Aprendizado Automático de Decomposições para a Previsão da Estrutura a Termo de Taxas de Juros com Redes Neurais

Piero Conti Kauffmann

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestre em Ciências

Programa: Ciência da Computação
Orientador: Prof. Dr. Julio Michael Stern

São Paulo
Julho de 2022

Aprendizado Automático de Decomposições para a Previsão da Estrutura a Termo de Taxas de Juros com Redes Neurais

Piero Conti Kauffmann

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em Julho de 2022.

Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Julio Michael Stern (orientador) – IME-USP
- Dr. Hellinton Hatsuo Takada
- Profª. Dra. Verónica Andrea González-López – IMECC-UNICAMP

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Agradecimentos

A minha família, pelo carinho e compreensão durante todos os anos de desenvolvimento deste trabalho. Aos meus amigos, por todo apoio e suporte nos momentos mais difíceis.

Resumo

Piero Conti Kauffmann. **Aprendizado Automático de Decomposições para a Previsão da Estrutura a Termo de Taxas de Juros com Redes Neurais**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

Este trabalho propõe um modelo para a previsão da estrutura a termo das taxas de juros que faz aprendizado automático de novas decomposições de curvas de taxas de juros a partir de um modelo linear Gaussiano de espaço de estados acoplado a uma rede neural geradora de decomposições. Para controlar a complexidade do modelo e garantir que as decomposições estimadas preservem propriedades desejáveis, como suavidade e ortogonalidade dos fatores latentes, uma distribuição Priori com efeito de regularização destas propriedades é definida para os parâmetros do modelo, e em seguida, é descrito um procedimento computacionalmente eficiente de estimação para todos os parâmetros do modelo em uma etapa. Uma avaliação empírica com 14 anos de dados históricos da curva de taxa de juros brasileira mostrou que a técnica proposta é capaz de obter melhores previsões fora-de-amostra que modelos tradicionais da literatura, como o modelo Nelson e Siegel dinâmico e variações.

Palavras-chave: Previsão da estrutura a termo das taxas de juros. Redes neurais. Aprendizagem de máquina. Modelagem bayesiana. Decomposição da curva de taxas de juros. Modelos de fatores dinâmicos.

Abstract

Piero Conti Kauffmann. **Learning Forecast-Efficient Yield Curve Factor Decompositions with Neural Networks**. Thesis (Masters). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2022.

This study proposes a term structure forecasting model that learns new yield curve decompositions directly from data, by combining a Gaussian linear state-space model with a neural network that generates smooth yield curve factor loadings. To reduce the complexity of the model and ensure that the estimated decompositions preserve desirable properties such as smoothness and orthogonality of the factors, Prior distributions with the regularization effect of these properties are defined for the model parameters. A computationally efficient estimation procedure based on the Kalman Filter and automatic differentiation algorithms for the state and space parameters is described. An evaluation of the model's performance on 14 years of historical data of the Brazilian yield curve shows that the proposed technique was able to obtain better overall out-of-sample forecasts than traditional approaches, such as the Dynamic Nelson and Siegel model and its extensions.

Keywords: Yield curve forecasting. Neural networks. Machine Learning. Bayesian modeling. Yield curve decomposition. Dynamic factor models.

Lista de Abreviaturas

AR	*Autoregressive Model*
BC	Banco Central
CDI	Certificado de Depósito Interfinanceiro
COPOM	Comitê de Política Monetária do Banco Central
CMN	Conselho Monetário Nacional
DI	Depósito Interfinanceiro
DNS	*Dynamic Nelson and Siegel Model model*
DSV	*Dynamic Nelson, Siegel and Svensson model*
ETTJ	Estrutura a Termo da Taxa de Juros
EM	*Expectation Maximization*
GLEE	Modelo Gaussiano Linear de Espaço de Estados
GPU	*Graphical Processing Unit*
LTN	Letra do Tesouro Nacional
LNEE	Modelo Linear-Neural de Espaço de Estados
MCMC	*Markov Chain Monte Carlo*
MLP	*Multilayer Perceptron*
PCA	*Principal Component Analysis*
ReLU	*Rectifier Linear Unit*
RMSE	*Root Mean Squared Error*
RW	*Random Walk*
SELIC	Sistema Especial de Liquidação e de Custódia
VAR	*Vector Autoregressive Model*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A estrutura a termo da taxa de juros (ETTJ), ou curva de taxa de juros, é um conceito fundamental da teoria econômica que quantifica a relação das taxas de juros de um determinado papel a diferentes prazos. As curvas de taxas de juros são estruturas ricas em informação que refletem as expectativas do mercado sobre as taxas de juros futuras e sobre a política monetária nacional, o que faz a análise e a previsão da ETTJ serem tópicos de suma importância para múltiplos agentes do mercado financeiro. Particularmente, o problema de previsão das curvas de taxas de juros se destaca pela sua importância nas atividades de precificação de ativos, gestão de ativos de renda fixa, gestão de risco financeiro e controle por órgãos reguladores.

O trabalho seminal de Diebold e Li (2006), que introduz o modelo Nelson e Siegel dinâmico, baseado na decomposição de Nelson e Siegel (1987), marca uma mudança importante na literatura de previsão da curva de taxa de juros que, antes concentrada em modelos de equilíbrio (Vasicek (1977), Cox *et al.* (2005) e Duffie e Kan (1996)) e de não-arbitragem (Hull e White (1990) e Heath *et al.* (1992)), passa agora a incorporar modelos estatísticos com enfoque na performance de previsão fora-de-amostra.

O sucesso do modelo proposto por Diebold e Li (2006) inspirou uma séries de extensões para o modelo Nelson e Siegel dinâmico, como a inclusão da hipótese de não-arbitragem (Christensen *et al.* (2009)), variáveis macroeconômicas (Diebold *et al.* (2006)), coeficientes variantes no tempo (Koopman *et al.* (2010)) e mudança de regime (Xiang e Zhu (2013)).

Grande parte do sucesso do modelo Nelson e Siegel dinâmico pode ser atribuído às propriedades vantajosas da decomposição de Nelson e Siegel (1987), que é capaz de decompor aproximadamente as curvas de taxas de juros em um conjunto fatores latentes de dimensão menor com boa performance. Com o objetivo de criar um modelo para a previsão da estrutura a termo, Diebold e Li (2006) usam uma abordagem de estimação de parâmetros em duas etapas, que primeiro decompõe de maneira independente cada curva de taxa de juros usando o modelo de Nelson e Siegel para formar uma série temporal de vetores de fatores latentes, e em seguida, estima os parâmetros de um modelo de séries temporais multivariadas usando os fatores latentes extraídos.

Trabalhos mais recentes mostram que modelos de decomposição independente de

curvas de juros, que podem ser adaptados para um problema de previsão usando a mesma abordagem de estimativa em duas etapas de Diebold e Li (2006), podem ser melhorados para considerar formatos mais complexos de curvas de taxas juros (Takada e Stern (2015), Faria e Almeida (2018), Mineo *et al.* (2020), por exemplo).

Alternativamente, outros trabalhos (como Bowsher e Meeks (2008) e Hays *et al.* (2012)) propõem a utilização de modelos de decomposição mais flexíveis para as curvas de taxas de juros em uma formulação de modelos de espaços de estados, o que possibilita que a estimação dos parâmetros temporais e de decomposição seja feita simultaneamente, e portanto, que o espaço obtido de fatores latentes seja mais eficiente para o problema de previsão.

Apesar da aparente superioridade teórica da abordagem de estimação em uma etapa, o número de trabalhos com esta abordagem é significativamente menor que a versão de duas etapas. Um dos principais motivos para esta discrepância é a potencial sucetibilidade desta metodologia ao problema de sobreajuste, o que pode dificultar a utilização destas técnicas efetivamente em alguns cenários. Por exemplo, Diebold e Rudebusch (2013) observam que o modelo Nelson e Siegel dinâmico com estimação em duas etapas produziu resultados superiores à versão com estimação em uma etapa, enquanto no caso da curva brasileira, a abordagem em uma etapa parece produzir resultados superiores (Caldeira *et al.* (2010), Caldeira *et al.* (2016b)).

De maneira análoga, este mesmo comportamento pode ser identificado em outras hipóteses de modelagem usuais da literatura também sujeitas à mesma dicotomia dos fenômenos de subajuste e sobreajuste, como por exemplo na hipótese de ortogonalidade do processo autoregressivo dos fatores extraídos, que também parece produzir resultados superiores para o caso americano (Diebold e Li (2006), De Pooter (2007)) e inferiores para o caso brasileiro (Caldeira *et al.* (2016a)). Tais fenômenos podem ser explicados pelo fato da ETTJ brasileira historicamente apresentar comportamento temporal (choques) de maior complexidade e uma diversidade maior de formatos de curva em relação a ETTJ americana.

Além disso, um possível segundo motivo para a escassez de trabalhos com decomposições estimáveis a partir de dados em uma etapa pode ser atribuído à dificuldade de adaptar e estender esta classe mais complexa de modelos à diferentes aplicações, diferentemente da abordagem de estimação em duas etapas que possuí alta flexibilidade, decorrente da hipótese de independência entre o modelo de decomposição e de evolução de fatores.

1.1 Objetivos

Os problemas de suscetibilidade variável ao fenômeno de sobreajuste e de dificuldade de extensão desta classe de modelos podem ser vistos, sob uma ótica unificada, como uma limitação de *flexibilidade* do processo de modelagem como um todo. Sob esta perspectiva, o presente trabalho tem como objetivo propor um modelo para previsão de curvas de taxas de juros com decomposições estimadas totalmente a partir de dados que seja capaz de generalizar algumas das hipóteses de modelagem da literatura e ofereça alta flexibilidade e adaptabilidade a múltiplos domínios.

Além disso, por se tratar de um modelo de previsão da ETTJ baseado em decomposições estimadas totalmente a partir de dados, este trabalho também tem como objetivo estudar e replicar organicamente as boas propriedades dos modelos tradicionais de decomposição das curvas de taxas de juros.

Portanto, os objetivos deste trabalho podem ser mais precisamente divididos em:

1. Revisar as principais propriedades desejáveis dos modelos de decomposição de curvas de taxas de juros
2. Propor um modelo geral e flexível para previsão da ETTJ que estime decomposições a partir dos dados para um número arbitrário de fatores com estimação de parâmetros em uma etapa
3. Propor mecanismos de regularização para o modelo que reflitam em propriedades de interesse da ETTJ
4. Avaliar empiricamente a performance do modelo proposto relativamente aos modelos tradicionais da literatura

1.2 Estrutura do Texto

O restante do texto deste trabalho é dividido em 5 seções. No Capítulo 2, os principais conceitos e propriedades sobre a estrutura a termo da taxas de juros são apresentados. No Capítulo 3, são revisados os principais modelos de previsão da ETTJ e suas propriedades. No Capítulo 4, a teoria básica dos modelos de redes neurais, estatística Bayesiana e modelos lineares Gaussianos de espaço de estados é revista, e em seguida, o modelo linear-neural de espaço de estados para previsão da ETTJ é apresentado. Por fim, nos capítulos 5 e 6, o modelo proposto é avaliado empiricamente e as conclusões finais e sugestões para trabalhos futuros são apresentadas.

Alguns dos elementos e resultados experimentais deste trabalho foram apresentadas previamente em Kauffmann *et al.* (2022), porém o presente trabalho contém muitos elementos novos, como uma revisão mais extensa da literatura (Capítulo 3), uma discussão mais profunda das limitações e motivações deste trabalho (Seção 4.2.1) e um novo experimento com dados artificiais (Seção 5.2).

Capítulo 2

Estrutura a Termo de Taxa de Juros (ETTJ)

Um título é uma obrigação contratual na forma de um empréstimo securitizado (instrumento de dívida), onde o comprador de um título empresta dinheiro ao emissor (Fortuna, 2010). O instrumento pode ser utilizado por governos (títulos públicos) ou empresas privadas (títulos privados) para financiar suas atividades ou alongar suas dívidas. Os rendimentos desses papéis podem ser reais, nominais ou indexados às taxas flutuantes e são recebidos em intervalos de tempo regulares definidos em documentos formais.

Um tipo importante de contrato para o estudo das taxas de juros é o denominado título pré-fixado zero-cupom, onde o comprador recebe todo o valor nominal na data de vencimento, atualizado por uma taxa de juros pré-definida no instante da compra.

No Brasil o título público pré-fixado zero-cupom é denominado letra do tesouro nacional (LTN), cuja estimativa de taxa de retorno apropriada é feita com base em projeções para as taxas de juros nominais para o período de fluência do título (entre a data de liquidação e resgate).

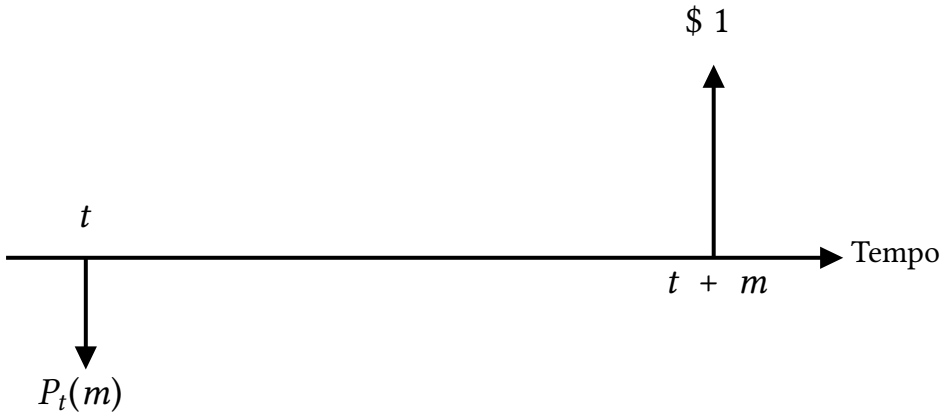

Figura 2.1: *Fluxo de caixa de um título pré-fixado zero-cupom de valor de face unitário e vencimento m, negociado em t pelo preço $P_t(m)$.*

A Figura 2.1 representa o fluxo de caixa sob a perspectiva de um comprador de um

título pré-fixado zero-cupom de valor de face unitário.

O preço $P_t(m)$ do título representado na Figura 2.1 de vencimento m e negociado no instante t, é dado por

$$P_t(m) = \frac{1}{(1 + y_t(m))^m} \tag{2.1}$$

sob um regime de capitalização composta, ou

$$P_t(m) = e^{-m y_t(m)} \tag{2.2}$$

sob um regime de capitalização contínua, onde o termo $y_t(m)$ é a taxa de juros de um instante t para um contrato de vencimento m.

A função y_t, que quantifica a relação medida em dado momento t entre um vencimento (ou vértice) e a taxa de juros de um título pré-fixado zero-cupom de mesmo vencimento, é denominada a estrutura termo da taxa de juros (ETTJ), ou curva de taxa de juros. Um diagrama desta relação é ilustrado na Figura 2.2.

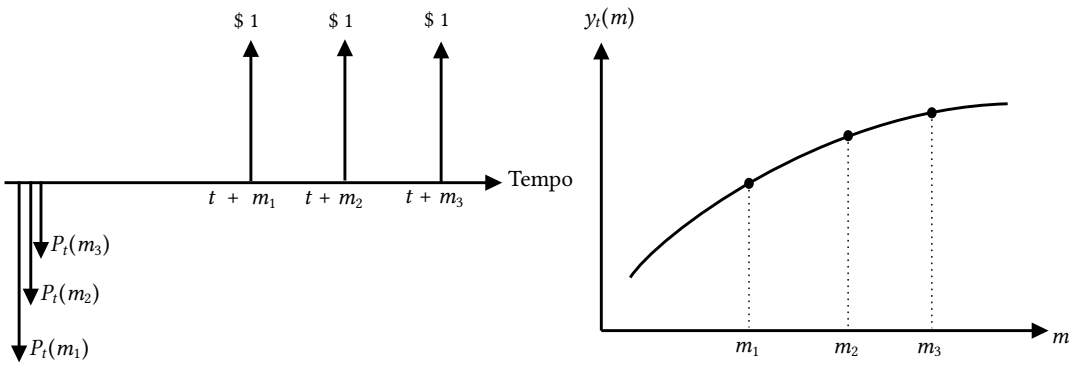

Figura 2.2: *Relação entre o fluxo de caixa de um título pré-fixado zero-cupom de valor de face unitário em três vencimentos m_1, m_2, m_3 e a respectiva curva de taxa de juros y_t do instante t.*

Na prática, a ETTJ é uma ferramenta importante que reflete as expectativas do mercado financeiro em relação as taxas de juros futuras em um dado momento, servindo como base para o apreçamento de instrumentos de renda fixa e de referência para o gerenciamento de risco de ativos financeiros.

2.1 Selic

A ETTJ também é um indicador importante para o controle da política monetaria nacional, exercida pelos bancos centrais (BC). Os juros são utilizados pelos bancos centrais como uma ferramenta de controle da inflação, uma vez que a alta ou a queda das taxas de juros impactam os hábitos de consumo da população, a tomada de crédito e o crescimento

do país. No Brasil, a taxa de juros básica da economia, que exerce influência sobre todas as demais taxas do país, é denominada taxa Selic.

A taxa Selic em si (taxa Selic efetiva) é medida como a taxa média ajustada das operações de financiamento no mercado interbancário de um dia, registradas no Sistema Especial de Liquidação e de Custódia (SELIC) (Fortuna (2010)). A cada 45 dias utéis, o Comitê de Política Monetária do Banco Central (COPOM), se reune para discutir e avaliar a atividade econômica do país e o cumprimento da meta geral de inflação, definida pelo Conselho Monetário Nacional (CMN). A partir desta discussão, uma meta para a taxa Selic é definida e o banco central atua no mercado afim de manter a taxa Selic efetiva próxima a da meta definida no COPOM.

Na hipótese de redução da meta da taxa Selic, a rentabilidade de títulos indexados a esta também diminui, reduzindo também o custos dos bancos e consequentemente os juros de empréstimos, o que tem efeito de aquecimento da economia nacional. Analogamente, quando a meta da Selic sobe, o custo dos bancos aumenta e há aumento nas taxas de juros causando uma desaceleração na economia.

2.2 Taxa DI

A taxa dos depósitos interfinanceiros (DI) é a taxa média de operações entre instituições financeiras nas quais um banco toma recursos de outra instituição financeira, usualmente por um dia útil, para cobrir necessidades de caixa. A taxa DI é calculada a partir de operações de certificados de depósito interfinanceiro (CDI), que lastreiam o mercado interbancário e só podem ser negociados entre instituições financeiras na central de custódia e liquidação financeira de títulos (CETIP).

Além disso, a taxa DI também é utilizada como referência de remuneração pós-fixada no mercado financeiro brasileiro e para indexação de outros produtos financeiros, como certificados de depósito bancário, empréstimos e contratos de derivativos.

2.3 Construção da ETTJ no Brasil

O termo ETTJ pode ser usado para se referir à uma variedade de estruturas financeiras. No presente trabalho, será considerada a ETTJ de taxa de juros *spot* nominal zero-cupom.

A construção da ETTJ depende da obtenção das taxas de juros de títulos públicos pré-fixados zero-cupom para diferentes prazos de vencimento, porém, como discutido anteriormente, a única taxa de juros explicitamente conhecida no presente é a taxa Selic. Portanto, os juros dos demais vencimentos são baseados unicamente na expectativa dos agentes participantes do mercado, observada através da negociação de títulos públicos federais prefixados (LTNs ou NTN-F) ou de contratos de derivativos (futuro de DI ou swap pré x DI).

Usualmente, o instrumento utilizado para construção da curva de juros são os contratos de futuro de DI, negociados na bolsa brasileira, devido a grande liquidez desses ativos no mercado. Os contratos futuros de DI possuem vencimento apenas no primeiro dia útil de cada mês subsequente, portanto a utilização de métodos de interpolação de vértices da curva de taxa de juros se torna necessária para a construção completa da curva em todos os vencimentos desejados. Alguns dos principais métodos de interpolação de vértices faltantes serão descritos adiante na Seção 2.6.1.

2.4 Formatos Teóricos para as Curvas de Taxas de Juros

Nesta seção, alguns formatos comuns das curvas de taxas de juros são apresentados, conjuntamente com possíveis justificativas econômicas do trabalho de Choudhry (2019) para estes formatos. Entretanto, na prática, as curvas de taxas de juros apresentam formas mistas, e pode ser difícil obter explicações plausíveis para todas as intenções dos agentes financeiros envolvidos.

2.4.1 Curva de Juros Normal

A curva de taxa de juros de formato ascendente côncavo com progressão suave (Figura 2.3, painel c.) é um dos formatos mais comuns para as curvas de taxas juros.

Uma interpretação possível para este formato é que o mercado espera que a economia nacional funcione com uma taxa usual de crescimento com a presença de pressão inflacionária. Portanto, investidores com investimentos de longo prazo esperam maiores rendimentos.

2.4.2 Curva de Juros Fortemente Inclinada

A curva de taxa de juros com formato ascendente acentuado (Figura 2.3, painel d.) pode representar dois cenários econômicos plausíveis:

1. A economia está no início de um período de crescimento, tal ciclo é associado a fortes pressões inflacionárias. O mercado espera que o governo atue mais agressivamente no controle da inflação (política monetária).
2. O governo está com dificuldade para financiar sua dívida de longo prazo, então os investidores precificam um prêmio maior para títulos de longo prazo.

