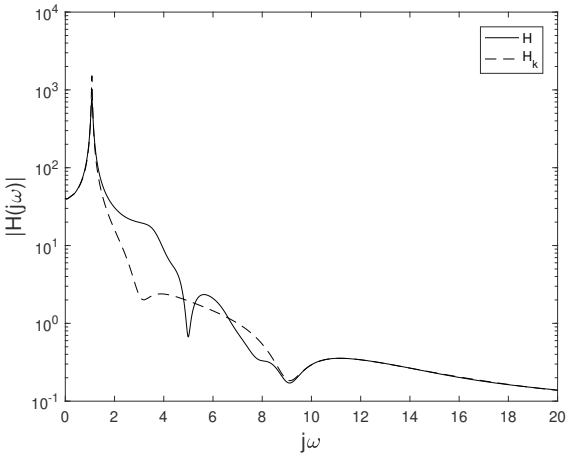
Agora, a fim de evidenciar as vantagens do uso de autovalores dominantes da equação de Lyapunov (Def. 3.13) na redução de modelo por balanceamento, vamos considerar o problema que consiste em calcular aproximações P_k e Q_k para o par de equações de Lyapunov:

$$\begin{aligned} AP + PA^T &= -BB^T \quad \text{e} \\ AQ + QA^T &= -CC^T. \end{aligned} \tag{3.46}$$

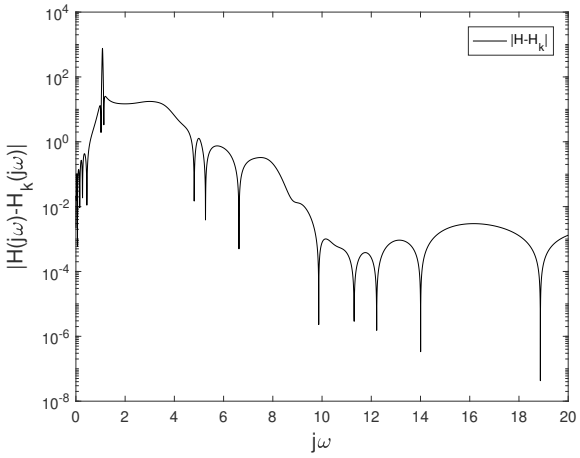
Para poder fazer um comparativo, vamos considerar o modelo `brasilsemtcsc` e aplicar o método SLRCF-ADI às equações (3.46) de duas maneiras, que se diferenciam apenas pelos parâmetros ADI escolhidos. No primeiro teste utilizamos o método heurístico de Penzl na íntegra,

com $k_+ = k_- = 30$ e um conjunto de 4 parâmetros ADI escolhidos de maneira a satisfazer o problema de minimização (3.43). No segundo teste, utilizamos 4 parâmetros ADI também. Porém, substituímos o conjunto $\mathcal{R}$ da função (3.43) por um conjunto de 10 autovalores dominantes (sem contabilizar pares complexos conjugados) que satisfazem a definição 3.13. Em ambos os testes utilizamos o método SLRCF-ADI com 20 iterações apenas, sem considerar nenhum outro critério de parada. Aplicamos o processo de redução por balanceamento descrito na seção 2.3 em ambos os testes e plotamos os respectivos gráficos da magnitude de $H(s)$ e $H_k(s)$, que são exibidos na figura 3.10.

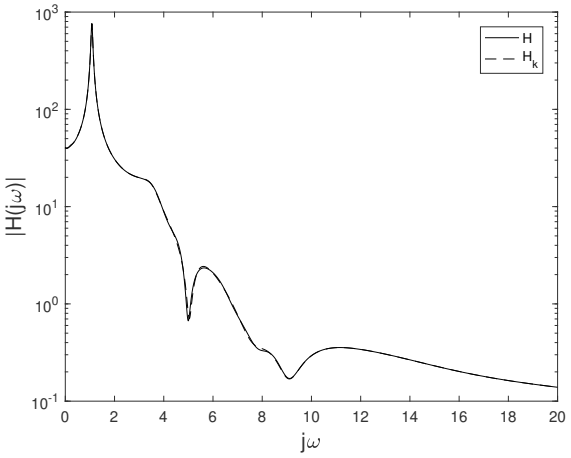
Figura 3.10: Redução de modelo utilizando SLRCF-ADI.



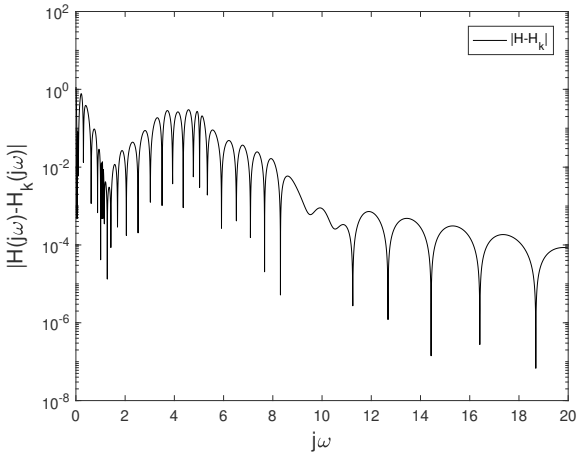
(a) Comparativo utilizando Penzl.



(b) Erro utilizando Penzl.



(c) Comparativo com autovalores dominantes.



(d) Comparação de desempenho por parâmetros ADI (Penzl × autovalores dominantes).

A precisão verificada nos gráficos 3.10c e 3.10d da figura 3.10 impressionam. Vale lembrar que essa precisão foi conseguida utilizando-se somente 4 decomposições LU da matriz A e apenas 20 iterações do método SLRCF-ADI, ou seja, o custo computacional dessa redução é significativamente baixo. Além disso, como foram só 20 iterações, o posto das aproximações para P e Q, geradas pelo

SLRCF-ADI, é de no máximo 20. Para determinar o tamanho do modelo reduzido consideramos o posto numérico da matriz PQ calculada com o comando `rank` do Matlab. Isso gerou um modelo reduzido de ordem 15 (lembre que o modelo original era de tamanho 1664). Portanto, para esse exemplo, conhecendo-se um conjunto de autovalores dominantes (Def 3.13), a redução de modelo por balanceamento utilizando o método SLRCF-ADI torna-se significativamente barata e eficiente.

3.3.5 Polos dominantes e o método RKSM

Conforme vimos na subseção 3.2.3, o método RKSM para equações de Lyapunov, assim como o método ADI, faz uso de parâmetros μ que influenciam na convergência do método. Druskin, Knizhnerman e Simoncini mostram em [6] que há uma relação de equivalência entre aproximações em subespaços de Krylov racionais e aproximações ADI para um conjunto ótimo de parâmetros μ. Por esse motivo propomos uma utilização dos polos dominantes definidos em 3.13 no método RKSM também.

Em [8] é proposta uma estratégia de escolha adaptativa dos parâmetros μ que ocorre durante a execução do método, aproveitando os valores de Ritz de A obtidos a cada iteração, isso torna o método robusto e prático. A função que calcula cada novo parâmetro para o método está implementada em Matlab na rotina `newpole` disponível em *http://www.dm.unibo.it/ simoncin/software.html*. A teoria envolvida nessa escolha adaptativa é extensa e por isso omitimos aqui. A função `newpole` depende de números positivos t_0 e t_1 que são aproximações para o maior e o menor autovalore de A, respectivamente, em módulo. Nesse mesmo trabalho é evidenciada uma grande sensibilidade na convergência do método em função da escolha desses parâmetros iniciais numa classe de métodos. Isso sugere que informações *a priori* da matriz A podem ser estudadas para uma melhor escolha de t_0 e t_1.

Acreditamos que, assim como observado no método ADI, a melhoria da convergência do método RKSM para equação de Lyapunov está atrelada à escolha de parâmetros que estejam associados a um subespaço invariante por A, dominante em P. Por isso, ao testar o método RKSM na equação de Lyapunov associada ao modelo `brasilsemtcsc`, fizemos modificações na escolha de t_0 e t_1 de modo a tentar fazer com que o intervalo $[t_0, t_1]$ contemple, em módulo, apenas um conjunto de autovalores de A. O espectro de A do modelo `brasilsemtcsc`, em módulo, está compreendido precisamente no intervalo $[2.56{\cdot}10^{-5}, 1.07{\cdot}10^{4}]$. Note que, pela equação (3.40), os polos dominantes tendem a estar concentrados nas proximidades do eixo imaginário.

A tabela 3.5 apresenta um comparativo entre os resultados obtidos da aplicação do método RKSM no problema `brasilsemtcsc` considerando diferentes valores para t_1.

Tabela 3.5: Comparativo do método RKSM no problema `brasilsemtcsc`.

	$t_1 = 1,0 \cdot 10^5$	$t_1 = 1,1 \cdot 10^4$	$t_1 = 200$	$t_1 = 100$
Erro Rel.	$2,3 \cdot 10^{-10}$	$2,3 \cdot 10^{-10}$	$2,7 \cdot 10^{-10}$	$2,1 \cdot 10^{-10}$
Erro Abs.	$1,3 \cdot 10^4$	$1,3 \cdot 10^4$	$1,6 \cdot 10^4$	$1,2 \cdot 10^4$
Nº de It.	37	37	25	35

Desenvolvemos o mesmos testes com o modelo `ww_vref_6405`, obtendo os resultados exibidos na tabela 3.6.

Tabela 3.6: Comparativo do método RKSM no problema `ww_vref_6405`.

	$t_1 = 1,0 \cdot 10^5$	$t_1 = 1,1 \cdot 10^4$	$t_1 = 200$	$t_1 = 100$
Erro Rel.	$6,7 \cdot 10^{-9}$	$2,4 \cdot 10^{-10}$	$1,8 \cdot 10^{-10}$	$4,0 \cdot 10^{-11}$
Erro Abs.	$104,0$	$38,2$	$27,9$	$6,3$
Nº de It.	26	26	24	25

Em ambas as situações estabelecemos uma tolerância $tol = 10^{-3}$ para o critério de parada definido em (3.23). O motivo para a escolha de uma tolerância pouco exigente é poder observar o comportamento do método nas primeiras iterações, pois, Como a escolha dos parâmetros μ é adaptativa e depende dos valores de Ritz gerados a cada iteração do método, a variação causada pela escolha inicial fica cada vez menos perceptível à medida que o número de iterações aumenta.

Tanto na tabela 3.5 como na tabela 3.6, verifica-se que o método RKSM necessita de menos iterações para atingir a tolerância exigida no critério de parada. O melhor desempenho foi obtido para $t_1 = 200$. Essa diferença no número de iterações é significativa em termos de custo computacional pois, no método RKSM, é preciso uma nova decomposição LU de $(A - \mu I)$ a cada iteração.

A título de curiosidade, tanto no modelo `brasilsemtcsc` como no sistema de teste `ww_vref_6405`, os 200 autovalores que resultam em maior magnitude de $\delta_i = \left|\frac{(w_i^H B)^2}{2Re(\lambda_i)}\right|$, dada na definição 3.13, possuem valor em módulo dentro do intervalo $[0, 100]$. Sendo assim, os resultados obtidos nos testes sugerem que há uma certa liberdade para a escolha t_1 menor do que o raio espectral de A no método RKSM, sem prejuízos à convergência. Mais do que isso, essa mudança na escolha pode até tornar o método mais eficiente se os autovalores de A dominantes na solução P estiverem devidamente compreendidos no intervalo $[t_0, t_1]$.

3.3.6 Uma aplicação da solução dada por similaridade na redução modal

Para esta seção, dado um sistema linear dinâmico (A, B, C), com $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$ e $C \in \mathbb{R}^{1\times n}$, vamos considerar o conjunto $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ de autovalores de A ordenados de acordo

com algum critério de dominância de polos em sistemas dinâmicos pré-estabelecido. Sabemos que, a partir de um conjunto de polos dominantes $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ e de seus respectivos autovetores, podemos construir um modelo reduzido H_k, conforme visto em (2.18) e (2.19). Além disso, podemos também calcular aproximações P_k e Q_k para as matrizes de Gram do sistema, que são soluções das equações de Lyapunov

$$\begin{aligned} &AP + PA^T = -BB^T \quad \text{e} \\ &A^TQ + QA = -C^TC, \end{aligned} \tag{3.47}$$

respectivamente, utilizando a fórmula definida em (3.32).

Na seção 2.1 vimos que a norma Hankel de $H(s)$ é dada por

$$\|H(s)\|_H := \tilde{\sigma}(H_G) = \max_{j=1,...n}\{\lambda_j^{1/2}(PQ)\}.$$

Mais do que isso, foi provado que os valores principais do operador de Hankel que representa o sistema são iguais aos autovalores do produto de matrizes PQ.

Vamos supor que um novo polo λ_{k+1} seja acrescentado ao conjunto $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ a fim de gerar um modelo reduzido H_{k+1}. Como os autovalores de PQ dependem continuamente das entradas de PQ, o acréscimo relativo gerado no produto de matrizes P_kQ_k, dado por

$$\mu(\lambda_{k+1}) := \frac{\|P_kQ_k - P_{k+1}Q_{k+1}\|}{\|P_kQ_k\|}, \tag{3.48}$$

nos dá uma estimativa para a significância do novo polo λ_{k+1} no modelo reduzido. Essa aferição é baseada no fato que P_kQ_k converge para PQ em n passos. Em teoria, a precisão dessa estimativa é bastante duvidosa, pois nada garante que P_kQ_k converge monotonamente para PQ, até porque essa convergência depende da escolha do polo acrescentado a cada termo da sequência. Além disso, pouco se sabe sobre perturbações em autovalores de uma matriz que é produto de duas matrizes de Cauchy. Mesmo assim, obtivemos resultados satisfatórios em aplicações numéricas que são apresentadas ao final dessa subseção.

Para problemas de grande porte, mesmo utilizando normas de baixo custo computacional, o cálculo de (3.48) pode ter uma complexidade computacional elevada. Isso se deve ao fato da matriz P_kQ_k ser densa e de tamanho $n \times n$. Por isso buscamos uma maneira mais prática de calcular essa estimativa. Vamos considerar a decomposição espectral de A dada por $A = W^H\Lambda V$, com $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ e $W^HV = I$. Note que a função de transferência $H(s)$ associada ao sistema linear dinâmico (A, B, C) é idêntica à função do sistema

$$\begin{cases} \dot{x}(t) = \Lambda x(t) + Ru(t) \\ y(t) = \zeta^T x(t), \end{cases} \tag{3.49}$$

com $R = (R_1, R_2, ..., R_n)^T$ sendo o vetor contendo os resíduos associados aos polos $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ e ζ sendo um vetor de "uns"de tamanho $n \times 1$. Mais detalhes sobre as representações da função

de transferência podem ser vistos em [41].

Perceba que, pela fórmula (3.32), as soluções das equações de Lyapunov $\Lambda\widehat{P} + \widehat{P}\Lambda^T = -RR^T$ e $\Lambda^T\widehat{Q} + \widehat{Q}\Lambda^T = -\zeta^T\zeta$, são

$$\begin{array}{c} \widehat{P} = \widehat{U}\widehat{X}C^{\lambda}\widehat{X}^H\widehat{U}^H \quad \text{e} \\ \widehat{Q} = \widehat{U}C^{\lambda}\widehat{U}^H, \end{array} \tag{3.50}$$

respectivamente, com $X = diag(R)$. Como Λ é uma matriz diagonal, a matriz $\widehat{U}$, cujas colunas são autovetores de Λ, é igual a matriz identidade. De acordo com a expressão (3.32), as projeções de $\widehat{P}$ e $\widehat{Q}$ no subespaço associado ao conjunto de k polos dominantes são:

$$\begin{array}{c} \widehat{P}_k = \widehat{U}_k\widehat{X}_kC_k^{\lambda}\widehat{X}_k^H\widehat{U}_k^H \quad \text{e} \\ \widehat{Q}_k = \widehat{U}_kC_k^{\lambda}\widehat{U}_k^H. \end{array} \tag{3.51}$$

Perceba que, embora as matrizes $\widehat{P}_k$ e $\widehat{Q}_k$ sejam de ordem $n \times n$, elas possuem apenas um bloco principal de tamanho $k \times k$ não nulo. Isso se deve ao fato que $\widehat{U}_k = [e_1|e_2|...|e_k]$, em que e_i denota um vetor canônico com a *i-esima* entrada não nula. Nessas condições, o cálculo do número $\|\widehat{P}_k\widehat{Q}_k\|_F$ pode ser reduzido a operações que envolvem apenas um bloco de tamanho $k \times k$ do produto $\widehat{P}_k\widehat{Q}_k$. Isso faz com que o cálculo de

$$\mu(\lambda_{k+1}) = \frac{\left\|\widehat{P}_k\widehat{Q}_k - \widehat{P}_{k+1}\widehat{Q}_{k+1}\right\|_F}{\left\|\widehat{P}_k\widehat{Q}_k\right\|_F} \tag{3.52}$$

tenha um custo computacional extremamente baixo quando o número de polos dominantes k é pequeno.

A primeira aplicação da estimativa (3.52) que propomos neste trabalho consiste em promover uma atuação mútua do critério $\varepsilon_{RMS}(k)$ e os valores de $\mu(\lambda_{k+1})$ para estabelecer um critério de parada para a redução modal. Nesse contexto, um conjunto de polos dominantes $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ pode ser considerado suficiente para o modelo reduzido se tanto $\mu(\lambda_{k+1})$ quanto $\varepsilon_{RMS}(k)$ estiverem apresentando valores menores do que alguma tolerância pré-estabelecida. Para que essa ideia fique mais clara, apresentamos um exemplo a seguir.

Exemplo 3.14 *Consideremos um sistema* (A, B, C) *com*

$$A = diag \begin{pmatrix} -0,05 + 0,5i \\ -0,05 - 0,5i \\ -0,10 + 1i \\ -0,10 - 1i \\ -0,11 + 20i \\ -0,11 - 20i \\ -0,20 + 21i \\ -0,20 - 21i \\ -20 + 25i \\ -20 - 25i \end{pmatrix}, \quad e \tag{3.53}$$

$$B = C = (1, 1, ..., 1)^T.$$

A escolha dos vetores B *e* C *em* (3.53) *se dá apenas para que a dominância efetiva dos polos esteja relacionada exclusivamente à forma como os polos estão distribuídos no plano complexo. Vamos considerar a ordem dos polos de* $H(s)$ *exatamente como estão dispostos em* (3.53) *(de cima para baixo), ou seja,*

$$\lambda_1 = -0.05 + 0.5i, \ \lambda_2 = -0.05 - 0.5i, \ ..., \ \lambda_{10} = -20 - 25i. \tag{3.54}$$

A figura a seguir mostra a distribuição desses polos no plano complexo.

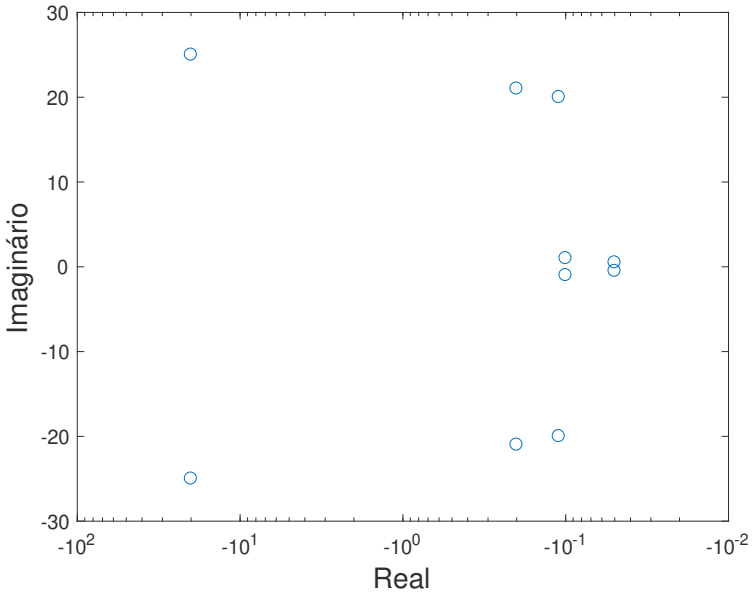


Figura 3.11: Autovalores de A (polos de $H(s)$).

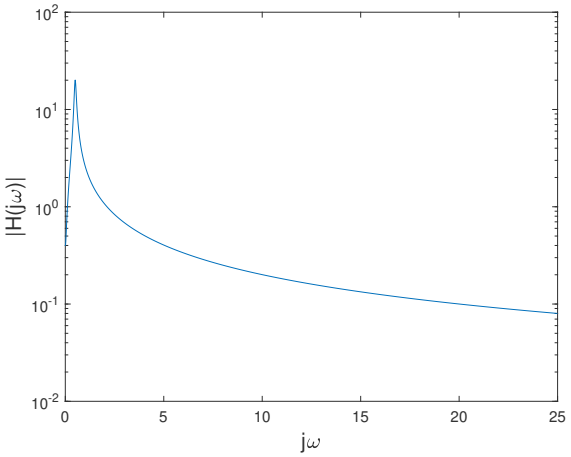
Utilizando o software Matlab, calculamos reduções modais $H_k(j\omega)$ *com base em subconjuntos de*

polos $\{\lambda_1, ..., \lambda_k\}$ *considerando a mesma ordem de* (3.54). *Os resultados dos testes estão dispostos na tabela 3.7 a seguir:*

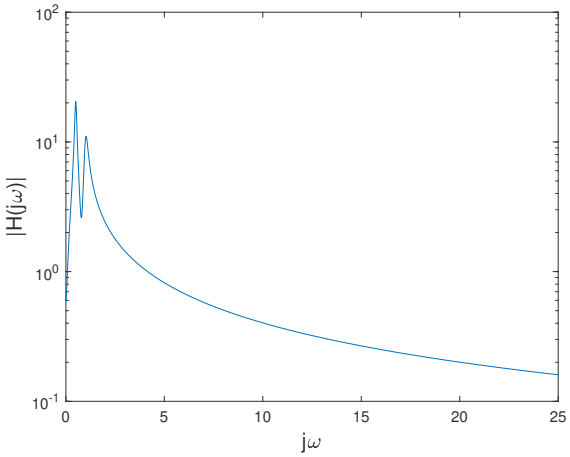
Tabela 3.7: Estimativas - Redução Modal

	$1/Re(\lambda_k)$	$\varepsilon_{RMS}(k)$	$\mu(\lambda_{k+1})$	$\|H(j\omega) - H_k(j\omega)\|_\infty$	$\|H(j\omega) - H_k(j\omega)\|_2$
$k = 2$	20	$0,0432$	$0,4459$	$10,0796$	$79,0495$
$k = 4$	10	$0,0017$	$0,1720$	$9,4065$	$61,4429$
$k = 6$	$9,09$	$0,0066$	$0,0988$	$5,0556$	$35,4978$
$k = 8$	5	$0,0014$	$0,0075$	$0,0595$	$2,1649$
$k = 10$	$0,05$	$3 \cdot 10^{-17}$	–	0	0

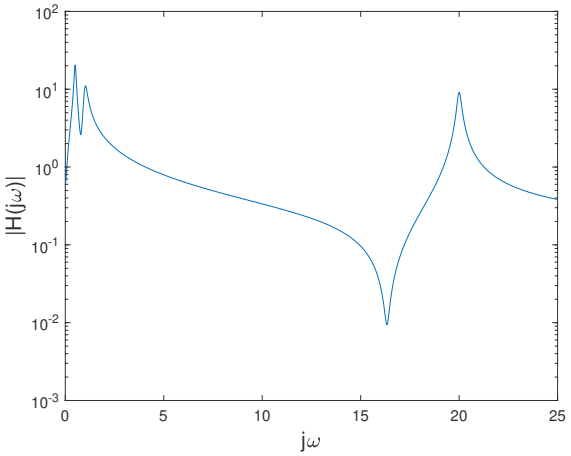
Os quatro primeiros pares de polos do sistema (3.53) *possuem uma dominância efetiva muito maior do que o último par. Isso pode ser evidenciado nos gráficos da magnitude versus frequência de* $H_k(j\omega)$ *apresentados na figura 3.12. Essa dominância coincide com a grandeza dos respectivos valores para* $1/Re(\lambda_k)$, *que são empregados na definição 2.7 de dominância. Note que a variação dos valores obtidos para* $\varepsilon_{RMS}(k)$, *a cada novo par de polos acrescentado no modelo, não parece proporcional à dominância efetiva dos polos. Isso se deve ao fato do terceiro par de polos estar distante dos anteriores no sentido do eixo imaginário, causando um efeito já mencionado na seção 2.2.1. A grandeza dos valores obtidos para* $\mu(\lambda_k)$, *por sua vez, condize muito mais com os efeitos que cada par de polos acrescentado na redução modal causa no gráfico de magnitude de* $H(j\omega)$.



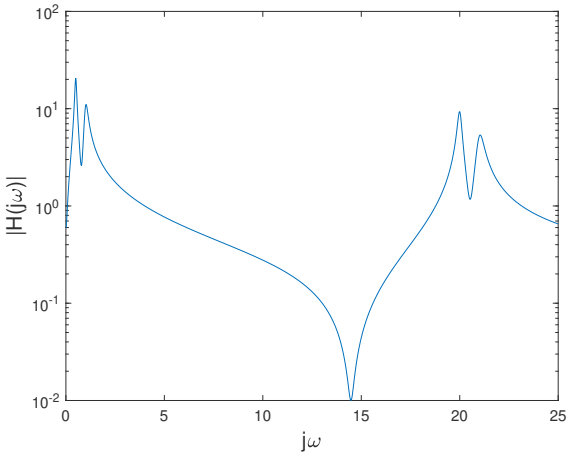
(a) k=2



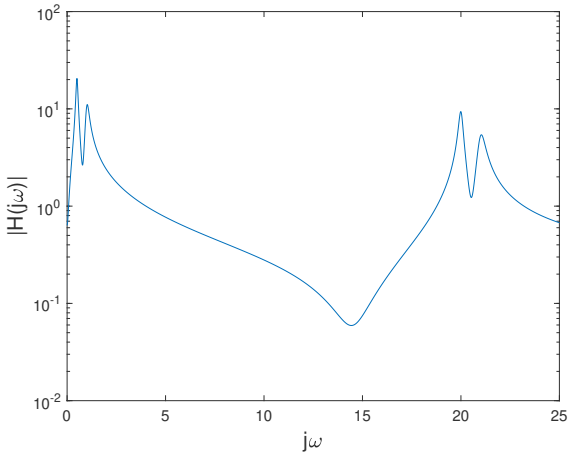
(b) k=4



(c) k=6



(d) k=8



(e) k=10

Figura 3.12: Magnitude das funções de transferência $|H_k(j\omega)|$ obtidas por redução modal com os polos dados em (3.53)

Optamos também por definir um critério de dominância que leve em consideração a estimativa (3.52). Essa definição está enunciada a seguir e é dividida em duas partes. A primeira parte é

idêntica à definição 2.7, porém serve para definir apenas um polo dominante de uma função de transferência H. A segunda parte serve para se construir um conjunto com mais de um polo dominante de $H(s)$.

Definição 3.15 (PQ-dominância Parte 1) *Um polo $\lambda \in \lambda(A)$ é dito PQ-dominante se*

$$\lambda = arg \max_{i=1,2,...,n} \left| \frac{R_i}{\alpha_i} \right|,$$

com R_i sendo o resíduo associado ao polo λ_i e α_i sendo a parte real de λ_i.

Para a segunda parte da definição, dado um subconjunto $\Omega \in \lambda(A)$, vamos considerar Ω^C como sendo o complementar de Ω em $\lambda(A)$.

Definição 3.16 (PQ-dominância Parte 2) *Dado um conjunto de polos $\Omega = \{\lambda_1, \lambda_2, ..., \lambda_k\} \subset \lambda(A)$, com $1 \leq k < n$, um polo $\lambda \in \omega^C$ é dito PQ-dominante perante Ω se*

$$\lambda = arg \max_{\lambda_{k+1} \in \Omega^C} \mu(\lambda_{k+1}),$$

com a função μ definida em (3.52).