2.4.3 Curva de Juros Invertida

Uma curva de juros é chamada de invertida se as taxas de longo prazo são menores que as taxas de curto prazo (Figura 2.3, painel a.). A curva de juros invertida usualmente

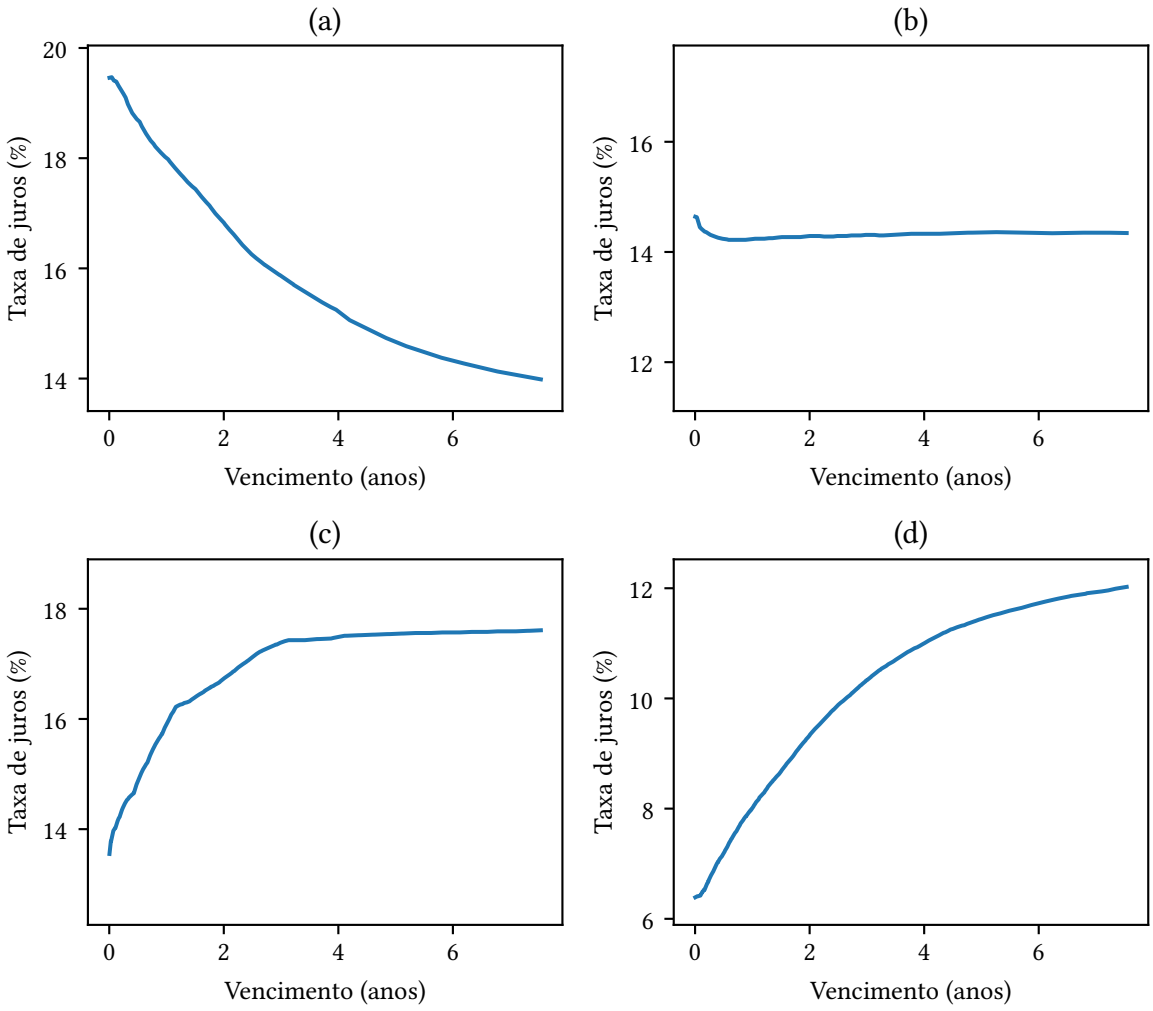

Figura 2.3: *Curvas de taxas de juros brasileiras, extraídas da base de dados descrita no Capítulo 5, para as datas: (a) 16/09/2005 (b) 22/08/2006, (c) 22/10/2008, e (d) 28/09/2018.*

significa que o mercado prevê uma desaceleração da economia, o que pode ser um indício de recessão ou estagnação econômica.

Uma possível justificativa para este fenômeno é que em economias recessivas ou em vias de recessão, o governo pode agir ativamente para reduzir as taxas de juros com o objetivo de aquecer a economia. Portanto, na presença desta expectativa, os investidores tendem a comprar títulos de longo prazo para travar as taxas antes que o governo as reduza, o que consequentemente reduz as taxas de longo prazo e resulta na inversão da curva da taxa de juros.

2.4.4 Curva de Juros Plana

Na curva de taxa de juros plana, ou *flat*, as taxas de longo prazo se assemelham às de curto prazo (Figura 2.3, painel b.), o que representa que o mercado está em ponto de inflexão, e portanto existem incertezas acerca do momento econômico.

2.5 Teorias de formação da ETTJ

Algumas teorias para explicar a formação da ETTJ foram propostas na literatura de teoria econômica. Segundo Hull (2003), as três principais teorias são

1. Teoria das expectativas: as taxas de juros de longo prazo refletem expectativas futuras das taxas de curto prazo. Portanto, o rendimento de um investidor que compra um papel com prazo de um ano e carrega até o vencimento deve ser equivalente ao rendimento esperado de um investidor que compra um papel semelhante com prazo de seis meses e no seu vencimento compra outro papel com prazo de seis meses novamente.

2. Teoria da segmentação do mercado: o mercado age de maneira segmentada e independente nas taxas de curto, médio e longo prazo. O que implica que as taxas de diferentes horizontes de vencimento devem ser vistas como componentes independentes.

3. Teoria da preferência por liquidez: os investidores possuem preferência por títulos com maior liquidez, e portanto exigem prêmio maior em títulos de vencimentos mais longos, que possuem risco de liquidez maior que títulos de vencimento mais curto.

As teorias das expectativas e da segmentação de mercado são menos aceitas na literatura, uma vez que não explicam bem empiricamente a dinâmica observada das curvas de taxas de juros. Apesar de também contestada, a teoria de preferência por liquidez é discutivelmente mais coerente com alguns resultados empíricos.

2.6 Propriedades Principais da ETTJ

As taxas de juros abordadas até agora, que representam as taxas de juros do mercado à vista para um título zero-cupom de dado vencimento, também são denominadas taxas *spott*. Um outro conceito útil para o estudo da ETTJ são as taxas de juros *forward*, que implicitamente indicam as expectativas das taxas de juros futuras.

Como visto anteriormente, sob um regime de capitalização contínua, o preço no instante t de um título de valor de face unitário, vencimento m e de taxa de juros *spot* $y_t(m)$ é dado por

$$P_t(m) = e^{-m y_t(m)} \tag{2.3}$$

A partir de P_t, a curva de taxa de juros *forward* instantânea f_t pode ser definida como

$$f_t(m) = -\frac{\partial \log P_t(m)}{\partial m} = y_t(m) + m y_t'(m) \tag{2.4}$$

A taxa *forward* instantânea pode ser interpretada como o valor esperado da taxa de juros *spot* para um instante infinitesimalmente no futuro, *se verdadeira hipótese da teoria*

das expecativas.

Também é possível expressar a relação entre a taxa *spot* $y_t(m)$ e a taxa *forward* instantânea $f_t(m)$ como

$$y_t(m) = -\log P_t(m) = \frac{1}{m}\int_0^m f_t(u)du \tag{2.5}$$

o que também implica que a taxa *spot* zero-cupom pode ser vista como uma média das taxas *forwards* instantâneas de todos os vencimentos até m.

2.6.1 Interpolação de Vértices Faltantes

O processo de interpolação de vértices faltantes é uma etapa importante da construção da ETTJ que pode impactar diretamente a qualidade dos resultados da aplicação de interesse. Particularmente, a metodologia de interpolação de vértices faltantes pode desempenhar um papel crucial em casos em que existe escassez de vértices observados.

A literatura deste tópico é vasta (veja Hagan e West (2006) e Floc'h (2013) para uma revisão bibliográfica extensiva e uma análise comparativa entre abordagens), porém como este não é um tópico principal do presente trabalho, apenas dois métodos mais simples de interpolação de vértices faltantes serão revisados.

2.6.2 Interpolação Linear

O método mais simples de interpolação de vértices faltantes é o método de interpolação linear aplicado diretamente na curva de taxa juros. Apesar de sua simplicidade atrativa, este método usualmente produz resultados ruins e é pouco utilizado na prática.

Ao aplicar o método de interpolação linear entre dois pares de vértices $(m_a, y_t(m_a))$ e $(m_b, y_t(m_b))$, a taxa de juros de um vértice faltante $m \in [m_a, m_b]$ é dada por

$$y_t(m) = y_t(m_a) + \left(\frac{y_t(m_b) - y_t(m_a)}{m_b - m_a}\right)(m - m_a) \tag{2.6}$$

2.6.3 Interpolação Flat-Forward

A metodologia *flat-forward* de interpolação (Maltz (2002)), assume uma taxa *forward* instantânea constante entre dois vértices adjacentes m_a e m_b.

$$f_t(m) = f_c, \quad \forall m \in [m_a, m_b] \tag{2.7}$$

Pelas equações 2.4 e 2.6, é possível verificar que

$$\int_{m_a}^{m_b} f_c \, du = m_b y_t(m_b) - m_a y_t(m_a)$$

$$f_c = \frac{m_b y_t(m_b) - m_a y_t(m_a)}{m_b - m_a} \tag{2.8}$$

Portanto, para qualquer $m \in [m_a, m_b]$, segue que

$$m y_t(m) = m_a y_t(m_a) + f_c(m - m_a)$$

$$m y_t(m) = \left(\frac{m_b y_t(m_b) - m_a y_t(m_a)}{m_b - m_a} \right) (m - m_a) \tag{2.9}$$

Da Equação 2.9, verifica-se o método de interpolação *flat-forward* é equivalente a um processo de interpolação linear na curva de log-preço, visto que $m y_t(m) = -\log P_t(m)$. Ainda que extremamente simples, a metodologia *flat-forward* é capaz de produzir resultados satisfatórios quando o número de vértices observados não é tão baixo.

Capítulo 3

Modelos Tradicionais de Previsão da ETTJ

Nesta seção são apresentados os principais modelos e metodologias para previsão da ETTJ. Decorrente da estrutura bi-dimensional da ETTJ (tempo x vencimento), as metodologias de modelagem tipicamente dependem da representação escolhida para a curva de taxa de juros $\boldsymbol{y}_t$. A opção mais direta para representar a curva de taxa de juros em um instante t é por meio de um vetor de taxas de juros

$$\boldsymbol{y}_t = \begin{bmatrix} y_{t,m_1} \\ y_{t,m_2} \\ ... \\ y_{t,m_M} \end{bmatrix} \tag{3.1}$$

onde $\{m_1, m_2, ..., m_M\}$ é um conjunto fixo de M vencimentos. A partir desta estrutura, uma abordagem tradicional de modelagem de séries temporais multivariadas pode ser empregada diretamente

$$\boldsymbol{y}_t = f_{\theta}(\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ..., \boldsymbol{y}_{t-L}) + \boldsymbol{\epsilon}_t, \quad t = (L+1), (L+2), ... \tag{3.2}$$

onde f é um modelo parametrizado por $\boldsymbol{\theta}$ e $\boldsymbol{\epsilon}_t$ é um termo de ruído. Apesar de sua conveniente simplicidade, esta metodologia possui algumas desvantagens. Tipicamente, a construção da série temporal $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \}$ depende da realização de um procedimento de interpolação ou extrapolação de vértices faltantes, visto que as taxas da ETTJ são observadas de maneira irregular ao longo do tempo. De maneira análoga, previsões das taxas de juros para vencimentos não presentes em $\{m_1, ..., m_M\}$ não podem ser obtidas diretamente.

Além disso, uma metodologia de modelagem baseada na representação multivariada pode não ser capaz de capturar algumas das propriedades fundamentais presentes nas curvas de taxa de juros, como o padrão de suavidade e estabilidade assintótica para vencimentos mais longos.

Uma opção alternativa é representar a curva de taxa de juros no instante t por meio de uma função continua $\boldsymbol{y}_t : \mathbb{R}^+ \longrightarrow \mathbb{R}$. Esta escolha de estrutura permite utilizar as observações de um instante t em qualquer conjunto de vencimentos, sem depender de um conjunto fixo de vencimentos. Além disso, as propriedades observadas nas curvas de taxas de juros, como suavidade e estabilidade assintótica, podem ser impostas limitando-se a família de funções utilizadas em um modelo para $\boldsymbol{y}_t$.

No restante deste capítulo, serão apresentados os principais modelos da literatura baseados em ambas metodologias, assim como seus respectivos métodos de estimação e previsão.

3.1 Passeio Aleatório

O modelo mais simples de previsão da ETTJ assume que as taxas variam de acordo com um passeio aleatório

$$\boldsymbol{y}_t = \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon} \overset{ind.}{\sim} N(0, I_M \sigma^2) \tag{3.3}$$

Apesar da simplicidade, o passeio aleatório é considerado um modelo extremamente competitivo, principalmente para previsão de horizontes curtos de tempo. Em alguns casos o modelo de passeio aleatório supera a performance de previsão de modelos mais complexos (Ang e Piazzesi (2003), Duffee (2002) e Hördahl *et al.* (2006)), servindo assim como um ponto de referência importante para o desenvolvimento de técnicas mais sofisticadas.

A previsão de um modelo de passeio aleatório no instante t para um horizonte h é dada por

$$\hat{\boldsymbol{y}}_{t+h|t} = \boldsymbol{y}_t \tag{3.4}$$

3.2 Modelo Autoregressivo Vetorial

O modelo autoregressivo vetorial (VAR) de ordem k para a ETTJ assume que as curvas de taxas de juros respeitam o processo estocástico

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{F}_1 \boldsymbol{y}_{t-1} + \boldsymbol{F}_2 \boldsymbol{y}_{t-2} + ... + \boldsymbol{F}_k \boldsymbol{y}_{t-k} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{iid}{\sim} N_M(0, \boldsymbol{P}) \tag{3.5}$$

onde $\boldsymbol{F}_1, ..., \boldsymbol{F}_k \in \mathbb{R}^{M \text{x} M}$ são k matrizes de parâmetros, $\boldsymbol{\mu} \in \mathbb{R}^M$ é um vetor de parâmetros de intercepto, $\boldsymbol{P} \in \mathbb{R}^{M \text{x} M}$ é uma matriz positiva definida e $\boldsymbol{\epsilon}_t$ é um vetor de ruído branco.

O modelo autoregressivo vetorial pode ser visto como uma extensão multivariada do processo autoregressivo univariado $x_t = \phi_1 x_{t-1} + ... + \phi_k x_{t-k} + \epsilon$. Particularmente, quando as matrizes $\boldsymbol{F}_1, ..., \boldsymbol{F}_k$ e $\boldsymbol{P}$ são diagonais, o processo autoregressivo vetorial de ordem k pode

ser escrito como M processos autoregressivos independentes para cada um dos vértices de $y_{t,m}$ de $\boldsymbol{y}_t$

$$y_{t,m} = \mu_m + \sum_{j=1}^{k} f_{j,m} y_{t-j,m} + \epsilon_{t,m} \tag{3.6}$$

No modelo autoregressivo vetorial VAR(k), o valor esperado de cada vértice da curva $\boldsymbol{y}_t$ é uma combinação linear de *todos* os vértices das k curvas dos instantes anteriores $\boldsymbol{y}_{t-1}$, ..., $\boldsymbol{y}_{t-k}$, o que permite incorporar dependencias temporais entre vértices diferentes de $\boldsymbol{y}_t$ e o histórico $\{\boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}\}$. Apesar disso, o número de parâmetros de um modelo VAR(k) varia quadraticamente com o número M de vértices da curva e linearmente com o valor de k, resultando em um número excessivo de parâmetros para o modelo.

Por este motivo, no problema de previsão da ETTJ é comum fixar $k = 1$, devido ao alto número de parâmetros das matrizes $\{\boldsymbol{F}_1, ..., \boldsymbol{F}_k\}$. Porém, esta simplificação muitas vezes não é o suficiente para previnir o problema de sobreajuste do modelo, visto que o número de vencimentos M é usualmente elevado.

Os parâmetros do modelo autoregressivo vetorial podem ser estimados a partir do método de mínimos quadrados multivariados (MQM). Para um modelo VAR(1), os estimadores MQM são dados por

$$\begin{bmatrix} \hat{\boldsymbol{\mu}} & \hat{\boldsymbol{F}}_1 \end{bmatrix} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1} \tag{3.7}$$

onde

$$\boldsymbol{X} = \begin{bmatrix} 1 & \cdots & 1 \\ y_{1,1} & \cdots & y_{T-1,1} \\ \vdots & \vdots & \vdots \\ y_{1,M} & \cdots & y_{T-1,M} \end{bmatrix}_{(M+1)\text{ x }(T-1)}, \quad \boldsymbol{Y} = \begin{bmatrix} y_{2,1} & y_{3,1} & \cdots & y_{T,1} \\ y_{2,2} & y_{3,2} & \cdots & y_{T,2} \\ \vdots & \vdots & \vdots & \vdots \\ y_{2,M} & y_{3,M} & \cdots & y_{T,M} \end{bmatrix}_{M\text{ x }(T-1)}$$

É possível verificar que os estimadores obtidos a partir da Equação 3.7 são os mesmos obtidos após aplicar o método de mínimos quadrados ordinários (OLS) M vezes em cada uma das equações de regressão de cada vértice m,

$$\begin{bmatrix} y_{2,m} \\ y_{3,m} \\ \vdots \\ y_{T,m} \end{bmatrix} = \boldsymbol{X}^T\boldsymbol{\beta}_m + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_M(0, I_M\sigma^2) \tag{3.8}$$

onde o vetor $\boldsymbol{\beta}_m$ corresponde a m-ésima linha da matriz $\boldsymbol{F}_1$.

Após a estimação dos parâmetros de um modelo VAR(1), uma previsão para a curva $\boldsymbol{y}_{t+1}$ pode ser obtida

$$\hat{\boldsymbol{y}}_{t+1|t} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{F}}_1\boldsymbol{y}_t \tag{3.9}$$

3.2.1 Modelos de Previsão Baseados em Métodos de Redução de Dimensionalidade

Uma alternativa popular para reduzir o número de parâmetros necessários para prever todos os vértices das curvas de taxas de juros é acoplar uma técnica de redução de dimensionalidade a um modelo de séries temporais. Modelos de redução de dimensionalidade são métodos consagrados na literatura de estudos da ETTJ, popularizados inicialmente neste contexto pelo trabalho de Litterman e Scheinkman (1991), que propõe o uso da análise de componentes principais para *hedging* de carteiras de renda fixa. Um dos motivos principais atribuídos a popularidade destas técnicas é o fenômeno de alta correlação entre taxas de vértices diferentes, que permite reduzir drasticamente a dimensionalidade deste tipo de informação.

No contexto de previsão da ETTJ, as técnicas de redução de dimensionalidade também são uteis efetivamente para dividir o problema original de previsão no espaço das curvas de taxas de juros a um problema de compressão de curvas de taxa de juros e um problema de previsão multivariado em um espaço de baixa dimensionalidade. Esta divisão é particularmente conveniente para o problema de previsão da ETTJ, pois permite modelar com níveis de complexidade diferentes a decomposição da curva em fatores comuns e a evolução temporal dos fatores.

Análise de Componentes Principais

A análise de componentes principais é uma técnica estatística que busca obter uma transformação linear ortogonal do espaço original de um conjunto de observações em um espaço de dimensão reduzida, de forma a preservar o máximo possível da variância total original dos dados. Dada matriz de dados $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \dots & \boldsymbol{y}_T \end{bmatrix}^T$, de vetor de médias $\boldsymbol{\mu}_Y$ e matriz de covariâncias Σ_Y, a *variância total* de $\boldsymbol{Y}$ é definida como a soma das variâncias de $\boldsymbol{Y}$, obtida pela soma da diagonal da matriz de covariâncias $tr(\Sigma_Y)$.

Considerando a transformação linear $\boldsymbol{Z} = \tilde{\boldsymbol{Y}}\boldsymbol{U}$, onde $\tilde{\boldsymbol{Y}}$ denota a matriz de dados $\boldsymbol{Y}$ com média centrada em zero e $\boldsymbol{U}$ uma matriz ortogonal de dimensão M x k, com $k < M$, o problema da análise de componentes principais pode ser escrito como o problema de otimização

$$\begin{aligned} \max_{\boldsymbol{U}} \quad & tr\left(\boldsymbol{U}^T \Sigma_Y \boldsymbol{U}\right) \\ \text{s.a.} \quad & \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \end{aligned} \tag{3.10}$$

decorrente de que $tr(\Sigma_Z) = tr\left(\boldsymbol{U}^T \Sigma_Y \boldsymbol{U}\right)$.

É possível demonstrar que a solução deste problema de maximização é a matriz formada pelo k autovetores $\boldsymbol{v}_{(1)}, ..., \boldsymbol{v}_{(k)}$ de Σ_Y de maiores autovalores associados $\lambda_{(1)}, ..., \lambda_{(k)}$

$$\boldsymbol{V}_k = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_k \end{bmatrix} \tag{3.11}$$

Além disso, também é possível verificar que a variância total máxima da projeção $\tilde{\boldsymbol{Y}}\boldsymbol{U}$ é dada por $tr(\Sigma_Z) = \sum_{i=1}^{k} \lambda_{(i)}$.

A partir da solução da análise de componentes principais, o problema de previsão da ETTJ pode ser traduzido ao problema de prever a série temporal de dimensão reduzida de fatores comuns $\{z_1, ..., z_T\}$ das linhas da matriz

$$Z = \tilde{\boldsymbol{Y}}\boldsymbol{V}_k \tag{3.12}$$

O espaço reduzido em k dimensões de $\{z_1, ..., z_T\}$ pode ser útil para aprimorar a performance de previsão de modelos autoregressivos vetoriais, visto que esta classe de modelos é usualmente sensível ao problema de sobreajuste em espaços de alta dimensionalidade.

Uma previsão $\hat{\boldsymbol{z}}_{t+1|t}$ do vetor $\boldsymbol{z}_{t+1}$ pode ser re-transformada no espaço original para uma previsão $\hat{\boldsymbol{y}}_{t+1|t}$ de $\boldsymbol{y}_{t+1}$

$$\hat{\boldsymbol{y}}_{t+1|t} = \boldsymbol{V}_k \hat{\boldsymbol{z}}_{t+1|t} \tag{3.13}$$

Apesar da simplicidade, a análise de componentes principais como método de redução de dimensionalidade para a ETTJ é tipicamente capaz de obter bons resultados com apenas três ou quatro fatores. Em Litterman e Scheinkman (1991), os autores propõe utilizar três fatores para descrever as curvas de taxa de juros que podem ser interpretados como nível, inclinação e curvatura. Além disso, Litterman e Scheinkman (1991) mostram que tais fatores são capazes de explicar cerca de 96% da variabilidade total dos retornos dos títulos de renda fixa.

Métodos Alternativos de Redução de Dimensionalidade

Uma variedade de técnicas de redução de dimensionalidade podem ser usadas no problema de previsão das curvas de taxas de juros de maneira análoga. Takada e Stern (2015), por exemplo, obtém bons resultados com técnicas de fatoração não negativa de matrizes e Suimon *et al.* (2020) com modelos baseados em *autoencoders*.

Alternativamente, o objetivo de redução de dimensionalidade também pode ser alcançado por meio de procedimentos de ajuste de funções contínuas parametrizadas, usualmente apelidados de métodos de *decomposição* da curva de taxa de juros. Tais métodos são usualmente úteis para modelar curvas de taxa de juros pois se aproveitam das propriedades de continuidade e suavidade da ETTJ. Nas próximas seções, alguns modelos famosos de decomposição da ETTJ são introduzidos.

3.2.2 Modelo Nelson e Siegel Dinâmico

Decomposição de Nelson e Siegel

O trabalho seminal de Nelson e Siegel (1987) propõe um modelo de quatro parâmetros para descrever a estrutura das curvas de taxas de juros como uma função continua $y(m)$.

A construção deste modelo se origina de uma aproximação da curva *forward* instantânea $f(m)$ (definida na Equação 2.4) por meio de polinômios de Laguerre, que são compostos de um produto de um polinômio e uma função de decaimento exponencial

$$f(m) = \beta_0 + \beta_1 e^{-\lambda m} + \beta_2 \lambda e^{-\lambda m} \tag{3.14}$$

onde β_0, β_1, β_2, λ são quatro parâmetros da aproximação.

Como visto anteriormente, a curva de taxa de juros $y(m)$ pode ser escrita em termos da curva *forward* por meio da equação integral

$$y(m) = \frac{1}{m} \int_0^m f(u)\, du \tag{3.15}$$

substituindo-se a curva *forward* $f(m)$ pela aproximação dada pela Equação 3.14, é possível obter o modelo de Nelson e Siegel para a curva de taxa de juros $y(m)$

$$y(m) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\lambda m}}{\lambda m} \right) + \beta_2 \left(\frac{1 - e^{-\lambda m}}{\lambda m} - e^{-\lambda m} \right) \tag{3.16}$$

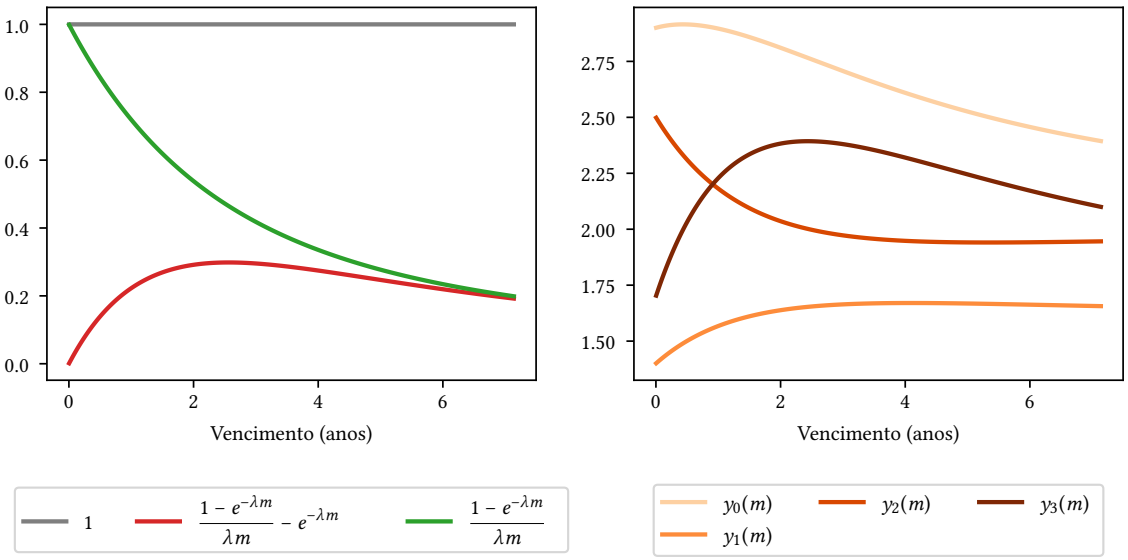

Figura 3.1: *Painel da esquerda: Funções de carga da decomposição de Nelson e Siegel com o parâmetro* $\lambda = 0.7$ *fixado. Painel da direita: Quatro curvas geradas pela decomposição do painel da esquerda,* y_1 $(\beta_0 = 2, \beta_1 = 1.12, \beta_2 = 0.9)$, y_2 $(\beta_0 = 1.6, \beta_1 = 0.5, \beta_2 = -0.2)$, y_3 $(\beta_0 = 2.0, \beta_1 = -0.8, \beta_2 = -0.5)$ *e* y_4 $(\beta_0 = 1.6, \beta_1 = 2.5, \beta_2 = 0.1)$

O trabalho de Nelson e Siegel (1987) mostrou que esta decomposição é capaz de ajustar adequadamente uma variedade conhecida de formatos de curvas de taxas de juros, como visto na Figura 3.1. Além disso, os parâmetros do modelo de Nelson e Siegel produzem interpretações economicamente relevantes.

- A função de carga do fator β_1, que decai exponencialmente com o vencimento da curva, permite interpretar β_1 como um fator que impacta os prazos mais curtos

- Analogamente, a função de carga do fator β_2, que começa em zero, atinge um pico e decai novamente para zero, permite interpretar β_2 como um fator de médio prazo
- β_0 pode ser interpretado como um fator de longo prazo, visto que a carga de β_0 é uma função constante que não decai com os vencimentos e dado o comportamento de decaimento exponencial das outras funções de carga

O parâmetro λ, também chamado de termo de decaimento exponencial, controla a taxa de decaimento da função de carga do fator de curto prazo e a localização do pico da função de carga do fator de médio prazo.

Apesar do número pequeno de parâmetros e da aparente similaridade entre o modelo de Nelson e Siegel e um modelo linear, a estimação dos parâmetros do modelo pode ser desafiadora devido a presença de múltiplos ótimos locais na função de erro quadrático médio do modelo. Um exemplo deste comportamento é ilustrado na Figura 3.2, que apresenta três ajustes do modelo de Nelson e Siegel para a curva de taxa de juros brasileira do dia 30 de agosto de 2007, obtidos através do algoritmo de Broyden–Fletcher–Goldfarb–Shanno (BFGS) com três pontos iniciais distintos.

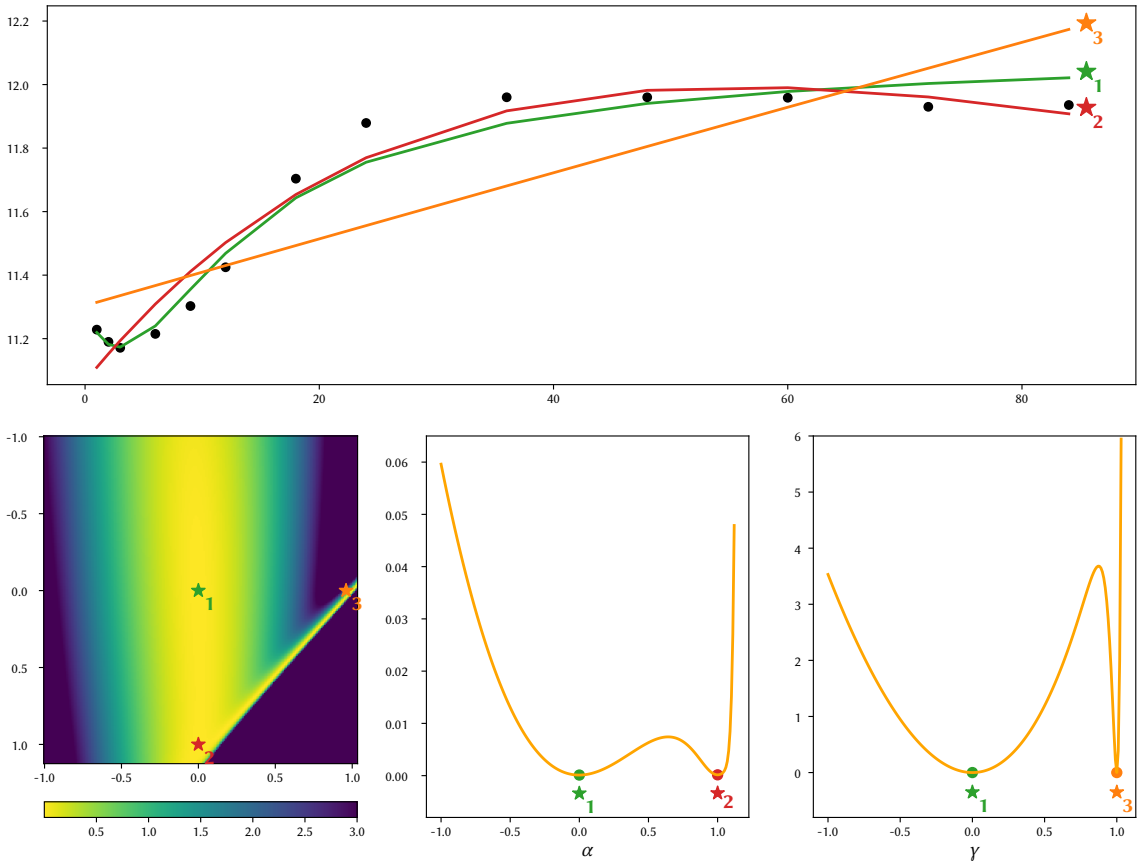

Figura 3.2: *Painel superior: Curva de taxa de juros brasileira em 30 de Agosto de 2007 (pontos no gráfico) e três ajustes do modelo de Nelson e Siegel* $\star_1$ ($\hat{\beta}_0 = 0.61, \hat{\beta}_1 = -0.04, \hat{\beta}_2 = -0.08, \hat{\lambda} = 3.24$), $\star_2$ ($\hat{\beta}_0 = 0.50, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 0.24, \hat{\lambda} = 0.29$) *e* $\star_3$ ($\hat{\beta}_0 = 0.71, \hat{\beta}_1 = -0.15, \hat{\beta}_2 = 2.78, \hat{\lambda} = 0.01$) *obtidos pelo algoritmo Broyden–Fletcher–Goldfarb–Shanno (BFGS). Painel inferior: Superfície de erro quadrático médio de* $EQM(\star_1 + \alpha(\star_2 - \star_1) + \gamma(\star_3 - \star_1))$, $EQM(\star_1 + \alpha(\star_2 - \star_1))$ *e* $EQM(\star_1 + \alpha(\star_3 - \star_1))$.

Da Figura 3.2, é possível verificar que as três estimativas obtidas para o vetor de parâmetros do modelo variam intensamente, especialmente na dimensão do parâmetro de decaimento exponencial λ. Este tipo de fenômeno pode prejudicar análises ou modelos que utilizem os fatores extraídos da decomposição de Nelson e Siegel de maneira longitudinal, devido a alta instabilidade nas séries temporais obtidas.

Uma solução sugerida em Nelson e Siegel (1987) para este problema é fixar o valor do parâmetro de decaimento exponencial λ, o que torna o modelo de Nelson e Siegel linear nos parâmetros β_0, β_1 e β_2 e garante solução única para o problema de estimação, além de possibilitar estimar os fatores do modelo eficientemente pelo método de mínimos quadrados. Para obter o valor λ, Nelson e Siegel (1987) sugerem um procedimento de busca exaustiva num conjunto finito Λ, que consiste em repetir a estimação dos fatores do modelo múltiplas vezes para cada elemento $\lambda \in \Lambda$, e escolher o valor que minimiza o erro quadrático entre a curva ajustada e os valores reais observados.

Modelo de Diebold e Li

O modelo Nelson e Siegel dinâmico, proposto por Diebold e Li (2006), sugere combinar a decomposição da curva de taxa de juros proposta por Nelson e Siegel (1987) à um modelo de séries temporais.

Com este objetivo, a evolução temporal das curvas de taxas de juros $\boldsymbol{y}_t = [y_t(m_1), ..., y_t(m_M)]^T$ em um determinado conjunto fixo de vencimentos $\{m_1, ..., m_M\}$ é modelada associando o vetor de parâmetros $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$ a um processo autoregressivo

$$\boldsymbol{\beta}_t = \boldsymbol{F}\boldsymbol{\beta}_{t-1} + \boldsymbol{w}_t, \quad w_t \overset{iid}{\sim} N_3(0, \boldsymbol{P}) \tag{3.17}$$

$$\boldsymbol{y}_t = \boldsymbol{H}_\lambda \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{iid}{\sim} N_M(0, \boldsymbol{Q}) \tag{3.18}$$

onde $\boldsymbol{P}$ e $\boldsymbol{F}$ são matrizes $3x3$, $\boldsymbol{Q}$ é uma matriz diagonal MxM, $\boldsymbol{Q}$ e $\boldsymbol{P}$ são matrizes semi positivas definidas e a matriz $\boldsymbol{H}_\lambda$ é dada por:

$$H_\lambda = \begin{bmatrix} 1 & \frac{1-e^{-\lambda m_1}}{\lambda m_1} & \left(\frac{1-e^{-\lambda m_1}}{\lambda m_1} - e^{-\lambda m_1}\right) \\ ... & ... & ... \\ 1 & \frac{1-e^{-\lambda m_M}}{\lambda m_M} & \left(\frac{1-e^{-\lambda m_M}}{\lambda m_M} - e^{-\lambda m_M}\right) \end{bmatrix} \tag{3.19}$$

O modelo Nelson e Siegel dinâmico pode ser visto como um modelo linear de espaço de estados Gaussiano, onde a Equação 3.17 descreve a evolução dos estados $\boldsymbol{\beta}_t$ e a Equação 3.18 descreve o mecanismo de medição do sistema.

Estimação em Duas Etapas

No artigo original de Diebold e Li (2006), os autores sugerem fixar $\boldsymbol{P} = diag(\sigma_1^2,\ \sigma_2^2,\ \sigma_3^3)$ e $\boldsymbol{F} = diag(f_1,\ f_2,\ f_3)$, reduzindo o processo estocástico da Equação 3.17 à três processos independentes do tipo AR(1). Apesar destas simplificações usualmente produzirem resultados melhores para a ETTJ americana, alguns trabalhos mostram que um modelo VAR(1) completo obtém resultados superiores no caso Brasileiro (Anisimov (2020), Caldeira *et al.* (2016a)).

Para estimar os fatores do modelo, os autores sugerem um procedimento de estimação simplificado baseado em duas etapas: extração dos fatores da decomposição da Nelson e Siegel e estimação dos parâmetros temporais do modelo.

Na etapa de extração dos fatores da decomposição, os vetores de fatores $\{\boldsymbol{\beta}_t\}_{t=1,...,T}$ são estimados como se fossem coeficientes de T modelos lineares distintos de cada instante de tempo $t = 1, ..., T$. Como visto anteriormente, ao assumir que o parâmetro de decaimento exponencial λ é fixo e conhecido, os fatores $\boldsymbol{\beta}_t$ podem ser obtidos pelo método de mínimos quadrados ordinários:

$$\hat{\boldsymbol{\beta}}_t = (\boldsymbol{H}_\lambda^T \boldsymbol{H}_\lambda)^{-1} \boldsymbol{H}_\lambda^T \boldsymbol{y}_t, \quad \text{para } t = 1, ..., T \tag{3.20}$$

A partir das estimativas obtidas, os parâmetros da componente temporal do modelo (Equação 3.17) são estimados seguindo o procedimento usual de estimação de modelos autoregressivos vetorias (VAR).

O parâmetro λ, tratado como fixo, pode ser obtido adaptando-se o procedimento de busca exaustiva sugerido em Nelson e Siegel (1987) para o conjunto completo de curvas consideradas:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \sum_{t=1}^{T} \|\boldsymbol{y}_t - \boldsymbol{H}_\lambda \hat{\boldsymbol{\beta}}_t\|_2^2 \tag{3.21}$$

onde $\hat{\boldsymbol{\beta}}_t$ é obtido a partir da Equação 3.20

Alternativamente, Diebold e Li (2006) propõem obter o valor de λ que maximiza função de carga do fator β_2 calculada no vencimento de médio amostral

$$\hat{\lambda} = \arg\min \left(\frac{1 - e^{-\lambda \overline{m}}}{\lambda \overline{m}} - e^{-\lambda \overline{m}} \right) \tag{3.22}$$

onde $\overline{m} = \frac{1}{M} \sum_{i=1}^{M} m_i$.

Estimação em Uma Etapa

Apesar da simplicidade do procedimento de estimação de duas etapas, esta abordagem sofre de algumas desvantagens. A mais evidente delas é que os fatores da decomposição são obtidos de maneira sub-ótima ao ignorar a dimensão temporal do problema. Além disso, a

incerteza do procedimento de estimação não é propagada, o que pode ser problemático caso seja de interesse obter intervalos de previsão.

Uma alternativa mais natural e direta para o procedimento de estimação do modelo Nelson e Siegel dinâmico consiste em obter a verossimilhança marginal exata a partir do filtro de Kalman, por meio do procedimento de decomposição do erro da previsão. O procedimento de estimação em uma etapa é descrito detalhadamente na Seção 4.1.4 para o caso geral de modelos lineares de espaço de estados Gaussianos.

Modelo de Nelson, Siegel e Svensson Dinâmico

Uma extensão da decomposição de Nelson e Siegel, proposta por Svensson (1994), inclui um quarto fator e um segundo parâmetro de decaimento exponencial no modelo:

$$y_t(m) = \beta_0 + \beta_1 \frac{1 - e^{-\lambda_1 m}}{\lambda_1 m} + \beta_2 \left(\frac{1 - e^{-\lambda_1 m}}{\lambda_1 m} - e^{-\lambda_1 m} \right) + \beta_3 \left(\frac{1 - e^{-\lambda_2 m}}{\lambda_2 m} - e^{-\lambda_2 m} \right) + \epsilon_t(m) \quad (3.23)$$

onde β_3 é um novo fator de médio prazo e λ_2 é um segundo parâmetro de decaimento exponencial. Svensson (1994) argumenta que a inclusão dos dois novos parâmetros melhora substancialmente o ajuste das curvas de taxas de juros diárias. O ganho de complexidade proporcionado por esta extensão pode ser visualizado a partir de uma análise comparativa entre curvas geradas pelo modelo de Svensson (1994) (Figura 3.3) e o modelo de Nelson e Siegel (1987) (Figura 3.1).

Dadas as semelhanças diretas com o modelo de Nelson e Siegel, uma versão dinâmica da extensão de Svensson (1994) também pode ser facilmente construída seguindo-se os passos descritos na seção anterior.

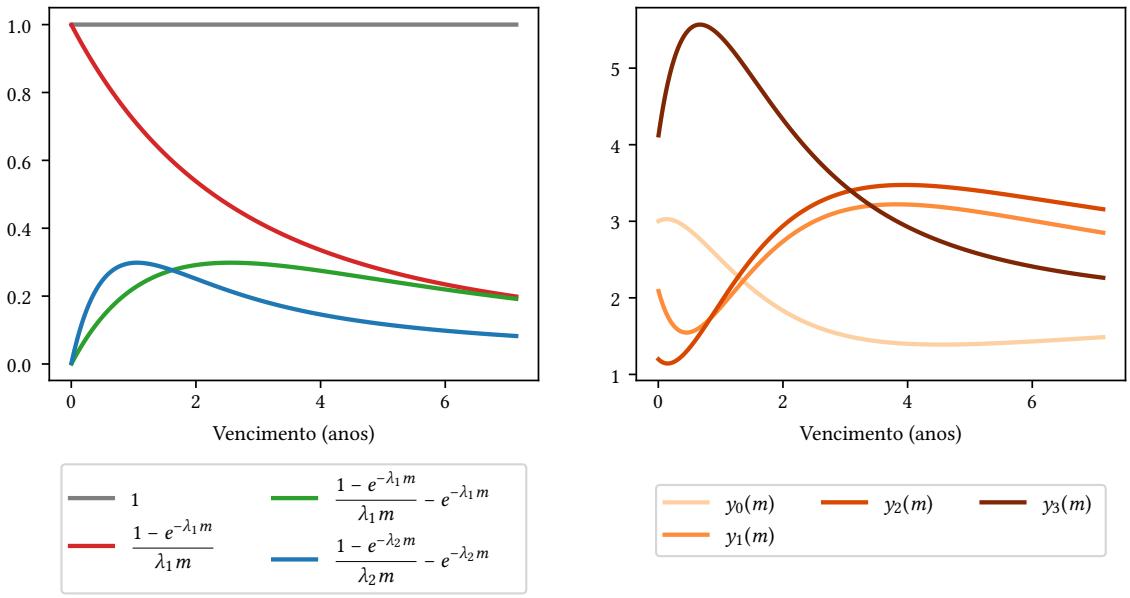

Figura 3.3: *Painel da esquerda: Funções de carga da decomposição de Nelson, Siegel e Svensson com os parâmetros* $\lambda_1 = 0.7$ *e* $\lambda_2 = 1.7$ *fixados. Painel da direita: Quatro curvas geradas pela decomposição do painel da esquerda,* y_1 $(\beta_0 = 2, \beta_1 = 1, \beta_2 = -5, \beta_3 = 3.)$*,* y_2 $(\beta_0 = 1.6, \beta_1 = 0.5, \beta_2 = 9, \beta_3 = -7.)$*,* y_3 $(\beta_0 = 2.0, \beta_1 = -0.8, \beta_2 = 9, \beta_3 = -5)$ *e* y_4 $(\beta_0 = 1.6, \beta_1 = 2.5, \beta_2 = -3, \beta_3 = 9)$

Capítulo 4

Aprendizado de Decomposições da ETTJ com Redes Neurais

4.1 Conceitos Teóricos

Neste sub-capítulo, são introduzidos os conceitos teóricos fundamentais para o principal objeto de estudo deste trabalho, o modelo linear-neural de espaço de estados para previsão da ETTJ.