Note que o polo λ que satisfaz a definição 3.15 também maximiza $\|\widehat{P}_1\widehat{Q}_1\|_F$, pois, de acordo com (3.51),

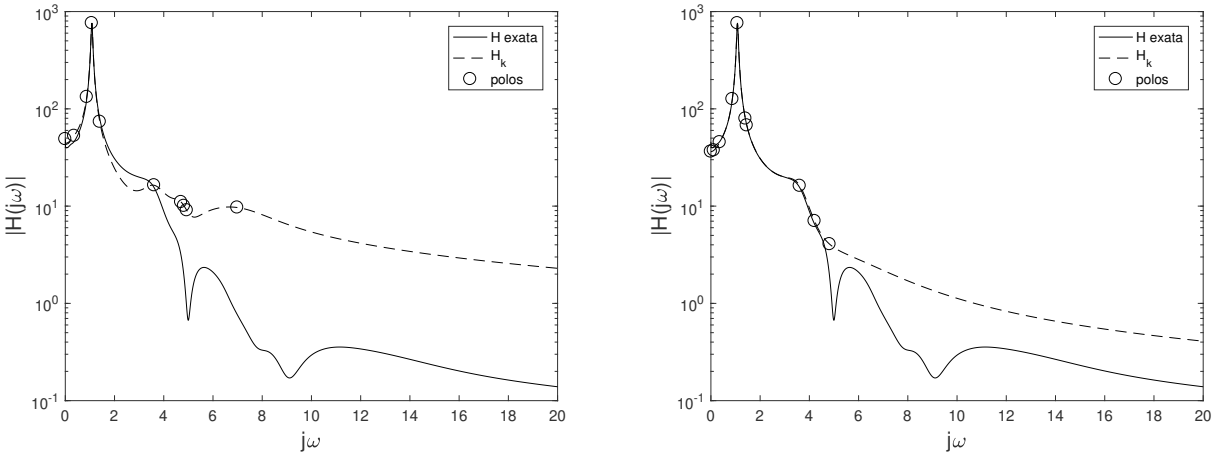
$$\widehat{P}_1\widehat{Q}_1 = \begin{bmatrix} \frac{1}{4}\left(\frac{R_1}{Re(\lambda_1)}\right)^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

Vamos, agora, desenvolver uma aplicação numérica considerando o sistema descritor `brasilsemtcsc`. Lembramos que, para fins didáticos, optamos por calcular explicitamente a matriz A, que é de ordem 1664, com o uso de Matlab. A partir disso foi possível obter todos os autovalores e autovetores de A por meio do comando `eig`. Esses cálculos têm um custo computacional alto que se justificam pela necessidade de entender o comportamento dos métodos. A tabela a seguir contém duas listas de polos. A primeira lista contém os 10 polos com maior valor para $\frac{|R_i|}{|Re(\lambda_i)|}$, obtidos a partir da decomposição espectral completa de A. A segunda lista contém os 10 primeiros polos que satisfazem as definições 3.15 e 3.16, também calculados a partir do espectro completo de A.

Tabela 3.8: Polos dominantes calculados a partir da decomposição espectral completa.

Dominância usual (def. 2.7)			PQ-Dominância			
Real	Imag	$\|R_i\|/\|Re(\lambda_i)\|$	Real	Imag	$\mu(\lambda_i)$	$\|R_i\|/\|Re(\lambda_i)\|$
-0,0335	1,0787	760,11	-0.0335	1.0787	–	760,11
-0,5567	3,6097	14,87	-0.6120	0.3587	0,0208	12,40
-0,6120	0,3587	12,40	-1.2936	1.4028	0,0138	5,59
-2,9445	4,8214	6,85	-2.9445	4.8214	0,0110	6,85
-0,4548	4,7054	5,78	-1.0829	0.8747	0,0093	3,78
-1,2936	1,4028	5,59	-1.4463	1.4565	0,0088	3,58
-0,7584	4,9367	5,11	-0.5567	3.6097	0,0081	14,87
-1,8415	6,9859	5,11	-2.2927	–	0,0068	2,92
$2{,}55 \cdot 10^{-7}$	–	4,65	-1.1727	0.1180	0,0054	2,76
-1,0829	0.8747	3,78	-4.0233	4.2124	0,0053	2,60

De acordo com a tabela 3.8, a significância de um polo perante a definição de PQ-dominância não é necessariamente proporcional à magnitude de $|R_i|/|Re(\lambda_i)|$. A figura 3.13 contém o gráfico do modelo reduzido sobreposto ao modelo original para cada um dos conjuntos de dez polos dominantes da tabela 3.8. Para isso utilizamos uma malha de 2000 pontos igualmente espaçados entre $\omega_0 = 0$ e $j\omega_f = 20j$, no domínio das frequências.



(a) Dominância usual (b) PQ-Dominância.

Figura 3.13: Gráficos de $H(j\omega)$ e $H_k(j\omega)$ com 10 polos dominantes.

Observando a figura 3.13 percebemos que, em comparação com os polos selecionados de acordo com a grandeza de $|R_i|/|Re(\lambda_i)|$, os 10 primeiros polos que satisfazem a condição de PQ-dominância estão concentrados em picos e pontos de inflexão compreendidos em regiões com maior magnitude de $\|H(j\omega)\|$.

Utilizando a malha criada para plotar os gráficos da figura 3.13, calculamos aproximações para $\|(H\text{-}H_k)(j\omega)\|_\infty$ e $\|(H\text{-}H_k)(j\omega)\|_2$ em ambos os cenários. Na tabela 3.9, os valores para k em cada linha indicam a quantidade de polos dominantes considerados em cada aproximação, obedecendo a ordem da tabela 3.8.

Tabela 3.9: Erro entre $H(j\omega)$ e $H_k(j\omega)$ a cada novo polo acrescentado.

	Dominância usual (def. 2.7)		PQ-Dominância	
k	$\|(H\text{-}H_k)(j\omega)\|_\infty$	$\|(H\text{-}H_k)(j\omega)\|_2$	$\|(H\text{-}H_k)(j\omega))\|_\infty$	$\|(H\text{-}H_k)(j\omega)\|_2$
1	17,96	69,59	17,96	69,59
2	8,23	57,41	20,70	98,09
3	12,35	89,82	16,62	60,21
4	10,58	111,23	12,91	73,04
5	11,53	127,08	12,91	72,89
6	9,90	93,16	14,16	81,78
7	7,23	70,49	8,53	79,59
8	9,41	112,90	5,72	62,87
9	9,41	112,85	8,51	51,17
10	9,22	110,41	9,48	22,73

Em resumo, a Tabela 3.9 indica que, ao acrescentarmos polos que satisfazem o critério de PQ-dominância ao modelo reduzido, ocorre uma diminuição do erro perante a norma $\|\cdot\|_\infty$, ao passo que o erro perante a norma $\|\cdot\|_2$ se mantém mais estável em comparação ao que acontece quando adicionamos polos seguindo a ordem de grandeza de $|R_i|/|Re(\lambda_i)|$. A título de curiosidade, fizemos testes considerando quantidades maiores de polos dominantes. No entanto, os valores obtidos para $\mu(\lambda_i)$ tendem a estabilizar-se muito rapidamente. Isso faz com que o teste fique inconclusivo.

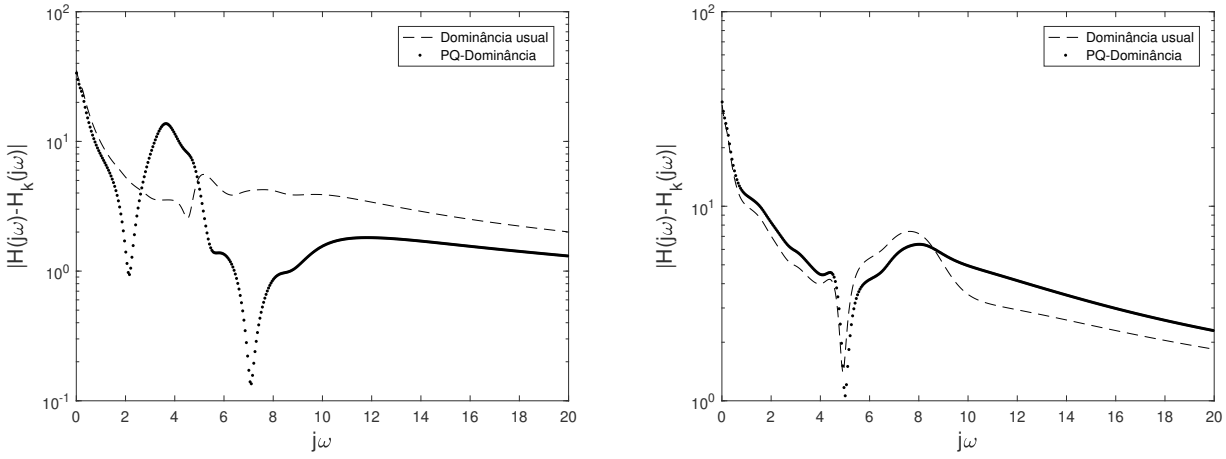
Na literatura existem vários métodos que calculam polos dominantes de um modelo $H(s)$, baseando-se na definição 2.7, sem necessitar conhecer todos os autovalores de A. Dentre os mais recentes estão o SADPA (*Subespace Acceleration Dominant Pole Algorithm*), introduzido por Rommes e Martins [34] e o DPSE (*Dominant Pole Spectrum Eigensolver*), apresentado por Martins [25]. Ambos os métodos tem em comum o uso de inversões de translações de A nas iterações, ou seja, passos do tipo $(\alpha I - A)^{-1}x$. Falando numa linguagem simples, o que ocorre nesses métodos é que, a cada novo polo dominante requerido, dá-se um chute inicial para o valor α, que é atualizado durante as etapas do método, levando-o a convergir para um polo dominante. Acreditamos que o critério de PQ-dominância pode auxiliar na escolha desses chutes iniciais em projetos futuros. Além disso, tanto o $DPSE$ quanto o $SADPA$ tem a capacidade de calcular mais de um autovalor de A simultaneamente. Após o método convergir para um conjunto de autovalores, é escolhido aquele que apresentar maior dominância. Nesse momento, acreditamos que o critério de PQ-dominância possa ser utilizado.

Para ilustrar, vamos considerar novamente o modelo `brasilsemtcsc` e, em seguida, calcular 50 polos dominantes (sem contabilizar pares complexos conjugados) utilizando o método DPSE. A partir desses polos, vamos construir dois subconjuntos de polos dominantes, um escolhido com base na definição 2.7 e outro escolhido em conformidade com a definição de PQ-dominância. Esses conjuntos de polos são exibidos na tabela a seguir:

Tabela 3.10: Polos dominantes calculados via DPSE.

Dominância usual (Def. 2.7)			PQ-dominância			
Real	Imag	$\lvert R_i\rvert/\lvert Re(\lambda_i)\rvert$	Real	Imag	$\mu(\lambda_i)$	$\lvert R_i\rvert/\lvert Re(\lambda_i)\rvert$
-0,0335	1.0787	760,109	-0,0335	1.0787	–	760,109
-0,5567	3,6097	14,87	-0,6120	0,3587	0,208	12.40
-0,6120	0,3587	12.40	-1,2936	1,4028	0,014	5,59
-2.9445	4.8214	6,85	-2.9445	4.8214	0,011	6,85
-1,2936	1,4028	5,59	-1.0829	0,8747	0,009	3,78
-1.8415	6,9859	5,11	-1.4463	1,4565	0,009	3,58
-1.0829	0,8747	3,78	-0,5567	3,6097	0,008	14,87
-1.4790	8.2551	3,68	-4,0233	4,2124	0,005	2,60
-1.4463	1,4565	3,58	-1.8415	6,9859	0,004	5,11
-4,0233	4,2124	2,60	-5.8148	4,8704	0,003	1,36

A figura 3.14, a seguir, exibe os erro entre a função de transferência original e a função $H_k(j\omega)$ do modelo reduzido para conjuntos de 5 e de 10 polos.



(a) Erro do modelo reduzido com 5 polos (b) Erro do modelo reduzido com 10 polos.

Figura 3.14: Gráficos de $|H(j\omega) - H_k(j\omega)|$ para Redução Modal via DPSE.

Observando os gráficos da figura 3.14, percebe-se que o erro da redução modal construída a partir de um conjunto de polos PQ-dominantes se mantém próximo do erro para o caso da

dominância usual. Esses resultados eram esperados pois, na tabela 3.10, nota-se que há vários polos em comum entre os conjuntos obtidos por cada critério de dominância, com uma ordenação ligeiramente diferente.

Capítulo 4

Um Novo Método do Tipo *Splitting* para Equações de Lyapunov

Neste capítulo, propomos uma estratégia para resolver equações de Lyapunov baseada em métodos do tipo *splitting* para sistemas lineares. Embora essa estratégia seja construída com base na representação vetorizada da equação, desenvolvemos uma técnica que evita resolução de sistemas da ordem n^2. Acreditamos que esse seja o principal ponto positivo do método.

Inicialmente vamos considerar um sistema linear genérico

$$Ax = b, \tag{4.1}$$

com $A \in \mathbb{R}^{n\times n}$ e $x, b \in \mathbb{R}^n$

Métodos iterativos como Jacobi e Gauss-Seidel são membros típicos da grande família de métodos que possuem iterações da forma

$$Mx_{k+1} = Nx_k + b \tag{4.2}$$

com $A = M - N$ sendo uma cisão (*splitting*) da matriz A [13]. Em geral, a matriz M é escolhida de modo que o sistema (4.2) seja relativamente fácil de resolver. A convergência (ou não) da sequência x_k para $A^{-1}b$ vai depender dos autovalores de $M^{-1}N$. Mais precisamente, o Teorema 4.1 estabelece que o sucesso do método (4.2) dependerá do *raio espectral* de $M^{-1}N$, que é definido por

$$\rho(F) = max\{\ |\lambda|\ :\ \lambda \in \lambda(F)\},$$

para uma matriz $F \in \mathbb{R}^{n\times n}$ qualquer.

Teorema 4.1 *[13] Suponha $b \in \mathbb{R}^n$ e $A = M - N \in \mathbb{R}^{n\times n}$ não-singular. Se M é não-singular e $\rho(M^{-1}N) < 1$, então os iterados x_k definidos por $Mx_{k+1} = Nx_k + b$ convergem para $x = A^{-1}b$ para qualquer vetor inicial x_0.*

Demonstração. [13, pg 511] □

Vamos considerar agora a equação de Lyapunov

$$AP + PA^T = -BB^T, \tag{4.3}$$

com $A \in \mathbb{R}^{n\times n}$, e $B \in \mathbb{R}^{n\times m}$, $m \leq n$. Vimos na seção 1.2 que a equação (4.3) é equivalente ao sistema linear

$$\tilde{A}\tilde{p} = \tilde{b}, \tag{4.4}$$

com $\tilde{A} = (I \otimes A + A \otimes I)$, $\tilde{p} = \text{vec}(P)$, e $\tilde{b} = \text{vec}(-BB^T)$. Perceba que o sistema (4.4) é da ordem n^2.

Dado um $\sigma > 0$, a equação (4.3) pode ser reescrita como

$$(A - \sigma I)P + P(A^T + \sigma I) = -BB^T. \tag{4.5}$$

A equação (4.5) é equivalente ao sistema

$$\tilde{A}_\sigma \tilde{p} = \tilde{b},$$

com $\tilde{A}_\sigma = [I \otimes (A - \sigma I) + (A + \sigma I) \otimes I]$. A partir disto, definimos a cisão $\tilde{A}_\sigma = M_\sigma - N_\sigma$, com $M_\sigma = I \otimes (A - \sigma I)$ e $N_\sigma = -(A + \sigma I) \otimes I$.

Vamos supor que a matriz $(A - \sigma I)$ é inversível. Então a matriz M_σ também é inversível e

$$M_\sigma^{-1} = I \otimes (A - \sigma I)^{-1}$$

Sendo assim, dado um vetor inicial $\tilde{p}_0 \in \mathbb{R}^{n^2}$, definimos a iteração do tipo *splitting* para equação (4.3) como sendo

$$\tilde{p}_{k+1} = M_\sigma^{-1} N_\sigma \tilde{p}_k + M_\sigma^{-1}\tilde{b} \tag{4.6}$$

Assumindo que a parte real de todos autovalores da matriz A é negativa, sabemos que a solução P da equação (4.3) é simétrica. Vamos escolher $P_0 = 0_{n\times n}$ em (4.6). Essa escolha, além de ser coerente com a simetria de P, também simplifica a implementação do algoritmo. Desse modo, $\tilde{p}_0 = 0_{n^2\times 1}$.

A cada iteração de (4.6) é preciso calcular $M_\sigma^{-1} N_\sigma \tilde{p}_k$. Perceba que

$$\begin{aligned} M_\sigma^{-1} N_\sigma &= -\left(I \otimes (A - \sigma I)^{-1}\right)((A + \sigma I) \otimes I) \\ &= -(A + \sigma I) \otimes (A - \sigma I)^{-1}. \end{aligned}$$

Note que, na primeira iteração do método (4.6) com $P_0 = 0_{n\times n}$, temos

$$\begin{aligned} M_\sigma^{-1} N_\sigma \tilde{p}_0 + M_\sigma^{-1}\tilde{b} &= \left[I \otimes (A - \sigma I)^{-1}\right]\tilde{b} \\ &= \text{vec}\left((A - \sigma I)^{-1} BB^T\right), \end{aligned}$$

ou seja,

$$P_1 = (A - \sigma I)^{-1} B B^T.$$

Para a segunda iteração do método, temos;

$$\begin{aligned} M_\sigma^{-1} N_\sigma \tilde{p}_1 + M_\sigma^{-1} \tilde{b} &= -\left[(A + \sigma I) \otimes (A - \sigma I)^{-1}\right] \mathrm{vec}\left((A - \sigma I)^{-1} B B^T\right) \\ &\quad + \left[I \otimes (A - \sigma I)^{-1}\right] \tilde{b} \\ &= \mathrm{vec}\left((A - \sigma I)^{-2} B B^T (A + \sigma I)\right) + \mathrm{vec}\left((A - \sigma I)^{-1} B B^T\right). \end{aligned}$$

portanto,

$$P_2 = -(A - \sigma I)^{-2} B B^T (A + \sigma I) + (A - \sigma I)^{-1} B B^T.$$

Repetindo esse processo sucessivas vezes, verificamos que a iteração (4.6) pode ser reescrita da seguinte forma:

$$P_{k+1} = \sum_{i=1}^{k+1} (-1)^{i+1} (A - \sigma I)^{-i} B B^T \left((A + \sigma I)^T\right)^{(i-1)}$$

Com isso chegamos ao seguinte algoritmo:

Algoritmo 5: SEL-Método do tipo Splitting para Equação de Lyapunov

Entrada: Matriz A, matriz B, escalar σ e um inteiro k_{max}.

Saída: Matriz P, solução aproximada para a equação (4.3).

```
início
    Y_1 = (A − σI)^{-1} B;
    Z_1 = B;
    para k=1,2,...,k_max faça
        P_k = (−1)^{k+1} Y_k Z_k^T + P_{k−1};
        se convergir então
            P = P_k;
            senão faça
                Y_k = (A − σI)^{-1} Y_{k−1};
                Z_k = (A + σI) Z_{k−1};
            fim
        fim
    fim
fim
retorna P
```

O algoritmo 4.2 aumenta o posto da matriz em, no máximo, uma unidade a cada iteração, acrescentando a matriz $Y_k Z_k^T$. Desta forma, o critério de parada pode ser definido a partir da estimativa do decaimento de P, fazendo o algoritmo parar quando

$$\frac{\|P_k - P_{k-1}\|}{\|P_{k-1}\|} < \varepsilon_{tol}.$$

para alguma tolerância ε_{tol} definida pelo usuário. Em outros termos,

$$\frac{\|Y_k Z_k^T\|}{\|P_{k-1}\|} < \varepsilon_{tol},$$

4.1 Convergência: a escolha do parâmetro σ

Nesta seção vamos definir um critério de convergência para o algoritmo 5, baseado no critério para métodos do tipo *splitting* para sistemas lineares.

Pelo Teorema 4.1, o algoritmo 5 converge se, e somente se, $\rho(M_\sigma^{-1} N_\sigma) < 1$. Note que os autovalores de $M_\sigma^{-1} N_\sigma = -(A + \sigma I) \otimes (A - \sigma I)^{-1}$ são da forma

$$\lambda_{ij} = -\frac{\lambda_i + \sigma}{\lambda_j - \sigma}, \tag{4.7}$$

com λ_i e λ_j sendo autovalores da matriz A. Portanto, para o caso da equação de Lyapunov, podemos reescrever o Teorema 4.1 da seguinte forma:

Teorema 4.2 *Os iterados P_k do algoritmo 5 convergem para a solução P da equação (4.3) se, e somente se,*

$$\left|\frac{\lambda_i + \sigma}{\lambda_j - \sigma}\right| < 1, \quad \forall\ \lambda_i, \lambda_j \in \lambda(A). \tag{4.8}$$

Uma consequência direta do Teorema 4.2 é que, sendo A uma matriz estável, a procura do parâmetro σ pode ser restringida ao conjunto $\{\sigma \in \mathbb{C} \ : \ Re(\sigma) > 0\}$.

Embora o algoritmo 5 se assemelhe muito às iterações do método ADI [48], ainda não tivemos evidências de convergência considerando múltiplos valores para σ. Por isso vamos manter σ fixo e buscar outras maneiras de acelerar a convergência do método. Como estamos considerando a matriz A real, seus autovalores complexos estão dispostos em pares complexos conjugados. Isso faz com que os autovalores de A estejam dispostos de maneira simétrica perante o eixo real e, por isso, restringimos nossa procura a um $\sigma > 0$.

Note que essa escolha de σ depende tanto do raio espectral como da relação que há entre a parte real e a parte imaginária de cada um dos autovalores de A. Para entender melhor, vejamos primeiro o caso em que autovalores de A são todos reais.

Proposição 4.3 *Sejam* $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ *os autovalores da matriz A do algoritmo 5, com* $Re(\lambda_i) < 0$, *para* $i = 1, ..., n$. *Sejam ainda*

$$\tau = \max_{i,j=1,...,n} |Re\lambda_i - Re\lambda_j| \quad e \quad \varsigma = \max_{j=1,...,n} \left| \frac{Im(\lambda_j)^2}{Re(\lambda_j)} \right|.$$

Se

$$\sigma > \frac{1}{2}(\tau + \varsigma), \tag{4.9}$$

Então o Algoritmo 5 converge.

Demonstração. Para cada $i = 1, .., n$, consideremos os números reais a_i e b_i tais que $\lambda_i = -a_i i + b_i$, com $a_i > 0$. Então, para todo $i, j = 1, , 2, ..., n$,

$$\begin{aligned}
\left| \frac{\lambda_i + \sigma}{\lambda_j - \sigma} \right| < 1 &\iff \frac{(\lambda_i + \sigma)^2}{(\lambda_j - \sigma)^2} < 1 \\
&\iff \frac{a_i^2 - 2a_i\sigma + \sigma^2 + b_i^2}{a_j^2 + 2a_j\sigma + \sigma^2 + b_j^2} < 1 \\
&\iff \sigma > \frac{a_j^2 + b_j^2 - a_i^2 - b_i^2}{2(a_i + a_j)} \\
&\iff \sigma > \frac{1}{2}\left[(a_j - a_i) + \frac{b_j^2 - b_i^2}{a_i + a_j} \right].
\end{aligned}$$

Como

$$\frac{b_j^2}{a_j} \geq \frac{b_j^2 - b_i^2}{a_i + a_j}, \quad \forall\ i, j = 1, , 2, ..., n$$

então, é suficiente termos

$$\sigma > \frac{1}{2}(\tau + \varsigma) \geq \max_{\{i,j=1,...,n\}} \frac{1}{2}\left[(a_j - a_i) + \frac{b_j^2}{a_j} \right]$$

□

Corolário 4.4 *Se os autovalores* $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ *da matriz A forem todos reais negativos tais que* $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n|$ *e se*

$$\sigma > \frac{1}{2}(\lambda_n - \lambda_1),$$

então o algoritmo 5 converge.

Corolário 4.5 *Sejam* $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ *os autovalores da matriz A do algoritmo 5, tais que* $\{|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n|\}$, *com* $Re(\lambda_i) < 0$ *e* $|Im\lambda_i| < |Re\lambda_i|$, *para* $i = 1, ..., n$. *Sejam ainda* a, b, c, d *números reais tais que* $\lambda_1 = -a + bi$ *e* $\lambda_n = -c + di$, *com* $a, c > 0$. *Se*

$$\sigma > \frac{1}{2}\left[(a - c) + (b - d) \right],$$

então o algoritmo 5 converge.

Demonstração. Para cada $i = 1, .., n$, consideremos os números reais a_i e b_i tais que $\lambda_i = -a_i i + b_i$, com $a_i > 0$. Então, da demonstração da Proposição 4.3, temos que a condição

$$\sigma > \max_{\{i,j=1,...,n\}} \frac{1}{2}\left[(a_j - a_i) + \frac{b_j^2 - b_i^2}{a_i + a_j}\right]$$

é suficiente para que o método convirja. Como $|\mathrm{Im}\lambda_i| < |\mathrm{Re}\lambda_i|$, para $i = 1, ..., n$, então

$$\frac{b_j^2 - b_i^2}{a_i + a_j} \geq b_j - b_i \quad \forall\ i, j = 1, , 2, ..., n.$$

Portanto, para que o método convirja, é suficiente que

$$\sigma > \frac{1}{2}\left[(a - c) + (b - d)\right] \geq \max_{\{i,j=1,...,n\}} \frac{1}{2}\left[(a_j - a_i) + (b_j - b_i)\right]$$

□

4.2 Calibragem com o uso de um parâmetro α

Note que a escolha do parâmetro σ depende de informações, *a priori*, da matriz A do sistema dinâmico. Uma matriz A cujo maior autovalor, em módulo, é muito maior do que o menor autovalor em módulo dessa matriz, exige um valor σ seja elevado para que haja convergência do método SEL. Isso pode tornar a razão do lado esquerdo de (4.8) muito próxima de um (principalmente nos casos em que houver autovalores de A com parte real relativamente pequena em módulo). Consequentemente, a convergência do método pode ficar muito lenta.

Exemplo 4.6 *Vamos supor que a matriz A do sistema (4.3) seja a matriz diagonal*

$$A = \begin{bmatrix} -100 & & & \\ & -10 & & \\ & & \ddots & \\ & & & -0,001 \end{bmatrix}$$

Pelo corolário 4.4, deve-se escolher $\sigma > \frac{99,999}{2}$. Se fixarmos $\sigma = 50$, então o módulo de um dos autovalores da matriz $M_\sigma^{-1}N_\sigma$, dados por (4.7), será

$$\left|\frac{0,001 + 50}{0,001 - 50}\right| = 0,99996 \approx 1.$$

Para contornar problemas com características similares ao do exemplo 4.6, propomos a estra-

tégia de escolher um $\alpha > 0$ para resolver a equação auxiliar

$$\tilde{A}\tilde{P} + \tilde{P}\tilde{A}^T = -BB^T, \tag{4.10}$$

com $\tilde{A} = (A - \alpha I)$, e, a partir disso, reconstruir a solução P de (4.3) a partir de $\tilde{P}$. Para isso, enunciamos o teorema a seguir, que faz uso da representação da solução P de (4.3) em termos do espectro de A, visto na seção 3.3.1.

Teorema 4.7 *Seja $\tilde{P}$ uma solução para a equação (4.10). Então, existem uma matriz de Cauchy generalizada C e uma matriz V, construídas a partir de $\tilde{P}$, tais que a solução da equação (4.3) é dada por*

$$P = VCV^H \tag{4.11}$$

Demonstração. Seja $\tilde{P}$ a solução da equação (4.10). Sabemos que

$$\tilde{P} = V\tilde{C}V^H, \tag{4.12}$$

em que $\tilde{C}$, é uma matriz de Cauchy generalizada cujas entradas são da forma $\frac{d_{ij}}{\lambda_i + \lambda_j^H - 2\alpha}$, com λ_i e λ_j sendo autovalores de A e V uma matriz inversível cujas colunas são autovetores de A. Os números d_{ij} denotam os denominadores das frações da soma (3.31). Como $\tilde{P}$ é simétrica semi-definida positiva, consideremos a decomposição SVD:

$$\tilde{P} = WSW^H. \tag{4.13}$$

É fato que a matriz W gera as colunas de V, ou seja, existe uma matriz Z tal que

$$V = WZ. \tag{4.14}$$

Definamos agora a matriz

$$R := W^H AW. \tag{4.15}$$

Como $AV = V\Lambda$, com $\Lambda = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_n\}$, então $RZ = Z\Lambda$ e, assim,

$$Z\Lambda Z^{-1} = R. \tag{4.16}$$

Logo, Z é matriz de autovetores de R e Λ é a matriz cuja diagonal contém os autovalores que são comuns a R e A. Agora, podemos calcular a matriz $\tilde{C}$, pois, substituindo (4.14) em (4.12) e multiplicando por $Z^{-1}W^H$ pela esquerda e por WZ^{-H} pela direita, temos

$$\tilde{C} = Z^{-1}W^H\tilde{P}WZ^{-H}.$$

Por outro lado, sabemos que a solução de (4.3) é dada por

$$P = VCV^H, \tag{4.17}$$

cujas entradas da matriz de Cauchy generalizada C são da forma $\frac{d_{ij}}{\lambda_i+\lambda_j^H}$. Podemos então construir os numeradores das entradas de $\tilde{C}$ (que são comuns às matrizes $\tilde{C}$ e C) a partir das entradas de $\tilde{C}$, utilizando a fórmula

$$d_{ij} = \tilde{C}_{ij}\left(\lambda_i + \lambda_j^H - 2\alpha\right).$$

isso conclui a construção das matrizes que aparecem no lado direito de (4.17).