Alguns dos principais objetos de estudo e conceitos da estatística Bayesiana são apresentados superficialmente na seção 4.1.1. O modelo *multilayer perceptron* clássico e sua versão Bayesiana são introduzidos na seção 4.1.2, com alguns comentários. Por fim, o modelo linear de espaço de estados Gaussiano e alguns resultados principais são descritos na seção 4.1.4.

4.1.1 Estatística Bayesiana

A estatística Bayesiana é uma metodologia para descrever incertezas associadas à um sistema por meio do teorema de Bayes, sob uma interpretação subjetiva do conceito de probabilidade. Ao contrário do paradigma frequentista, que define a probabilidade de um evento como o limite de sua frequência relativa em múltiplos ensaios, a interpretação Bayesiana define probabilidade como uma medida subjetiva dos graus de crença (ou incerteza) associados a ocorrência de um evento por um observador.

Seja X uma variável aleatória observável parametrizada pelo vetor de parâmetros θ, cuja função de densidade de probabilidade para um conjunto de parâmetros é representada pela forma funcional $f_{X|\theta}(x) = p(x|\theta)$. Usualmente, denota-se uma amostra da variável aleatória X usando-se símbolo x, portanto $p(x|\theta)$ também pode ser usado para se referir à verossimilhança referente à amostra x.

Diferentemente da estatística frequentista, a metodologia Bayesiana trata o vetor de parâmetros de interesse θ como uma variável aleatória latente do sistema, cuja função de densidade de probabilidade é representada pela forma funcional $p(\theta)$. A distribuição de

θ, conhecida como distribuição *priori* do sistema, representa a incerteza do observador acerca de θ antes de observar qualquer realização de X.

A partir do teorema de Bayes, é possível quantificar o conhecimento do observador acerca de θ após observar x por meio da distribuição *posteriori*, representada por $f_{\theta|X=x}(\theta) = p(\theta|x)$ e dada por

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int_{\theta \in S_\theta} p(\theta)p(x|\theta)d\theta} \tag{4.1}$$

partindo do pressuposto que θ é uma variável aleatória contínua de suporte S_θ. Como o denominador da expressão de $p(\theta|x)$ é constante em relação à θ, a equação anterior pode ser expressada mais succintamente como

$$p(\theta|x) \propto p(\theta)p(x|\theta) \tag{4.2}$$

O teorema de Bayes excerce o papel fundamental de ligar o conhecimento inicial do observador frente à θ, descrito pela distribuição *priori*, ao conhecimento atualizado frente à uma evidência empírica, descrito pela distribuição *posteriori*. Assim sendo, a distribuição *posteriori* é considerada o objeto de estudo central da inferência Bayesiana e suas aplicações.

A especificação da distribuição *priori* é de suma importância por permitir a inclusão de conhecimento de domínio do observador de maneira sólida do ponto de vista teórico. Parte da extensa literatura da estatística Bayesiana se concentra em estudar escolhas de distribuições *priori* para uma grande diversidade de modelos estatísticos, tornando mais simples o trabalho de inclusão de conhecimento de domínio na forma de distribuições *priori* do praticante da modelagem Bayesiana.

A estimação de parâmetros de modelos Bayesianos é realizada de maneira pontual ou completa. Usualmente se dá o nome de estimação completa o processo de obtenção da distribuição *posteriori* ou de algum tipo de aproximação para a mesma. A obtenção de uma estimativa fixa para o vetor de parâmetros θ é chamada de estimação pontual.

Algumas das principais metodologias de obtenção de estimativas pontuais de modelos Bayesianos são baseadas em medidas de tendência central da distribuição *posteriori* $\theta|(X = x)$. Alguns dos exemplos mais notáveis são

- Moda da distribuição *posteriori*, também conhecido como método de estimação *maximum a posteriori* (MAP)

$$\theta^*_{MAP} = \underset{\theta \in S_\theta}{\operatorname{argmax}}\, p(\theta|x) \tag{4.3}$$

- Valor esperado da distribuição *posteriori*

$$\theta^*_{EP} = \mathbb{E}(\theta|X = x) \tag{4.4}$$

Um problema comum enfretado na estatística Bayesiana origina-se da possibilidade da distribuição *posteriori* não ser conhecida, devido à complexidade da chamada constante de normalização $\int_{\theta \in S_\theta} p(\theta)p(x|\theta)d\theta$, presente no denominador do teorema de Bayes. Tal constante frequentemente não possui forma analítica conhecida ou não pode ser expressa com funções elementares, o que impossibilita o procedimento de estimação completa exata e de estimação baseada em alguns estimadores pontuais, como o estimador da esperança da distribuição *posteriori* (Equação 4.4). Porém, em muitos casos, este problema pode ser contornado por meio de métodos de inferência Bayesiana aproximada.

O estimador MAP (Equação 4.3), equivalente à moda da distribuição posteriori, é um dos estimadores pontuais computacionalmente mais eficientes, visto que pode ser obtido pela solução de um problema de otimização e portanto independe do valor da constante de normalização, o que permite simplificar a função objetivo do problema de maximização para o produto $p(\theta)p(x|\theta)$.

4.1.2 Modelo *Multilayer Perceptron*

Redes neurais são uma classe ampla de modelos não-lineares de aprendizado de máquina que possuem uma gama extensa de aplicações em diversas áreas do conhecimento. Uma subclasse bastante estudada dos modelos de redes neurais são os modelos *multilayer perceptron* (MLP), que possuem a propriedade de aproximação universal em conjuntos compactos (Cybenko (1989)), assim como funções de base radial (RBF) e polinômios aproximadores. Uma vantagem dos modelos de redes neurais são sua natureza hierarquica, que permite a geração de representações em multiplos níveis de granularidade e complexidade.

Um modelo *multilayer perceptron* $f : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ de K camadas, aplicado à um vetor de observações $x \in \mathbb{R}^n$ pode ser definido pelas equações

$$f(x) = h_K \tag{4.5}$$

$$h_i = \begin{cases} x, & i = 0 \\ g_i\left(W_{i-1}h_{i-1} + b_{i-1}\right), & i = 1, ..., K \end{cases} \tag{4.6}$$

onde W_{i-1} é uma matriz de parâmetros (também conhecida como 'pesos' da camada i) e b_{i-1} é um vetor de parâmetros (chamados de 'interceptos' da camada i) de mesma dimensão de $W_{i-1}h_{i-1}$. As representações intermediárias h_i calculadas para o vetor de entrada x são chamadas de unidades ocultas da rede neural.

As funções $g_i : \mathbb{R} \longrightarrow \mathbb{R}$ são chamadas de funções de ativação e são aplicadas ao final de cada camada da rede neural. Todas as funções de ativação da rede neural são funções contínuas, semi-diferenciáveis e usualmente não-lineares. Além disso, a maioria das funções de ativação são aplicadas elemento à elemento no vetor $W_{i-1}h_{i-1} + b_{i-1}$.

Portanto, o modelo *multilayer perceptron* pode ser entendido como uma sequência de transformações lineares intercaladas por funções de ativação que são usualmente não-lineares. Na Figura 4.1, um diagrama para uma rede neural *multilayer perceptron* $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ com três camadas é representada. Para o exemplo, os parâmetros estimáveis

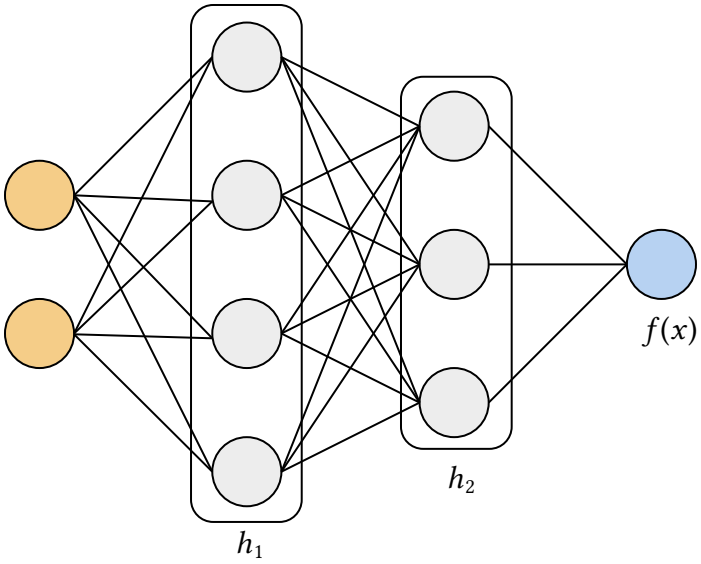

Figura 4.1: *Diagrama de um modelo Multilayer Perceptron com $K = 2$ camadas, dimensão de entrada $n = 2$ e dimensão de saída $m = 1$ e tamanhos de camada $N_1 = 4$ e $N_2 = 3$.*

da rede neural são os elementos das matrizes $\{W_0, W_1, W_2\}$ de dimensões 4x2, 3x4 e 1x3 e os elementos dos vetores $\{b_0, b_1, b_2\}$ de dimensões 4, 3 e 1.

A função de ativação g_i tem papel importante na rede neural resultante e sua escolha pode depender da natureza do problema. Nas camadas intermediárias de uma rede neural MLP, as funções de ativação usualmente cumprem o papel de quebrar a linearidade das transformações consecutivas, fazendo assim que o modelo *multilayer perceptron* seja um modelo não linear em relação ao vetor de entradas x. Na última camada da rede neural, a função de ativação final g_K tem também o papel de mapear a última unidade oculta h_{K-1} para o domínio de saída de interesse.

Algumas das funções de ativação clássicas da literatura para modelos MLP são: a função sigmóide (ou logística) $\sigma : \mathbb{R} \longrightarrow [0, 1]$, cuja função inversa é conhecida como a função de ligação *logit*, a função tangente hiperbólica $tanh : \mathbb{R} \longrightarrow [0, 1]$ e a função unidade linear retificada $ReLU : \mathbb{R} \longrightarrow \mathbb{R}^+$.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4.7}$$

$$ReLU(x) = max(0, x)$$

Diferentemente de outros modelos de aprendizado de máquina baseados em métodos de *kernel*, como *support vector machines* (SVM) e regressão via processos Gaussianos (GP), a complexidade computacional de prever uma nova observação com uma rede neural independe do tamanho da amostra utilizada na fase de treinamento, o que torna atrativo o

uso dessa classe de modelos em grandes conjuntos de dados.

Além disso, a arquitetura de uma rede neural *multilayer perceptron* (MLP) é extremamente compatível com algoritmos de computação paralela, o que permite diminuir considerávelmente o tempo efetivo de execução e treinamento do modelo com o uso de placas de processamento gráfico (GPUs).

Na estimação de modelos *(multilayer perceptron)* para problemas de regressão ($y_i \in \mathbb{R}^k$) é usual estimar os parâmetros do modelo θ a partir do método de mínimos quadrados. Para uma amostra de observações $y_1, ..., y_n$, a perda quadrática é dada por

$$L(\theta) = \sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 \tag{4.8}$$

onde $\hat{y}_i = f_\theta(x_i)$. De maneira equivalente, podemos associar as observações $y_1, ..., y_n$ a um modelo probabilístico gerador dos dados

$$y_i \overset{iid}{\sim} N_k(f_\theta(x_i), \sigma_\epsilon^2 I_k), \quad i = 1, ..., n \tag{4.9}$$

É fácil verificar que minimizar a função de perda quadrática é equivalente à maximizar a log-verossimilhança do processo gerador descrito, se tratarmos o parâmetro σ_ϵ^2 como fixo e conhecido.

$$p(y_1, ..., y_n|\theta, \sigma_\epsilon^2) = (2\pi)^{-\frac{kn}{2}} \det(I_k\sigma_\epsilon^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \exp\left[-\frac{1}{2}(y_i - \hat{y}_i)^T(I_k\sigma_\epsilon^2)^{-1}(y_i - \hat{y}_i)\right] \tag{4.10}$$

$$\log p(y_1, ..., y_n|\theta, \sigma_\epsilon^2) = -\frac{kn}{2}\log(2\pi\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(y_i - \hat{y})^T(y_i - \hat{y}) \tag{4.11}$$

$$\log p(y_1, ..., y_n|\theta, \sigma_\epsilon^2) \propto -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\|y_i - \hat{y}_i\|_2^2 \tag{4.12}$$

Os parâmetros do modelo podem ser estimados maximizando-se a verossimilhança da amostra $y_1, ..., y_n$ a partir de métodos de otimização de primeira ordem. Os gradientes $\nabla_\theta f(x)$ do vetor de parâmetros θ do modelo podem ser obtidos por meio de algoritmos de diferenciação automática (Rall (1986) e Wengert (1964)), que permitem calcular os gradientes de maneira eficiente para qualquer arquitetura diferencíavel de modelo.

Devido ao número elevado e alta simetria dos parâmetros estimáveis, há uma preocupação natural de que os algoritmos de otimização empregados fiquem presos em máximos locais da verossimilhança de performance subótima. Essa preocupação é legítima em regimes onde o número de parâmetros é baixo em relação ao tamanho da amostra, mas como alguns trabalhos observam (Allen-Zhu *et al.* (2019a), Safran e Shamir (2018), Oymak e Soltanolkotabi (2020) e Allen-Zhu *et al.* (2019b)), este problema desaparece em um regime de super-parametrização, onde a convergência para soluções ótimas é rápida e estável, mesmo empregrando-se apenas métodos de otimização de primeira ordem.

4.1.3 Redes Neurais Bayesianas

Variações bayesianas para os modelos de redes neurais foram primeiro estudadas pelos trabalhos de Buntine e Weigend (1991), MacKay (1992) e Neal (1992). Segundo MacKay (1992), o ponto de vista Bayesiano pode ser benéfico para os modelos de redes neurais pois trata o problema de estimação a partir do ponto de vista de distribuições de probabilidade definidas sobre o espaço paramétrico, o que permitira estudar as incertezas associadas ao modelo de maneira mais geral e direta.

MacKay (1992) também argumenta que o uso de distribuições priori nos modelos formaliza e facilita a utilização de mecanismos de regularização, o que previne o problema de sobreajuste dos dados de treinamento, que ocorrem quando o modelo memoriza os dados de treinamento e perde capacidade de generalização em novos contextos.

Uma versão Bayesiana para o modelo discutido na seção anterior (Eq. 4.9) pode ser construído associando distribuições priori ao vetor de parâmetros da rede neural θ e a variância do termo de ruído σ_ϵ^2. Como discutido na Seção 4.1.1, o log da função densidade de probabilidade posteriori do modelo para um conjunto de observações $(x_1, y_1), ..., (x_n, y_n)$ é dado por

$$\log p(\theta, \sigma_\epsilon^2 | y_1, ..., y_n) \propto \log p(\theta, \sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \| y_i - f_\theta(x_i) \|_2^2 \tag{4.13}$$

No contexto de redes neurais bayesianas, é comum assumir que $p(\theta, \sigma_\epsilon^2) = p(\theta)\, p(\sigma_\epsilon^2)$. Uma extensa variedade de distribuições priori para o vetor de parâmetros θ foram estudadas na literatura de Redes Neurais Bayesianas. Em particular, a distribuição normal isotrópica se destaca como uma alternativa simples e efetiva para atenuar o fenômeno de sobreajuste, além de possuir conexões com trabalhos amplamente explorados como a regressão *ridge* (Hoerl e Kennard (1970)), *weight decay* (Plaut (1986)) e a teoria de regularizadores de Tikhonov (Tikhonov (1943), Tikhonov (1963)).

Sob a priori normal isotrópica, a distribuição de probabilidade priori do vetor de parâmetros θ é dada por

$$\theta \sim N(0, \sigma_{NN}^2 I), \quad \sigma_{NN}^2 > 0 \tag{4.14}$$

$$\log p(\theta) \propto -\frac{1}{\sigma_{NN}^2} \| \theta \|_2^2 \tag{4.15}$$

que alternativamente pode ser vista como um termo de penalidade L_2 (ou penalidade *ridge*, Hoerl e Kennard (1970)) para a log-densidade da distribuição posteriori $p(\theta, \sigma_\epsilon^2 | y_1, ..., y_n)$ resultante, onde o hiperparâmetro σ_{NN}^2 controla a força do efeito de regularização nos parâmetros do modelo. Um exemplo do efeito de regularização em modelos *multilayer perceptron* é ilustrado na Figura 4.2.

Alguns trabalhos estudam outras possíveis escolhas de distribuição priori para redes neurais Bayesianas, como a distribuição Laplace (Williams (1995)), distribuição *horseshoe*

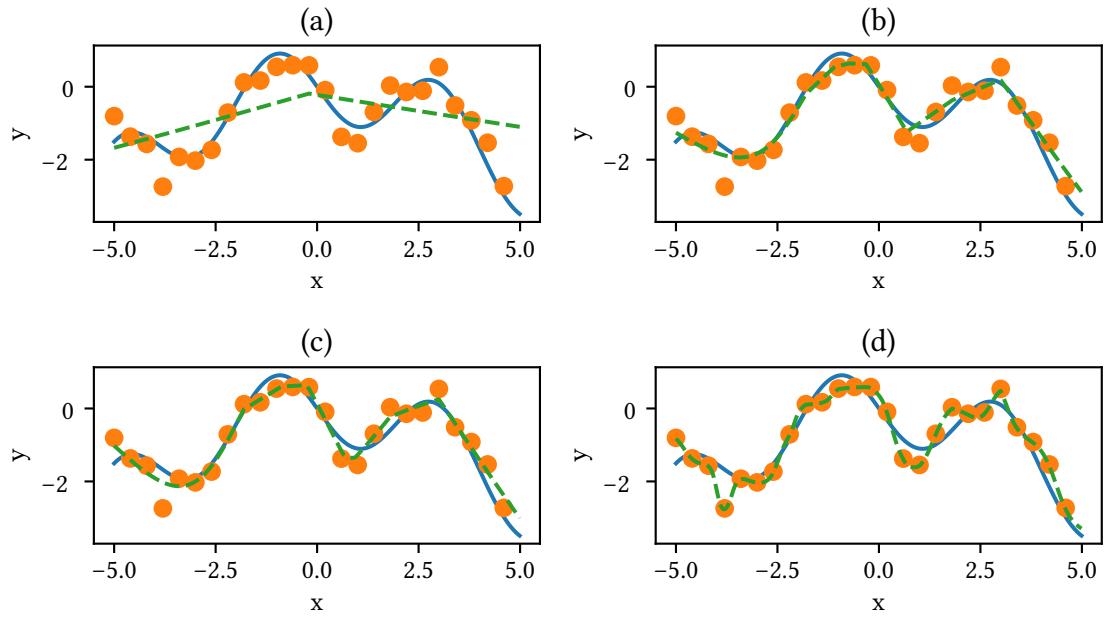

Figura 4.2: *Aproximações da função* $f(x) = -\frac{1}{10}x^2 - sen(5x/\pi)$ *(linha sólida azul nos gráficos) a partir de uma amostra de* 25 *pontos acrescidos de ruído* $N(0, \sqrt{0.4})$ *(pontos em laranja) com modelos de redes neurais Multilayer Perceptron de duas camadas (linha pontilhada, verde) estimados na amostra com diferentes valores para* $\tau = \sigma_{NN}^{-2}$*: (a)* $\tau = 10^{-1}$*, (b)* $\tau = 10^{-2}$*, (c)* $\tau = 5 \cdot 10^{-3}$ *e (d)* $\tau = 0$*.*

(Carvalho *et al.* (2009)) e log-uniforme (Kingma *et al.* (2015)) que induzem esparsidade no vetor de parâmetros.

Como estudado em Neal (1996), outra propriedade importante das redes neurais Bayesianas de uma camada oculta, sob distribuições priori independentes e identicamente distribuídas de variância finita, convergem para um processo gaussiano no limite em que a dimensão da camada oculta tende ao infinito, devido ao teorema do limite central. Outros trabalhos (Lee *et al.* (2017), Matthews *et al.* (2018)) generalizam esse resultado para redes com profundidade arbitrária sob outras restrições, e encontram *kernels* aproximadamente equivalentes a modelos de redes neurais profundas.

Dado o carácter não-linear e super-parametrizado dos modelos de redes neurais, a obtenção, mesmo que aproximada, da distribuição posteriori para esta classe de modelos pode ser uma tarefa extremamente desafiadora. Métodos tradicionais de inferência baseados em *Markov Chain Monte Carlo* (MCMC) são pouco efetivos para esta classe de modelos devido ao problema de alta dimensionalidade e baixa identificabilidade dos parâmetros do modelo. Além disso, aproximações baseadas em processos Gaussianos com funções de *kernel* equivalente podem ser pouco efetivas computacionalmente quando o tamanho do conjunto de dados é elevado.

4.1.4 Modelo Linear Gaussiano de Espaço de Estados

A classe de modelos lineares Gaussianos de espaço de estados, como vista em Durbin e Koopman (2012), pode ser escrita como

$$\boldsymbol{\beta}_t = \boldsymbol{F}\boldsymbol{\beta}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim N_k(0, \boldsymbol{P}), \quad t = 1, 2, ... \tag{4.16}$$

$$\boldsymbol{x}_t = \boldsymbol{H}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N_m(0, \boldsymbol{Q}) \tag{4.17}$$

onde $\boldsymbol{x}_t$ é um vetor observável de dimensão m x 1 e $\boldsymbol{\beta}_t$ é um vetor latente de dimensão k x 1, chamado de *vetor de estados* no instante t. Além disso, assume-se também que $\boldsymbol{\beta}_0 \sim N(\boldsymbol{\mu}_0, \Sigma_0)$.

O modelo de espaço de estados descreve um sistema dinâmico onde cada quantidade observável $\boldsymbol{x}_t$ depende exclusivamente de um vetor latente de estados $\boldsymbol{\beta}_t$ que governa a evolução temporal do sistema. O ato de realizar uma medição no sistema é descrito por meio da Equação 4.17, chamada de *equação de medição* do sistema, que transforma linearmente o vetor de estados $\boldsymbol{\beta}_t$ usando a matriz $\boldsymbol{H}$ de dimensão m x k e o vetor de ruído de observação $\boldsymbol{\epsilon}_t$. De maneira complementar, a equação 4.16 que descreve a dinâmica do sistema por meio de um modelo autoregressivo vetorial (3.2) é chamada de *equação de evolução* do sistema.

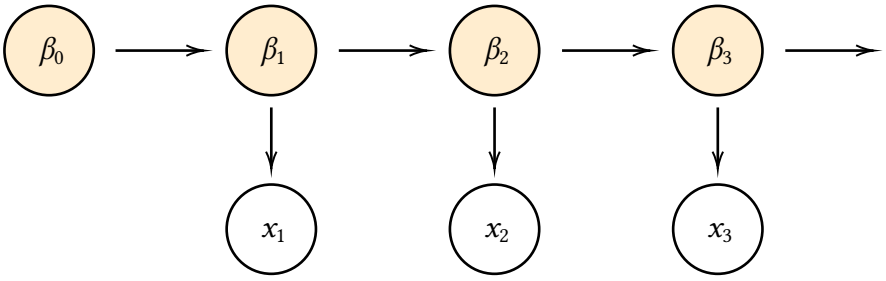

Figura 4.3: *Diagrama de um modelo de espaço de estados de variáveis latentes* $\{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, ...\}$ *(em laranja) e observações* $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...\}$ *(em branco).*

Modelos lineares Gaussianos de espaços de estados são amplamente utilizados em diversas áreas do conhecimento, como por exemplo em Engenharia (Chan *et al.* (1979), Weng *et al.* (2006)), Economia (Athans (1974), Pasricha (2006)) e Metereologia (Delle Monache *et al.* (2011)). Em muitas aplicações, um dos principais interesses é obter a distribuição do vetor de estados dado um conjunto de observações $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Para este objetivo, tipicamente, utilizam-se procedimento de *filtração* ou *suavização*.

Os procedimentos de filtração buscam obter as distribuições dos estados do sistema condicionados nas observações disponíveis até o tempo presente t, $p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t)$, enquanto os procedimentos de suavização tem como objetivo calcular as distribuições dos estados condicionados a todas as n observações disponíveis $p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$.

Decorrente do fato das distribuições de $\boldsymbol{\beta}_0$, $\boldsymbol{\epsilon}_t$ e $\boldsymbol{w}_t$ serem normais multivariadas e da linearidade do sistema dinâmico, a distribuição conjunta de todos os estados e medidas $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_n)$ também é normal multivariada, o que garante que a distribuição condicional de qualquer subconjunto de variáveis da distribuição completa também será normal multivariada. A partir deste resultado, o procedimento de filtração iterativo conhecido como Filtro de Kalman pode ser usado para obter a distribuição exata de $(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t)$.

Filtro de Kalman

O Filtro de Kalman é um algoritmo iterativo para obter $p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t)$ que depende da execução de duas etapas para todo $t = 0, 1, ..., n$

- **Atualização temporal**: Obter $p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1})$ usando $p(\boldsymbol{\beta}_{t-1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1})$
- **Etapa de medição**: Obter $p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t)$ usando $p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1})$

Atualização Temporal

Partindo de uma distribuição conhecida $(\boldsymbol{\beta}_{t-1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}) \sim N_k(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{P}_{t-1|t-1})$ e decorrente da equação de evolução de estados $\boldsymbol{\beta}_t = \boldsymbol{F}\boldsymbol{\beta}_{t-1} + \boldsymbol{w}_{t-1}$, é fácil verificar pelas propriedades de linearidade da distribuição normal multivariada que

$$\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1} \sim N_k(\boldsymbol{F}\boldsymbol{\mu}_{t-1|t-1},\ \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}^T + \boldsymbol{P}) \tag{4.18}$$

ou mais resumidamente

$$\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1} \sim N_k(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{P}_{t|t-1}) \tag{4.19}$$

onde $\boldsymbol{\mu}_{t|t-1} = \boldsymbol{F}\boldsymbol{\mu}_{t-1|t-1}$ e $\boldsymbol{P}_{t|t-1} = \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}^T + \boldsymbol{P}$. O que conclui a etapa de atualização temporal.

Etapa de Medição

A partir da distribuição obtida na etapa anterior $(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1})$, podemos obter a distribuição conjunta de $(\boldsymbol{\beta}_t, \boldsymbol{x}_t)$ a partir da transformação linear dada pela equação de medição $\boldsymbol{x}_t = \boldsymbol{H}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t$ e usando novamente propriedades de linearidade da distribuição normal multivariada

$$\left(\begin{bmatrix}\boldsymbol{\beta}_t \\ \boldsymbol{x}_t\end{bmatrix} \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}\right) \sim N_{k+m}\left(\begin{bmatrix}\boldsymbol{\mu}_{t|t-1} \\ \boldsymbol{H}\boldsymbol{\mu}_{t|t-1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{t|t-1} & \boldsymbol{P}_{t|t-1}\boldsymbol{H}^T \\ \boldsymbol{H}\boldsymbol{P}_{t|t-1} & \boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}^T + \boldsymbol{Q}\end{bmatrix}\right) \tag{4.20}$$

Ao condicionar a distribuição conjunta de $(\boldsymbol{\beta}_t, \boldsymbol{x}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t)$ na variável $\boldsymbol{x}_t$ usando o lema da distribuição condicional de distribuições normais multivariadas, obtem-se que

$$(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t) \sim N_k(\boldsymbol{\mu}_{t|t}, \boldsymbol{P}_{t|t}) \tag{4.21}$$

onde

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}^T + \boldsymbol{Q})^{-1}(\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{\mu}_{t|t-1}) \tag{4.22}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}^T + \boldsymbol{Q})^{-1}\boldsymbol{H}\boldsymbol{P}_{t|t-1} \tag{4.23}$$

o que conclui a etapa de medição.

Procedimentos de filtração são particularmente úteis na estimação de parâmetros de um modelo linear Gaussiano de espaço de estados. Dado um sub-conjunto desconhecido de parâmetros do sistema $\boldsymbol{\theta} \subseteq (\boldsymbol{H}, \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{P})$, a verossimilhança marginal $p(\boldsymbol{x}_1, ..., \boldsymbol{x}_n|\boldsymbol{\theta})$ é dada por

$$\begin{aligned} p(\boldsymbol{x}_1, ..., \boldsymbol{x}_n|\boldsymbol{\theta}) &= \prod_{t=1}^{n} \int p(\boldsymbol{x}_t, \boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}, \boldsymbol{\theta}) d\boldsymbol{\beta}_t \\ &= \prod_{t=1}^{n} \int p(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}, \boldsymbol{\theta}) p(\boldsymbol{x}_t|\boldsymbol{\beta}_t, \boldsymbol{\theta}) d\boldsymbol{\beta}_t \end{aligned} \tag{4.24}$$

Repetindo-se as etapas de atualização temporal e medição para cada $t = 0, ..., n$, é possível computar as distribuições $p(\boldsymbol{x}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t-1}, \boldsymbol{\theta})$, necessárias para avaliar a verossimilhança marginal do sistema $p(\boldsymbol{x}_1, ..., \boldsymbol{x}_n|\boldsymbol{\theta})$ no problema de estimação dos parâmetros de um modelo linear Gaussiano de espaço de estados.

Além disso, a distribuição do último estado filtrado do sistema $(\boldsymbol{\beta}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{\theta})$ também pode ser usada para prever $\boldsymbol{x}_{t+1}$, por meio das equações de evolução e medição, no caso em que $\boldsymbol{\theta}$ é conhecido.

$$(\boldsymbol{x}_{t+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_t, \boldsymbol{\theta}) \sim N_m(\boldsymbol{H}\boldsymbol{F}\boldsymbol{\mu}_{t|t}, \boldsymbol{H}(\boldsymbol{F}\boldsymbol{P}_{t|t}\boldsymbol{F}^T + \boldsymbol{P})\boldsymbol{H}^T + \boldsymbol{Q}) \tag{4.25}$$

Para $\boldsymbol{\theta}$ desconhecido, a distribuição preditiva de $\boldsymbol{x}_{t+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_t$ é dada por

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_t) = \int p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_{t+1}, \boldsymbol{\theta})\ p(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_t) d\boldsymbol{\theta} \tag{4.26}$$

No caso em que a distribuição posteriori $p(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_t)$ é desconhecida, mas é possível obter amostras desta através de um procedimento de inferência aproximado, amostras para a distribuição preditiva também podem ser obtidas através da equação 4.25.

4.2 Modelo Linear-Neural de Espaço de Estados para previsão da ETTJ

4.2.1 Motivação e Revisão Bibliográfica

Como verificado em diversos trabalhos da literatura (Cajueiro *et al.* (2009), Caldeira *et al.* (2010), Caldeira *et al.* (2016b), Caldeira *et al.* (2016a)), os modelos tradicionais de previsão da ETTJ brasileira de melhor performance usualmente se baseiam nos conceitos de redução de dimensionalidade ou de decomposição das curvas das de taxas de juros. Em especial, os modelos Nelson e Siegel dinâmico (Diebold e Li (2006)) e Nelson, Siegel e Svensson dinâmico se destacam graças aos inúmeros resultados empíricos que indicam a superioridade desta classe de modelos em relação às abordagens tradicionais alternativas.

A partir de uma análise comparativa entre a classe de modelos Nelson e Siegel dinâmico (com estimação em uma etapa) e abordagens alternativas vistas no Capítulo 3, uma lista de possíveis motivos para o sucesso desta classe de modelos pode ser construída:

(i) Possibilidade de estimação conjunta dos componentes temporais e locais

(ii) Continuidade e suavidade das curvas preditas

(iii) Estabilidade assintótica das curvas preditas

A característica (i) garante que a decomposição ou redução de dimensionalidade seja coerente com a componente temporal especificada, e portanto tem alta importância para o problema de previsão de curvas, apesar de causar sucetibilidade ao problema de sobreajuste em alguns contextos. A característica (ii) pode ser vista como uma inclusão de conhecimento de domínio que garante que taxas de juros de vencimentos parecidos também devem ser parecidas. O item (iii) favorece formatos de curvas com taxas que se estabilizam em vencimentos mais longos, o que garante que as curvas geradas pelo modelo sejam mais plausíveis teoricamente.

No entanto, como visto no Capítulo 3, a classe de modelos Nelson e Siegel dinâmico depende de um conjunto de funções de carga pré-determinadas e fixadas (ao escolher o parâmetro λ), limitando a capacidade expressiva do modelo apenas a combinações lineares de funções deste conjunto. Com o objetivo de relaxar esta hipótese, novos modelos que permitem estimar decomposições dinamicamente a partir do conjunto de dados foram propostos nos últimos anos.

Algumas destas abordagens buscam melhorar o ajuste de decomposições de curvas de taxas de juros de maneira independente (desconsiderando a componente temporal do modelo), mas que depois podem ser adaptadas ao contexto de previsão da ETTJ usando-se uma abordagem de estimação baseada em duas etapas (como descrito no Capítulo 3). Faria e Almeida (2018), por exemplo, sugerem um modelo misto para a curva de taxa de juros usando um modelo paramétrico baseado na extensão de Svensson (1994) para a extremidade mais longa e B-Splines cúbicas para as extremidade mais curtas, Takada e Stern (2015) propõe a utilização de algoritmos de fatoração não-negativa de matrizes

e Mineo *et al.* (2020) sugerem uma decomposição livre de arbitragem com B-Splines restritas.

	Decomposição dinâmica	(i)	(ii)	(iii)
Diebold e Li (2006)		✓	✓	✓
Bowsher e Meeks (2008)		✓	✓	
Hays *et al.* (2012)	✓	✓	✓	
Takada e Stern (2015)	✓			
Faria e Almeida (2018)	✓		✓	✓
Suimon *et al.* (2020)	✓			
Mineo *et al.* (2020)	✓		✓	

Tabela 4.1: *Comparativo entre trabalhos revisados quanto às características: Decomposição dinâmica (estimada a partir de dados), Estimação conjunta dos componentes temporais e locais (i), Continuidade e suavidade (ii) e Estabilidade assintótica (iii)*

Outros autores propõe a utilizaçãode Splines cúbicas em um modelo de espaço de estados, que permite estimar as duas componentes do modelo em única etapa. Bowsher e Meeks (2008) sugerem o modelo *Functional Signal Plus Noise* que descreve a curva de taxa de juros com uma única Spline cúbica de k nós que interpola um conjunto de coordenadas latentes da curva, que são modeladas com um processo VAR(1). Hays *et al.* (2012) sugere a utilização de uma Spline cúbica para cada função de carga de um modelo de espaço de estados.

Na Tabela 4.1, os trabalhos revistos nessa seção são representados em relação à aderência às quatro propriedades listadas. É possível verificar que nenhum dos modelos propostos atingem todas as propriedades listadas simultaneamente, o que estimula naturalmente a pergunta da possibilidade de um modelo com todas estas propriedades atingir melhores resultados.

Como uma tentativa de preencher esta lacuna, na seção a seguir, é apresentada uma extensão do modelo linear Gaussiano de espaço de estados, com o objetivo de gerar novas decomposições a partir de dados e preservar as características descritas neste capítulo.

4.2.2 Modelo

Com o objetivo de obter novas decomposições para as curvas de taxas de juros em um modelo linear gaussiano de k fatores, uma abordagem intuitiva seria estimar toda a matriz de medição $\boldsymbol{H}$ de M x k parâmetros livres conjuntamente com os demais parâmetros do modelo ($\boldsymbol{F}$, $\boldsymbol{P}$, $\boldsymbol{Q}$).

No entanto, esta abordagem sofre de duas principais desvantagens. As previsões não podem ser calculadas para alguma maturidade não observada $m \notin \{m_1, ..., m_M\}$ sem um procedimento de interpolação ou extrapolação. Além disso, o modelo não induz que taxas

de juros com vencimentos parecidos também devem ser parecidas (como discutido em 4.2.1).

Com o objetivo de atenuar estes dois problemas, a seguinte extensão ao modelo linear de espaço de estados Gaussiano é proposta:

$$\boldsymbol{\beta}_t = \boldsymbol{F}\boldsymbol{\beta}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \overset{ind.}{\sim} N_k(\boldsymbol{0}, \boldsymbol{P}) \tag{4.27}$$

$$y_t(m) = \boldsymbol{g}_{\boldsymbol{\theta}}(m)^T \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t(m) \tag{4.28}$$

onde m é um vencimento, $\boldsymbol{\beta}_t$ é um vetor de k fatores do instante de tempo t e $\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{m}) : \mathbb{R} \longrightarrow \mathbb{R}^k$ é uma rede neural de vetor de parâmetros θ. A função $\boldsymbol{g}_{\boldsymbol{\theta}}$ pode ser vista como um gerador parametrizado de k funções de base do espaço de curvas de taxas de juros:

$$\boldsymbol{g}_{\boldsymbol{\theta}}(m) = \begin{bmatrix} \phi_\theta^{(0)}(m) \\ \phi_\theta^{(1)}(m) \\ \vdots \\ \phi_\theta^{(k-1)}(m) \end{bmatrix} \tag{4.29}$$

onde $\phi_\theta^{(0)}, ..., \phi_\theta^{(k-1)}$ são os termos da última camada da rede neural $\boldsymbol{g}_{\boldsymbol{\theta}}$.

Alternativamente, o modelo também pode ser interpretado como uma única rede neural $f_\theta(m) : \mathbb{R} \longrightarrow \mathbb{R}$ cuja última camada possui coeficientes variantes no tempo (segundo o processo estocástico de 4.27) e função de ativação identidade.

A estrutura da rede neural $\boldsymbol{g}_{\boldsymbol{\theta}}$ garante cargas continuas e diferenciáveis, permitindo avaliar a $y_t(m)$ em qualquer vencimento, assim como a decomposição de Nelson e Siegel. Entretanto, para simplificar os cálculos adiante, $\boldsymbol{g}_{\boldsymbol{\theta}}(m)$ é avaliada apenas no vetor de vencimentos fixos $[m_1, ..., m_M]^T$.

Com esta simplificação, a equação de medição (4.28) do modelo de espaços de estados pode ser então representada em termos de $\boldsymbol{y}_t = [y_t(m_1), ..., y_t(m_M)]^T$ e a matriz de medições $\boldsymbol{H}_\theta = [g_\theta(m_1), ..., g_\theta(m_M)]^T$, produzindo a familiar expressão da equação de medição:

$$\boldsymbol{y}_t = \boldsymbol{H}_\theta \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{ind.}{\sim} N_M(\boldsymbol{0}, \boldsymbol{Q}) \tag{4.30}$$

onde $\boldsymbol{Q}$ é uma matriz positiva definida.

Para garantir flexibilidade e boas propriedades do modelo, é proposta uma arquitetura para a rede neural g_θ composta por duas camadas escondidas, com funções de ativação tangente hiperbólica (4.7), que são suaves e possuem boas propriedades de convergência (LeCun *et al.*, 2012).

Para facilitar a interpretação dos fatores do modelo, a primeira função de carga de $\boldsymbol{g}_\theta$ é fixada como $\phi_\theta^{(0)}(m) = 1$. Além disso, para induzir representações similares ao modelo de Nelson e Siegel, uma função de ativação sigmóide é incluída na camada final da rede neural, limitando assim as funções de carga geradas no intervalo $[0, 1]$ e facilitando que

a rede neural seja capaz de recriar padrões de decaimento exponencial para 0 ou para 1, assim como a decomposição de Nelson e Siegel.

Com a flexibilidade que a inclusão do modelo de redes neurais proporciona, é possível obter decomposições contínuas para um número arbitrário de fatores desejados, ao mesmo tempo que o acoplamento com modelo de evolução de estados (4.27) garante que estas decomposições produzam fatores relevantes para a descrever o comportamento temporal. Entretanto, como os modelos de redes neurais são aproximadores universais de funções, restringir a complexidade do modelo proposto pode ter suma importância para garantir a suavidade das decomposições obtidas e prevenir o problema de sobreajuste.

Além disso, como Diebold e Rudebusch (2013) observam, induzir independência entre fatores na equação de evolução de estados, efetivamente reduzindo o modelo VAR(1) à múltiplos modelos AR(1) independentes, tem efeito benéfico nas previsões obtidas para a curva americana. Entretanto, no caso brasileiro, os trabalhos de Anisimov (2020) e Caldeira *et al.* (2016a) obtém resultados superiores utilizando um modelo VAR(1) sem estas simplificações. O que também sugere que a presença de um método de regularização calibrável para o modelo de evolução de estados pode ser benéfico em alguns casos.

Com estes objetivos, um mecanismo de regularização para o modelo proposto baseado na escolha de distribuições priori para ambas componentes do modelo é descrito a seguir na Seção 4.2.3. Em seguida, um procedimento de estimação em uma etapa é descrito na Seção 4.2.4.

4.2.3 Especificação da Priori

Modelo de Decomposição

Os modelos de redes neurais são usualmente suscetíveis ao problema de sobreajuste em de conjuntos de dados pequenos ou na ausência de mecanismos de regularização apropriados (Barron (1991)). Como visto anteriormente na seção 4.1.3, o *framework* Bayesiano pode ser aplicado de maneira bem sucedida no contexto de redes neurais, fornecendo uma variedade de técnicas de regularização baseados na escolha de prioris para os parâmetros do modelo.

Como discutido no capítulo 4.1.3, a distribuição normal isotrópica é uma escolha popular de priori para redes neurais Bayesianas com efeito de regularização. Para uma rede neural g_θ de $|\theta|$ parâmetros, a priori normal isotrópica é dada por

$$\theta \sim N(0, \sigma_{NN}^2 I), \sigma_{NN}^2 > 0 \tag{4.31}$$

$$\log p(\theta) = -\frac{|\theta|}{2} log(2\pi) - \frac{|\theta|}{2} log(\sigma_{NN}^2) - \frac{1}{2\sigma_{NN}^2} \sum_{w \in \theta} ||w||_2^2 \tag{4.32}$$

O grau de regularização de g_θ é controlado pelo hiperparâmetro σ_{NN}^2, onde valores mais baixos para σ_{NN}^2 induzem penalidades mais altas no modelo, como visto no exemplo

da Figura 4.4.