□

Vejamos agora uma interpretação para o resultado que acabamos de mostrar. Na seção 3.3.1 vimos que a matriz C da fórmula (4.11) pode ser escrita como

$$C = XC^{\Lambda}X^H,$$

com $X := V^{-1}B = Z^{-1}W^HB$ e C^{Λ} sendo a matriz de Cauchy cujas entradas são da forma

$$C^{\lambda(A)} = \frac{1}{\lambda_i + \lambda_j^H}.$$

Sendo assim, com base na demonstração do teorema 4.7, sabe-se que

$$P = WZXC^{\Lambda}X^HZ^HW^H. \tag{4.18}$$

Como P é solução de $AP + PA^T = -BB^T$ então, pela ortogonalidade de W, a matriz $\widehat{P} = ZXC^{\Lambda}X^HZ^H$ é solução da equação

$$W^HAW\widehat{P} + \widehat{P}W^HA^TW = -W^HBB^TW. \tag{4.19}$$

A reformulação (4.19) da equação de Lyapunov (4.3) permite que pensemos numa maneira de aproximarmos a solução P por meio de uma projeção que descrevemos a seguir.

Perceba que a solução P, dada pela fórmula (4.11), depende da decomposição (4.13) e da decomposição espectral (4.16). Em termos de implementação, essas decomposições acabam sendo inviáveis para sistemas de tamanho elevado. Supondo que tenhamos uma aproximação $\tilde{P}_k$, de posto $k << n$, para a solução da equação (4.10) e consideremos a decomposição SVD econômica $\tilde{P}_k = W_kS_kW_k^H$ em (4.13). Deste modo, a expressão (4.15) dá lugar a

$$R_k = W_k^HAW_k.$$

Agora, os valores da matriz diagonal

$$\Lambda_k = Z_k^{-1} R_k Z_k,$$

para alguma matriz inversível Z_k, são os valores de Ritz da matriz A associados aos vetores da matriz W_k. Portanto, por (4.18) e (4.19), a matriz

$$\widehat{P}_k = Z_k X_k C^{\Lambda_k} X_k^H Z_k^H,$$

com $X_k = diag(Z_k^{-1} W_k^H B)$, é solução da equação

$$W_k^H A W_k \widehat{P}_k + \widehat{P}_k W_k^H A^T W_k = -W_k^H B B^T W_k,$$

que é a projeção da equação de Lyapunov (4.4) no espaço gerado por W_k. Além disso, como W_k é ortonormal, o resíduo $E = AP_k + P_k A^T + BB^T$ gerado pela aproximação $P_k = W_k^H \widehat{P}_k W_k$ satisfaz a condição de ortogonalidade

$$W_k^H E W_k = 0,$$

ou seja, P_k é a projeção da solução P em $span\{W_k\}$.

É importante salientar que, em geral, a unicidade da solução P_k da equação (4.19) só é garantida para os casos em que a matriz A é dissipativa, ou seja, quando $A + A^T$ é uma matriz definida negativa [43]. Na maioria dos problemas de grande porte, essa propriedade é de difícil verificação.

Em resumo, podemos aplicar o Algoritmo 5 no sistema transladado (4.10) para encontrar um subespaço de projeção gerado por W_k e, neste espaço, encontrar uma aproximação P_k para a solução P da equação de Lyapunov (4.3). Os passos desse processo estão descritos no algoritmo 6 a seguir.

Algoritmo 6: PSEL - Método de projeção com Splitting para Equação de Lyapunov

Entrada: Matriz A, matriz B, escalares σ e α, um número inteiro i_{max} e uma tolerância tol.

Saída: Matriz P, solução aproximada para a equação $AP + PA^T = -BB^T$.

início
- $\tilde{A} \leftarrow A - \alpha I$;
- $\tilde{P} \leftarrow SEL(\tilde{A}, B, \sigma, i_{\max}, tol)$
- Calcular a decomposição svd econômica $U_k S_k W_k^H$ de $\tilde{P}$ considerando apenas os maiores valores principais.
- $R \leftarrow U_k^H A U_k^H$
- Calcular decomposição espectral ZDZ^{-1} de R. ($D = diag\{d_1, d_2, ..., d_k\}$)
- $V_k \leftarrow U_k Z$
- $\tilde{C}^{\Lambda_k} \leftarrow Z^{-1} U_k^H \tilde{P} U_k Z^{-H}$ (matriz de Cauchy generalizada do sistema transladado)
- **para** *i,j=1,2,...,k* **faça**
  - $C_{ij}^{\Lambda_k} = \tilde{C}_{ij}^{\Lambda_k} \left(d_i + d_j^H - 2\alpha\right)$;
- **fim**
- $P = V_k C^{\Lambda_k} V_k^H$

fim

retorna P

Uma vez definido o parâmetro σ, uma escolha para o valor α poderia ser feita resolvendo o seguinte problema de minimização a seguir, que é baseado na expressão (4.8).

$$\alpha = \arg\min_{\tilde{\alpha} \in \mathbb{R}_+^*} \left\{ \left| \frac{\lambda_i - \tilde{\alpha} + \sigma}{\lambda_j - \tilde{\alpha} - \sigma} \right| \right\}. \tag{4.20}$$

Porém, é preciso conhecer todos os autovalores de A para fazer esse cálculo, o que torna essa escolha inviável. Além disso, pode não ser vantajoso utilizar um valor de α elevado apenas para aumentar a convergência do método. Quando o maior (em módulo) autovalor de A é muito maior do que o menor (em módulo) autovalor de A, o sistema $AP + PA^T = -BB^T$ torna-se muito sensível à translações. Veja o exemplo a seguir:

Exemplo 4.8 *Vamos considerar a equação $AP+PA^T = -BB^T$ com a matriz A dada no exemplo 4.6 e $B = (1,1,1,1,1,1)^T$. A partir disso, definamos o sistema transladado $(A - \alpha I)P + P(A - \alpha I)^T = -BB^T$. A figura 4.1 a seguir mostra o decaimento dos autovalores da solução P para alguns valores de α. Para o cálculo da solução P utilizamos o comando "`lyap`"do Matlab.*

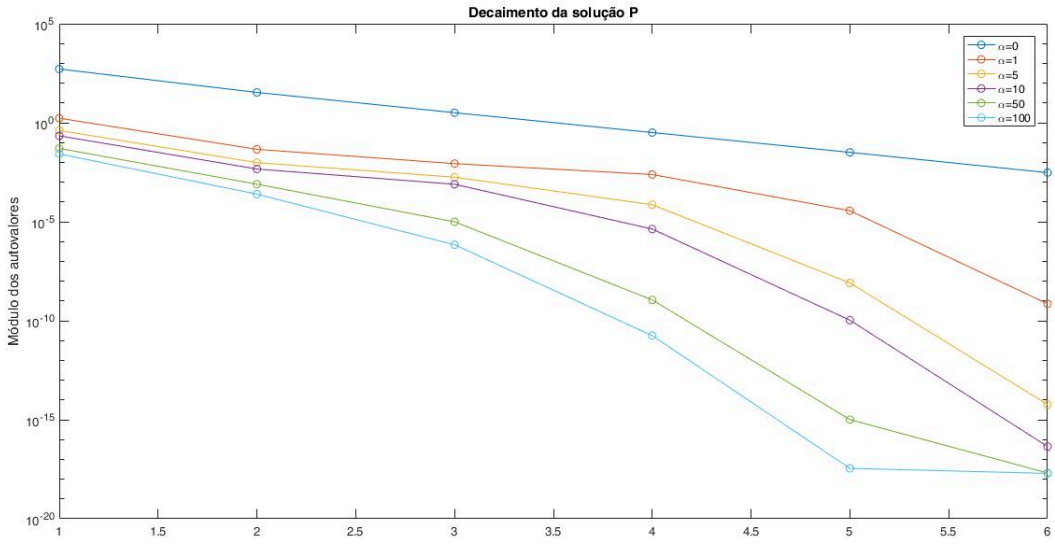


Figura 4.1: Decaimento dos autovalores de P para sistemas transladados $(A-\alpha I)P+P(A-\alpha I)^T = -BB^T$.

Portanto, na escolha do parâmetro α, é conveniente considerar como principal objetivo o de afastar ligeiramente a razão $\left|\frac{\lambda_i+\sigma}{\lambda_j-\sigma}\right|$ de 1 quando λ_i e λ_j são próximos à origem.

4.3 Testes Numéricos

Consideremos a equação de Lyapunov

$$AP + PA^T = -BB^T$$

com $A = T \otimes I + I \otimes T$ de ordem n^2 e T sendo uma matriz tridiagonal de ordem n dada por

$$T = \begin{bmatrix} 3 & -1 & & & \\ -1 & 3 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 3 & -1 \\ & & & -1 & 3 \end{bmatrix} \tag{4.21}$$

e $B = (1, 0, ..., 0)^T$. A distribuição dos autovalores de A é exibida na Figura 4.2.

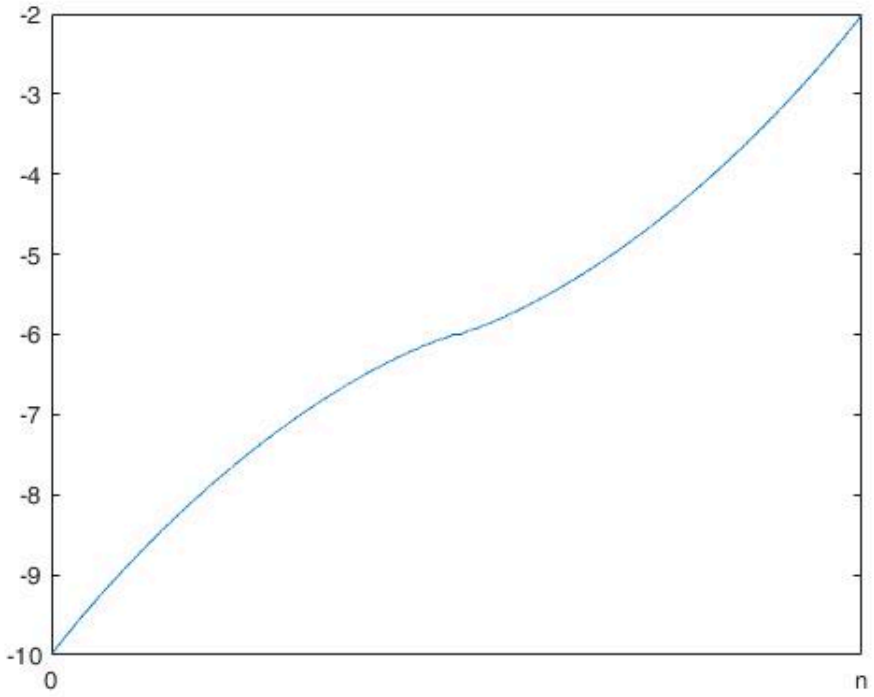


Figura 4.2: Distribuição de autovalores da matriz $A = T \otimes I + I \otimes T$,

Nesses moldes, construímos uma matriz A de ordem $n^2 = 900$ e aplicamos os algoritmos 5 e o algoritmo 6. A título de comparação, aplicamos dois métodos de projeção em subespaçoes de Krylov KPIK e RKSM [40] e também o método SLRCF-ADI [11]. A tabela a seguir apresenta alguns resultados obtidos com a aplicação desses métodos.

	SEL	PSEL	KPIK	RKSM	SLRCF-ADI
Erro abs.	$6.8 \cdot 10^{-3}$	$8,6 \cdot 10^{-4}$	$2.6 \cdot 10^{-07}$	$2.9 \cdot 10^{-07}$	$1,2 \cdot 10^{-11}$
Erro rel.	$4.2 \cdot 10^{-7}$	$5,4 \cdot 10^{-7}$	$1,6 \cdot 10^{-09}$	$8,1 \cdot 10^{-10}$	$1,3 \cdot 10^{-14}$
Dim. subspaço	—	—	12	8	—
Nº it.	20	20	6	8	10
Nº decomp. LU	1	1	1	8	12
Posto de P	10	10	9	7	9
Norma de P (Fro)	80.16156	80.16160	80.1616	80.16160	80.16160

Consideremos agora um problema similar, porém trocando a matriz T de (4.21) por

$$T = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

de modo que, para $n^2 = 900$, o autovalor de A de maior magnitude é $\lambda_1 = -7.9795$ e o menor autovalor de A em módulo é $\lambda_n = -0.0205$. Os resultados obtidos para esse novo sistema são mostrados na tabela a seguir.

	SEL	PSEL	KPIK	RKSM	SLRCF-ADI
Erro abs.	$15.6 \cdot 10^2$	$23,8$	$9.6 \cdot 10^{-5}$	$1.4 \cdot 10^{-3}$	$5.2 \cdot 10^{-4}$
Erro rel.	$1.6 \cdot 10^{-3}$	$4,4 \cdot 10^{-5}$	$4,3 \cdot 10^{-8}$	$8,1 \cdot 10^{-10}$	$1,6 \cdot 10^{-8}$
Dim. subspaço	—	—	20	10	—
Nº it.	20	20	10	10	10
Nº decomp. LU	1	1	1	8	12
Posto de P	10	11	17	10	10
Norma de P (Fro)	$7.29 \cdot 10^3$	$4.03 \cdot 10^3$	$4.06 \cdot 10^3$	$4.06 \cdot 10^3$	$4.06 \cdot 10^3$

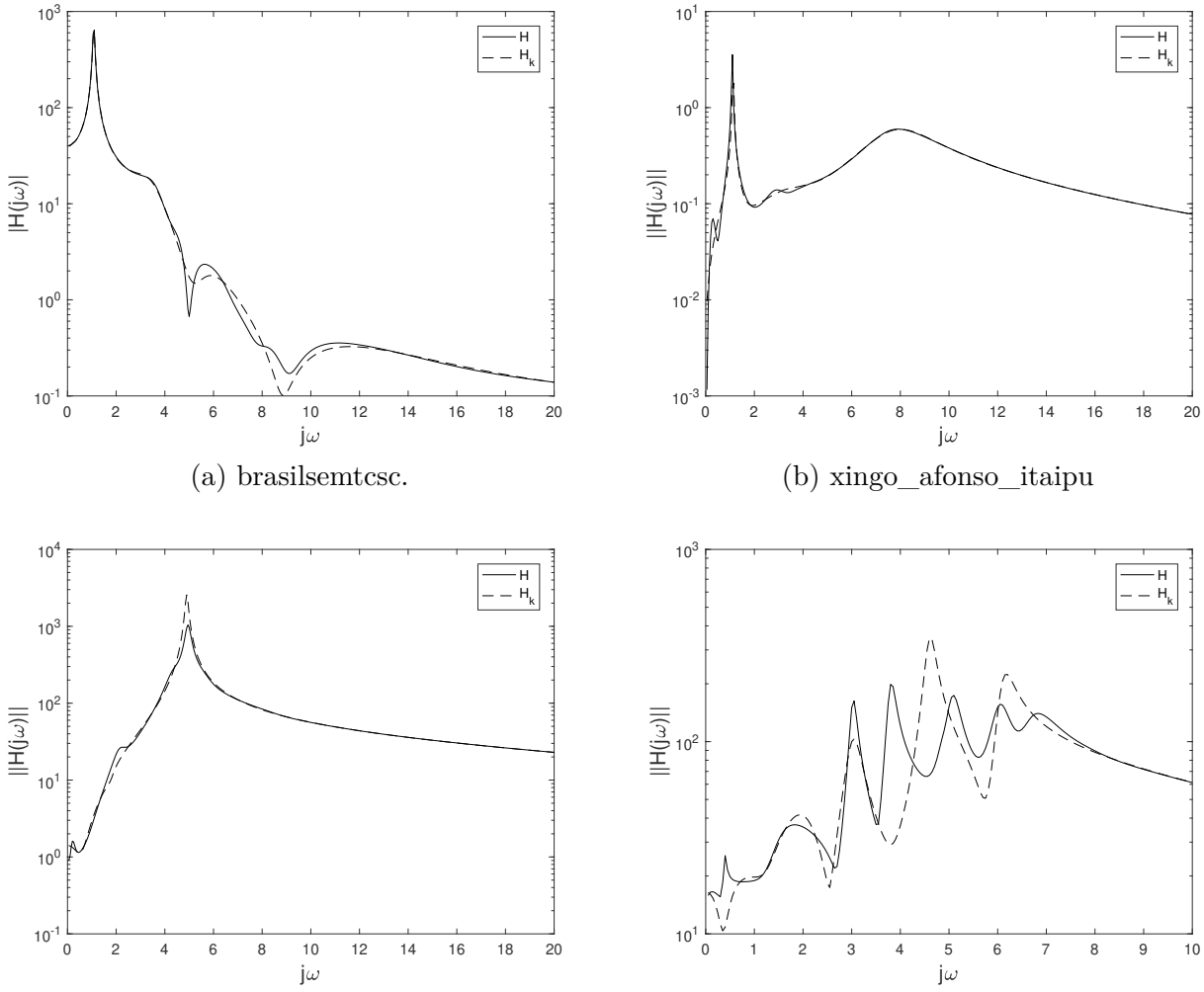
Nota-se que os métodos SEL e PSEL apresentam mais dificuldades em calcular a solução P quando a razão $\frac{\lambda_1}{\lambda_n}$ é grande. Acreditamos que essa deficiência seja ocasionada pela necessidade de utilizar um parâmetro σ que está distante de autovalores de A que estão próximos da origem e que, por isso, exercem grande influência na solução P . Nessas situações, os métodos KPIK, RKSM e SLRCF-ADI são mais robustos. No entanto, em casos em que isso não ocorre, os métodos SEL e PSEL tornam-se atrativos por exigir apenas uma decomposição LU da matriz A, além de serem métodos de fácil implementação. O método RKSM, por exemplo, necessita do cálculo das projeções $V_k^T A V_k$, que podem ser caras em sistemas descritores de grande porte. Além disso, não há garantia de convergência para os casos em que a matriz A é não dissipativa. O método SLRCF-ADI, por sua vez, além de necessitar de uma decomposição LU para cada iteração, também depende de um conjunto de parâmetros calculados previamente.

Por último, vamos analisar o comportamento do método SEL no cálculo de soluções de equações de Lyapunov necessárias para a redução por balanceamento de um modelo descritor. Repare que a estrutura do algoritmo SEL permite aplicar o método tanto em sistemas SISO como em sistemas MIMO. No entanto, a dificuldade que surge é justamente relacionada à forma como os autovalores da matriz A estão distribuídos no plano complexo. No modelo `brasilsemtcsc`, por exemplo, a magnitude dos autovalores de A vai da ordem de 10^{-5} até a ordem 10^4. Isso praticamente inviabiliza a utilização dos métodos SEL e PSEL nessa situação, pois o valor de σ teria que ser

muito grande, distante demais dos autovalores de A, fazendo com que a convergência seja muito lenta.

Porém, nós obtivemos resultados bastante curiosos aplicando o método SEL da seguinte forma: ao invés de buscar um valor de σ que satisfaça o Teorema 4.2, escolhemos um valor de σ, ainda positivo, porém mais próximo da origem. Nesse caso, o método SEL não converge. Para controlar a norma das matrizes iteradas durante a execução do método, decidimos normalizar as matrizes Z_k e Y_k a cada iteração do algoritmo 5. Além dos modelos `brasilsemtcsc`, já apresentado anteriormente, aplicamos o nosso método nos modelo SISO `xingo_afonso_itaipu` e nos modelos `bips98_606` `xingo3012` que são do tipo MIMO, ambos retirados da página *https://sites.google.com/site/ron*

Figura 4.3: Magnitude da função $H_k(j\omega)$ em comparação com a função original $H(j\omega)$.



(a) brasilsemtcsc.

(b) xingo_afonso_itaipu

(c) xingo3012.

(d) bips98_606.

O sistema `xingo_afonso_itaipu` possui uma matriz jacobiana J de ordem 13250, enquanto a matriz A tem tamanho $n = 1664$. O modelo `bips98_606` é de ordem 7135 e tem $n = 606$ variáveis de estado. As matrizes B e C desse sistema são de tamanho $n \times m$, com $m = 4$. Por fim, no modelo `xingo3012`, temos uma matriz jacobiana de ordem 20944 com $n = 3012$ variáveis

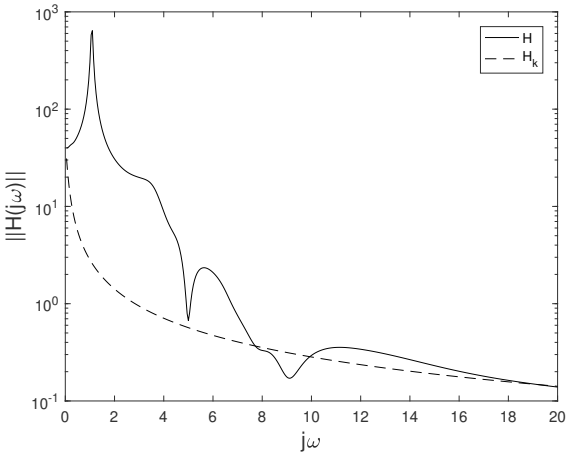
de estado e cada uma das matrizes B e C possuem $m = 2$ colunas.

Os modelos reduzidos plotados na figura 4.3 foram construídos utilizando redução por balanceamento. As soluções P e Q das equações de Lyapunov relacionadas a cada sistema foram calculadas utilizando o método SEL com $\sigma = 5, 5$ e 20 iterações apenas. Conforme mencionamos anteriormente, foi adicionada uma etapa auxiliar no método, que consiste em normalizar as matrizes Z_k e Y_k do algoritmo 5 a cada passo.

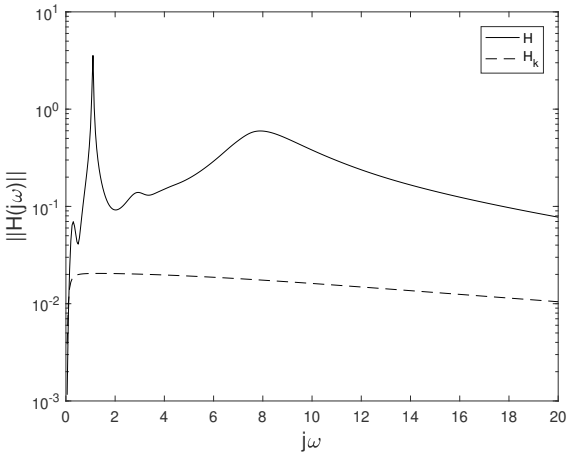
Note que, de acordo com a figura figura 4.3, principalmente nos modelos em que a função de transferência possui apenas um pico significativo, os gráficos dos modelos reduzidos possuem traços bastante fidedignos aos do modelo original no que diz respeito a magnitude de $H(j\omega)$. Isso é interessante por dois motivos: primeiro que não há convergência do método SEL nesse caso e segundo que foram necessárias poucas iterações do algoritmo 5, que necessita apenas de uma decomposição LU da matriz A em cada modelo testado. Relembre que a redução de um sistema por balanceamento consiste em projetar as matrizes do sistema numa base construída a partir das soluções P e Q das equações de Lyapunov associadas. Por isso, os testes apresentados na figura 4.3 nos induzem a acreditar que a base que está sendo construída pelo método SEL nessas aplicações está correta no ponto de vista geométrico. No entanto, é preciso ainda descobrir a maneira correta de controlar a norma das matrizes iteradas a fim de promover a convergência do método.

Um outro ponto importante, que provavelmente seja uma curiosidade do leitor nesse momento, é que o parâmetro σ positivo continua sendo necessário para obter bons resultados, mesmo que ele não seja escolhido visando a convergência do método SEL. Para ilustrar esse fato, repetimos os testes com o SEL nos sistemas `brasilsemtcsc` e `xingo_afonso_itaipu` utilizando $\sigma = 0$ e $\sigma = -5$. Os resultados são exibidos na figura a seguir:

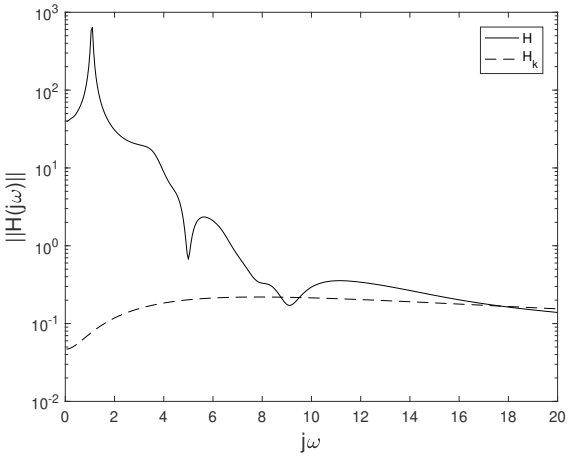
Figura 4.4: Magnitude da função $H_k(j\omega)$ em comparação com a função original $H(j\omega)$ (outros valores de σ).



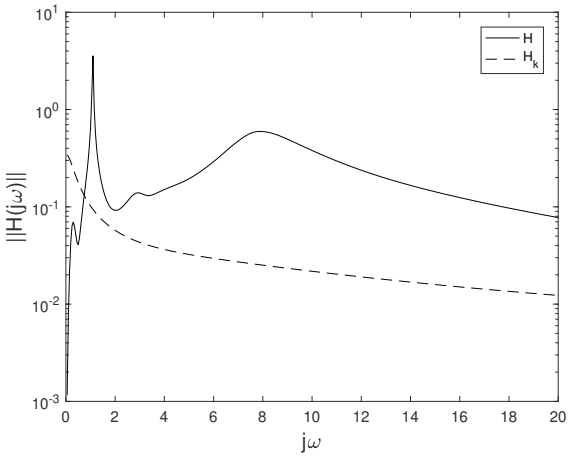
(a) brasilsemtcsc com $\sigma = 0$.



(b) xingo_afonso_itaipu com $\sigma = 0$



(c) brasilsemtcsc com $\sigma = -5$.



(d) xingo_afonso_itaipu com $\sigma = -5$

De acordo com o Teorema 4.3, a convergência do método SEL é garantida quando $\sigma > 0$ é grande o suficiente para satisfazer a condição 4.9. Essa análise leva em consideração o espectro todo da matriz A. No entanto, os testes que acabamos de exibir nos levam a acreditar que essa condição possa ser substituída por alguma outra que leve em consideração apenas um conjunto de polos dominantes do sistema dinâmico que, em geral, está concentrado numa região próxima a origem do plano complexo.

Capítulo 5

Comparativo entre os métodos de redução de modelo com exemplos numéricos.

Embora os capítulos anteriores já contenham algumas aplicações numéricas dos métodos abordados nesse trabalho, optamos por destinar um capítulo exclusivamente para esse fim.

Optamos por focar na eficácia dos métodos em redução de modelos descritores esparsos. Isso permite que façamos uma comparação ampla entre métodos. Além de comparar, entre si, métodos de resolução da equação de Lyapunov, podemos comparar métodos de redução por balanceamento com métodos de redução modal. Essa comparação não é muito comum na literatura pois, como mencionado em [46], o custo computacional do método de redução por balanceamento consiste num processo computacionalmente caro se comparado com a redução modal. No entanto, os aprimoramentos que propomos para métodos como o SLRCF-ADI e o RKSM, bem como a inserção de um novo método do tipo *splitting*, tornam essa diferença menos significativa. Além disso, essa comparação permite também testar critérios de dominância (como a PQ-dominância) de polos em sistemas dinâmicos.

Para este capítulo escolhemos alguns dos sistemas de teste disponíveis na página *https://sites.google.com* Esses sistemas são oriundos do Centro de Pesquisas de Energia Elétrica (Eletrobras Cepel). A tabela a seguir contém os nomes e as principais características dos modelos que escolhemos para este trabalho.

Tabela 5.1: Modelos Descritores

Modelo	# Estados	# Var. algébricas	# Entradas	# Saídas
`xingo_afonso_itaipu`	1664	11586	1	1
`ww_vref_6405`	1664	11587	1	1
`nopss_11k`	1257	10428	1	1
`bips97_1676`	1676	11633	8	8

Recordemos que, quando o sistema dinâmico possui apenas uma entrada e uma saída, é chamado de SISO. Quando há múltiplas entradas e múltiplas saídas, chamamos o sistema dinâmico de MIMO. Em todos os sistemas da tabela 5.1, a matriz D é nula e, por isso, foi desconsiderada na execução dos métodos implementados.

Os sistemas `xingo_afonso_itaipu`, `ww_vref_6405` e `bips97_1676` são estáveis, enquanto que a matriz A do sistema `nopss_11k` possui alguns autovalores com parte real positiva, ou seja, trata-se de um sistema que não é estável. Embora todos os estudos relacionados a redução por balanceamento descritos nesse trabalho dependam da hipótese de que o sistema é estável, tivemos a curiosidade de realizar testes em um sistema que não satisfaz essa condição. Para nossa surpresa, como pode-se verificar nos resultados exibidos na próxima seção, esse fato parece não ter afetado o desempenho de nenhum dos métodos. No entanto, acreditamos que isso seja algo esporádico.

Nos exemplos numéricos desse capítulo, tal como nos testes feitos anteriormente, a eficiência dos métodos de redução de modelo é analisada com base no gráfico da magnitude da função de transferência H no domínio das frequências. Recorde que, pela identidade de Parseval (2.16), o comportamento da função H no domínio das frequências está intimamente ligado à performance do sistema dinâmico no domínio do tempo.

5.1 Testes em sistemas SISO

Nesta seção apresentamos testes realizados apenas com os quatro primeiros sistemas descritos na tabela 5.1. Num primeiro momento expomos resultados obtidos com a redução de modelo utilizando cada um dos algoritmos: SLRCF-ADI, RKSM e SEL. Em seguida, exibimos resultados obtidos com a utilização da redução modal.

5.1.1 Redução por Balanceamento via Método SLRCF-ADI

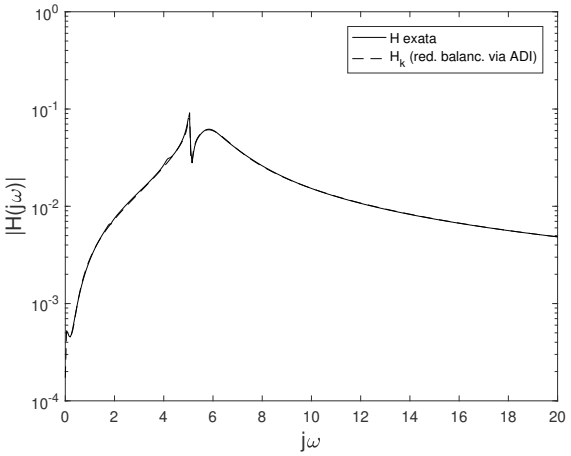
Ao aplicar o método SLRCF-ADI para encontrar o par de soluções das equações de Lyapunov para a redução de modelo, optamos em limitar apenas o número de iterações do método, sem utilizar o critério de parada definido em na expressão (3.18). Fizemos isso para poder tecer algumas observações relacionadas ao critério de parada do método no contexto da redução de

modelo.

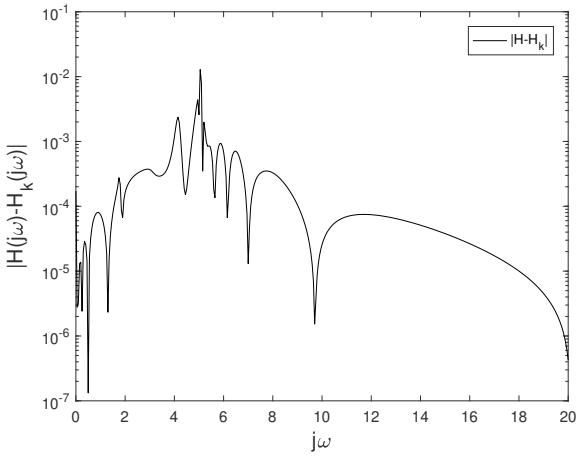
Durante a execução do método, foi utilizado um conjunto de apenas quatro parâmetros μ_i (parâmetros ADI). Esses parâmetros são os quatro autovalores dominantes (definição 3.13) na solução da equação de Lyapunov. Para calcular esses autovalores dominantes foi necessário obter a decomposição espectral completa da matriz A de cada modelo, o que encarece o método. No entanto, acreditamos que os resultados apresentados nesse trabalho podem motivar a construção, em projetos futuros, de um método que calcule esses autovalores de maneira mais eficiente.

A figura 5.1, a seguir, apresenta um comparativo entre a função de transferência $H(j\omega)$, original do sistema `nopss_11k`, e a função de transferência $H_k(j\omega)$ do sistema reduzido. Devido ao método ADI ser cíclico sobre um conjunto com quatro parâmetros, o número de iterações do método em cada teste é sempre um múltiplo do número quatro.

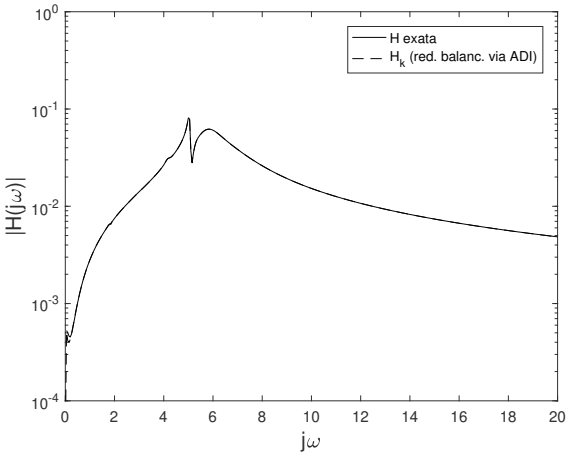
Figura 5.1: Redução por balanceamento utilizando método SLRCF-ADI no modelo `nopss_11k`.



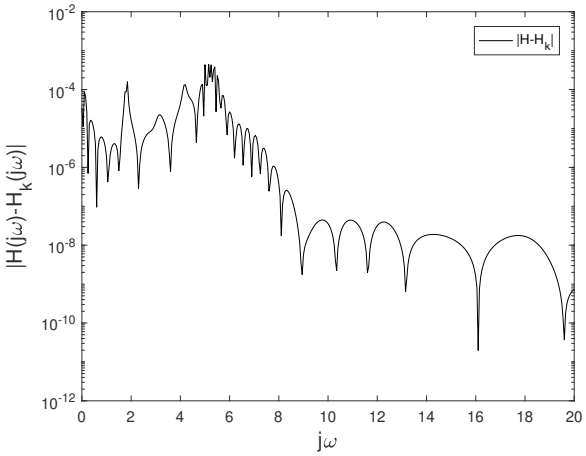
(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (12 it. do mét. SLRCF-ADI).



(b) Erro $|H(j\omega) - H_k(j\omega)|$ (12 it. do mét. SLRCF-ADI).



(c) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (80 it. do mét. SLRCF-ADI).



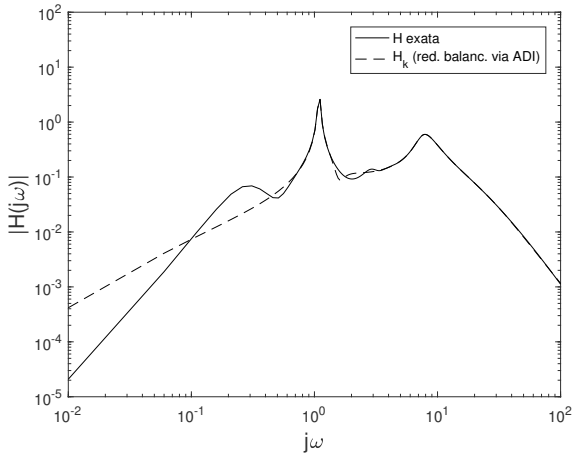
(d) Erro $|H(j\omega) - H_k(j\omega)|$ (80 iterações do método SLRCF-ADI).

Antes de qualquer comparação, algo que nos chama a atenção logo nesse primeiro teste é que,

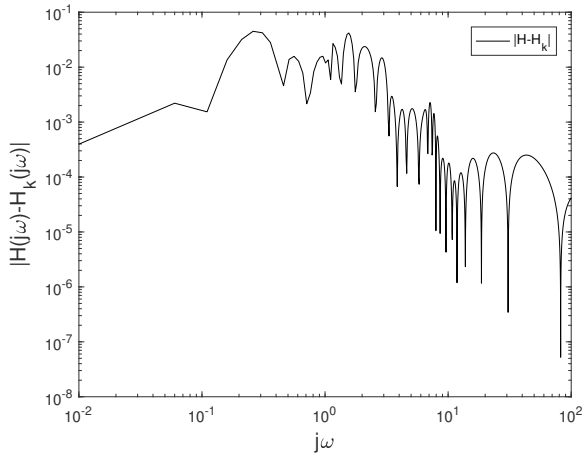
com base na figura 5.1, percebemos que 20 iterações do método SLRCF-ADI já são suficientes para obter um modelo reduzido que retrata o modelo `nopss_11k` de maneira bastante fiel.

A figura 5.2, a seguir, exibe os resultados obtidos com a utilização do método SLRCF-ADI no modelo `xingo_afonso_itaipu`.

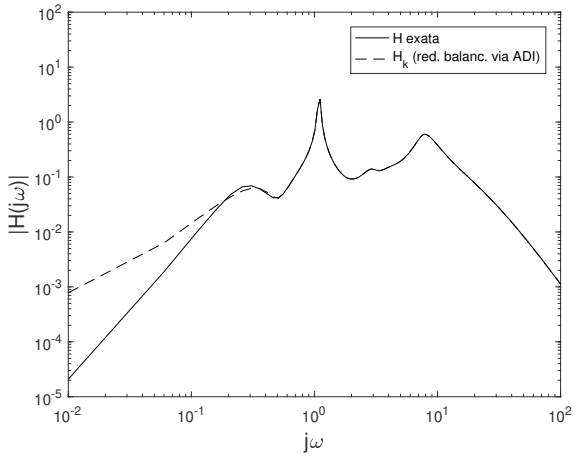
Figura 5.2: Redução por balanceamento utilizando método SLRCF-ADI no modelo `xingo_afonso_itaipu`.



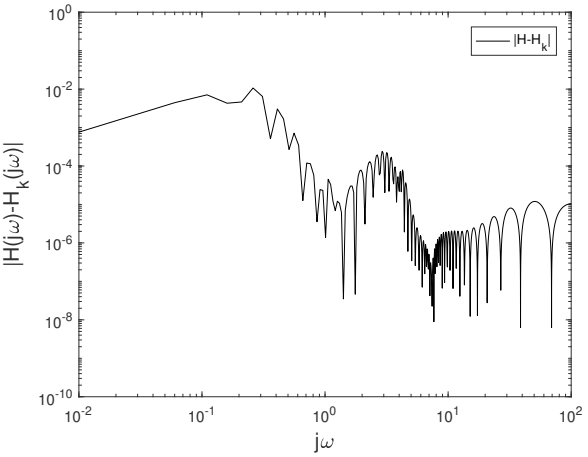
(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (12 it. do mét. SLRCF-ADI).



(b) Erro $|H(j\omega) - H_k(j\omega)|$ (12 iterações do método SLRCF-ADI).



(c) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (80 it. do mét. SLRCF-ADI).

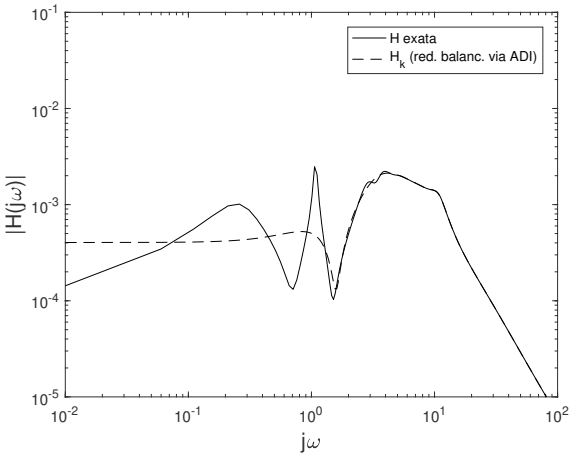


(d) Erro $|H(j\omega) - H_k(j\omega)|$ (80 iterações do método SLRCF-ADI).

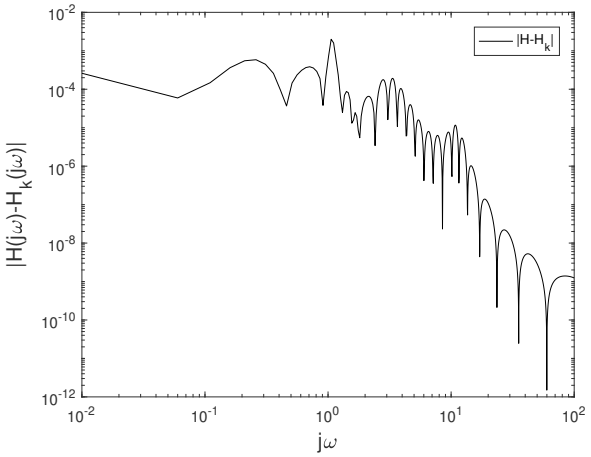
De maneira similar ao que ocorreu no exemplo anterior, foi possível obter um sistema reduzido bastante fidedigno ao original `xingo_afonso_itaipu` logo nas primeiras iterações.

Por fim, exibimos os gráficos da comparação entre $H(j\omega)$ e $H_k(j\omega)$ no modelo `ww_vref_6405`.

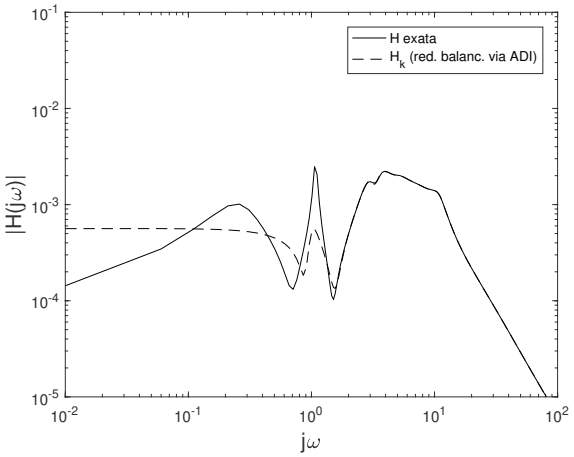
Figura 5.3: Redução por balanceamento utilizando método SLRCF-ADI no modelo `ww_vref_6405`.



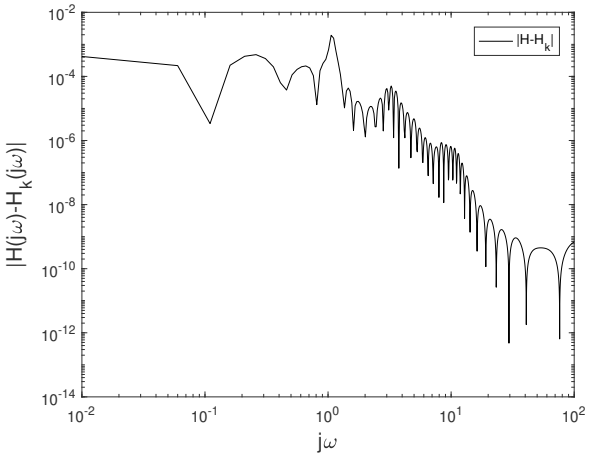
(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (12 it. do mét. SLRCF-ADI).



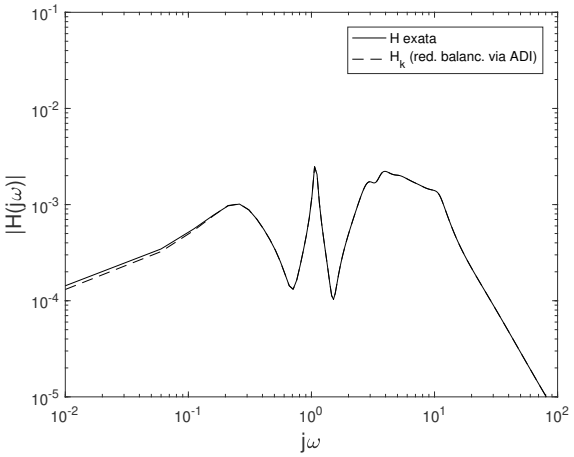
(b) Erro $|H(j\omega) - H_k(j\omega)|$ (12 it. do mét. SLRCF-ADI).



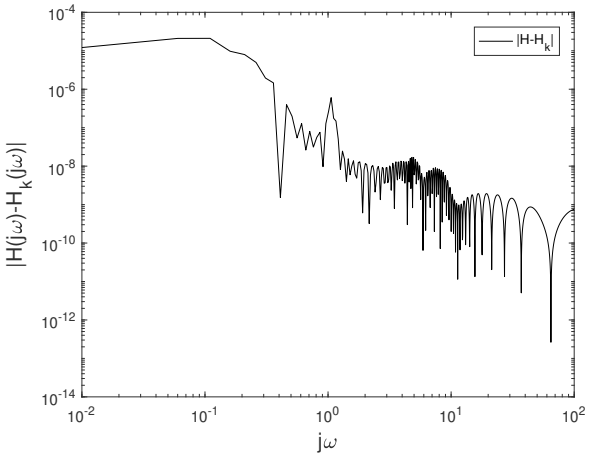
(c) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (28 it. do mét. SLRCF-ADI).



(d) Erro $|H(j\omega) - H_k(j\omega)|$ (28 it. do mét. SLRCF-ADI).



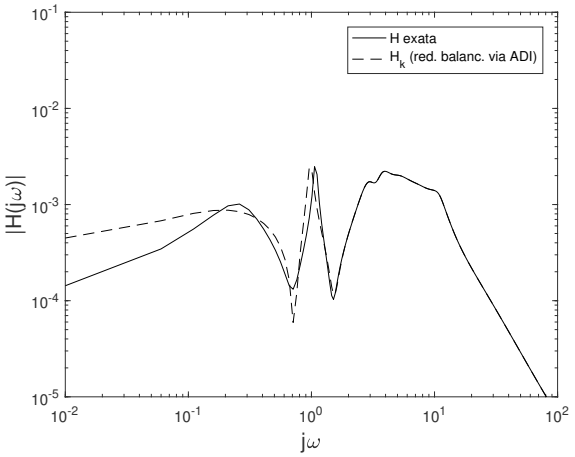
(e) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$ (80 it. do mét. SLRCF-ADI).



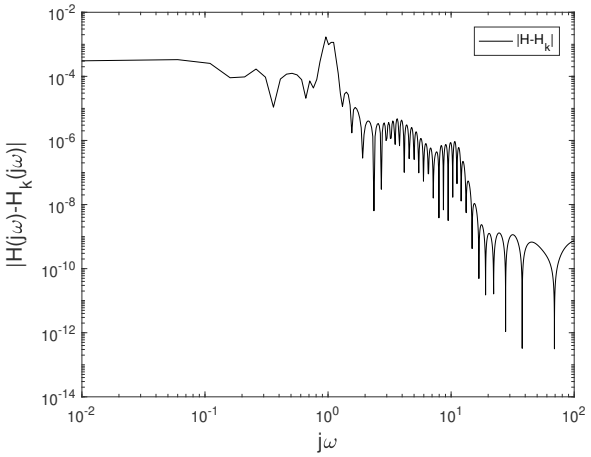
(f) Erro $|H(j\omega) - H_k(j\omega)|$ (80 it. do mét. SLRCF-ADI).

Diferentemente do que ocorreu nos dois exemplos numéricos anteriores, os resultados numéricos obtidos com o modelo `ww_vref_6405` e mostrados na figura 5.4 mostram que, nesse caso, é preciso mais do que 20 iterações do método para que o gráfico do modelo reduzido seja coerente com o original nos principais picos de $|H(j\omega)|$. Acreditamos que isso se deve ao fato desse último modelo possuir uma quantidade maior de polos com dominância efetiva significativa, como podemos ver nos resultados apresentados na próxima subseção. Acreditamos que, para um bom aproveitamento do método SLRCF-ADI, a quantidade de parâmetros utilizados deve ser, de alguma forma, proporcional a quantidade de polos que exercem dominância efetiva no sistema dinâmico. Para ilustrar essa afirmação, veja a seguir os resultados obtidos para o mesmo modelo, utilizando o método SLRCF-ADI com 10 parâmetros ao invés de 4 e com 20 iterações apenas.

Figura 5.4: Redução por balanceamento utilizando 20 iterações do método SLRCF-ADI (com 10 parâmetros) no modelo `ww_vref_6405`.



(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.



(b) Erro $|H(j\omega) - H_k(j\omega)|$.

Vamos tratar agora sobre as questões relacionadas ao critério de parada mencionado no início da subseção. Veja, na figura 5.5 o comportamento dos erros relativos $\frac{\|P_{k+1}-P_k\|_F}{\|P_k\|_F}$ e $\frac{\|Q_{k+1}-Q_k\|_F}{\|Q_k\|_F}$, calculados separadamente durante a execução do método de redução do método SLRCF-ADI no sistema de testes `xingo_afonso_itaipu`.

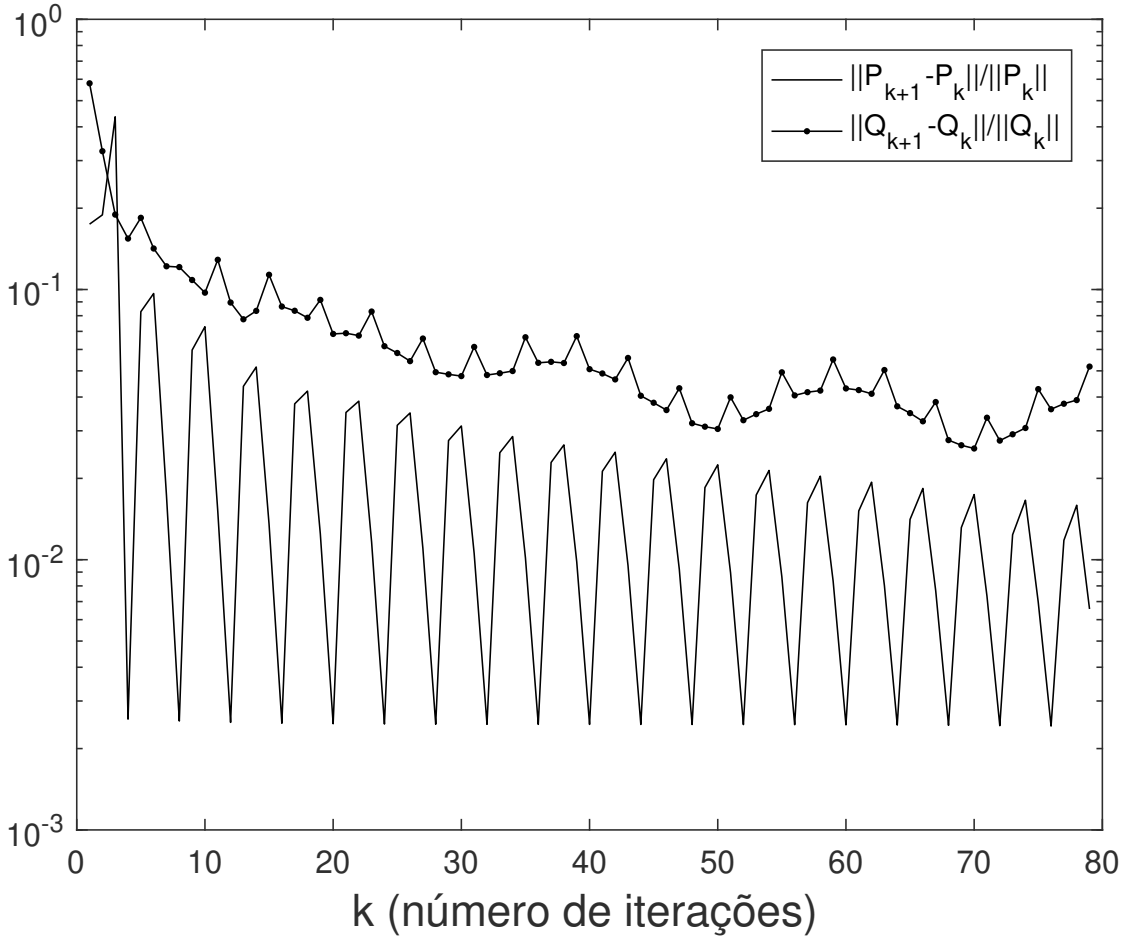


Figura 5.5: Erro relativo calculado com as as aproximações P_k e Q_k separadamente (modelo `xingo_afonso_itaipu`)

Como era de se esperar, a diminuição do erro relativo em ambas as situações apresentadas na figura 5.5 segue um padrão cíclico. Isso sugere que, num possível critério de parada, essa aferição deveria ser feita uma vez a cada ciclo apenas, visto que há discrepância entre os acréscimos relativos causados por cada um dos parâmetro utilizados. Vale observar também que, pelo que foi visto na seção 2.1, a precisão do modelo reduzido não depende necessariamente de se ter boas aproximações para as soluções P e Q, separadamente, para cada uma das equações de Lyapunov referentes. Na verdade, a eficiência da redução por balanceamento depende de se ter uma boa aproximação para o produto de matrizes PQ. O gráfico da figura 5.6, a seguir, exibe o comportamento do acréscimo relativo $\frac{\|P_{4k}Q_{4k}-P_{4(k-1)}Q_{4(k-1)}\|_F}{\|P_{4(k-1)}Q_{4(k-1)}\|_F}$, calculado a cada ciclo de 4 iterações durante o cálculo simultâneo das aproximações P_k e Q_k no sistema de testes `xingo_afonso_itaipu`.

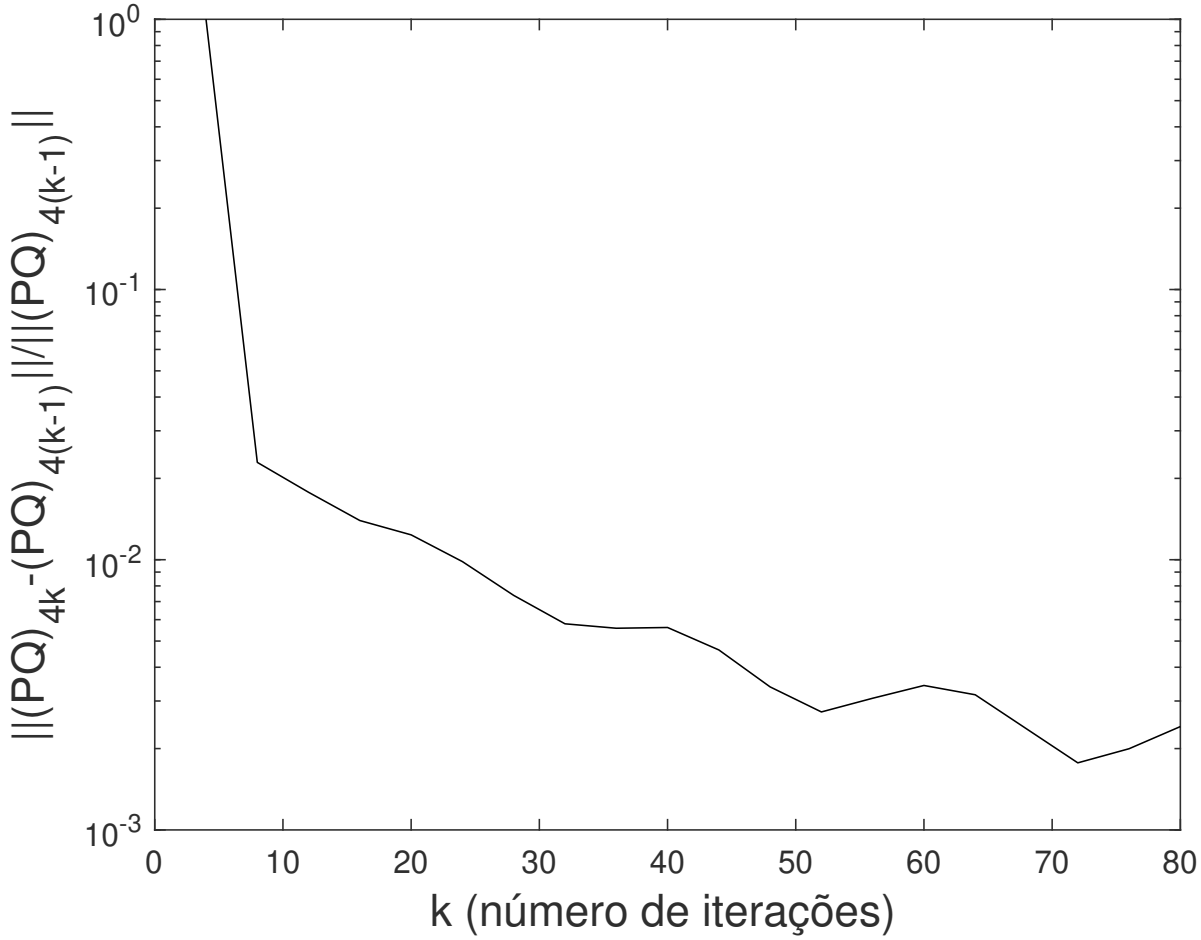


Figura 5.6: Acréscimo relativo de P_kQ_k a cada 4 iterações (modelo `xingo_afonso_itaipu`)

Acreditamos que, na redução de modelo por balanceamento utilizando o método SLRCF-ADI, é conveniente definir um critério de parada baseado em uma tolerância para os acréscimos relativos apresentados no gráfico da figura 5.6. Além das justificativas que já mencionamos, imagine a seguinte situação: suponha que, com base num critério de parada estabelecido para cada equação de Lyapunov, sejam calculadas aproximações P_i e Q_j de modo que $rank(P_i) << rank(Q_j)$. Como $rank(P_iQ_j) \leq min\{rank(P_i), rank(Q_j)\}$, ao multiplicar essas duas matrizes, estaríamos, de certa forma, "desperdiçando"uma parte da solução Q_i sem antes saber da relevância dela nos sistema dinâmico como um todo. Sabemos que o processo de calcular os acréscimos relativos a P_kQ_k é caro para ser feito durante a execução do método. Isso nos motiva a pensar em estratégias eficientes para gerar estimativas para esses acréscimos em projetos futuros.