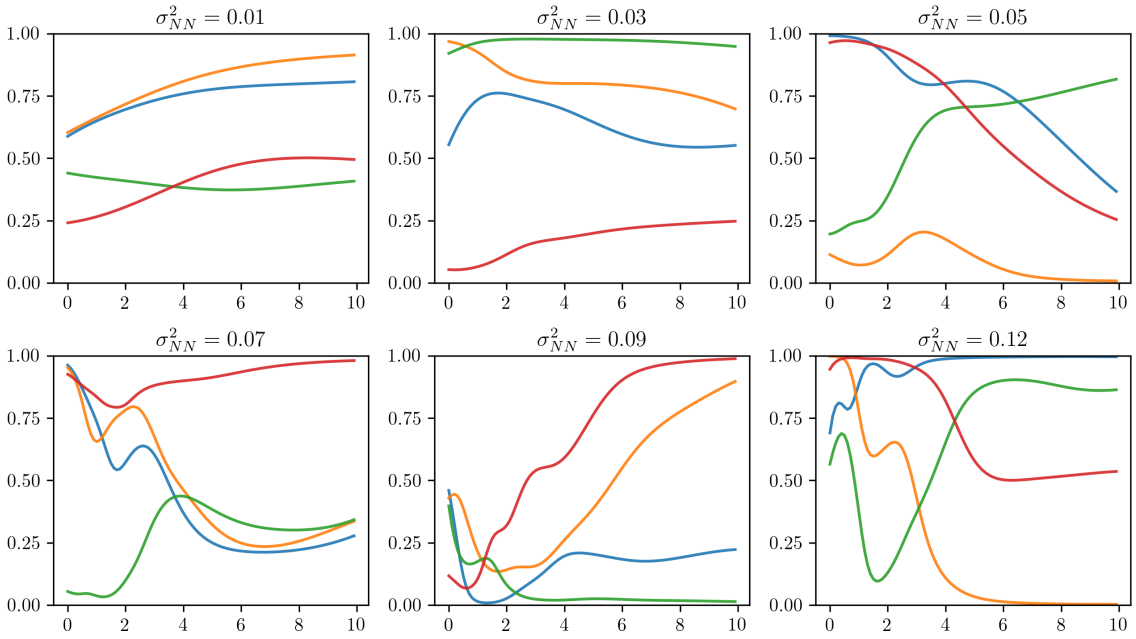

Figura 4.4: *Seis amostras da rede neural* $g_\theta : \mathbb{R} \longrightarrow \mathbb{R}^4$ *(utilizada no experimento do capítulo 5.3) obtidas sob a priori normal isotrópica* $\theta \sim N(0, \sigma^2_{NN} I)$ *para diferentes valores de* σ^2_{NN}.

Uma técnica possível para escolher de maneira subjetiva o hiperparâmetro σ^2_{NN} consiste em analisar as características das decomposições de curvas geradas sob a distribuição priori de $\boldsymbol{\theta}$ para múltiplos valores de σ^2_{NN}. Conforme ilustrado na Figura 4.4, a partir da análise gráfica das decomposições geradas pelo modelo sob a priori especificada, é possível escolher o valor do hiperparâmetro σ^2_{NN} em termo da suavidade desejada das decomposições obtidas.

Parâmetros de Transição de Estados

Outro caso de uso popular de distribuições priori com efeito de regularização é a classe de modelos modelos autoregressivos vetoriais (VAR), que possuem número de parâmetros excessivo quando utilizados em espaços de alta dimensionalidade.

Uma escolha comum de priori com efeito de regularização para esta classe de modelos é a priori de Minnesota (Litterman (1986)), que associa a matriz de transição de um processo VAR a uma distribuição normal centrada ao redor de um passeio aleatório. Essa hipótese é especialmente útil no contexto de previsão de curvas de taxas de juros, pois como discutido anteriormente, o passeio aleatório é usualmente um modelo difícil de se superar, especialmente em curtos horizontes de tempo.

Diferentemente da abordagem *empirical bayes* sugerida no trabalho original de Litterman (1986), uma variação da priori de Minnesota baseada na conjugação entre as distribuições normal e gama inversa (Giannone *et al.* (2015)) pode ser construída a partir das relações:

$$vec(\boldsymbol{P}) \overset{iid}{\sim} IG(a, b), \quad a > 0, \quad b > 0 \tag{4.33}$$

$$vec(\boldsymbol{F})|\boldsymbol{P} \sim N_k(vec(\boldsymbol{I}_k), \boldsymbol{V}) \tag{4.34}$$

onde $\boldsymbol{P} = diag(P_1, P_2, ..., P_k)$ é uma matriz diagonal de covariâncias do termo de ruído $\boldsymbol{w}_t$ e $\boldsymbol{V}$ é uma matriz de covariâncias de termos v_{ij} dados por

$$v_{ij} = \begin{cases} \lambda^2, & \text{se } i = j \\ (\lambda\gamma)^2 P_i/P_j, & \text{se } i \neq j \end{cases}, \quad \lambda > 0, \quad 0 \leq \gamma \leq 1 \tag{4.35}$$

O hiperparâmetro λ controla o grau de certeza na hipótese de passeio aleatório, enquanto γ controla a intensidade do relacionamento de fatores distintos, relativo ao valor de λ. Ao reduzir o valor do hiperparâmetro γ, a influência entre fatores distintos também diminui. Particularmente, quando $\gamma = 0$, o modelo VAR(1) pode ser escrito como k processos AR(1) independentes. Na Figura 4.5, o efeito do parâmetro γ é ilustrado em quatro amostras das matrizes de transição geradas pela priori de Minnesota, com $a = 2$, $b = 10^{-3}$ e $\lambda = 0.2$.

Assim como na seção anterior, a seleção dos hiperparâmetros da distribuição priori pode ser feita de maneira subjetiva através da análise das matrizes de transição e fatores gerados em diferentes configurações de hiperparâmetros. A presença de um modelo ou estudo de referência também pode ser usada nesta etapa como base para escolha dos valores dos hiperparâmetros (a, b, λ, γ).

4.2.4 Estimação

Condicional aos parâmetros do modelo $\boldsymbol{\Theta} = [\boldsymbol{\theta}, \boldsymbol{F}, \boldsymbol{P}, \boldsymbol{Q}]$, o caráter linear do modelo é preservado, e portanto o Filtro de Kalman pode ser utilizado para obter a distribuição exata dos fatores β_t dadas as últimas $(t-1)$ observações. Portanto, como visto no Capítulo 4.1.4, é possível avaliar a função de densidade da distribuição posteriori, dada por

$$\begin{aligned} p(\Theta|y) &\propto p(\Theta) \prod_{t=1}^{T} p(y_t|y_1, ..., y_{t-1}, \Theta) \\ &= p(\Theta) \prod_{t=1}^{T} \int_{\mathbb{R}} p(\beta_t|y_1, ..., y_{t-1}, \Theta) p(y_t|\beta_t, \Theta) d\beta_t \\ &= p(\Theta) \prod_{t=1}^{T} \mathcal{N}_k(y_t; H_\theta \beta_{t|t-1}, H_\theta P_{t|t-1} H_\theta^T + Q) \end{aligned} \tag{4.36}$$

onde $\mathcal{N}$ é a função de densidade gaussiana e os termos $\beta_{t|t-1} = E(\beta_t|y_1, ..., y_{t-1}, \Theta)$ e $P_{t|t-1} = Var(\beta_t|y_1, ..., y_{t-1}, \Theta)$ são obtidos diretamente das equações de recorrência de Kalman.

Para obter as estimativas pontuais dos parâmetros do modelo a função de densidade da distribuição posteriori pode ser maximizada utilizando métodos de otimização baseados

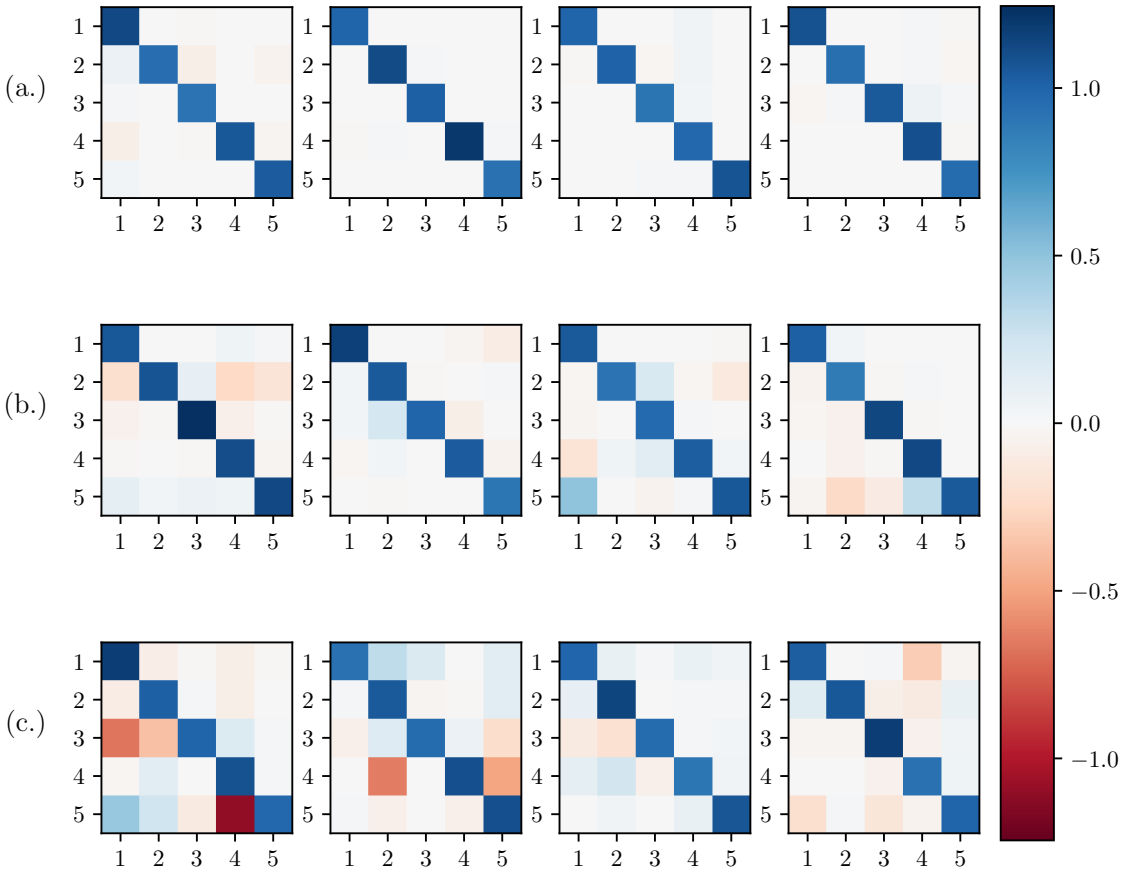

Figura 4.5: *Quatro amostras da matriz F com $k = 5$ fatores sob a priori de Minnesota em três configurações diferentes de hiperparâmetros. Em todos os painéis: $a = 2$, $b = 10^{-3}$, $\lambda = 0.2$. Para as linhas: (a) $\gamma = 0.33$, (b) $\gamma = 0.66$ e (c) $\gamma = 0.99$.*

em gradiente. Através do algoritmos de diferenciação automática, o gradiente $\nabla_\Theta p(\Theta|y)$ pode ser calculado de maneira eficiente para qualquer arquitetura diferenciável de rede neural. O custo computacional de se calcular o Filtro de Kalman em cada avaliação da densidade $p(\Theta|y)$ pode ser reduzido usando-se o algoritmo de computação paralela proposto por Särkkä e García-Fernández (2019) que, assintoticamente ao número de unidades de processamento disponíveis, alcança complexidade logaritmica de tempo efetivo de processamento.

Apesar da obtenção de estimativas pontuais que maximizam $p(\Theta|y)$ ser direta, inferir a distribuição $p(\Theta|y)$ pode ser uma tarefa extremamente complexa. Como visto na Seção 4.1.3, os modelos de redes neurais são super parametrizados e possuem baixa identificabilidade, o que dificulta a utilização de métodos de inferência baseados em *Markov Chain Monte Carlo* (MCMC), que se provam necessários na ausência de soluções analíticas para a posteriori, como no caso do modelo proposto. Além disso, como cada avaliação de $p(\Theta|y)$ exige a aplicação do Filtro de Kalman, gerar um número elevado de amostras da posteriori pode se tornar computacionalmente desafiador.

Recentemente, diversos métodos de inferência aproximada foram propostos para quantificar a incerteza de estimação em modelos de redes neurais Bayesianas (ver (Filos *et al.*,

2019) para uma discussão comparativa entre as abordagens mais utilizadas), a quantificação da incerteza pode ser crucial para aplicações com o objetivo de gerenciamento de risco ou precificação de derivativos, porém estes temas fogem do escopo deste trabalho.

4.2.5 Previsão

A partir das estimativas pontuais para as matriz F e os parâmetros θ, uma previsão da ETTJ para o instante $t + h$ pode ser obtida evoluindo a equação de estados a partir do último estado filtrado $\beta_{t|t}$ e fazendo-se a medição segundo a Equação 4.28:

$$\hat{y}_{t+h}(m) = g_{\hat{\theta}_{MAP}}(m)^T \hat{F}^h_{MAP} \beta_{t|t} \tag{4.37}$$

Capítulo 5

Avaliação Empírica

5.1 Metodologia

A qualidade das previsões dos modelos ajustados é avaliada numa amostra de teste para múltiplos horizontes de tempo. Para cada instante de tempo t da amostra de validação, a previsão para $t + h$ é obtida e comparada com a curva efetiva y_{t+h}, os resultados são registrados em uma base de dados, onde são calculadas métricas de performance agregadas.

Para os modelos baseados na abordagem de dois passos, a previsão para h-passos à frente no momento t são obtidas através da evolução dos fatores extraídos disponíveis em t para o passo $t + h$, utilizando o modelo VAR estimado. A previsão resultante é então convertida para o formato de curva de juros usando a equação de decomposição.

Para os modelos baseados na abordagem de estimativa em uma etapa, o procedimento é semelhante, mas requer uma etapa extra de atualização do Filtro Kalman no final de cada iteração. O procedimento completo é descrito abaixo, para cada período de tempo t:

1. Obter a previsão $\hat{\boldsymbol{F}}^h \hat{\boldsymbol{\beta}}_{t|t}$ para os fatores da curva $t + h$, a partir do vetor de fatores filtrados $\hat{\boldsymbol{\beta}}_{t|t} = E(\boldsymbol{\beta}|\boldsymbol{y}_1, ..., \boldsymbol{y}_t)$ e a matriz de transição estimada $\hat{\boldsymbol{F}}$.

2. Converter os fatores previstos usando a matriz de medição estimada $\hat{\boldsymbol{H}}$ para obter a previsão da curva de taxa de juros completa de $t + h$:

$$\hat{\boldsymbol{y}}_t^{t+h} = \hat{\boldsymbol{H}} \hat{\boldsymbol{F}}^h \hat{\boldsymbol{\beta}}_{t|t} \tag{5.1}$$

3. Atualizar as equações de recorrência de Kalman usando a próxima observação $\boldsymbol{y}_{t+1}$ para obter $\hat{\boldsymbol{\beta}}_{t|t}$.

As previsões são então avaliadas utilizando a métrica *Root Mean Squared Error* (RMSE), calculada para cada vencimento m e horizonte h considerados:

$$RMSE(m,h) = \sqrt{\frac{1}{T-t_0}\sum_{t=t_0}^{T}\left(y_{t+h}(m) - \hat{y}_t^{t+h}(m)\right)^2} \tag{5.2}$$

onde t_0 é o índice da primeira observação fora da amostra utilizada na estimação, $\hat{y}_t^{t+h}(m)$ é a previsão (calculada em t) da taxa de juros de vencimento m para o instante $t+h$, e T é o número total de observações do conjunto de dados.

Outra métrica de interesse para medir a qualidade das previsões de cada modelo é a média aritmética e desvio padrão das métricas RMSE, calculadas ao longo dos vencimentos considerados

$$mRMSE(h) = \frac{1}{M}\sum_{m=1}^{M} RMSE(m,h) \tag{5.3}$$

$$sRMSE(h) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(RMSE(m,h) - mRMSE(h)\right)^2} \tag{5.4}$$

Os modelos baseados em estimação em uma etapa (como Nelson e Siegel dinâmico e o modelo proposto no capítulo 4.2.2) são estimados usando métodos de otimização baseada em gradiente da biblioteca *Pytorch* (Paszke *et al.* (2019)), com a implementação do algoritmo *parallel-scan* do Filtro de Kalman (Särkkä e García-Fernández (2019)) da biblioteca *Pyro* (Bingham *et al.* (2019)), na linguagem de programação *Python*.

5.2 Experimento com Dados Artificiais

5.2.1 Descrição do Experimento

Nesta seção, um experimento com dados artificiais com objetivo de testar a viabilidade em termos práticos do modelo linear-neural de espaço de estados é conduzido. Para isso, um processo gerador baseado em um caso particular do modelo LNEE é criado:

$$\boldsymbol{z}_t = \boldsymbol{F}\boldsymbol{z}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \overset{iid}{\sim} N_2(\boldsymbol{0}, \boldsymbol{P}) \tag{5.5}$$

$$\boldsymbol{y}_t(m) = \boldsymbol{\phi}(m)^T\boldsymbol{z}_t + \boldsymbol{\epsilon}_t(m), \quad \epsilon_t(m) \overset{iid}{\sim} N(\boldsymbol{0}, \sigma_\epsilon^2) \tag{5.6}$$

onde $\boldsymbol{F} = \begin{bmatrix} 0.88 & -0.80 \\ 0.88 & 0.33 \end{bmatrix}$, $\boldsymbol{P} = \begin{bmatrix} 0.04 & 0.00 \\ 0.00 & 0.04 \end{bmatrix}$, $\sigma_\epsilon^2 = 0.04$ e as funções de carga são dadas por

$$\boldsymbol{\phi}(m) = \begin{bmatrix} \phi_0(m) \\ \phi_1(m) \end{bmatrix} = \begin{bmatrix} e^{-(m-0.25)^2/0.25} \\ e^{-(m-0.75)^2/0.04} \end{bmatrix} \tag{5.7}$$

Deste processo, são geradas um total de 500 observações, sendo as primeiras 400 destinadas à estimativa dos parâmetros dos modelos testados e as últimas 100 para avaliação. Além disso, para facilitar a manipulação dos dados gerados, as curvas são avaliadas num conjunto restrito de 20 pontos $\{m_1, ..., m_{20}\}$ igualmente espaçados no intervalo $[0, 1]$.

Os valores escolhidos para $\boldsymbol{F}$ e $\boldsymbol{P}$ garantem que a série temporal gerada dos fatores $\boldsymbol{z}_t$ do modelo tenham comportamento cíclico e relativamente previsível, como visto na Figura 5.1.

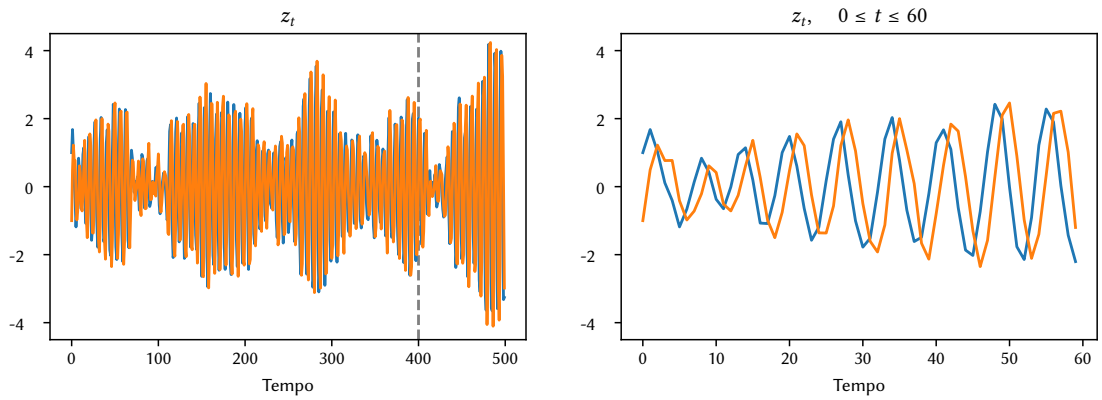

Figura 5.1: *Série temporal do primeiro (linha azul) e segundo fator (linha laranja) de $\boldsymbol{z}_t$. Painel da esquerda: série completa. Painel da direita: primeiras 60 observações.*

Na Figura 5.2, as funções de cargas da Equação 5.7 são representadas conjuntamente com as primeiras 4 curvas geradas pelo modelo, retratadas com e sem a presença do termo de ruído de medição $\boldsymbol{\epsilon}_t$. O valor de $\sigma_\epsilon^2 = 0.04$ é escolhido de forma a introduzir uma quantiade significativa de ruído, sem distorcer completamente o formato das curvas geradas, para testar a robustez das decomposições obtidas pelo modelo proposto na presença de ruído.

Neste experimento, o modelo Linear-Neural de Espaço de Estados (LNEE) é inicializado com uma rede neural g_θ de três camadas com 64, 32 e 32 unidades ocultas, respectivamente. Assim como discutido em 4.2.2, são utilizadas funções de ativação do tipo tangente hiperbólica para as camadas intermediárias e sigmóide para a camada final.

Dado o número reduzido de fatores, uma distribuição priori pouco informativa é atribuída aos parâmetros da componente temporal do modelo. Para isso, são usados os hiperparâmetros $(\lambda, \gamma) = (0.5, 0.9)$, e $(a, b) = (0.1, 0.001)$, obtidos através do procedimento de simulação descrito em 4.2.3. Com o intuito de analisar as decomposições obtidas pela rede neural g_θ na ausência de mecanismos de regularização, a variância para os parâmetros de g_θ é fixada como $\sigma_{NN}^2 = 10^5$.

Os parâmetros do modelo LNEE são estimados a partir de métodos de maximização de primeira ordem para a densidade da distribuição posteriori $p(\Theta|y_1, ..., y_{400})$, assim como descrito no Capítulo 4.2.4.

Para verificar a importância na escolha de pontos iniciais dos algoritmos de otimização, a estimação do modelo LNEE é feita duas vezes, a primeira com inicialização aleatória para todos os parâmetros do modelo, e a segunda, usando as componentes principais da matriz

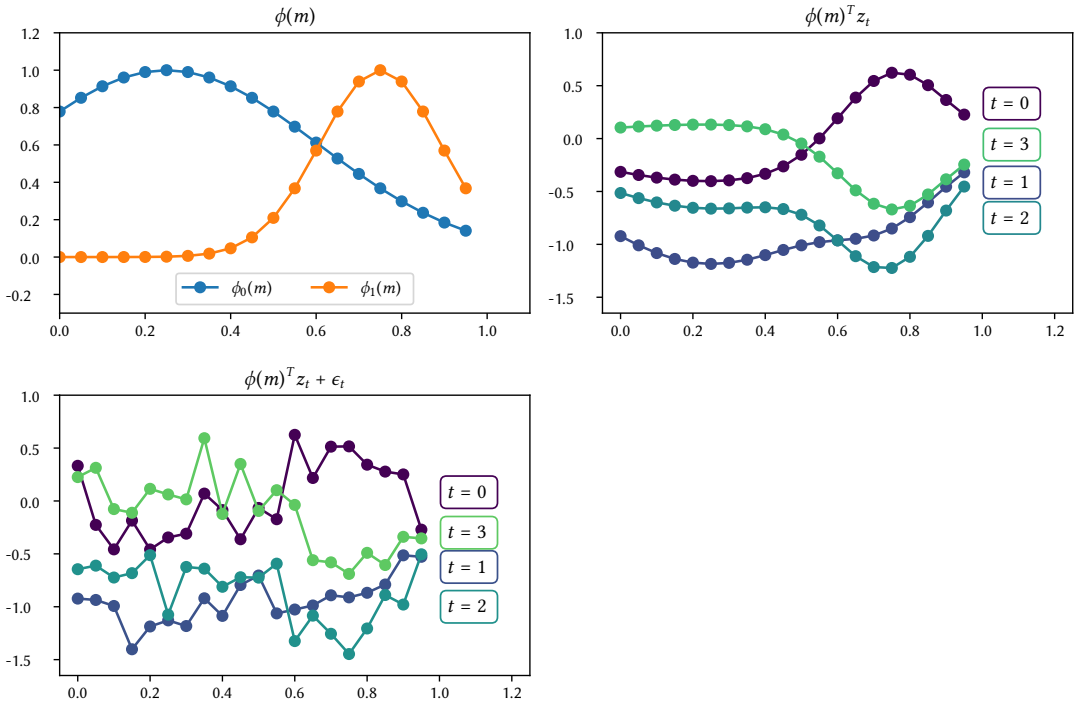

Figura 5.2: *Painel superior esquerdo: funções de carga $\boldsymbol{\phi}(m)$ calculadas para $m \in \{m_1, ..., m_{20}\}$. Painel superior direito: quatro primeiras curvas geradas no conjunto de dados antes da inclusão do termo de ruído de medição. Painel inferior: quatro primeiras curvas geradas no conjunto de dados após a inclusão do termo de ruído de medição ϵ_t.*

de dados para inicializar para as funções de carga g_θ do modelo. Na segunda execução, os parâmetros iniciais θ_0 de g_θ são obtidos a partir de um procedimento de pré-treinamento da rede neural com o objetivo de minimizar o erro quadrático médio entre as funções de carga do modelo e a matriz de componentes principais de Y.

5.2.2 Resultados

Na Figura 5.3, são representadas as decomposições obtidas ao final do procedimento de estimação dos parâmetros do modelo. É possível verificar que ambos os modelos LNEE foram capazes de recuperar grande parte do formato das funções de carga originais do processo gerador dos dados. Em especial, nota-se que o modelo LNEE com inicialização aleatória capturou melhor a segunda função de carga enquanto o outro modelo capturou melhor a primeira função de carga.

Como esperado, o procedimento de estimação do modelo LNEE, com funções de carga inicializadas a partir das duas componentes principais, atingiu convergência mais rapidamente que o modelo com incialização aleatória, reduzindo pela metade o número total de iterações necessárias.

A série temporal dos fatores filtrados $\hat{\boldsymbol{\beta}}_{t|t}$, obtida através das equações de recorrência do Filtro de Kalman (4.1.4), é apresentada na Figura 5.4 para ambos os modelos. Nota-se que

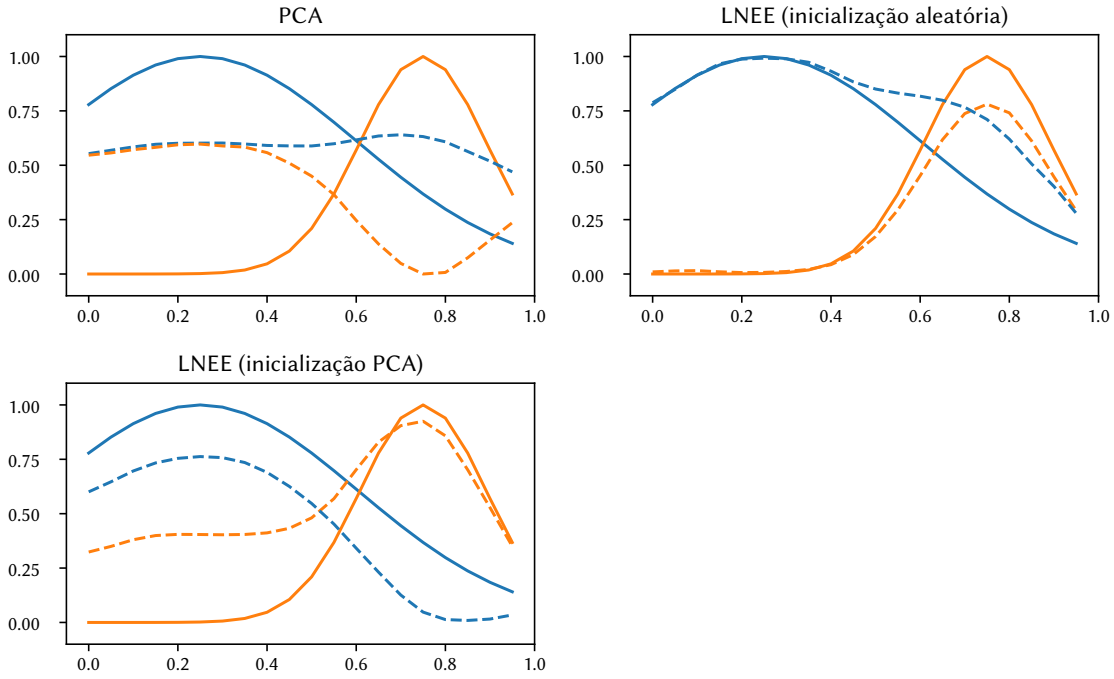

Figura 5.3: *Funções de carga reais da Equação 5.7 (em linhas sólidas) e funções de carga estimadas (em linhas pontilhadas). Painel superior esquerdo: duas primeiras componentes principais. Painel superior direito: decomposição estimada do modelo LNEE com θ inicializado aleatoriamente. Painel inferior: decomposição estimada do modelo LNEE, com θ inicializado a partir das duas primeiras componentes principais.*

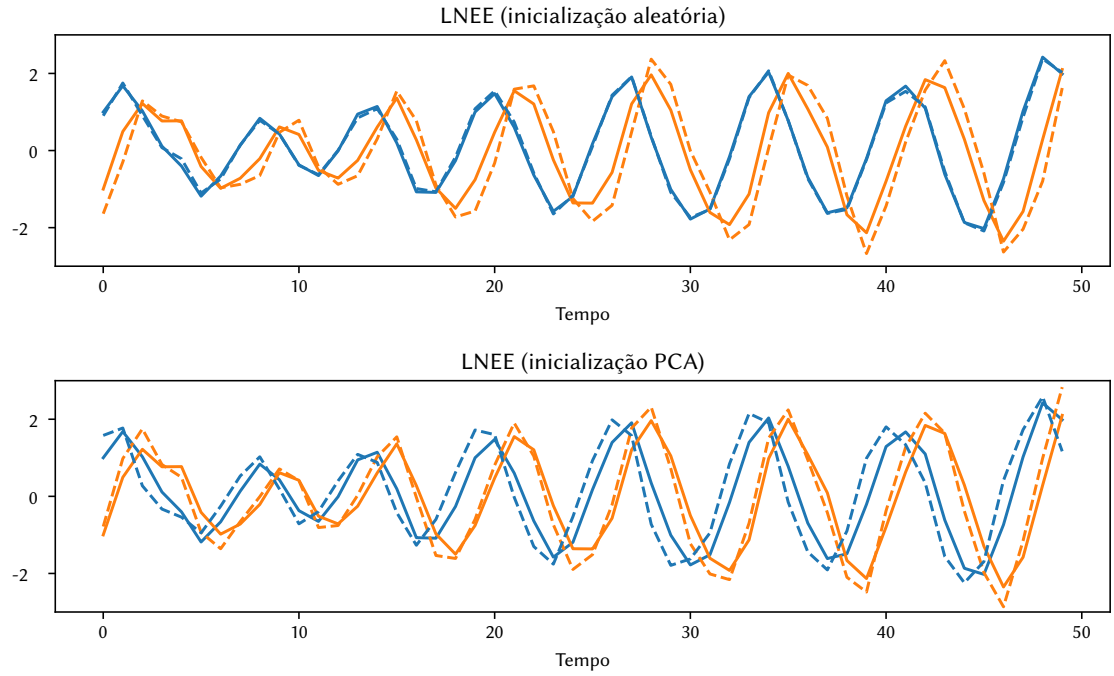

Figura 5.4: *Primeiras 50 observações dos fatores filtrados $\hat{\beta}_{t|t} = E(\beta_t|y_1, ..., y_t)$ (linhas tracejadas) dos modelos LNEE com as duas inicializações (aleatória e componentes principais) e a série temporal dos fatores z_t gerados pelo modelo verdadeiro (linhas sólidas).*

ambos os modelos foram capazes de reproduzir com precisão o padrão cíclico do processo gerador dos dados, entretanto, os fatores filtrados do modelo com inicialização aleatória obtiveram resultados mais próximos dos fatores originais.

A performance dos dois modelos LNEE são avaliadas usando o processo descrito na seção 5.1, para os horizontes de 1, 5, 10, 15 unidades de tempo. Um exemplo das previsões obtidas pelo modelo LNEE com inicialização aleatória é apresentado na Figura 5.5. Dos resultados das métricas RMSE médias, apresentados na Tabela 5.1, é possível concluir que apesar das diferenças aparentes entre as decomposições obtidas, ambos os modelos obtém resultados semelhantes na tarefa de previsão.

Horizonte (dias)	Inicialização	$\overline{RMSE}$	d.p
1	aleatória	**0.299**	0.035
	PCA	0.300	0.034
5	aleatória	**0.510**	0.086
	PCA	0.510	0.086
10	aleatória	0.744	0.125
	PCA	**0.744**	0.125
15	aleatória	**0.981**	0.171
	PCA	0.982	0.171

Tabela 5.1: *Média e desvio-padrão da métrica RMSE das previsões does modelos LNEE com as duas inicializações (aleatória e componentes principais) para os horizontes de 1, 5, 10 e 15 unidades de tempo.*

A partir dos resultados do experimento realizado, é possível concluir que o modelo proposto foi capaz de recuperar com precisão a decomposição original do processo gerador criado, mesmo sob a presença de ruído. Além disso, foi possível que uma inicialização cuidadosa dos parâmetros da rede neural g_θ pode reduzir significativamente o tempo de convergência na estimação dos parâmetros do modelo sem afetar significativamente os resultados obtidos na tarefa de previsão.

Além disso, nota-se que mesmo sob a ausência de regularização e presença de ruído de medição no conjunto de dados, a rede neural do modelo proposto produziu funções de carga com um padrão suave e estável. No entanto, por se tratar de um experimento controlado, os resultados podem ser pouco conclusivos. Para investigar melhor a influência da distribuição priori nas decomposições obtidas e performance de previsão, um experimento com este propósito é conduzido na Seção 5.3 usando os dados da ETTJ brasileira.

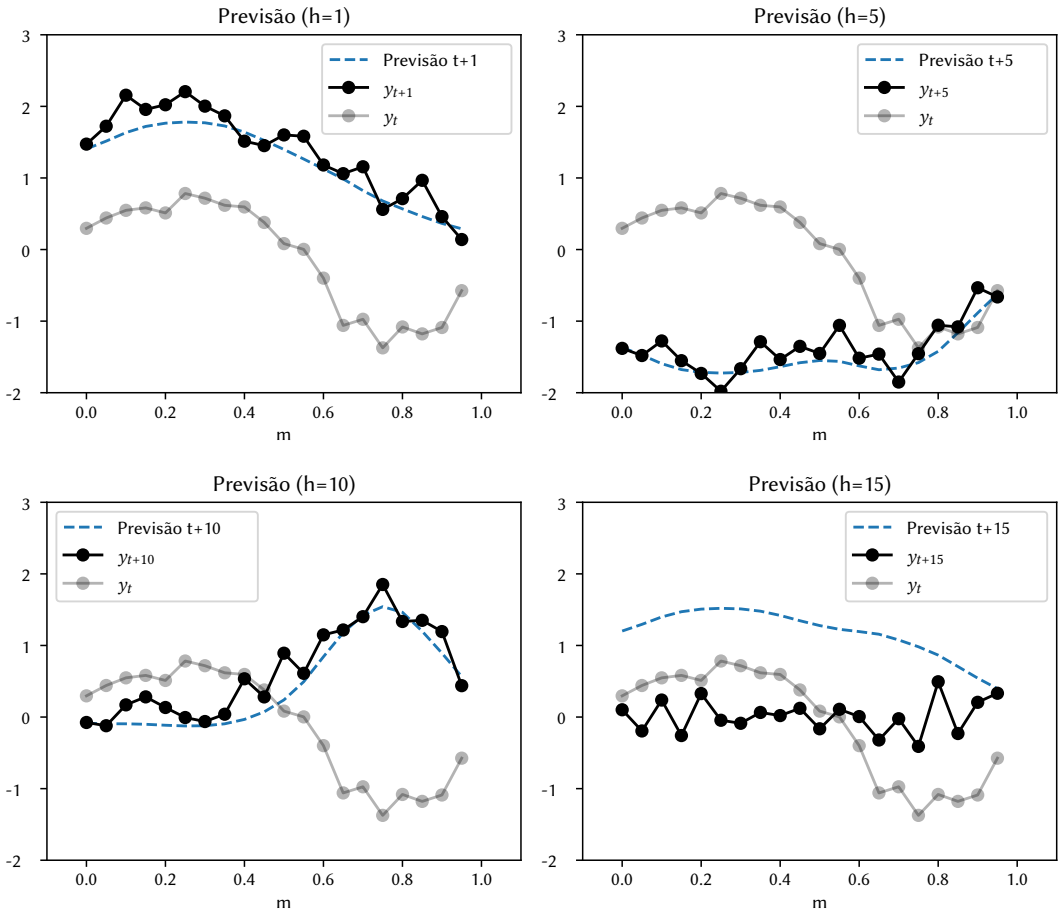

Figura 5.5: *Quatro previsões do modelo LNEE (inicialização aleatória) para os horizontes de 1, 5, 10 e 15 unidades de tempo, partindo do instante de tempo* $t = 400$.

5.3 Experimentos com a ETTJ Brasileira

5.3.1 Descrição do Experimento

A fim de avaliar empiricamente o desempenho da técnica proposta, um experimento é realizado com a curva de juros brasileira construída e divulgada pela bolsa de valores brasileira B3 a partir de contratos futuros de depósitos interfinanceiros (Futuro DI)[1], como usualmente usado na literatura brasileira de modelagem da estrutura a termo das taxas de juros (ver Caldeira *et al.* (2010), Vicente e Tabak (2008) e Cajueiro *et al.* (2009), por exemplo). Diferentemente das taxas de contratos de *swap*, os contratos futuros de DI tem maior diversidade de vencimentos e alta liquidez, o que torna atraente a utilização desta classe de ativos para a construção das curvas de taxas de juros, ainda que a série histórica disponível para este tipo de ativo seja menor que a de contratos de *swap*.

A base de dados extraída compreende o período de Agosto de 2003 à Agosto de 2018 para os vencimentos de 1, 2, 1, 2, 3, 6, 12, 24, 36, 48, 60, 72 e 84 meses de 21 dias úteis.

[1]Informações obtida a partir do Manual de Curvas do site oficial da B3.

Os vértices faltantes da base de dados são interpolados usando-se o método *flat-forward* (Maltz (2002), descrito na seção 2.6.3). As principais estatísticas descritivas do conjunto de dados são apresentadas na Tabela 5.2.

Vencimento (meses)	Média	Desvio Padrão	Mínimo	Máximo	$\hat{\rho}(21)$	$\hat{\rho}(252)$
1	11.60	2.97	6.45	21.01	0.99	0.56
2	11.59	2.96	6.43	20.56	0.99	0.55
3	11.58	2.95	6.41	20.16	0.99	0.55
6	11.59	2.90	6.41	19.53	0.99	0.53
12	11.68	2.80	6.59	19.30	0.98	0.52
24	11.94	2.56	7.10	19.84	0.97	0.52
36	12.10	2.44	7.65	20.53	0.96	0.54
48	12.21	2.40	7.98	21.34	0.96	0.56
60	12.27	2.42	8.21	22.10	0.96	0.57
72	12.32	2.45	8.38	22.82	0.96	0.57
84	12.37	2.48	8.53	23.33	0.96	0.58

Tabela 5.2: *Estatísticas descritivas para taxas de juros anuais, referentes ao período de Agosto de 2003 à Agosto de 2018 da base de dados descrita em 5.3.1, em escala de pontos percentuais. As autocorrelações amostrais para 21 e 252 dias úteis das taxas de juros em cada vencimento são reportadas respectivamente nas colunas $\hat{\rho}(21)$ e $\hat{\rho}(252)$.*

São comparadas as previsões fora da amostra obtidas pelo modelo linear-neural de espaço de estados (LNEE) em relação as seguintes abordagens clássicas: passeio aleatório (RW), o modelo Nelson e Siegel dinâmico (DNS) e o modelo Nelson, Siegel e Svensson (DSV). Ambos os modelos DNS e DSV são testados com as abordagens de estimação de parâmetros em uma etapa e duas etapas.

Para avaliar se a inclusão da rede neural é benéfica no modelo proposto (LNEE), um modelo Gaussiano linear de espaço de estados (GLEE) com matriz de carga $\boldsymbol{H}$ estimável também é incluso no experimento. Além disso, para verificar se os modelos com 4 fatores são suficientes, o modelo LNEE e GLEE são ambos testados com 4 e 5 fatores latentes.

Na abordagem de estimação em duas etapas para os modelos DNS e DSV, seguindo o método descrito por Nelson e Siegel (1987), um procedimento de otimização combinatória é usado para escolher os parâmetros de decaimento exponencial que minimizam a soma total dos quadrados da curva de juros observada e a curva reconstruída a partir dos fatores de decomposição, como descrito em 3.2.2. Os valores obtidos por este procedimento para os parâmetros de decaimento exponencial também são re-utilizados na abordagem de estimação em uma etapa.

Para aumentar a estabilidade numérica e acelerar a convergência, os parâmetros estimáveis dos modelos DNS e DSV de uma etapa são inicializados com as estimativas obtidas na abordagem em duas etapas. Com o mesmo objetivo, a matriz de medição $\boldsymbol{H}$ do modelo GLEE é inicializada com a matriz de carga $\boldsymbol{H}_\lambda$ do modelo DSV. Além disso, o vetor de parâmetros $\boldsymbol{\theta}$ da rede neural do modelo LNEE também é inicializado com uma solução que reproduz aproximadamente a matriz de cargas do modelo DSV, usando o mesmo

procedimento de pré-treinamento da rede neural g_θ utilizado no experimento anterior (5.2.1).

Dada a complexidade computacional do modelo LNEE, todas as abordagens sugeridas são avaliadas usando os primeiros 12 anos de dados para estimar uma vez os parâmetros e os 3 anos restantes para avaliar as previsões fora da amostra para 1 semana, 1 mês, 3 meses e 6 meses à frente.

Em todos os modelos LNEE, são utilizadas redes neurais com duas camadas intermediárias, ambas com 300 unidades ocultas intercaladas de funções de ativação tangente hiperbólica e funções de ativação sigmóide para a última camada, assim como descrito em 4.2.2.

Os hiperparâmetros da distribuição priori do modelo LNEE são escolhidos com base nos resultados obtido do modelo DSV com estimação em uma etapa e com a ajuda de procedimentos de simulação. Assim como no experimento anterior, distribuições priori pouco informativas são usadas a fim de avaliar o comportamento do modelo na ausência de fortes influências externas. Uma análise mais abrangente do impacto da distribuição priori no desempenho do modelo é apresentada a seguir na Seção 5.3.3.

Para os hiperparâmetros da distribuição gama inversa, são selecionados $(a, b) = (0.1, 0.001)$ tais que a priori resultante seja pouco informativa e tenha pico próximo à estimativa da variância de transição dos estados obtida no modelo DSV. Como descrito em 4.2.3, os demais hiperparâmetros da priori de Minnesota são escolhidos de forma qualitativa através de uma inspeção manual das matrizes de transição e das séries temporais de fatores geradas. Deste procedimento, os valores $(\lambda, \gamma) = (0.5, 0.9)$ são escolhidos por gerarem observações visualmente plausíveis e que ainda mantém boa parcela de variabilidade.

A partir de um procedimento análogo de simulação para as funções de carga g_θ (Figura 4.4), $\sigma^2_{NN} = 0.05$ é escolhido por gerar funções de carga suaves, mas que ainda apresentam complexidade superior às funções de carga da decomposição de Nelson e Siegel.

5.3.2 Resultados

As tabelas 5.3 a 5.6 reportam os resultados das métricas RMSE para as previsões fora de amostra dos nove modelos considerados e quatro horizontes previsão. Em cada modelo, as métricas de performance são reportadas nos 11 vencimentos considerados, assim como a média geral e desvio-padrão da métrica RMSE.

Na previsão para 1 semana (Tabela 5.3), todos os modelos, exceto o modelo 5-GLEE e o modelo Nelson e Siegel dinâmico obtiveram melhores pontuações do que o modelo de passeio aleatório (RW), que é tipicamente um modelo difícil de ser superado em horizontes curtos de previsão. O modelo LNEE de cinco fatores alcançou o menor valor de RMSE médio, seguido de perto pelo modelo LNEE de quatro fatores.

Para as previsão de 1 mês (Tabela 5.4), os modelos LNEE ainda alcançaram o menor valor de RMSE médio geral e melhor performance nas maturidades de longo prazo. Entretanto, os modelos DSV também apresentaram bom desempenho, superando os demais modelos nos vencimentos de curto e médio prazo.