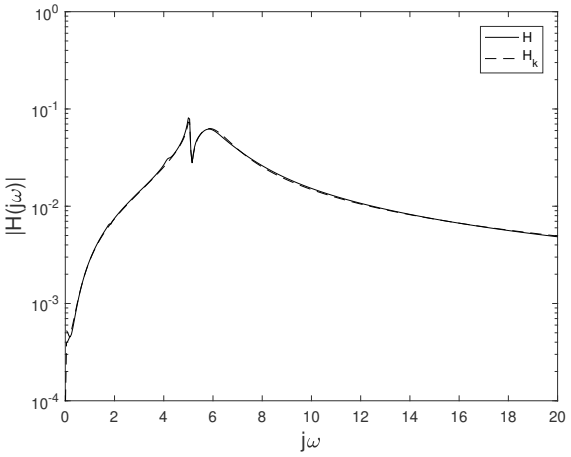
5.1.2 Redução por Balanceamento via RKSM

Nesta seção exibimos os resultados obtidos com a redução por balanceamento utilizando o método RKSM, que é baseado em subespaços de Krylov racionais. Nesses testes, utilizamos como critério de parada uma tolerância de 10^{-4} para uma estimativa da expressão (3.23). No que diz

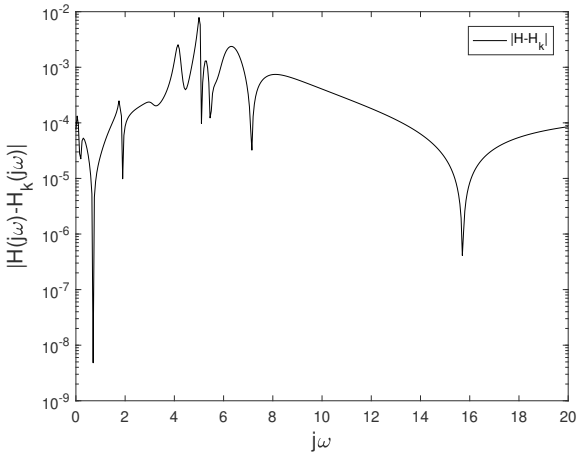
respeito aos parâmetros μ_i utilizados na construção da base do subespaço de Krilov racional, método RKSM depende de uma estimativa inicial (t_0, t_1) para o intervalo que contém os valores absolutos dos autovalores de A. Nos testes exibidos nesta seção, utilizamos $t_0 = 3 \cdot 10^{-5}$ e $t_1 = 200$. Essa escolha é baseada na discussão, feita na subseção 3.3.5, sobre autovalores de A que são dominantes na solução da equação de Lyapunov.

Nas figuras 5.7, 5.8 e 5.9, a seguir, são mostrados os resultados utilizando RKSM na redução dos modelos `nopss_11k RKSM`, `xingo_afonso_itaipu` e `ww_vref_6405`, respectivamente.

Figura 5.7: Comparativo entre a função de transferência H original do sistema `noops_11k` e o modelo H_k reduzido por balanceamento via RKSM.

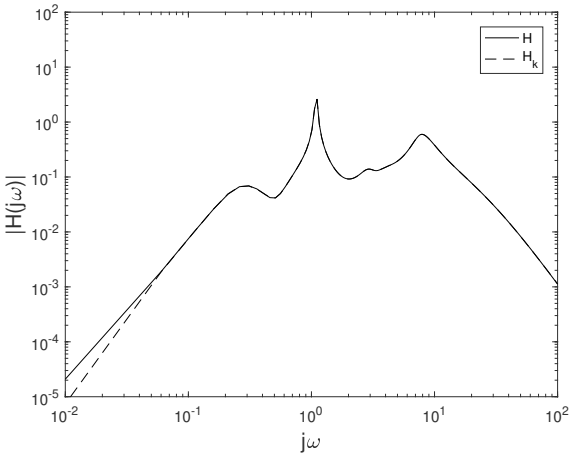


(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.

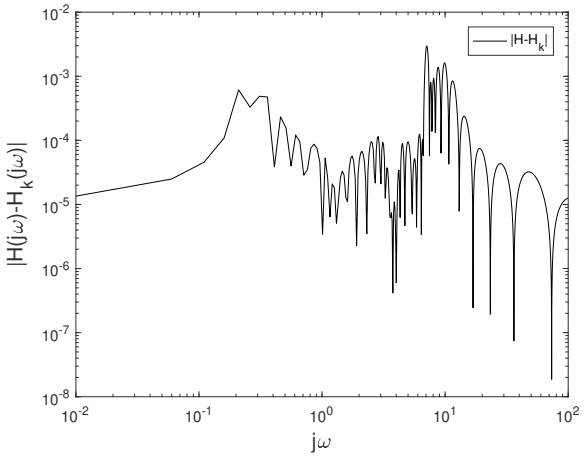


(b) Erro $|H(j\omega) - H_k(j\omega)|$.

Figura 5.8: Comparativo entre a função de transferência H original do sistema `xingo_afonso_itaipu` e o modelo H_k reduzido por balanceamento via RKSM.

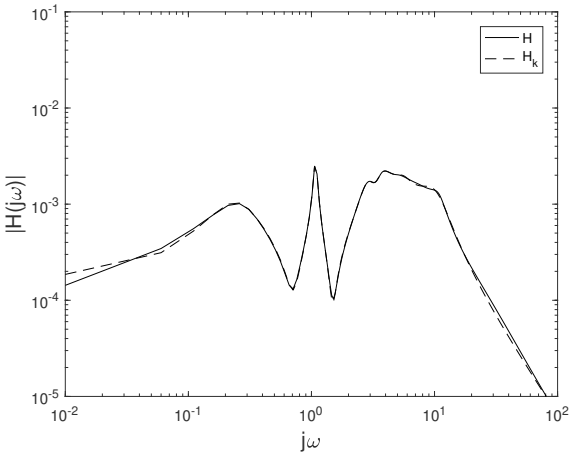


(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.

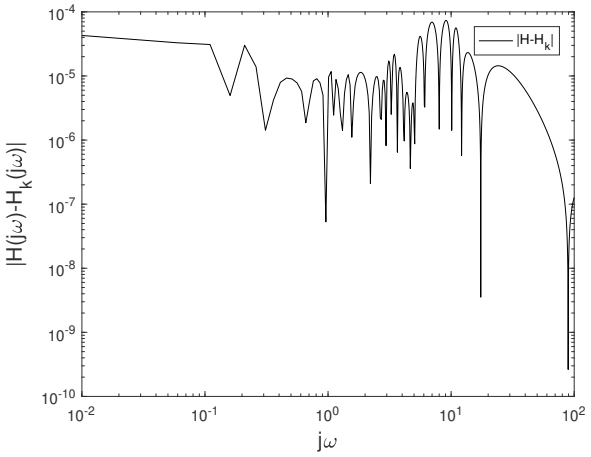


(b) Erro $|H(j\omega) - H_k(j\omega)|$.

Figura 5.9: Comparativo entre a função de transferência H original do sistema `ww_vref_6405` e o modelo H_k reduzido por balanceamento via RKSM.



(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.



(b) Erro $|H(j\omega) - H_k(j\omega)|$.

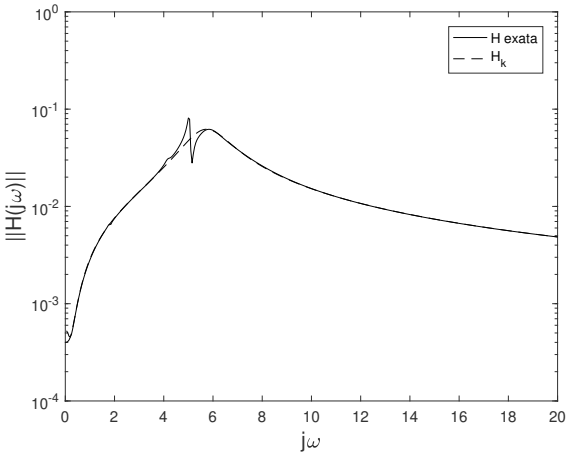
O método de redução utilizando o algoritmo RKSM, além de produzir modelos reduzidos significativamente fiéis, mostra-se como uma alternativa bastante robusta e autônoma. Para que o método seja inicializado, o usuário precisa fornecer apenas as matrizes A e B, um par de estimativas t_0 e t_1 para os autovalores de A, além da tolerância desejada para o critério de parada. No entanto, essa autonomia é compensada por um alto custo computacional, pois, por se tratar de um método adaptativo, requer uma nova decomposição LU da matriz A a cada iteração. Essa constatação fica mais evidente na tabela exibida ao final desta seção.

5.1.3 Redução por Balanceamento via Método SEL

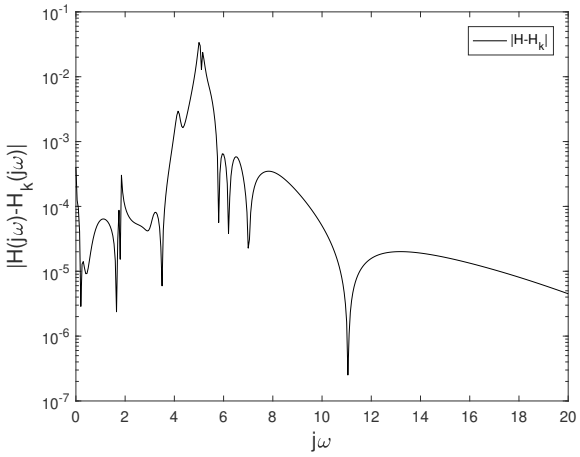
Nesta subseção ilustramos as experimentações numéricas feitas com o nosso método para resolução da equação de Lyapunov baseado em métodos splitting para sistemas lineares. Mais especificamente, testamos o algoritmo 5, que chamamos de SEL.

Como já foi mencionado no capítulo anterior, ainda não conseguimos compreender perfeitamente o funcionamento do método SEL em equações de Lyapunov cuja matriz A possui alguns autovalores muito próximos à origem e outros com valor absoluto muito grande. Nos testes feitos com o SEL para esta subseção, consideramos o parâmetro $\sigma = 5,5$ e adicionamos a etapa auxiliar que consiste em normalizar, a cada iteração, as matrizes Y_k e Z_k que aparecem nas linhas 9 e 10 do algoritmo 5. Como já mencionamos, essa etapa auxiliar se deve ao fato de que o método não converge em situações em que o valor absoluto de algum autovalor é muito grande comparado com σ (veja o Teorema 4.3). Cada um dos resultados apresentados nas figuras a seguir foi obtido com 20 iterações do método SEL.

Figura 5.10: Comparativo entre a função de transferência H original do sistema `noops_11k` e o modelo H_k reduzido por balanceamento via SEL.

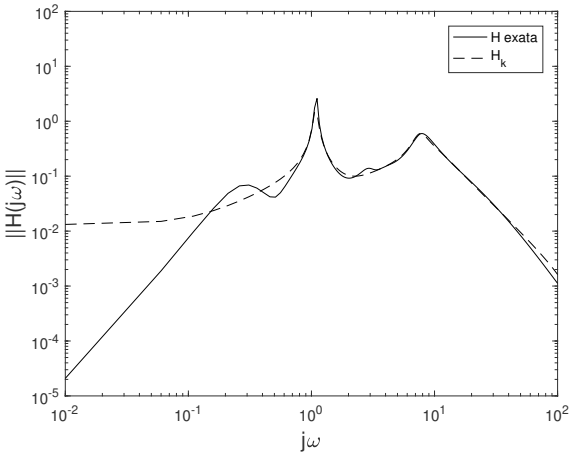


(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.

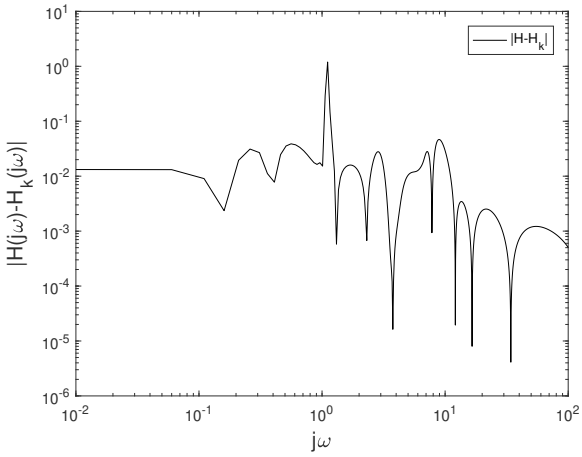


(b) Erro $|H(j\omega) - H_k(j\omega)|$.

Figura 5.11: Comparativo entre a função de transferência H original do sistema `xingo_afonso_itaipu` e o modelo H_k reduzido por balanceamento via SEL.

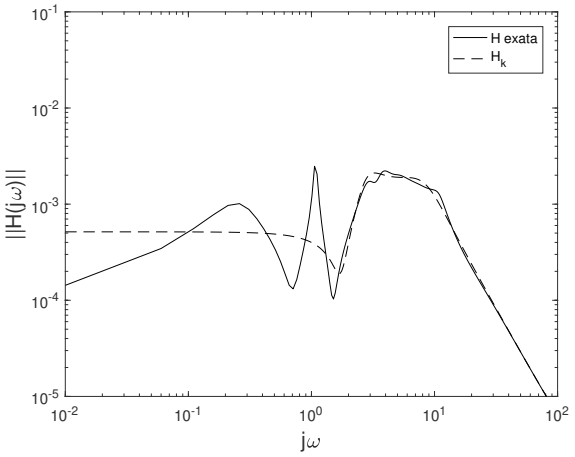


(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.

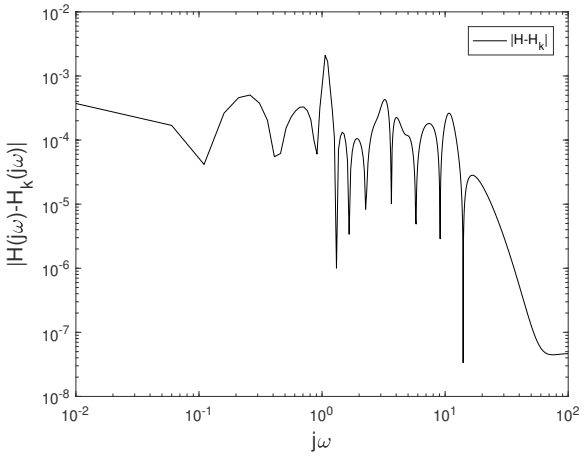


(b) Erro $|H(j\omega) - H_k(j\omega)|$.

Figura 5.12: Comparativo entre a função de transferência H original do sistema `ww_vref_6405` e o modelo H_k reduzido por balanceamento via SEL.



(a) Magnitude de $H(j\omega)$ e de $H_k(j\omega)$.



(b) Erro $|H(j\omega) - H_k(j\omega)|$.

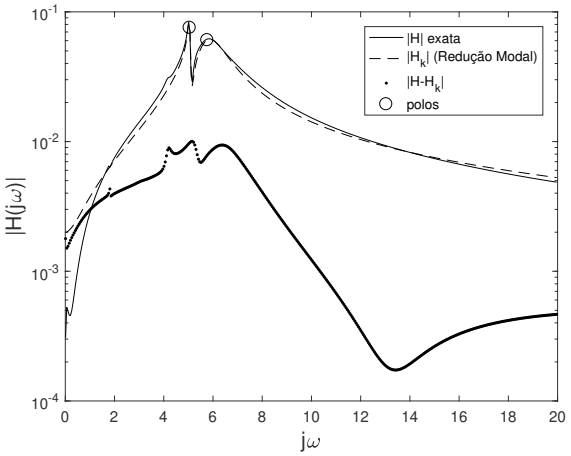
Esses testes preliminares utilizando o método SEL em redução de modelo mostram que essa estratégia é bastante promissora, principalmente em sistemas dinâmicos cuja função de transferência possui poucos "picos"de magnitude. Vale lembrar que, utilizando o método SEL, a resolução de uma equação de Lyapunov necessita de apenas uma decomposição LU da matriz A. Isso deixa o algoritmo SEL em uma posição de destaque. Ao final dessa seção são exibidos mais detalhes como, por exemplo, o tamanho de cada um dos sistemas reduzidos obtidos.

5.1.4 Redução Modal

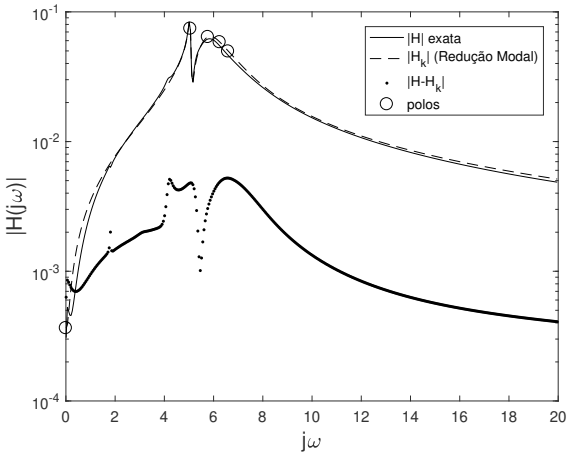
Nesta subseção exibimos os resultados que obtivemos utilizando a redução modal. Separamos os testes em duas partes. Na primeira parte, utilizamos o método DPSE [34] para calcular um conjunto de 50 polos dominantes para cada modelo e, em seguida, utilizamos a definição de PQ-dominância (definições 3.15 e 3.15) para ordená-los. Na segunda parte, para termos uma noção de como seria a redução modal se os polos dominantes fossem calculados de maneira ótima, utilizamos a decomposição espectral completa da matriz A de cada sistema para calcular os polos dominantes no contexto da definição de PQ-dominância.

A figura 5.13 mostra a evolução da versão reduzida do modelo `nopss_11k` ao passo que aumentamos o número de polos dominantes considerados.

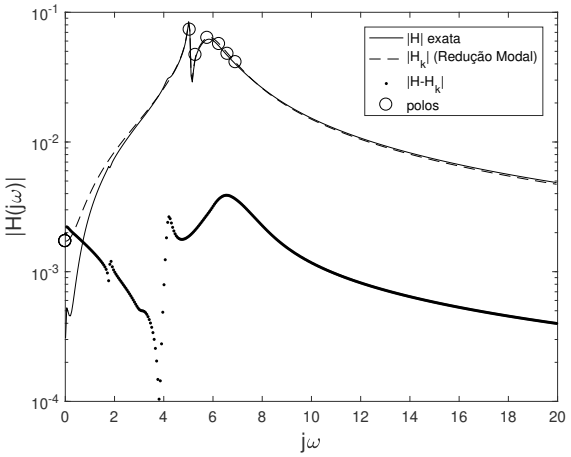
Figura 5.13: Função original versus modelo reduzido (Redução Modal) e erro absoluto para o sistema `noops_11k` (polos dominantes calculados com DPSE e ordenados por PQ-dominância.



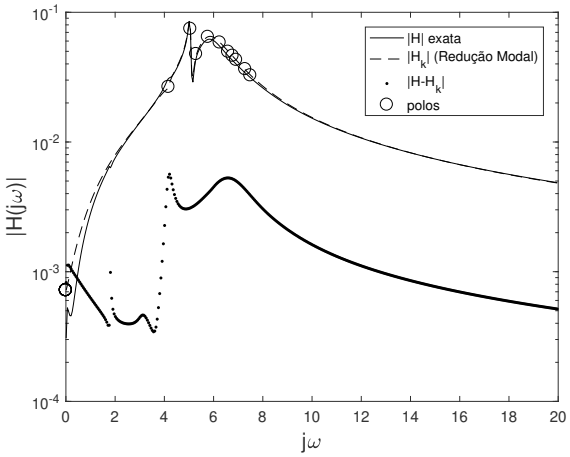
(a) Com 2 polos.



(b) Com 5 polos.



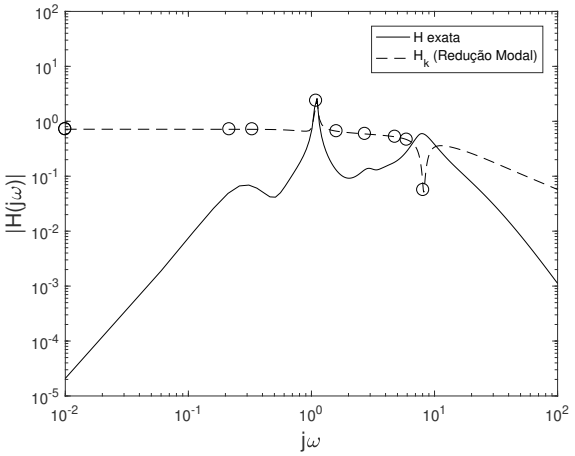
(c) Com 10 polos.



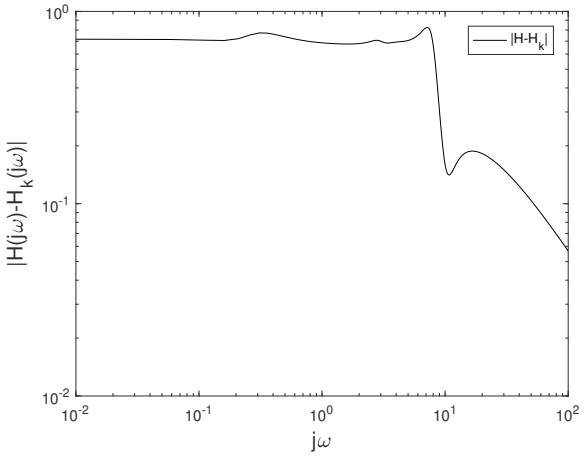
(d) Com 20 polos.

De maneira similar ao que foi feito na figura anterior, a figura 5.14 apresenta reduções modais feitas sobre o modelo `xingo_afonso_itaipu`.

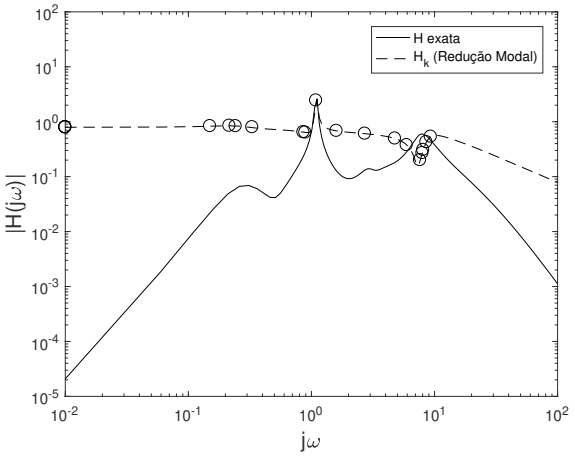
Figura 5.14: Comparativo entre a função H original do sistema `xingo_afonso_itaipu` com a versão reduzida H_k (Redução Modal) e erro absoluto para (polos dominantes calculados com DPSE e ordenados por PQ-dominância.



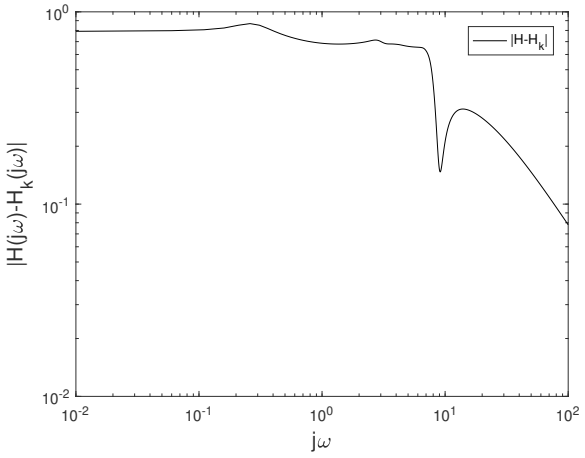
(a) $H(j\omega)$ e $H_k(j\omega)$ com 10 polos.



(b) Erro $H(j\omega) - H_k(j\omega)$ com 10 polos.



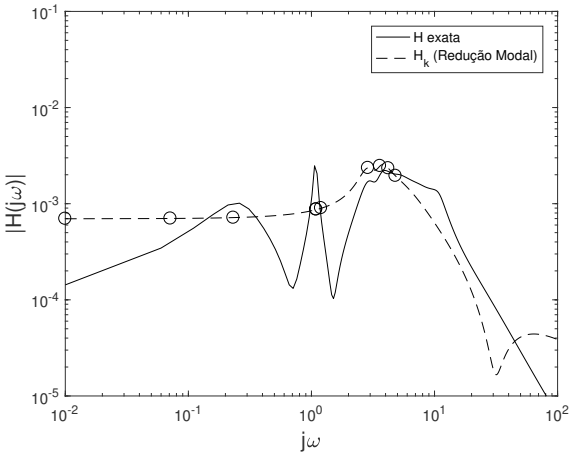
(c) $H(j\omega)$ e $H_k(j\omega)$ com 20 polos.



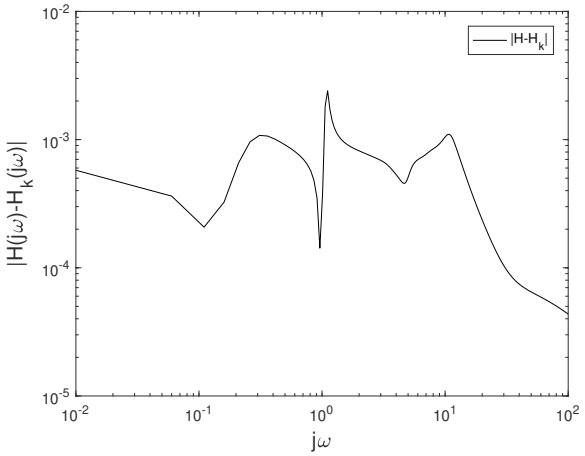
(d) Erro $H(j\omega) - H_k(j\omega)$ com 20 polos.

Por último, exibimos resultados obtidos ao aplicar a redução modal com os polos dominantes calculados pelo método DPSE no sistema de teste `ww_vref_6405`.

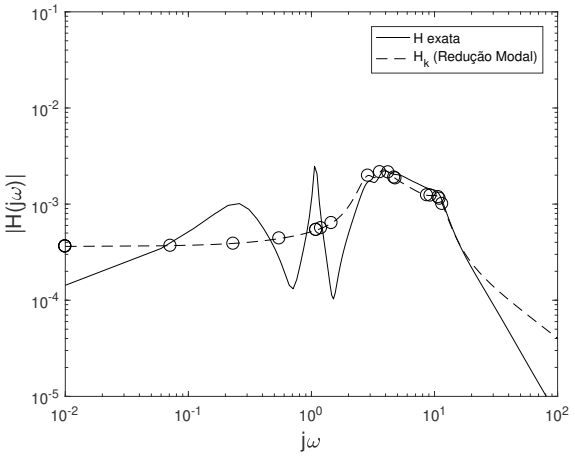
Figura 5.15: Comparativo entre a função H original do sistema `ww_vref_6405` com a versão reduzida H_k (Redução Modal) e erro absoluto para (polos dominantes calculados com DPSE e ordenados por PQ-dominância.



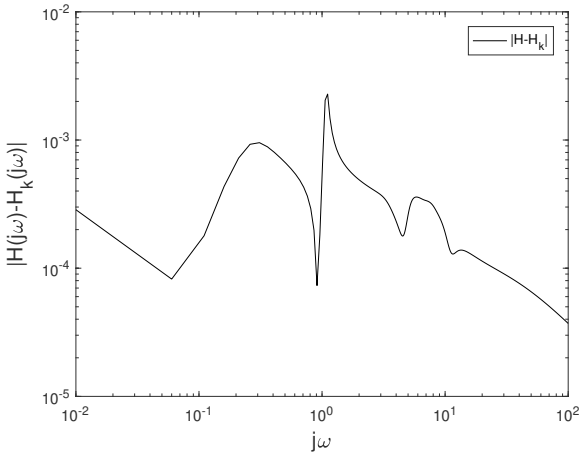
(a) $H(j\omega)$ e $H_k(j\omega)$ com 10 polos.



(b) Erro $H(j\omega) - H_k(j\omega)$ com 10 polos.



(c) $H(j\omega)$ e $H_k(j\omega)$ com 20 polos.



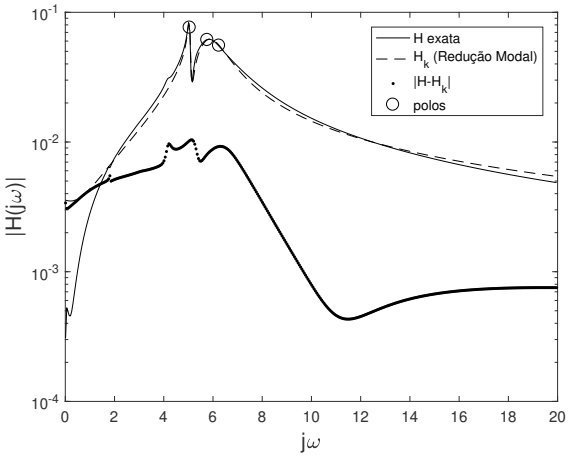
(d) Erro $H(j\omega) - H_k(j\omega)$ 20 polos.

Observando a figura 5.13 percebemos que, com poucos polos dominantes, é possível obter um modelo reduzido relativamente próximo do original. O mesmo não ocorre com os sistemas `xingo_afonso_itaipu` e `ww_vref_6405`. Note que, nesses dois últimos sistemas dinâmico, a magnitude de cada uma das funções de transferência apresenta vários picos, o que indica que existe uma quantidade maior de polos dominantes em comparação com o sistema sistema `noops_11k`. Além disso, há o fato de que esses polos dominantes estão aparentemente mais afastados entre si. Isso provavelmente interfere negativamente na eficiência do método DPSE, fazendo com que ele convirja para polos que não possuem dominância efetiva no modelo mas que estão, de alguma maneira, mais próximos aos chutes iniciais dados ao algoritmo.

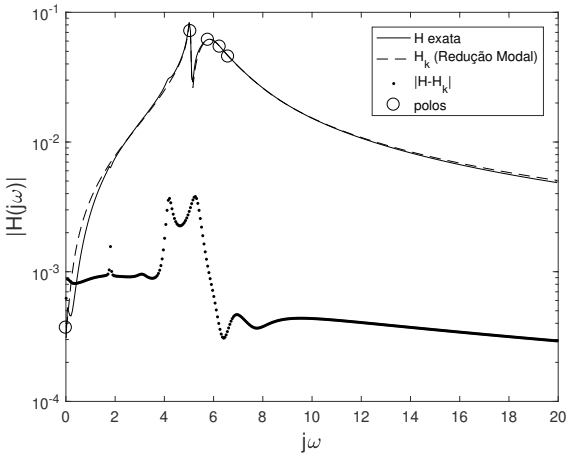
Visando proporcionar uma melhor compreensão sobre o comportamento de cada um desses modelos, bem como a verificação da relevância da definição de PQ-dominância, optamos por calcular todos os autovalores e todos os autovetores da matriz A de cada um desses sistemas. Em

seguida, ordenamos os polos dominantes de acordo com a definição de PQ-dominância para, a partir disso, realizar reduções modais com alguns desses polos dominantes, selecionados de acordo com essa mesma ordem. Os resultados obtidos por essa construção são exibidos nas figuras 5.16, 5.17 e 5.18, a seguir.

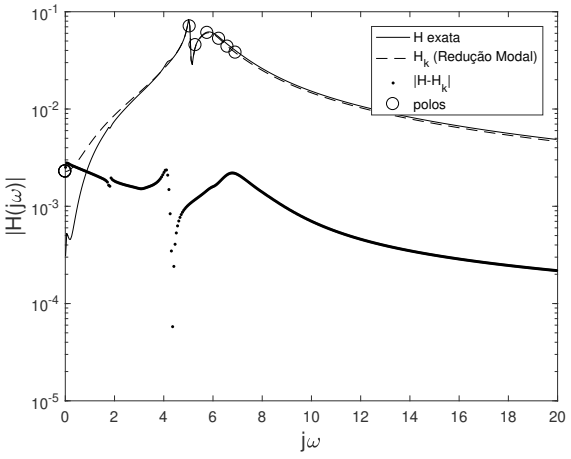
Figura 5.16: Função original versus modelo reduzido (Redução Modal) e erro absoluto para o sistema `noops_11k` via PQ-dominância a partir da decomposição espectral completa de A.



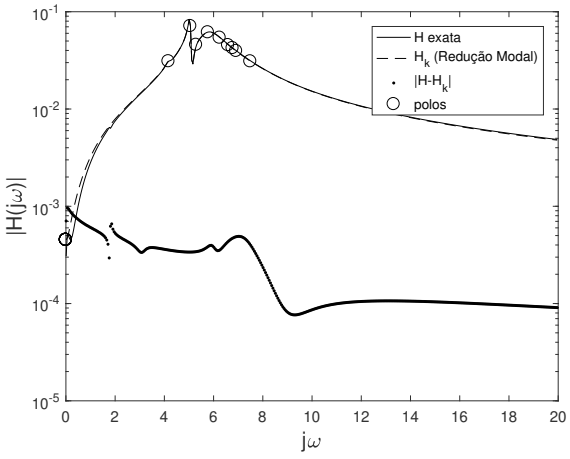
(a) Com 3 polos.



(b) Com 5 polos.

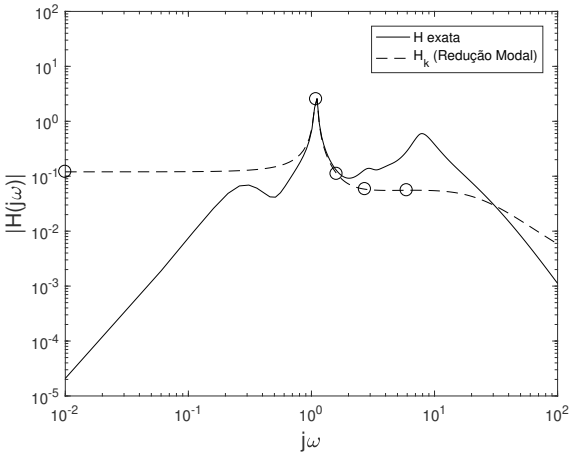


(c) Com 10 polos.

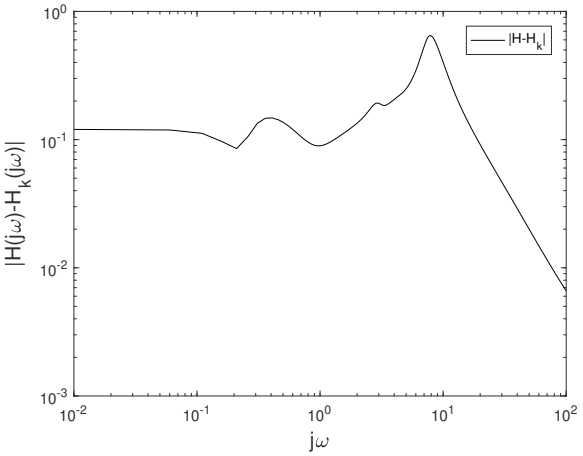


(d) Com 20 polos.

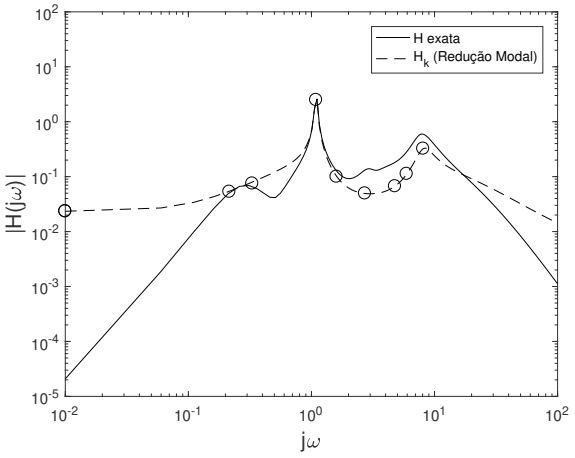
Figura 5.17: Função original versus modelo reduzido (Redução Modal) e erro absoluto para o sistema `xingo_afonso_itaipu` via PQ-dominância a partir da decomposição espectral completa de A.



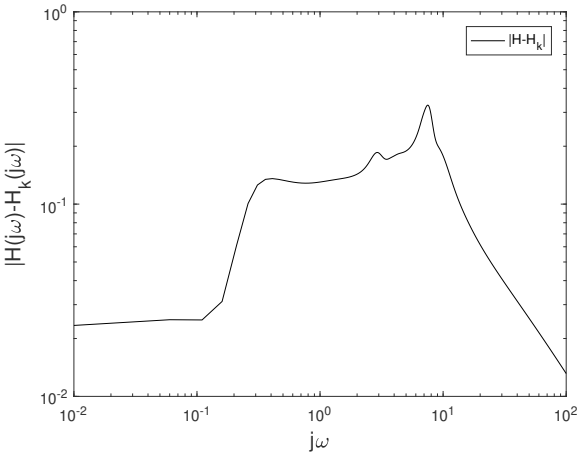
(a) $H(j\omega)$ e $H_k(j\omega)$ com 5 polos.



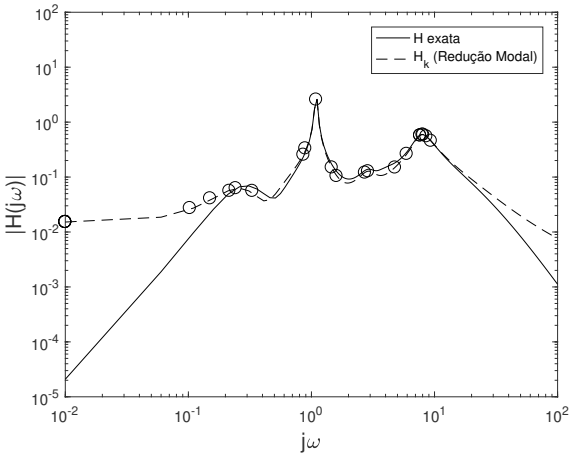
(b) Erro $H(j\omega) - H_k(j\omega)$ com 10 polos.



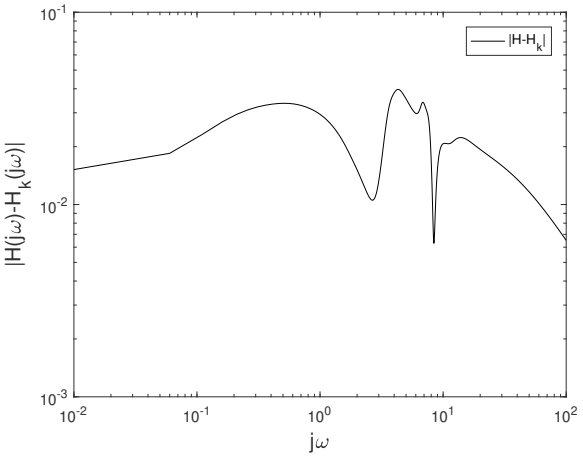
(c) $H(j\omega)$ e $H_k(j\omega)$ com 10 polos.



(d) Erro $H(j\omega) - H_k(j\omega)$ com 10 polos.

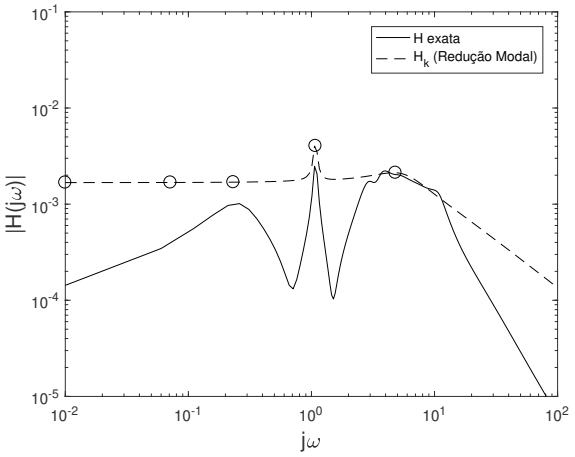


(e) $H(j\omega)$ e $H_k(j\omega)$ com 25 polos.

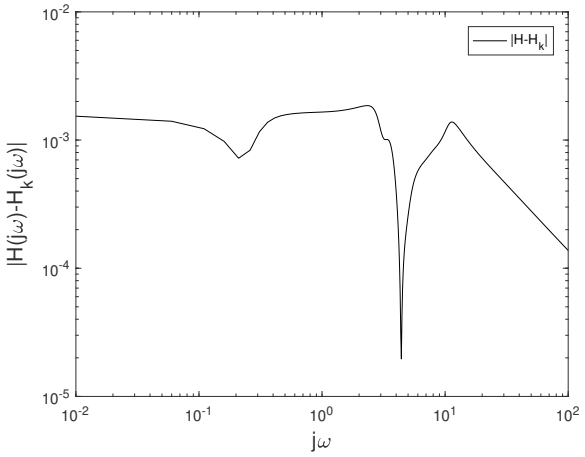


(f) Erro $H(j\omega) - H_k(j\omega)$ 25 polos.

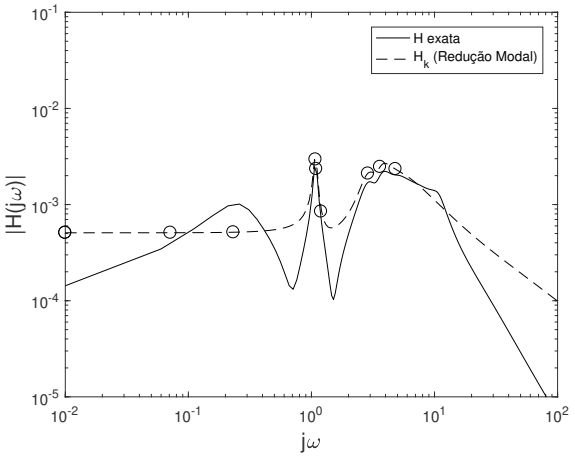
Figura 5.18: Função original versus modelo reduzido (Redução Modal) e erro absoluto para o sistema `ww_vref_6405` via PQ-dominância a partir da decomposição espectral completa de A.



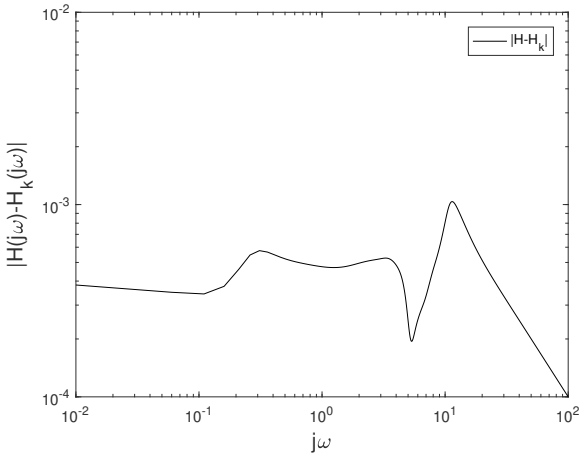
(a) $H(j\omega)$ e $H_k(j\omega)$ com 5 polos.



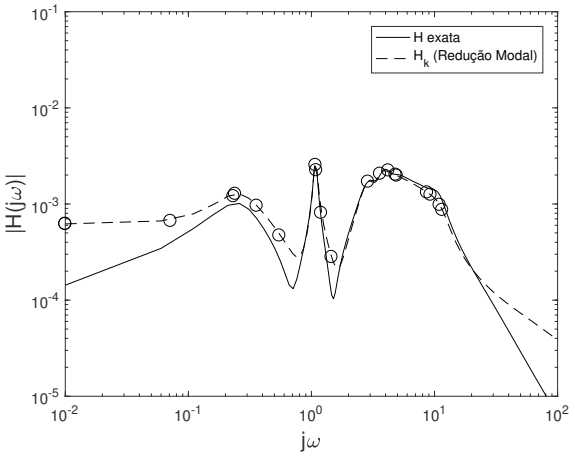
(b) Erro $H(j\omega) - H_k(j\omega)$ com 10 polos.



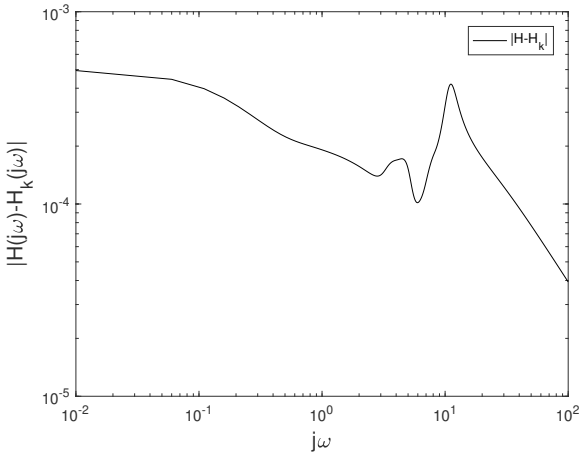
(c) $H(j\omega)$ e $H_k(j\omega)$ com 10 polos.



(d) Erro $H(j\omega) - H_k(j\omega)$ com 10 polos.



(e) $H(j\omega)$ e $H_k(j\omega)$ com 20 polos.

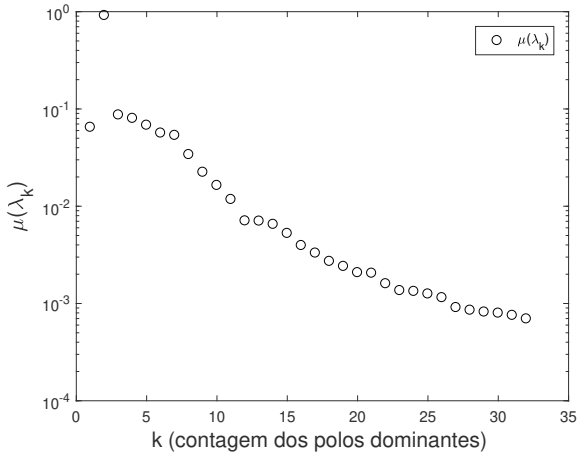


(f) Erro $H(j\omega) - H_k(j\omega)$ 20 polos.

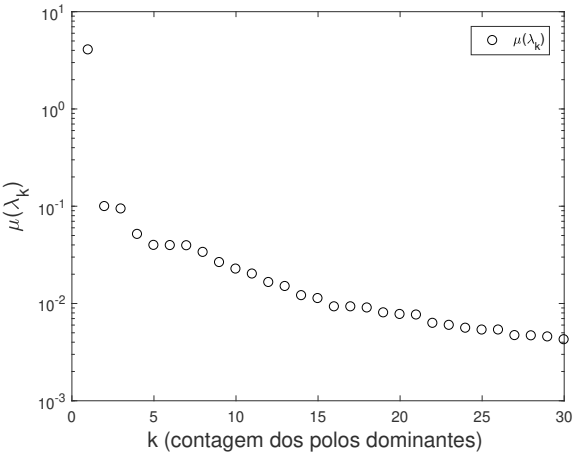
Como era de se esperar, nos modelos `xingo_afonso_itaipu` e `ww_vref_6405`, é preciso uma

quantidade maior de polos dominantes para se ter uma boma aproximação via redução modal. Essa característica também é perceptível nos gráficos da figura 5.19, que ilustram os valores $\mu(\lambda_k)$ (definidos na expressão (3.52)) que mensura a dominância relativa de cada polo.

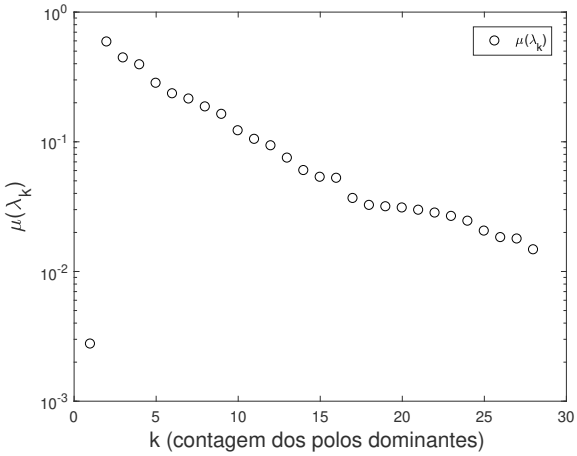
Figura 5.19: Decaimento dos valores $\mu(\lambda_k)$ da definição de polos PQ-dominantes.



(a) `nopss_11k`.



(b) `xingo_afonso_itaipu`.



(c) `ww_vref_6405`.

Perceba que o decaimento de $\mu(\lambda_k)$ é mais significativo no sistema `nopss_11k` do que nos demais. Isso condiz com as percepções que tivemos ao plotar o gráfico da magnitude da função de transferência H desse sistema. No caso do sistema `xingo_afonso_itaipu`, bem como o sistema `ww_vref_6405`, a discrepância entre os valores $\mu(\lambda_k)$ é menor, isso sugere a necessidade de um número maior de polos dominantes para se obter uma boa aproximação por redução modal, algo que é coerente com os gráficos exibidos anteriormente também.

Nesse momento, diante dos testes realizados, não podemos deixar de destacar a coerência entre o critério de PQ-dominância, definido por nós, e a dominância efetiva dos polos na função de transferência do sistema dinâmico. É perceptível que os polos escolhidos por esse critério tendem a estar localizados em regiões mais “acidentadas"do gráfico da magnitude de $H(j\omega)$. Além

disso, o decaimento dos valores $\mu(\lambda_k)$ condiz com o efeito que esses polos causam nos modelos reduzidos.

5.1.5 Algumas Considerações Sobre os Testes Realizados com sistemas SISO

A fim de comparar os métodos aplicados nas subseções anteriores, decidimos reunir aqui algumas informações pertinentes a todas essas estratégias. No que se refere à redução por balanceamento, as tabelas a seguir apresentam o número de decomposições LU, da matriz A, utilizadas por cada um dos métodos de resolução da equação de Lyapunov. Essa contagem já leva em consideração a resolução das duas equações de Lyapunov em cada problema, por exemplo, o método do tipo splitting, por utilizar um único parâmetro σ para a resolução de cada equação de Lyapunov, necessita de apenas uma decomposição Lu de A a cada implementação, contabilizando duas decomposições a cada aplicação do método de redução de modelo. Além disso, é exibido o tamanho (ou dimensão) do modelo reduzido obtido em cada situação. Esse tamanho está representado pela sigla MOR (*Modelo de Ordem Reduzida*).

Tabela 5.2: Número de dec. LU e ordem do modelo reduzido (sist. `noops_11k`)

	SLRCF-ADI (12 it.)	SLRCF-ADI (80 it.)	RKSM	SEL
# LU	8	8	51	2
MOR	10	22	7	9

Tabela 5.3: Número de dec. LU e ordem do modelo reduzido (sist. `xingo_afonso_itaipu`)

	SLRCF-ADI (12 it.)	SLRCF-ADI (80 it.)	RKSM	SEL
# LU	8	8	60	2
MOR	9	26	24	13

Tabela 5.4: Número de dec. LU e ordem do modelo reduzido (sist. `ww_vref_6405`)

	SLRCF-ADI (28 it.)	SLRCF-ADI (80 it.)	RKSM	SEL
# LU	8	8	77	2
MOR	20	35	13	11

Levando em consideração que os modelos originais testados são da ordem de milhar, pode-se dizer que os três métodos descritos nas tabelas 5.2, 5.3 e 5.4 foram eficientes no cálculo dos modelos reduzidos.

O método RKSM se mostra significativamente mais caro do que os demais, justamente por ser um método adaptativo, que calcula e utiliza um novo parâmetro μ a cada nova inversão da

matriz $A - \mu I$. Outra desvantagem do método baseado em subespaços de Krylov racionais é que, além das inversões das translações da matriz A, é preciso calcular a projeção $V_k^T A V_k$ a cada nova iteração (V_k é a base do subespaço de Krylov gerado a cada iteração k). Isso encarece ainda mais o processo. A vantagem desse método é que ele depende de pouca informação *a priori* do problema.

O método SLRCF-ADI possui um custo computacional significativamente baixo quando o número de parâmetros cíclicos utilizados é pequeno. No entanto, a análise de complexidade do método deve levar em conta a necessidade de se calcular esse conjunto de parâmetros previamente. Nesse sentido, a constatação da eficiência plena do método ainda depende da criação de um algoritmo específico para o cálculo desses parâmetros e que ainda não foi desenvolvido até o momento da confecção desse trabalho.

Por fim, o método SEL é o mais surpreendente em termos de complexidade numérica, pois exige apenas duas decomposições LU (uma para cada equação de Lyapunov) a cada iteração e possui a implementação mais simples de todos os três métodos testados.

Quanto aos métodos de redução modal, a ordem do modelo reduzido é sempre igual ao número de polos dominantes mais os número de possíveis pares complexos conjugados desses polos. O método DPSE tem traços que assemelham-se ao método da potência inversa para autovalores e, para calcular 10 polos dominantes de um sistema como os que testamos aqui, utiliza, em média, 60 decomposições LU.

O método de redução modal é de uma natureza completamente diferente dos métodos de redução por balanceamento. É injusto comparar a eficiência dessas duas estratégias com base apenas na diferença entre as funções de transferência e o considerando apenas o custo computacional das decomposições LU. Enquanto o esforço computacional empregado no método de redução modal está concentrado quase que só no cálculo dos polos dominantes, a redução por balanceamento é composta por um processo que, além de exigir as soluções P e Q das duas equações de Lyapunov associadas ao sistema dinâmico (é nessa parte que contabilizamos as decomposições LU), necessita de uma decomposição SVD do produto $U^T L$, em que U é o fator de Cholesky de Q e L é o fator de Cholesky de Q. Enquanto uma decomposição LU necessita de uma quantidade de operações de pontos flutuantes na ordem de $\frac{2n^3}{3}$, enquanto uma decomposição SVD pode exigir bem mais do que isso. Entretanto, os aprimoramentos que propomos para o método SLRCF-ADI fazem com que esse método calcule a solução da equação de Lyapunov de maneira rápida e eficiente. Além disso, esse método faz bom uso da esparsidade dos sistemas descritores e calcula diretamente o fator de Cholesky da solução desejada. Sendo assim, uma vez calculados os fatores L e U por meio do algoritmo SLRCF-ADI, a parte mais intensiva do trabalho restante está concentrada apenas na decomposição em valores singulares do produto $U^T L$. Essa última parte, por sua vez, é compensada pela fidelidade dos modelos reduzidos calculados nesse processo. Seguindo nessa linha de raciocínio, o método SEL se mostra bastante promissor, principalmente por exigir um custo computacional baixíssimo para calcular as soluções P e Q.

5.2 Testes em um sistema MIMO

Nesta seção abordamos testes feitos com o sistema `bips97_1676` que possui múltiplas variáveis de entrada e de saída. Nessa situação, a função de transferência $H(j\omega)$ é uma matriz de tamanho 8×8. Por isso, a representação gráfica do valor absoluto dessa função não faz sentido. Ao invés disso, plotamos os gráficos do maior e do menor valor singular dessa função (como é feito também em [11])

A fórmula definida pela expressão (3.52), definida para avaliar a função $\mu(\lambda_k)$ utilizada para selecionar polos dominantes perante a definição de PQ-dominância, não pode ser empregada em sistemas do tipo MIMO. Por isso, para que possamos, de alguma forma, comparar a redução modal com a redução por balanceamento, optamos por calcular a redução modal diretamente a partir de um conjunto de 20 polos dominantes calculados pelo algoritmo SAMDP (*Subespace Accelerated MIMO Dominant Pole Algorithm*), apresentado em [32] e disponível em *https://sites.google.com/site/rommes/software*. Os resultados obtidos nesse teste são exibidos na figura a seguir:

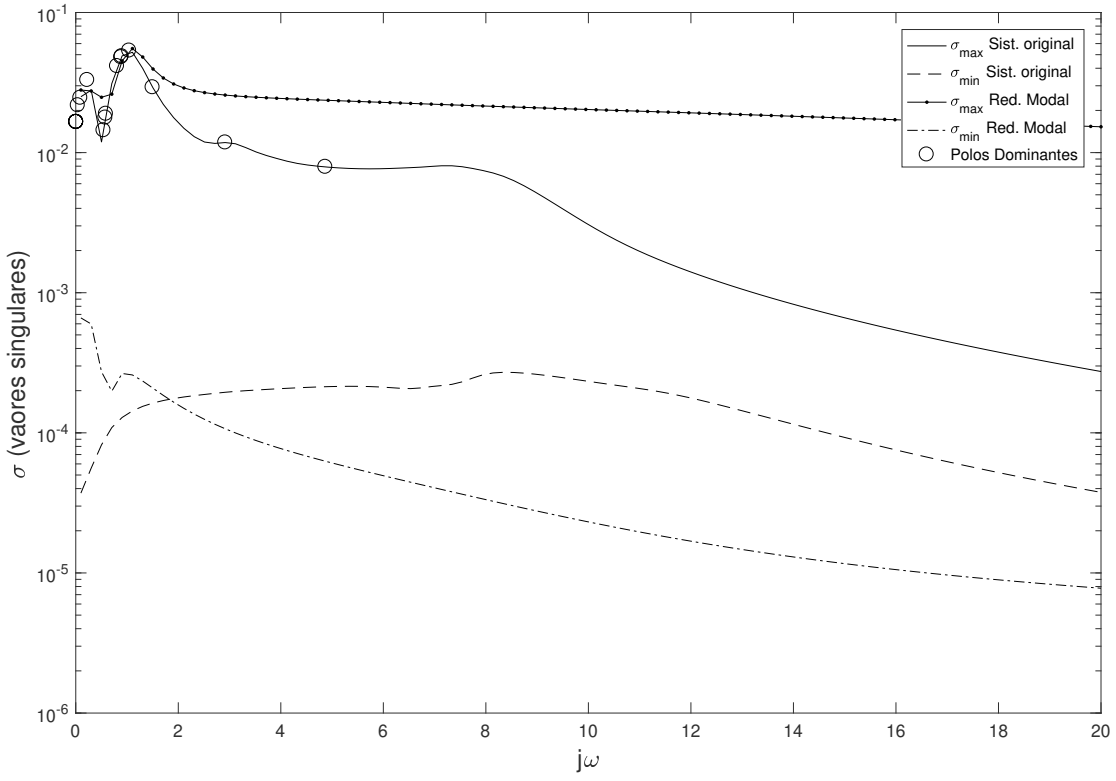


Figura 5.20: Redução modal do sistema `bips97_1676` via SAMDP.

A figura 5.21 exibe os gráficos obtidos a partir da redução por balanceamento realizada por meio do método SLRCF-ADI cujos parâmetros utilizados são os 4 autovalores de A dominantes na solução da equação de Lyapunov. Em cada equação de Lyapunov, utilizamos 20 iterações do

algoritmo SLRCF-ADI.

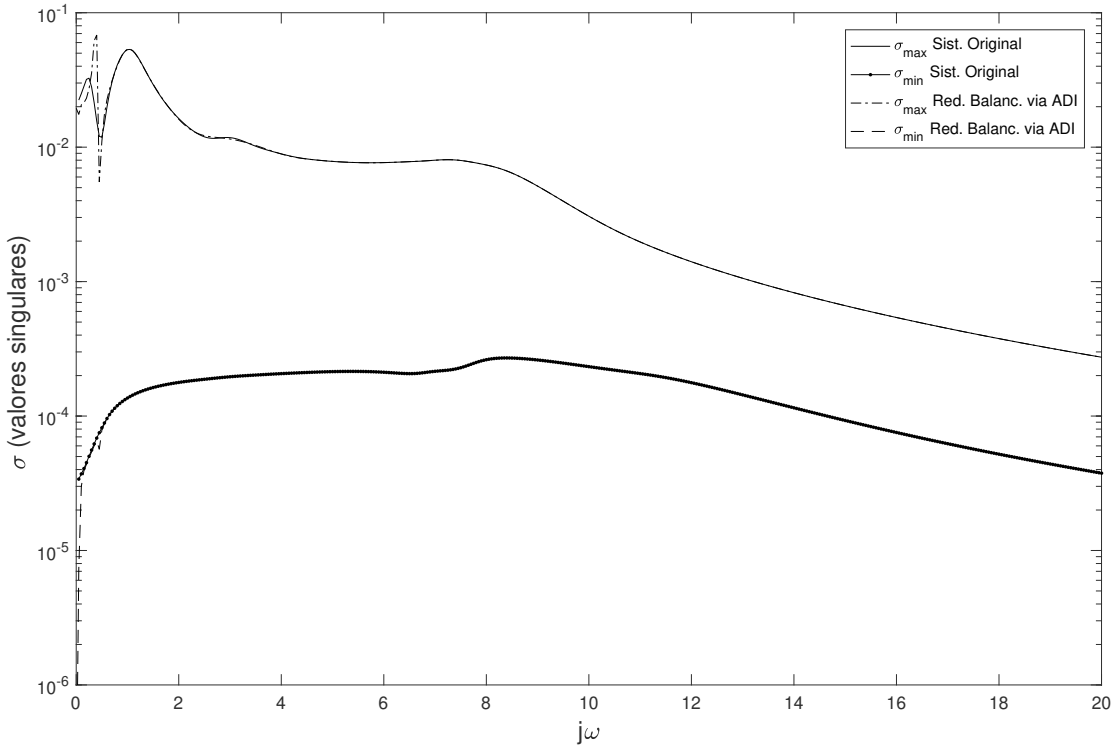


Figura 5.21: Redução por balanceamento do sistema `bips97_1676`, via SLRCF-ADI.

Na figura 5.21, destaca-se a precisão e a fidelidade do modelo reduzido com relação à função de transferência do sistema original. Mais uma vez, essa precisão vem acompanhada de um custo da execução do método SLRCF-ADI significativamente baixo. O mesmo não ocorre com o método RKSM, exibido na figura 5.22 que, além de exigir um esforço computacional maior, não retorna um modelo reduzido tão preciso. Acreditamos que essa discrepância possa ser minimizada se a estrutura adaptativa do método que calcula os parâmetros do RKSM for modificada a fim de contemplar melhor os autovalores dominantes na solução de Lyapunov.

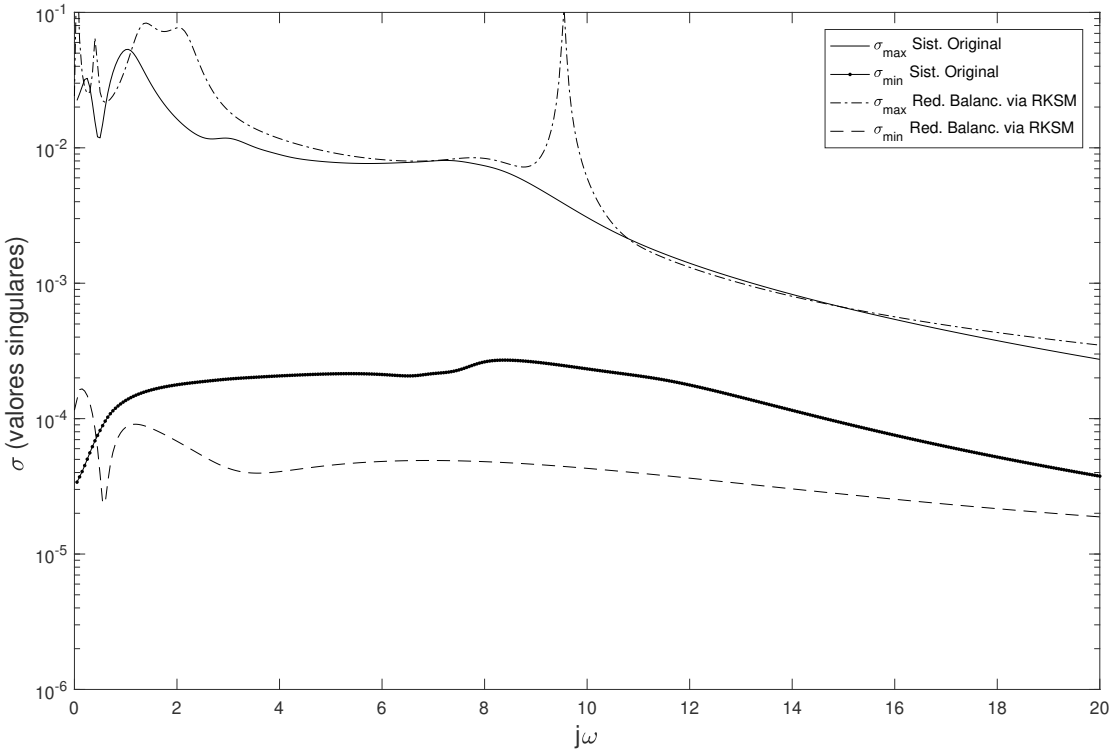


Figura 5.22: Redução por balanceamento do sistema `bips97_1676`, via RKSM com $s_1 = 200$.

Aproveitamos esse exemplo para atestar a influência provocada pela sutil mudança de escolha dos parâmetros iniciais t_0 e t_1, requeridos pelo método RKSM, que propomos nesse trabalho. Para o teste apresentado na figura 5.23, utilizamos $t_1 = 200$. Já no teste exibido na figura 5.23, foi fornecido o parâmetro inicial $t_1 = 10^4$ que está próximo ao raio espectral de A, conforme sugerido em [8], na versão original do método. A título de curiosidade, o espectro da matriz A do sistema dinâmico `bips97_1676` é exibido no gráfico da figura 5.24.

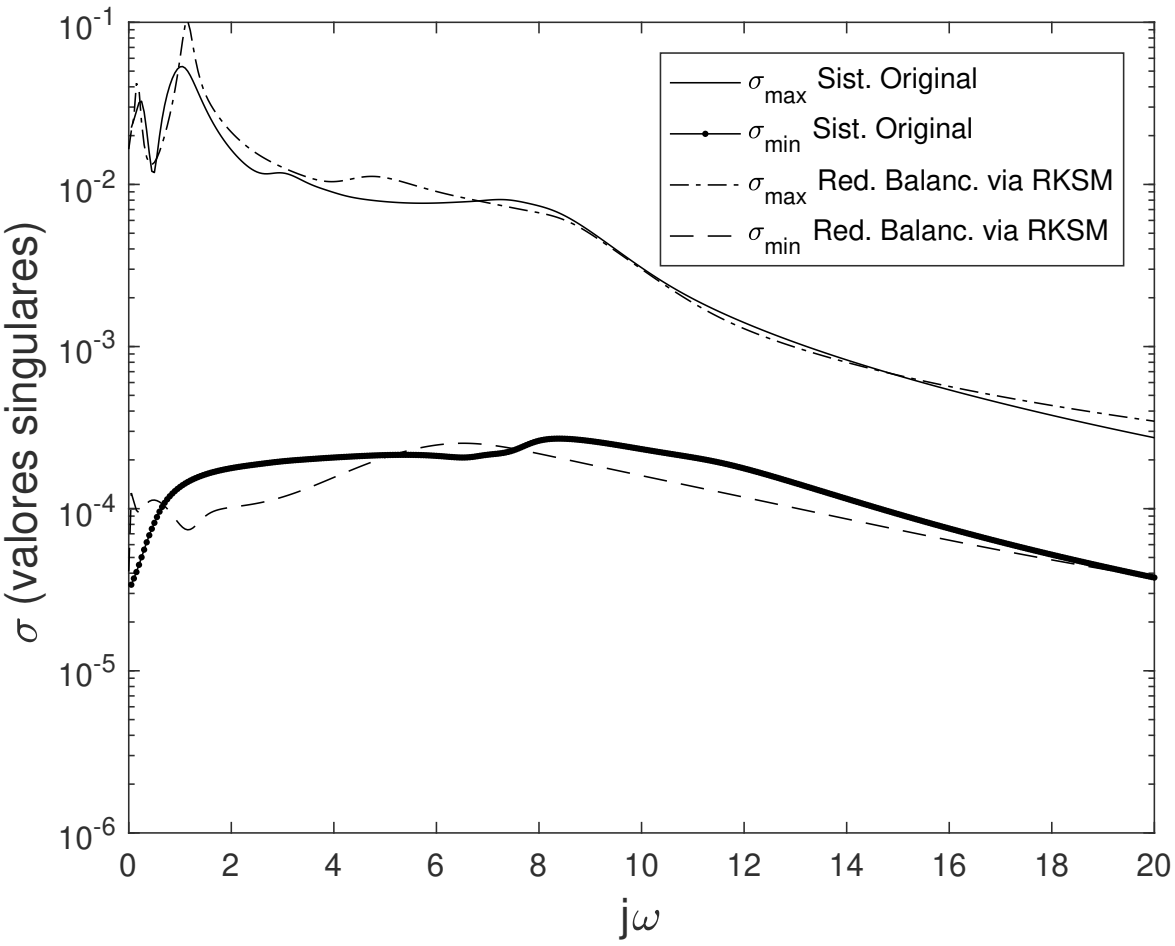
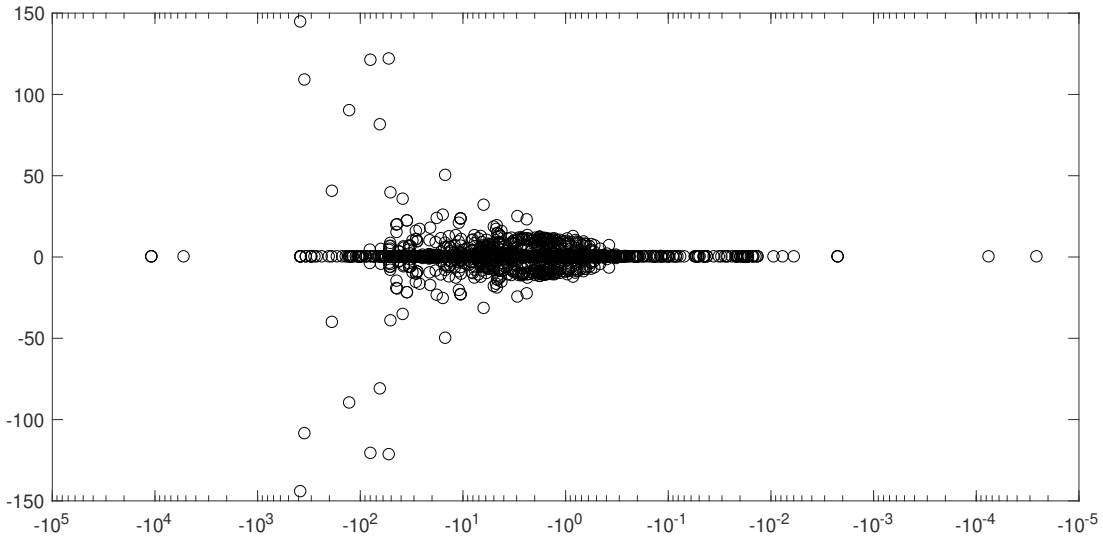


Figura 5.23: Redução por balanceamento do sistema `bips97_1676`, via RKSM com $s_1 = 10^4$.

Figura 5.24: Esectro da matriz A do sistema `bips97_1676`.



Por fim, utilizamos o método SEL com 20 iterações (com a etapa adicional de normalização das matrizes Y_k e Z_k) para cada equação de Lyapunov para gerar o modelo reduzido apresentado na figura 5.25. Novamente fomos surpreendidos com a eficácia do método que, mesmo estando numa versão ainda precária, gerou um modelo mais preciso até do que o construído a partir do algoritmo RKSM.

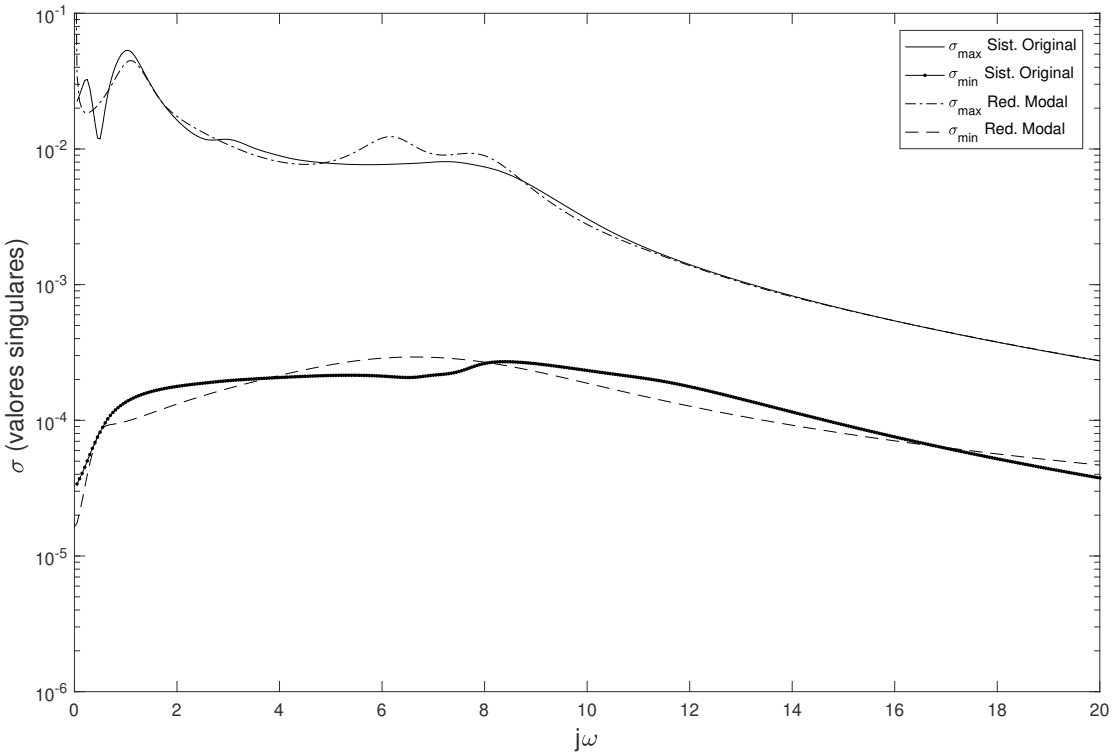


Figura 5.25: Redução por balanceamento do sistema `bips97_1676`, via SEL.

Considerações Finais

Analisando as construções feitas nesse trabalho, percebemos uma espécie de "via de mão dupla"entre as técnicas de redução de modelo e os métodos de resolução da equação de Lyapunov. Por um lado, a definição de polos dominantes em sistemas dinâmicos motivou-nos a buscar uma definição similar no contexto da equação de Lyapunov vista de maneira isolada. Por outro lado, a compreensão que tivemos sobre a equação de Lyapunov, permitiu-nos contribuir de maneira significativa com conceitos relacionados à dominância de polos em sistemas dinâmicos. Esse vínculo tão harmonioso entre as teorias e suas aplicações nos parece muito interessante.

Falando mais especificamente sobre cada método, podemos dizer a robustez do método ADI, bem como o método baseado em subespaços de Krylov racionais, são fatos inegáveis. Mesmo assim, a teoria e os testes numéricos apresentados nesse trabalho mostram que esses métodos podem ser significativamente melhorados se seus parâmetros forem escolhidos a partir de um conjunto de autovalores dominantes na solução P da equação de Lyapunov. A principal dificuldade que temos em aplicar tal estratégia em situações práticas da engenharia é que ainda não existe um método que calcule, de maneira eficiente, esse conjunto de autovalores dominantes. Isso é, sem dúvida, algo que pode ser explorado em projetos de pesquisa futuros. Em suma, dada a equação de Lyapunov $AP + PA = -BB^T$, com A diagonalizável, é preciso desenvolver um método que calcule um autovalor λ de A, com autovetores associados w^H e v, cuja magnitude do número $\frac{\|w_i^H B\|_2^2}{|2Re(\lambda)|}$ seja maior do que a dos demais autovalores. Acreditamos que a possível construção de um método como esse quebraria as barreiras impostas pelos altos custos computacionais impostos atualmente pelo método de redução por balanceamento.

Neste trabalho pudemos nos aprofundar no estudo da relação entre os polos dominantes e as matrizes de Gram P e Q de um sistema dinâmico. Esse estudo originou a definição de PQ-dominância. Em testes numéricos verificamos que esse novo critério tem potencial para contribuir com o aprimoramento de métodos de redução modal.

Por fim, o método SEL que introduzimos nesse trabalho, além de impressionar pela sua simplicidade em termos de implementação e pelo baixo custo computacional, representa um passo importante no resgate do uso de métodos para a equação de Lyapunov baseados na representação por produto de Kronecker. Ultimamente, esses métodos não tem sido muito explorados. Isso ocorre, muito provavelmente, pelo receio de se ter que operar com matrizes da ordem de n^2. No entanto, acreditamos que muitas outras técnicas de resolução para sistemas lineares podem ser

adaptadas para a equação de Lyapunov sem haver a necessidade de iterar matrizes tão grandes. Uma prova disso é o método de GC proposto recentemente por Mikkelsen em [28].

Os testes numéricos realizados com o método SEL, para fins de redução de modelo via balanceamento, mostraram que o método é bastante promissor nesse tipo de aplicação. Foi possível gerar boas aproximações para a função de transferência em alguns exemplos testados com um custo computacional relativamente baixo. Isso mostra que os espaços de projeção gerados pelas matrizes P e Q, calculadas por esse método, possuem boas qualidades em termos de redução de modelo. No entanto, ainda é preciso compreender melhor os aspectos que interferem na convergência do método para poder torná-lo mais robusto e preciso.

Referências Bibliográficas

[1] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems (Advances in Design and Control)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[2] A. C. Antoulas, D. Sorensen, and Y. Zhou. On the decay rate of Hankel singular values and related issues. *Systems and Control Letters*, 46(5):323 – 342, 2002.

[3] R. W. Brockett. *Finite dimensional linear systems*. Series in decision and control. New York, Wiley, 1970.

[4] A. M. Cohen. *Numerical Methods for Laplace Transform Inversion*. Springer Publishing Company, Incorporated, 1st edition, 2007.

[5] D. de Figueiredo. *Equações diferenciais aplicadas*. IMPA, 1997.

[6] V. Druskin, L. Knizhnerman, and V. Simoncini. Analysis of the Rational Krylov Subspace and ADI Methods for Solving the Lyapunov Equation. *SIAM Journal on Numerical Analysis*, 49(5):1875–1898, 2011.

[7] V. Druskin, C. Lieberman, and M. Zaslavsky. On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. *SIAM Journal on Scientific Computing*, 32(5):2485–2496, 2010.

[8] V. Druskin and V. Simoncini. Adaptive rational krylov subspaces for large-scale dynamical systems. *Systems and Control Letters*, 60(8):546(15), 2011.

[9] D. F. Enns. Model reduction with balanced realizations: An error bound and a frequency weighted generalization. In *The 23rd IEEE Conference on Decision and Control*, pages 127–132, Dec 1984.

[10] M. Fiedler. Notes on Hilbert and Cauchy matrices. *Linear Algebra and its Applications*, 432(1):351 – 356, 2010.

[11] F. D. Freitas, J. Rommes, and N. Martins. Gramian-based reduction method applied to large sparse power system descriptor models. *IEEE Transactions on Power Systems*, 23(3):1258–1270, Aug 2008.

[12] K. Glover. All optimal hankel-norm approximations of linear multivariable systems and their $\mathcal{L}^\infty$-error bounds. *International Journal of Control*, 39(6):1115–1193, 1984.

[13] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[14] S. Gomes, S. L. Varricchio, N. Martins, and C. Portela. Results on modal analysis to speed-up electromagnetic transient simulations. In *IEEE Power Engineering Society General Meeting, 2005*, pages 1132–1139 Vol. 2, June 2005.

[15] U. Graf. *Applied Laplace Transforms and z-Transforms for Scientists and Engineers*. Birkhäuser Basel, 2004.

[16] A. Graham. *Kronecker products and matrix calculus: with applications*. Ellis Horwood series in mathematics and its applications. Horwood, 1981.

[17] R. A. Horn and C. R. Johnson, editors. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986.

[18] K. Jbilou, A. Messaoudi, and H. Sadok. Global FOM and GMRES algorithms for matrix equations. *Applied Numerical Mathematics*, 31(1):49 – 63, 1999.

[19] T. Kailath. *Linear Systems*. Information and System Sciences Series. Prentice-Hall, 1980.

[20] L. A. Knizhnerman and V. Simoncini. A new investigation of the extended Krylov subspace method for matrix function evaluations. *Numerical Linear Algebra with Applications.*, 17:615–638, 2009.

[21] D. A. Kolesnikov and I. V. Oseledets. From low-rank approximation to a rational krylov subspace method for the lyapunov equation. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1622–1637, 2015.

[22] J. R. Leigh. *Functional analysis and linear control theory.(Mathematics in science and engineering)*. Academic Press Inc,(London) LTD, New York, USA, 1980.

[23] A. Lu and E. Wachspress. Solution of Lyapunov equations by alternating direction implicit iteration. *Computers and Mathematics with Applications*, 21(9):43 – 58, 1991.

[24] N. Martins. Efficient eigenvalue and frequency response methods applied to power system small-signal stability studies. *IEEE Power Engineering Review*, PER-6(2):47–47, Feb 1986.

[25] N. Martins. The dominant pole spectrum eigensolver [for power system stability analysis]. *IEEE Transactions on Power Systems*, 12(1):245–254, Feb 1997.

[26] N. Martins, L. T. G. Lima, and H. J. C. P. Pinto. Computing dominant poles of power system transfer functions. *IEEE Transactions on Power Systems*, 11(1):162–170, Feb 1996.

[27] A. Michel, L. Hou, and D. Liu. *Stability of Dynamical Systems: Continuous, Discontinuous, and Discrete Systems*. Systems and Control: Foundations and Applications. Birkhäuser Boston, 2008.

[28] C. C. K. Mikkelsen. *Numerical methods for large Lyapunov equations*. PhD thesis, Purdue University, 2009.

[29] V. Peller. *Hankel Operators and Their Applications*. 3Island Press, 2003.

[30] T. Penzl. A cyclic low-rank smith method for large sparse lyapunov equations. *SIAM Journal on Scientific Computing*, 21(4):1401–1418, 1999.

[31] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Systems and Control Letters*, 40(2):139 – 144, 2000.

[32] J. Rommes. Methods for eigenvalue problems with applications in model order reduction. *Universiteit Utrecht, Nederland*, 01 2007.

[33] J. Rommes. Arnoldi and jacobi-davidson methods for generalized eigenvalue problems $ax = \lambda bx$ with singular b. *Math. Comp.*, 77(262):995 – 1015, 2008.

[34] J. Rommes and N. Martins. Efficient computation of multivariable transfer function dominant poles using subspace acceleration. *IEEE Transactions on Power Systems*, 21(4):1471–1483, Nov 2006.

[35] A. Ruhe. The rational krylov algorithm for nonsymmetric eigenvalue problems. iii: Complex shifts for real matrices. *BIT Numerical Mathematics*, 34(1):165–176, Mar 1994.

[36] A. Ruhe. Rational krylov algorithms for nonsymmetric eigenvalue problems. In G. Golub, M. Luskin, and A. Greenbaum, editors, *Recent Advances in Iterative Methods*, pages 149–164, New York, NY, 1994. Springer New York.

[37] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003.

[38] T. Shores. *Applied Linear Algebra and Matrix Analysis*. Undergraduate Texts in Mathematics. Springer New York, 2007.

[39] V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM Journal on Scientific Computing*, 29(3):1268–1288, 2007.

[40] V. Simoncini. Computational methods for linear matrix equations. *SIAM Review*, 58(3):377–441, 2016.

[41] J. R. Smith, F. Fatehi, C. S. Woods, J. F. Hauer, and D. J. Trudnowski. Transfer function identification in power system applications. *IEEE Transactions on Power Systems*, 8(3):1282–1290, Aug 1993.

[42] R. A. Smith. Matrix equation $xa + bx = c$. *SIAM Journal on Applied Mathematics*, 16(1):198–201, 1968.

[43] T. Stykel and V. Simoncini. Krylov subspace methods for projected Lyapunov equations. *Appl. Numer. Math.*, 62(1):35–50, Jan. 2012.

[44] M. S. Tombs and P. I. Truncated balanced realization of a stable non-minimal state-space system. *International Journal of Control*, 46(4):1319–1330, 1987.

[45] M. M. Uddin, M. S. Hossain, and M. F. Uddin. Rational Krylov subspace method (rksm) for solving the lyapunov equations of index-1 descriptor systems and application to balancing based model reduction. In *2016 9th International Conference on Electrical and Computer Engineering (ICECE)*, pages 451–454, Dec 2016.

[46] S. L. Varricchio. Hybrid modal-balanced truncation method based on power system transfer function energy concepts. *IET Generation, Transmission and Distribution*, 9:1186–1194(8), August 2015.

[47] S. L. Varricchio, F. D. Freitas, N. Martins, and F. C. Véliz. Computation of dominant poles and residue matrices for multivariable transfer functions of infinite power system models. *IEEE Transactions on Power Systems*, 30(3):1131–1142, May 2015.

[48] E. Wachspress. *The ADI model problem*. Springer, New York, NY, 2013.

[49] E. L. Wachspress. Extended application of alternating direction implicit iteration model problem theory. *Journal of the Society for Industrial and Applied Mathematics*, 11(4):994–1016, 1963.

[50] E. L. Wachspress. Iterative solution of the Lyapunov matrix equation. *Applied Mathematics Letters*, 1(1):87 – 90, 1988.

[51] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Feher/Prentice Hall Digital and. Prentice Hall, 1996.

Apêndice A

Demonstrações dos teoremas do Capítulo 1

Demonstração do Teorema 1.2:

Sabemos que $x(t) = e^{A(t-t_0)}x(t_0)$ para um instante inicial t_0. Consideremos a decomposição de Jordan de A:

$$A = SJS^{-1},$$

com S sendo uma matriz inversível e J sendo uma matriz de Jordan da forma:

$$J = \begin{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} & & & \\ & \begin{bmatrix} \lambda_{r+1} & 1 & \\ & \ddots & 1 \\ & & \lambda_{r+1} \end{bmatrix} & & \\ & & \ddots & \\ & & & \begin{bmatrix} \lambda_{r+p} & 1 & \\ & \ddots & 1 \\ & & \lambda_{r+p} \end{bmatrix} \end{bmatrix}, \tag{A.1}$$

em que os escalares $\lambda_1, \lambda_2, ..., \lambda_r, \lambda_{r+1}, ...\lambda_{r+p}$ são os autovalores de A. Então $x(t) = e^{SJtS^{-1}}x(0)$. Pela representação de $e^{SJtS^{-1}}$ por série de potências, pode-se perceber que

$$e^{SJtS^{-1}} = Se^{Jt}S^{-1}.$$

Definamos $J_0, J_1, ...J_p$ como sendo os blocos de Jordan da expressão (A.1). O bloco

$$J_0 = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}$$

contém os autovalores associados aos autoespaços de dimensão 1 e, para cada $k = 1, 2, ..., p$, a matriz J_k representa o bloco de Jordan, de ordem $m_k \geq 2$, associado ao autovalor λ_{r+k}. Vamos denotar $J_k = \lambda_k I_{m_k} + Z_{m_k}$, para $k = 1, 2, ..., p$, com m_k representando a ordem do bloco J_k e

$$Z_{m_k} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Uma vez que a estrutura de uma matriz de Jordan é preservada sobre a adição e potenciação, então

$$e^{Jt} = \begin{bmatrix} e^{J_0 t} & 0 & \cdots & 0 \\ 0 & e^{J_1 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{J_p t} \end{bmatrix},$$

ou seja, para cada $k = 1, 2, ...p$, $e^{J_k t} = e^{(\lambda_k I_{m_k} + Z_{m_k})t} = e^{\lambda_k I_{m_k} t} e^{Z_{m_k} t} = e^{\lambda_k t} e^{Z_{m_k} t}$, pois as matrizes I_{m_k} e Z_{m_k} comutam entre si. Note que $Z_{m_k}^M = 0$ para todo k natural, sempre que $M > m_k$. Neste caso, a série de potências que representa $e^{Z_{k_m} t}$ é finita. Desta forma,

$$e^{Z_{m_k} t} = \begin{bmatrix} 1 & t & \cdots & t^{m_k - 1}/(m_k - 1)! \\ 0 & 1 & \cdots & t^{m_k - 2}/(m_k - 2)! \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Dada uma matriz $F \in \mathbb{R}^{n \times n}$, considere

$$\|F\|_1 := \max_{j=1,...,n} \sum_{i=1}^{n} |f_{ij}|,$$

com f_{ij} representando a entrada ij da matriz F. Nessas condições, é evidente que

$$\|e^{At} x_0\|_1 \leq \|S\|_1 \|S^{-1}\|_1 \|x_0\|_1 \max_{k=0,1,...,p} \|e^{J_k t} x_0\|_1.$$

Para J_0, tem-se que $\|e^{J_0 t} x_0\|_1 \leq \beta e^{\alpha_0 t}$, com $\alpha_0 = \max_{i=1,...,r} Re(\lambda_i)$ e $\beta = \|S\|_1 \|S^{-1}\|_1 \|x_0\|_1$. Ademais,

temos

$$\|e^{J_k t}x_0\|_1 \leq \beta e^{\alpha_k t}\left(1 + t + ... + \frac{t^{m_k}}{m_k}\right),$$

com $\alpha_k = Re(\lambda_{r+k})$. Em geral, é válido que

$$\|e^{At}x_0\|_1 \leq \beta e^{\alpha t}\left(1 + t + ... + \frac{t^{M_k}}{M_k}\right),$$

com $\alpha = \max_{i=1,..,r+p} Re(\lambda_i)$ e M sendo algum número inteiro entre 0 e n que depende da dimensão do autoespaço generalizado associado ao autovalor com maior parte real.

Portanto, $\|e^{At}x_0\|_1 \longrightarrow 0$ quando $t \to \infty$ se, e somente se, $\alpha < 0$. Além disso, se $\alpha = 0$, existe uma constante $\delta \in \mathbb{R}$ tal que $\|e^{At}x_0\|_1 \leq \delta$ para todo t se, e somente se, $M = 0$, ou seja, quando o autovalor de maior parte real tiver multiplicidades algébrica e geométrica iguais. No restante dos casos, $\lim_{t\longrightarrow\infty} \|e^{At}x_0\|_1 \longrightarrow \infty$. □

Demonstração do Teorema 1.4:

Considere $\varepsilon > 0$ e

$$l = \min_{\|x\|=\varepsilon} \Theta(x). \tag{A.2}$$

Como Θ é estritamente positiva no conjunto compacto $\{x : \|x\| = \varepsilon\}$, então $l > 0$. Como Θ é contínua e $\Theta(0) = 0$, existe $\delta = \delta(\varepsilon) > 0$ tal que

$$\|x\| < \delta \Rightarrow \Theta(x) < \frac{l}{2}. \tag{A.3}$$

Sabendo que $\frac{d}{dt}\Theta(x(t)) \leq 0$, pode-se afirmar que a função $t \longrightarrow \Theta(x(t))$ é não crescente, ou seja, se escolhermos um instante $\bar{t}$ tal que $\|x(\bar{t})\| < \delta$, temos que

$$\Theta(x(t)) \leq \Theta(x(\bar{t})) \leq \frac{l}{2}, \quad \forall t \geq \bar{t}. \tag{A.4}$$

Isso significa que, se uma solução $x(t)$ começar dentro da bola de raio δ em $\bar{t}$, ela deve permanecer dentro da bola de raio ε, pois, caso contrário, se existisse $t^* \geq \bar{t}$ tal que $\|x(t^*)\| \geq \varepsilon$, então, por (A.2), teríamos que $\Theta(t^*) \geq l$, fato esse que contradiz (A.4), mostrando que o sistema é estável.

Vamos assumir agora que $\frac{d}{dt}\Theta(x(t)) < 0$, ou seja, que a função é de Lyapunov estrita. Provemos que esta solução tende a zero quando $t \longrightarrow \infty$. Dado $\varepsilon > 0$, vamos mostrar que existe $\tilde{t} = \tilde{t}(\varepsilon) > 0$ tal que

$$\|x(t)\| < \varepsilon, \quad \forall t \geq \tilde{t}.$$

Vimos anteriormente que, para um determinado valor $\bar{t}$, existe $\delta(\varepsilon)$ tal que

$$\|x(\bar{t})\| < \delta(\varepsilon) \Rightarrow \|x(t)\| < \varepsilon, \quad \forall t \geq \bar{t}. \tag{A.5}$$

Se existir $\tilde{t} = \tilde{t}(\varepsilon) > 0$ tal que $\|x(t)\| < \delta(\varepsilon)$, segue de (A.5) que $\|x(t)\| < \varepsilon, \quad \forall t \geq \tilde{t}$, e o Teorema estaria demonstrado. Suponhamos, por contradição, que não existe um tal $\tilde{t}$, isto é,

$$\|x(t)\| \geq \delta(\varepsilon), \quad \forall t \geq \bar{t}.$$

Tomando $R > \varepsilon$, sabe-se que $\|x(\bar{t})\| < \delta(R) \Rightarrow \|x(t)\| < R, \quad \forall t \geq \bar{t}$. Isso significa que a solução permanece por todo tempo na coroa circular

$$\delta(\varepsilon) \leq \|x(t)\| \leq R, \quad \forall t \geq \bar{t}.$$

Como $\frac{d}{dt}\Theta(x(t))$ é estritamente negativa, e a coroa é um conjunto compacto, segue que

$$\frac{d}{dt}\Theta(x(t)) \leq -c, \quad \text{com } c > 0, \quad \forall t > \overline{t}.$$

Então, integrando ambos os lados de $\bar{t}$ a t, obtemos

$$\Theta(x(t)) \leq \Theta(\bar{t}) - ct, \quad \forall t \geq 0,$$

que é uma contradição, pois Θ teria que assumir valores negativos. □

Demonstração do Lema 1.29:

Assumamos que existe uma solução $P > 0$ para (1.13). Sejam $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ os autovalores da matriz A com autovetores associados $\{v_1, v_2, ..., v_n\}$, respectivamente. Multiplicando (1.13) por v_i^H pela esquerda e por v_i pela direita, temos

$$(\overline{\lambda_i} + \lambda_i) v_i^H P v_i = v_i^H E v_i < 0, \quad \textit{para} \quad 1 \leq i \leq n.$$

Uma vez que $v_i^H P v_i > 0$, então $(\overline{\lambda_i} + \lambda_i) < 0$. Disto segue que $Re(\lambda_i) < 0$ para $1 \leq i \leq n$.

Vamos supor agora que A é estável. Definimos a matriz

$$P = -\int_0^\infty e^{A\tau} E e^{A^T \tau} d\tau,$$

que existe graças à estabilidade de A. Como $E < 0$, então $P > 0$. Além disso, P é solução da

equação (1.13), pois

$$AP + PA^T = -\int_0^\infty \left[Ae^{A\tau}Ee^{A^T\tau} + e^{A\tau}Ee^{A^T\tau}A^T\right] d\tau = -\int_0^\infty \frac{d}{d\tau}\left[e^{A\tau}Ee^{A^T\tau}\right] d\tau = E.$$

Por fim, provemos a unicidade da solução. Suponhamos que $\tilde{P}$ é uma outra solução de (1.13). Então,

$$A^T(P - \tilde{P}) + (P - \tilde{P})A = 0.$$

Se definirmos $M(t) = e^{A^T t}(P - \tilde{P})e^{At}$, temos que

$$\lim_{t \longrightarrow \infty} M(t) = 0,$$

em virtude da estabilidade de A. Por outro lado,

$$\dot{M}(t) = e^{A^T t}A^T(P - \tilde{P})e^{At} + e^{A^T t}(P - \tilde{P})Ae^{At} = e^{A^T t}\left[A^T(P - \tilde{P}) + (P - \tilde{P})A\right]e^{At} = 0.$$

Assim, podemos afirmar que $M(t) = e^{A^H t}(P - \tilde{P})e^{At} = 0$, para todo $t > 0$. Disto segue que $P = \tilde{P}$.

Demonstração do Teorema 1.30: Assumamos inicialmente que $\delta(A) \neq 0$; isso implica que existe $x^H \neq 0$ tal que $x^H A = i\lambda x^H$, para algum autovalor λ. Multiplicando ambos os lados de (1.13) por x^H e x, respectivamente, obtemos que $x^H Ex = 0$, que é uma contradição com a positividade de E. Disto segue que $\delta(A) = 0$, isto é, $in(A) = (k, 0, r)$ para $k > 0$ e $r + k = n$, com n sendo a ordem da matriz A. Assumamos, sem perda de generalidade, que A está particionada na forma

$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \quad com \quad in(A_{11}) = (0, 0, r) \quad e \quad in(A_{11}) = (k, 0, 0).$$

Se particionarmos P e E em blocos de mesma ordem que os blocos de A, verificamos que (1.13) implica em

$$A_{11}P_{11} + P_{11}A_{11}^H = E_{11} > 0 \quad e \quad A_{22}P_{22} + P_{22}A_{22}^H = E_{22} > 0.$$

O Lema 1.29 garante que $P_{11} > 0$ e $P_{22} < 0$. Agora, note que podemos escrever

$$P_{11} = V_{11}V_{11}^T \in \mathbb{R}^{r \times r} \quad e \quad P_{22} = -V_{22}V_{22}^T \in \mathbb{R}^{k \times k},$$

com $det(V_{11}) \neq 0$ e $det(V_{22}) \neq 0$. Assim,

$$P = diag(V_{11}, V_{22})Y diag(V_{11}^T, V_{22}^T), \quad com \quad Y = \begin{pmatrix} I_r & Y_{12} \\ Y_{12}^T & -I_k \end{pmatrix},$$

para alguma matriz $Y_{12} \in \mathbb{R}^{k\times r}$. *Finalmente, assumindo que* $r \leq k$, *seja* $Y_{12} = U\Sigma W^T$, *com* $UU^T = I_k$, $WW^T = I_r$ *e* $\Sigma = \left(\overline{\Sigma}\ 0_{kr}\right)$, *em que* $\overline{\Sigma} = diag\{\sigma_1, ..., \sigma_r\}$, $\sigma_i \geq \sigma_{i+1} \geq 0$. *Então*

$$Y = \begin{pmatrix} U & 0 \\ 0 & W \end{pmatrix} \Pi \begin{pmatrix} U^T & 0 \\ 0 & W^T \end{pmatrix}, \quad com \quad \Pi = \begin{pmatrix} I_r & \overline{\Sigma} & 0 \\ \overline{\Sigma}^T & -I_r & 0 \\ 0 & 0 & -I_{k-r} \end{pmatrix}.$$

Claramente a matriz Π *possui autovalores* $\pm\sqrt{1+\sigma_i^2}$, $i = 1, .., r$, *e* -1 *com multiplicidade* $k - r$, *isto é,* Π *possui* r *autovalores positivos e* k *autovalores negativos. Assim, pela lei de Silvester da inércia, tanto* Y *quanto* P *possuem a mesma inércia que* Π. *Isso completa a prova.* □

Demonstração do Teorema 1.31: Verifiquemos inicialmente que a controlabilidade implica em $\delta(A) = 0$ *e* $\delta(P) = 0$. *Com efeito, suponhamos que* $\delta(A) \neq 0$, *ou seja, existe um valor real* α *e um vetor* $v \neq 0$, *tais que* $A^T v = i\alpha v$. *Além disso,* $v^H(AP + PA^T)v = v^H Ev$. *O lado esquerdo da expressão é igual a* $(-i\alpha + i\alpha)v^H Pv = 0$. *Sendo assim,* $x^H E = 0$. *Isso contradiz o Lema (1.14). Supondo agora* $\delta(P) = 0$, *existe um vetor* $w \neq 0$ *e um valor* β *tais que* $Pw = i\beta w$. *Como o sistema é controlável, pelo Lema (1.13),* $P > 0$. *Portanto,* $\beta = 0$ *e* $Pw = 0$. *Temos também que* $w^H(AP + PA^T)w = w^H HQw$. *Como* $w^H(AP + PA^T)w = 0$, *então* $x^H Ew = 0$, *que implica em* $w^H E = 0$. *Sendo assim,* $w^H(AP + PA^T) = w^H AP = w^H E = 0$. *Além disso,*

$$A(AP + PA^T)A^T = AEA^T \Rightarrow w^H A(AP + PA^T)Aw = w^H AEAw = 0 \Rightarrow w^H AE = 0.$$

Repetindo esse processo por sucessivas vezes, concluímos que $w^H A^{k-1}E = 0$ *para qualquer* k *natural. Isso contradiz o fato de* (A, E) *ser controlável.*

Apêndice B

Demonstrações dos teoremas do Capítulo 2

Demonstração do Teorema 2.8:

Inicialmente, consideremos o caso especial em que $k = n$. As transformações (2.29) satisfazem

$$TT^{\dagger} = T^{\dagger}T = I_n$$

e as matrizes de controlabilidade e observabilidade $\widehat{P}$ e $\widehat{Q}$, respectivamente, de $(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$, são dadas por

$$\widehat{P} = TPT^T = \Sigma^{-1/2}V^TL^T(UU^T)LV\Sigma^{-1/2} = \Sigma$$
$$\widehat{Q} = (T^T)^{-1}QT^{-1} = \Sigma^{-1/2}W^TU^T(LL^T)UW\Sigma^{-1/2} = \Sigma.$$

Isso mostra que o sistema resultante está na forma balanceada.

Agora, vamos supor que $\Sigma_2 \geq 0$. Multiplicando a equação de Lyapunov (1.25) por W^TU^T pela esquerda e por UW pela direita e substituindo a decomposição de Cholesky $Q = LL^T$ obtem-se que

$$W^TU^TA^TLL^TUW + W^TU^TLL^TAUW = -W^TU^TC^TCUW.$$

Substituindo a decomposição SVD de U^TL,

$$\begin{aligned}\begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix} U^TAL \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} L^TAU \begin{bmatrix} W_1 & W_2 \end{bmatrix} \\ = -\begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix} U^TC^TCU \begin{bmatrix} W_1 & W_2 \end{bmatrix}.\end{aligned} \tag{B.1}$$

A equação (B.1) pode ser escrita como um conjunto de quatro equações de Lyapunov, sendo uma delas

$$W_1^TU^TALV_1\Sigma_1 + \Sigma_1V_1^TL^TAUW_1 = -W_1^TU^TC^TCUW_1.$$

Operando de maneira análoga com a equação (1.24), é possível obter a equação

$$V_1^T L^T A U W_1 \Sigma_1 + \Sigma_1 W_1^T U^T A^T L V_1 = -V_1^T L^T B B^T L V_1.$$

Como, por (2.32),

$$V_1^T L^T A U W_1 = \Sigma_1^{1/2} \widehat{A} \Sigma_1^{1/2}, \quad \Sigma_1^{1/2} \widehat{B} = V_1^T L^T B \quad e \quad \widehat{C} \Sigma_1^{1/2} = C U W_1,$$

então

$$\begin{aligned} \widehat{A}^T \Sigma_1 + \Sigma_1 \widehat{A} &= -\widehat{C}^T \widehat{C} \\ \widehat{A} \Sigma_1 + \Sigma_1 \widehat{A}^T &= -\widehat{B} \widehat{B}^T, \end{aligned}$$

Mostrando assim que o sistema transformado por T_L e T_R dadas em (2.31) está na forma balanceada. □

Demonstração do Teorema 2.9:

Consideremos

$$E(s) := H(s) - H_k(s).$$

Note que

$$\begin{aligned} E(s) &= C(sI - A)^{-1} B - D - \widehat{C}(sI_k - \widehat{A})^{-1} \widehat{B} - D \\ &= \begin{bmatrix} C & -\widehat{C} \end{bmatrix} \begin{bmatrix} (sI - A)^{-1} & 0 \\ 0 & (sI_k - \widehat{A})^{-1} \end{bmatrix} \begin{bmatrix} B \\ \widehat{B} \end{bmatrix} \\ &= \begin{bmatrix} C & -\widehat{C} \end{bmatrix} \left(sI_{n+k} - \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix} \right)^{-1} \begin{bmatrix} B \\ \widehat{B} \end{bmatrix}. \end{aligned}$$

Portanto, pela Definição de função de transferência (2.5), a função $E(s)$ representa o sistema (A_e, B_e, C_e, D_e) com

$$A_e = \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix}, \quad B_e = \begin{bmatrix} B \\ \widehat{B} \end{bmatrix}, \quad C_e = \begin{bmatrix} C & -\widehat{C} \end{bmatrix} \quad e \quad D_e = 0.$$

Seja

$$P_e = \begin{bmatrix} P_{11} & P_{12} \\ P_{21}^T & P_{22} \end{bmatrix}$$

uma matriz graminiana, particionada conforme os blocos de $E(s)$, que satisfaz

$$\begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{21}^T & P_{22} \end{bmatrix} + \begin{bmatrix} P_{11} & P_{12} \\ P_{21}^T & P_{22} \end{bmatrix} \begin{bmatrix} A^T & 0 \\ 0 & \widehat{A}^T \end{bmatrix} = - \begin{bmatrix} B \\ \widehat{B} \end{bmatrix} \begin{bmatrix} B^T & \widehat{B}^T \end{bmatrix}. \tag{B.2}$$

Claramente $P_{11} = P$, em que P é a matriz graminiana de (A, B, C, D), e $P_{22} = \Sigma_1$. Além disso, P_{12} é a solução da equação

$$AP_{12} + P_{12}\widehat{A}^T = -B\widehat{B}^T,$$

que, pelo Teorema 1.26, possui única solução desde que

$$\lambda_i(A) + \lambda_j(\widehat{A}^T) \neq 0$$

para todo $i, j = 1, 2, ..., n$. Portanto, para provar a unicidade basta verificar que

$$Re(\lambda_i(A)) < 0 \quad e \quad Re(\lambda_i(\widehat{A})) \leq 0.$$

Sabemos que $Re(\lambda_i(A)) < 0$ por hipótese. Além disso, como

$$\begin{aligned} \widehat{A}\Sigma_1 + \Sigma_1\widehat{A}^T &= -\widehat{B}\widehat{B}^T, \\ \widehat{A}^T\Sigma_1 + \Sigma_1\widehat{A} &= -\widehat{C}^T\widehat{C}, \end{aligned}$$

então $Re(\lambda_i(\widehat{A})) \leq 0$. Portanto, a equação (B.2) admite única solução. Substituindo (2.32) na equação (B.2) obtém-se

$$AP_{12} + P_{12}(T_L A T_R^{\dagger}) = -BB^T T_L.$$

Substituindo, agora, BB^T pela expressão (1.24), utilizando as definições de T_L e T_R e decomposições de Cholesky $P = UU^T$ e $Q = LL^T$, chegamos à expressão

$$A(P_{12} - UU^T R V_1 \Sigma_1^{-1/2}) + (P_{12}\Sigma^{-1/2}W_1^T - UA^T L V_1 \Sigma_1^{1/2} = 0. \tag{B.3}$$

Note que, pela ortogonalidade de V_1,

$$U^T L V_1 = W_1 \Sigma_1 V_1^T V_1 = W_1 \Sigma_1 \tag{B.4}$$

e, pela ortogonalidade de W_1,

$$U = U(2W_1 W_1^T),$$

Além disso,

$$UW_2 W_2^T U^T A^T L V_1 \Sigma_1^{-1/2} = 0.$$

A última expressão pode ser obtida substituindo $\Sigma_2 = 0$ na expressão (B.1) e verificando que

$W_2^T U^T A^T L V_1 = 0$.

Vamos agora rearranjar (B.3), obtendo

$$A(P_{12} - UW_1\Sigma_1^{1/2}) + (P_{12} - UW_1\Sigma^{1/2})\Sigma^{-1/2}U_1^T U^T A^T L V_1 \Sigma_1^{-1/2} = 0. \tag{B.5}$$

A equação (B.5) deriva da equação (B.2) que, por sua vez, admite única solução. Então podemos afirmar que $P_{12} = UW_1\Sigma_1^{1/2}$. *Disto segue que a matriz de controlabilidade de* $E(s)$ *é dada por*

$$P_e = \begin{bmatrix} UU^T & UW_1\Sigma_1^{1/2} \\ \Sigma_1^{1/2}W_1^T U^T & \Sigma \end{bmatrix}.$$

Similarmente é possível verificar que a matriz de observabilidade de $E(S)$ *é*

$$Q_e = \begin{bmatrix} LL^T & -LV_1\Sigma_1^{1/2} \\ -\Sigma_1^{1/2}V_1^T L^T & \Sigma \end{bmatrix}.$$

Finalmente, calculamos o produto P_eQ_e *a fim de examinar a norma Hankel do sistema construído a partir do erro* $E(s)$.

$$\begin{aligned} P_eQ_e &= \begin{bmatrix} UU^T LL^T - UW_1\Sigma_1 V_1^T L^T & -UU^T LV_1\Sigma_1^{1/2} + UW_1\Sigma_1^{3/2} \\ \Sigma_1^{1/2}W_1^T U^T LL^T - \Sigma_1^{3/2}V_1 L^T & -\Sigma_1^{1/2}W_1^T U^T LV_1\Sigma_1^{1/2} + \Sigma_1^2 \end{bmatrix} \\ &= \begin{bmatrix} U(U^T L - W_1\Sigma_1 V_1^T)L^T & U(-U^T LV_1 + W_1\Sigma_1^{1/2})\Sigma_1^{1/2} \\ \Sigma_1^{1/2}(W_1^T U^T L - \Sigma_1 V_1)L^T & \Sigma_1^{1/2}(-W_1^T(U^T LV_1 + \Sigma_1) + \Sigma_1^{1/2}) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Os blocos superiores, direito e esquerdo, além do bloco inferior direito de P_eQ_e, *são nulos em virtude da decomposição SVD de* $U^T L$. *O bloco inferior esquerdo se anula por causa de (B.4).*

Como P_e *e* Q_e *são as matrizes de controlabilidade e observabilidade, respectivamente, do sistema gerado pelo erro* $E(s)$, *segue da Definição da norma Hankel que*

$$\|H(s) - H_k(s)\|_H = \|E(s)\|_H = 0,$$

como queríamos provar. A controlabilidade e a observabilidade do sistema reduzido seguem por construção, levando em consideração o Corolário 1.13 e o Teorema 1.22. □

Demonstração do Teorema 2.10:

Seja $E_\infty := \|H(s) - H_k(s)\|_\infty$. *Definamos as matrizes*

$$\begin{aligned}
\varphi(s) &:= (sI_k - A_{11})^{-1} \\
\Delta(s) &:= sI_{n-k} - A_{22} - A_{21}\varphi(s)A_{12} \\
\tilde{B}(s) &:= A_{21}\varphi(s)B_1 + B_2 \\
\tilde{C}(s) &:= C_1\varphi(s)A_{12} + C_2,
\end{aligned}$$

então

$$\begin{aligned}
H(s) - H_k(s) &= C(sI_k - A)^{-1}B - C_1\varphi(s)B_1 \\
&= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} (sI_k - A_{11}) & -A_{12} \\ -A_{21} & (sI_{n-k} - A_{22}) \end{bmatrix}^{-1} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}.
\end{aligned}$$

Pela fórmula de inversão para matrizes em blocos [19], tem-se que

$$\begin{bmatrix} (sI_k - A_{11}) & -A_{12} \\ -A_{21} & (sI_{n-k} - A_{22}) \end{bmatrix}^{-1} = \begin{bmatrix} \varphi(s) + \varphi(s)A_{12}\Delta^{-1}(s)A_{21}\varphi(s) & \varphi(s)A_{12}\Delta^{-1}(s) \\ \Delta^{-1}(s)A_{21}\varphi(s) & \Delta^{-1}(s) \end{bmatrix}.$$

Então

$$\begin{aligned}
H(s) - H_k(s) &= C_1\left[\left(\varphi(s) + \varphi(s)A_{12}\Delta^{-1}(s)A_{21}\varphi(s)\right)B_1 + \left(\varphi(s)A_{12}\Delta^{-1}(s)\right)B_2\right] \\
&\quad + C_2\left[\Delta^{-1}(s)A_{21}\varphi(s)B_1 + \Delta^{-1}(s)B_2\right] \\
&= \left(C_1\varphi(s)A_{12} + C_2\right)\Delta^{-1}(s)\left(A_{21}\varphi(s)B_1 + B_2\right) \\
&= \tilde{C}(s)\Delta^{-1}(s)\tilde{B}(s).
\end{aligned}$$

Então, para um $j\omega$ *fixado,*

$$\begin{aligned}
\max_{1\le i\le n} \sigma_i\left(H(j\omega) - H_k(j\omega)\right) &= \max_{1\le i\le n} \lambda_i^{1/2}\left(\tilde{B}^*(j\omega)\Delta^{-*}(j\omega)\tilde{C}^*(j\omega)\tilde{C}(j\omega)\Delta^{-1}(j\omega)\tilde{B}(j\omega)\right) \\
&= \max_{1\le i\le n} \lambda_i^{1/2}\left(\Delta^{-1}(j\omega)\tilde{B}(j\omega)\tilde{B}^*(j\omega)\Delta^{-*}(j\omega)\tilde{C}^*(j\omega)\tilde{C}(j\omega)\right).
\end{aligned}$$

Particionando as equações (1.24) e (1.25), conforme (2.30) e (2.33) obtemos as seguintes equações:

$$A_{11}\Sigma_1 + \Sigma_1 A_{11}^T + B_1 B_1^T = 0 \tag{B.6}$$

$$A_{12}\Sigma_2 + \Sigma_1 A_{21}^T + B_1 B_2^T = 0 \tag{B.7}$$

$$A_{22}\Sigma_2 + \Sigma_2 A_{22}^T + B_2 B_2^T = 0 \tag{B.8}$$

$$A_{11}^T\Sigma_1 + \Sigma_1 A_{11} + C_1^T C_1 = 0 \tag{B.9}$$

$$A_{21}^T\Sigma_2 + \Sigma_1 A_{12} + C_1^T C_2 = 0 \tag{B.10}$$

$$A_{22}^T\Sigma_2 + \Sigma_2 A_{22} + C_2^T C_2 = 0. \tag{B.11}$$

Manipulando a definição de $\tilde{B}(s)$ *juntamente às as equações (B.6), (B.7) e (B.8), é possível chegar*

à expressão

$$\tilde{B}(j\omega) = \Delta(j\omega)\Sigma_2 + \Sigma_2\Delta^*(j\omega).$$

De maneira análoga, obtém-se

$$\tilde{C}(j\omega) = \Sigma_2\Delta(j\omega) + \Delta^*(j\omega)\Sigma_2.$$

Substituímos, então, estas duas últimas expressões em (B) e realizamos algumas simplificações a fim de obter

$$\max_{1\leq i\leq n} \sigma_i\left(H(j\omega) - H_k(j\omega)\right) = \max_{1\leq i\leq n} \lambda_i^{1/2}\left(\left[\Sigma_2 + \Delta^{-1}(j\omega)\Sigma_2\Delta^*(j\omega)\right]\left[\Delta^{-*}(j\omega)\Sigma_2\Delta(j\omega) + \Sigma_2\right]\right)$$

Supondo que $k = n-1$, então $\Sigma_2 = \sigma_n$ e $\Delta(j\omega)$ é de ordem um. Assim, podemos escrever

$$\max_{1\leq i\leq n} \sigma_i\left(H(j\omega) - H_k(j\omega)\right) = \sigma_n\sqrt{(1+\mathcal{A}(j\omega))(1+\mathcal{A}^{-1}(j\omega))},$$

com $\mathcal{A}(j\omega) = \Delta^*(j\omega)\Delta^{-1}(j\omega) = \overline{\Delta(j\omega)}\Delta^{-1}(j\omega)$. Note que $|\mathcal{A}(j\omega)| = 1$ para todo ω no domínio das frequências. Sendo assim,

$$\max_{1\leq i\leq n} \sigma_i\left(H(j\omega) - H_k(j\omega)\right) = \sigma_n\sqrt{\frac{(1+\mathcal{A}(j\omega))^2}{\mathcal{A}(j\omega)}} = \sigma_n|1+\mathcal{A}(j\omega)|. \leq \sigma_n\left(1+|\mathcal{A}(j\omega)|\right) = 2\sigma_n.$$

Isso conclui a demonstração para o caso $k = n-1$.

Seja agora $E_r(s) := H(s)_{r+1} - H(s)_r$, para $r = 1, 2, ..., n-1$. Pelo que acabamos de provar,

$$\max_{1\leq i\leq n} \sigma_i\left(H_{r-1}(j\omega) - H_r(j\omega)\right) \leq 2\sigma_{r+1}.$$

Como $H(j\omega) - H_k(j\omega) = \sum_{r=k}^{n-1} E_r(s)$, então, pela definição de $E_r(s)$,

$$\begin{aligned}
\max_{1\leq i\leq n} \sigma_i\left(H(j\omega) - H_k(j\omega)\right) &= \max_{1\leq i\leq n} \sigma_i\left(\sum_{r=k}^{n-1} E_r(s)\right) \\
&= \sum_{r=k}^{n-1} \max_{1\leq i\leq n} \sigma_i\left(E_r(s)\right) \\
&= 2\sum_{r=k}^{n-1} \sigma_r,
\end{aligned}$$

como queríamos demonstrar. □