	Vencimentos (meses úteis)												
Modelo	1	2	3	6	12	24	36	48	60	72	84	**Média**	**d.p**
RW	0.95	0.95	1.07	1.47	**1.99**	2.66	2.90	2.98	3.02	3.03	3.03	2.19	0.87
DNS $^{(2)}$	1.02	0.79	1.14	2.15	2.54	2.66	2.99	3.03	3.00	3.15	3.34	2.34	0.89
DNS $^{(1)}$	1.10	0.81	1.17	2.20	2.53	2.80	2.95	2.94	**2.95**	3.05	3.20	2.34	0.84
DSV $^{(2)}$	0.77	0.79	1.05	1.46	2.16	2.68	2.90	3.01	2.98	3.04	3.15	2.18	0.93
DSV $^{(1)}$	0.75	0.78	0.98	**1.35**	2.18	2.81	2.81	2.90	2.95	3.01	3.12	2.15	0.93
4-GLEE	0.71	0.75	0.97	1.51	2.16	2.79	**2.81**	**2.89**	2.96	3.01	3.08	2.15	0.93
5-GLEE	0.69	0.89	1.28	1.89	2.10	**2.60**	2.87	3.06	3.10	3.02	3.01	2.23	0.87
4-LNEE	0.75	0.73	0.89	1.36	2.08	2.80	2.81	2.90	2.96	3.00	3.05	2.12	0.94
5-LNEE	**0.66**	**0.72**	**0.89**	1.37	2.07	2.65	2.85	2.96	3.00	**2.98**	**2.96**	**2.10**	0.95

Tabela 5.3: *Métrica RMSE ($\times 10^3$) das previsões fora de amostra para horizonte de tempo de 1 semana. Os símbolos $^{(1)}$ e $^{(2)}$ são usados para diferenciar estimação em uma e duas etapas, respectivamente.*

	Vencimentos (meses úteis)												
Modelo	1	2	3	6	12	24	36	48	60	72	84	**Média**	**d.p**
RW	3.27	3.24	3.34	3.75	4.40	5.34	5.76	5.91	5.98	6.00	6.00	4.82	1.16
DNS $^{(2)}$	1.57	1.98	2.57	3.86	4.72	5.09	5.60	5.92	6.02	6.25	6.47	4.55	1.70
DNS $^{(1)}$	1.58	1.94	2.51	3.75	4.65	5.19	**5.36**	**5.56**	5.73	5.88	6.04	4.38	1.58
DSV $^{(2)}$	**1.31**	**1.69**	2.11	2.93	**4.00**	**4.98**	5.48	5.77	5.85	6.03	6.25	4.22	1.80
DSV $^{(1)}$	1.61	1.79	2.02	**2.72**	4.18	5.41	5.49	5.62	5.76	5.87	6.02	4.23	1.74
4-GLEE	1.75	2.00	2.32	3.06	4.23	5.40	5.48	5.61	5.74	5.80	5.88	4.30	1.61
5-GLEE	2.24	2.70	3.12	3.76	4.24	5.19	5.71	6.06	6.09	5.92	5.86	4.63	1.40
4-LNEE	1.63	1.75	**2.00**	2.72	4.03	5.32	5.42	5.56	5.70	5.78	5.85	**4.16**	1.70
5-LNEE	1.77	1.98	2.24	2.94	4.07	5.16	5.49	5.65	**5.67**	**5.59**	**5.53**	4.19	1.56

Tabela 5.4: *Métrica RMSE ($\times 10^3$) das previsões fora de amostra para horizonte de tempo de 1 mês. Os símbolos $^{(1)}$ e $^{(2)}$ são usados para diferenciar estimação em uma e duas etapas, respectivamente.*

	Vencimentos (meses úteis)												
Modelo	1	2	3	6	12	24	36	48	60	72	84	**Média**	**d.p**
RW	9.35	9.33	9.34	9.46	9.87	11.09	11.89	12.05	12.00	11.99	11.96	10.76	1.21
DNS $^{(2)}$	4.51	5.20	5.90	7.52	9.05	10.27	11.27	11.77	11.92	12.20	12.45	9.28	2.87
DNS $^{(1)}$	4.42	5.03	5.63	7.14	8.93	10.37	10.80	11.03	11.17	11.27	11.37	8.83	2.63
DSV $^{(2)}$	**3.64**	**4.24**	4.80	6.05	7.50	**9.24**	10.42	11.08	11.36	11.76	12.11	8.38	3.09
DSV $^{(1)}$	4.20	4.42	4.68	5.77	8.14	10.46	10.87	11.01	11.10	11.15	11.24	8.46	2.93
4-GLEE	4.98	5.29	5.57	6.32	8.12	10.32	10.69	10.83	10.90	10.89	10.93	8.62	2.47
5-GLEE	6.49	6.80	7.01	7.21	7.91	9.92	10.83	11.26	11.27	11.08	10.99	9.16	1.95
4-LNEE	4.14	4.40	**4.67**	**5.60**	7.70	10.09	10.58	10.79	10.91	10.96	11.02	8.26	2.85
5-LNEE	4.81	5.04	5.24	5.89	**7.50**	9.58	**10.16**	**10.34**	**10.34**	**10.25**	**10.21**	**8.12**	2.32

Tabela 5.5: *Métrica RMSE ($\times 10^3$) das previsões fora de amostra para horizonte de tempo de 3 meses. Os símbolos $^{(1)}$ e $^{(2)}$ são usados para diferenciar estimação em uma e duas etapas, respectivamente.*

	Vencimentos (meses úteis)												
Modelo	1	2	3	6	12	24	36	48	60	72	84	**Média**	**d.p**
RW	18.08	18.14	18.19	18.17	17.89	18.37	19.04	18.94	18.71	18.64	18.58	18.43	0.36
DNS $^{(2)}$	8.82	9.55	10.30	12.27	14.39	16.28	17.49	17.89	17.90	18.10	18.31	14.66	3.59
DNS $^{(1)}$	8.47	9.04	9.64	11.59	14.39	16.64	17.07	16.98	16.87	16.78	16.77	14.02	3.42
DSV $^{(2)}$	8.03	8.86	9.65	11.53	13.33	15.20	16.54	17.17	17.39	17.77	18.12	13.96	3.67
DSV $^{(1)}$	8.49	8.86	9.34	11.24	14.38	16.89	17.06	16.85	16.70	16.56	16.54	13.90	3.47
4-GLEE	9.75	10.15	10.51	11.77	14.23	16.52	16.60	16.39	16.23	16.06	15.99	14.02	2.73
5-GLEE	10.67	10.78	10.80	11.09	12.63	15.20	15.90	16.03	15.96	15.78	15.70	13.69	2.34
4-LNEE	**8.03**	**8.43**	**8.82**	10.39	13.36	16.11	16.48	16.40	16.34	16.24	16.23	13.35	3.49
5-LNEE	8.45	8.69	8.87	**9.91**	**12.38**	**15.00**	**15.48**	**15.43**	**15.34**	**15.23**	**15.21**	**12.73**	2.97

Tabela 5.6: *Métrica RMSE ($\times 10^3$) das previsões fora de amostra para horizonte de tempo de 6 meses. Os símbolos $^{(1)}$ e $^{(2)}$ são usados para diferenciar estimação em uma e duas etapas, respectivamente.*

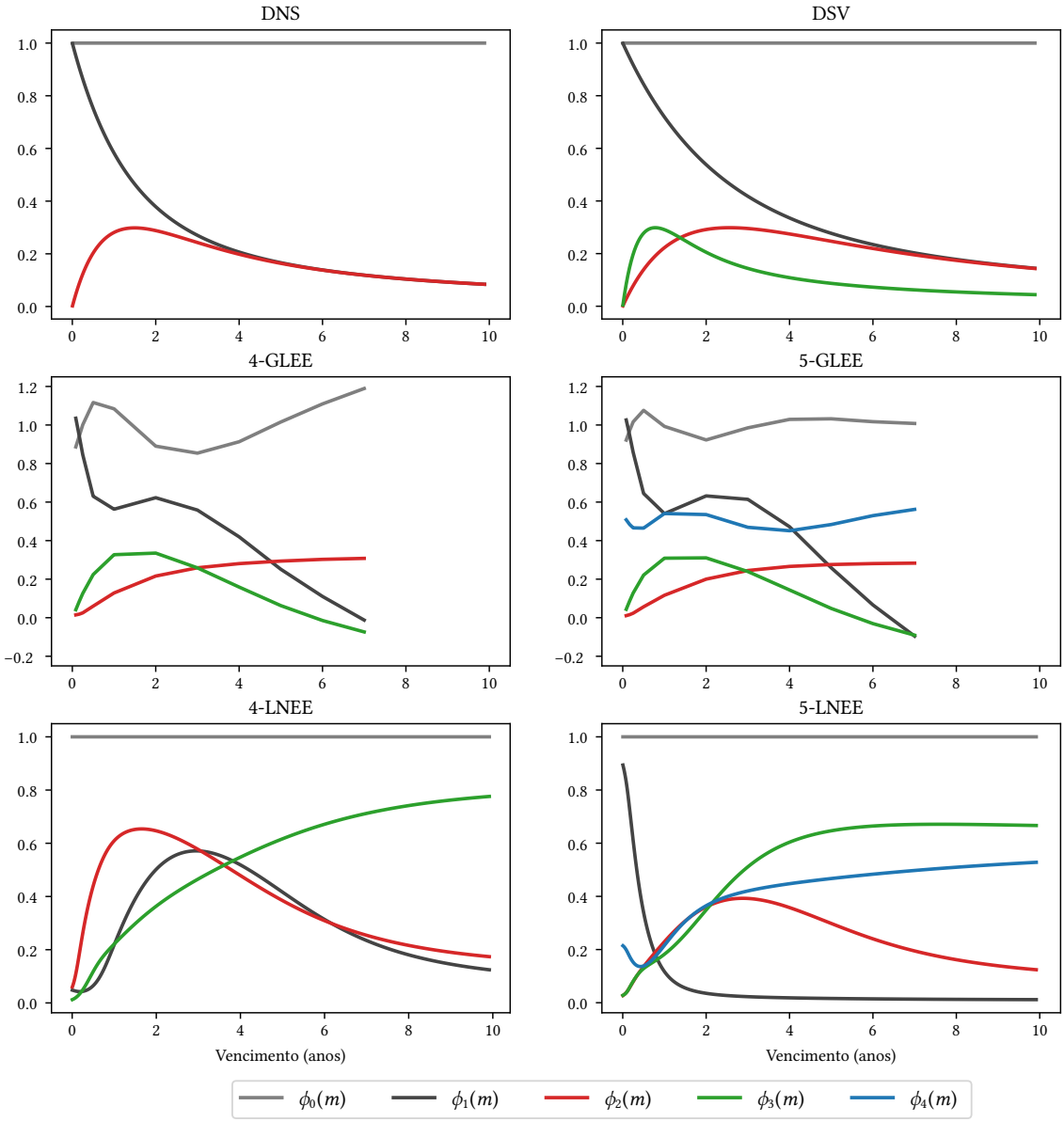

Figura 5.6: *Funções de carga estimadas dos modelos DNS, DSV, GLEE e LNEE. As cargas dos modelos GLEE são linearmente interpoladas dentro intervalo dos 11 vencimentos considerados, entretanto, não são extrapoladas para nenhum vencimento fora deste intervalo.*

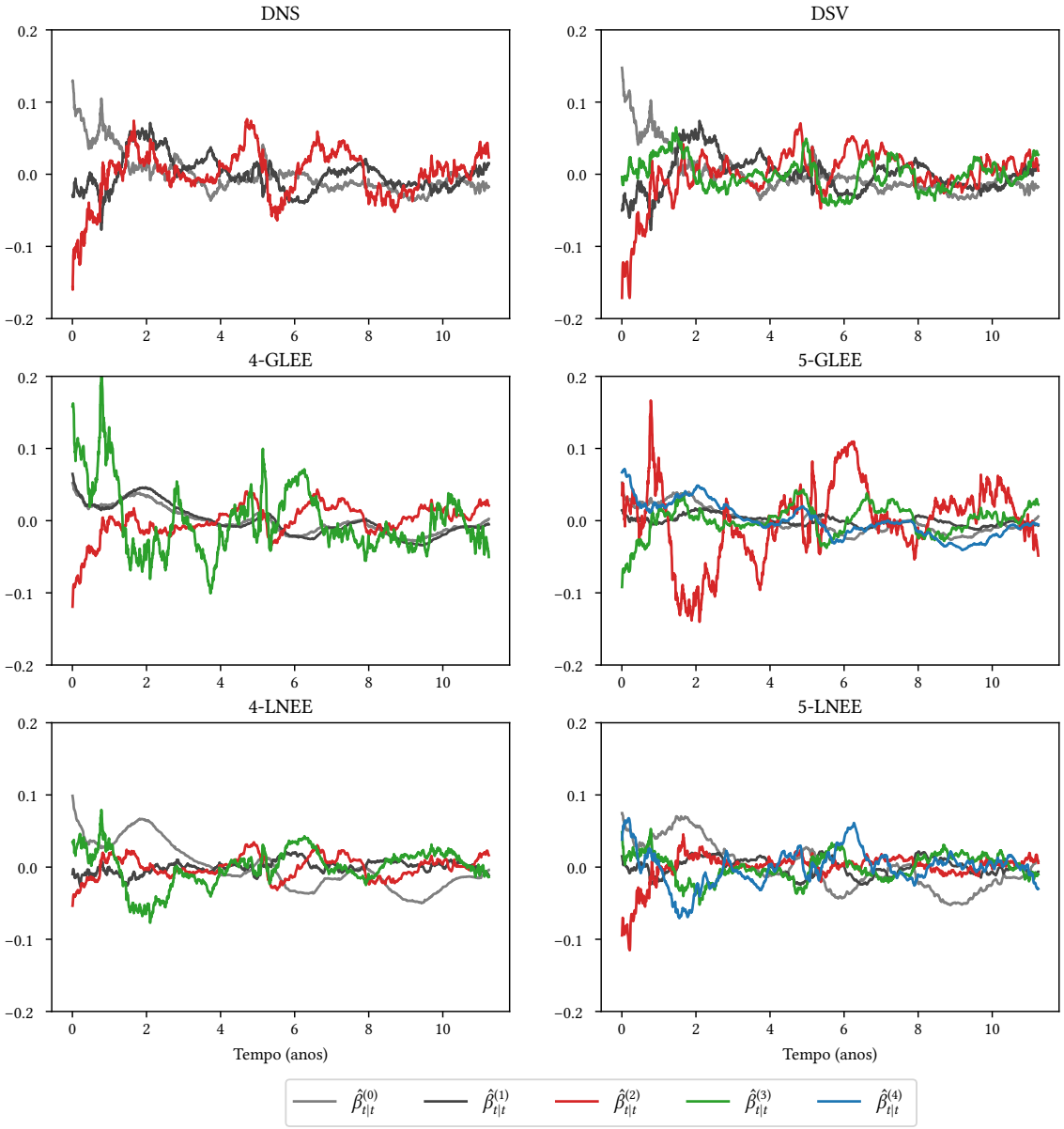

Figura 5.7: *Séries temporais dos fatores filtrados* $\hat{\beta}_{t|t} = E(\beta_t|y_1, ..., y_t)$ *dos modelos DNS e DSV, com estimação em uma etapa, e dos modelos GLEE e LNEE.*

O mesmo padrão observado nas previsões de 1 mês também é encontrado nos resultados do horizonte de 3 meses (Tabela 5.5). No entanto, a diferença de desempenho médio entre os modelos LNEE e os outros candidatos torna-se mais perceptível. Surpreendentemente, para o modelo DSV, a abordagem de estimação em duas etapas obteve RMSE visivelmente menor do que a versão em uma etapa. Observa-se também que o desempenho dos modelos GLEE diminuiu significativamente, permanecendo apenas acima dos modelos DNS e do modelo de passeio aleatório.

No horizonte de previsão de 6 meses (Tabela 5.6), o desempenho relativo dos modelos LNEE em relação aos outros candidatos foi muito superior, especialmente no modelo de cinco fatores. Observa-se também que o modelo DSV de duas etapas e o modelo de 5-GLEE

alcançaram um bom desempenho.

As funções de carga estimadas dos modelos DNS, DSV, GLEE e LNEE são mostradas na Figura 5.6. Da figura, nota-se que funções de carga estimadas dos modelos GLEE apresentam comportamento não suave e instável. Além disso, apesar do modelo GLEE ser inicializado com a decomposição do modelo DSV, as cargas estimadas dos modelos GLEE mostraram ter perdido o comportamento assimptótico original da decomposição de Nelson e Siegel, o que pode dificultar o processo de extrapolação da curva para vencimentos fora do intervalo originalmente considerado.

Ao contrário dos modelos GLEE, as funções de carga dos modelos LNEE de 4 e 5 fatores apresentaram comportamento suave e assimptóticamente estável, que são propriedades importantes das curvas de taxas de juros, como discutido na seção 4.2.1.

Em alguns aspectos, a interpretação dos fatores da decomposição estimada do modelo 4-LNEE é semelhante à do modelo de Nelson, Siegel e Svensson. Por exemplo, os fatores $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ podem ser interpretados como dois fatores de médio prazo com picos em torno dos vencimentos de 18 e 36 meses, embora ϕ_1 apresente uma leve saliência em torno de vencimento mais curtos e taxas de crescimento e decaimento mais rápidas. Outra diferença notável entre os fatores do modelo 4-LNEE e DSV pode ser vista na função de carga ϕ_3, que se assemelha a uma versão espelhada do padrão de decaimento exponencial do modelo DSV.

Na versão de 5 fatores do modelo LNEE, a interpretação torna-se mais complexa. Entretanto, da Figura 5.6 é possível identificar um fator de curto prazo ($\boldsymbol{\beta}_1$) e três tipos diferentes de fatores de médio e longo prazo ($\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ e $\boldsymbol{\beta}_4$). De maneira similar à versão do modelo LNEE de 4 fatores, umas das função de carga de médio prazo também apresenta uma saliência na extremidade curta da curva, o que indica uma correlação entre movimentos de médio e curto prazo neste fator.

Na Figura 5.7, a série temporal dos fatores filtrados de cada modelo é apresentada. É possível verificar que os fatores extraídos dos modelos 4-GLEE e 5-GLEE apresentam grande variabilidade, especialmente no segundo e terceiro fator. Além disso, das decomposições avaliadas, os fatores filtrados dos modelos LNEE exibiram maior estabilidade, o que indica que as decomposições aprendidas pela rede neural do modelo são úteis para a tarefa de previsão.

5.3.3 Análise de Impacto dos Hiperparâmetros da Distribuição Priori

Com o objetivo de avaliar empiricamente o impacto da distribuição priori especificada no modelo LNEE, o experimento de previsão descrito na seção anterior é repetido usando um conjunto variável de hiperparâmetros para o modelo LNEE de quatro fatores.

Devido ao alto número de hiperparâmetros disponíveis, apenas um subconjunto dos hiperparâmetros mais importantes são considerados na análise. Assumindo uma perspectiva prática, os principais hiperparâmetros do modelo LNEE são a variância dos coeficientes da rede neural σ^2_{NN}, que controla diretamente o grau de regularização das funções de carga

geradas, e o hiperparâmetro γ da priori de Minnesota, que controla a força da crença na hipótese de passeio aleatório.

		σ^2_{NN}			
Horizonte (dias)	γ	5×10^{-3}	1×10^{-2}	5×10^{-2}	1×10^{-1}
5	0.1	2.16	2.15	2.14	2.14
	0.5	2.12	2.12	**2.12**	2.13
	0.9	2.13	2.13	2.15	2.12
20	0.1	4.44	4.36	4.30	4.29
	0.5	**4.09**	4.10	4.13	4.12
	0.9	4.14	4.16	4.19	4.16
60	0.1	10.15	9.75	9.52	9.47
	0.5	**8.12**	8.13	8.16	8.15
	0.9	8.20	8.24	8.28	8.27
120	0.1	18.63	17.70	17.16	17.07
	0.5	**13.11**	13.11	13.18	13.17
	0.9	13.18	13.27	13.35	13.37

Tabela 5.7: *RMSE médio das previsões fora de amostra do modelo 4-LNEE de cada combinação dos hiperparâmetros da distribuição Priori* $\gamma \in [0.1, 0.5, 0.9]$ *e* $\sigma^2_{NN} \in [5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}]$, *fixando-se* $(a, b) = (0.1, 0.001)$ *e* $\lambda = 0.5$.

Na Tabela 5.7, os valores da métrica RMSE médio de todos os vencimentos são reproduzidos para cada combinação considerada de hiperparâmetros. Dos resultados experimentais, é possível verificar que a distribuição priori de Minnesota impactou significativamente o desempenho geral das previsões dos modelos, especialmente em horizontes de tempo maiores. Observa-se que a qualidade das previsões obtidas é significativamente menor na penalidade mais forte testada ($\gamma = 0.1$). Entretanto, verifica-se que um valor médio de penalidade ($\gamma = 0.5$) produziu resultados superiores aos obtidos no experimento anterior com uma priori pouco informativa ($\gamma = 0.9$).

Da Tabela 5.7, também é possível identificar que o parâmetro de regularização da rede neural produziu baixo impacto no RMSE médio das previsões. Porém, verifica-se que a penalização mais forte testada $\sigma^2_{NN} = 5 \times 10^{-3}$ produziu ganhos consistentes para os casos em que $\gamma \in \{0.5, 0.9\}$, o que indica que uma regularização mais agressiva pode ser benéfica para o desempenho geral do modelo.

Na Figura 5.8, são representadas as funções de carga estimadas para todos modelos testados. É possível identificar que as decomposições de hiperparâmetros $\gamma \in \{0.5, 0.9\}$ e mesma variância σ^2_{NN} são altamente similares. Além disso, nota-se que o hiperparâmetro de regularização da rede neural σ^2_{NN} afetou principalmente a altura dos picos das funções de carga e a velocidade de decaimento nos vencimentos mais longos.

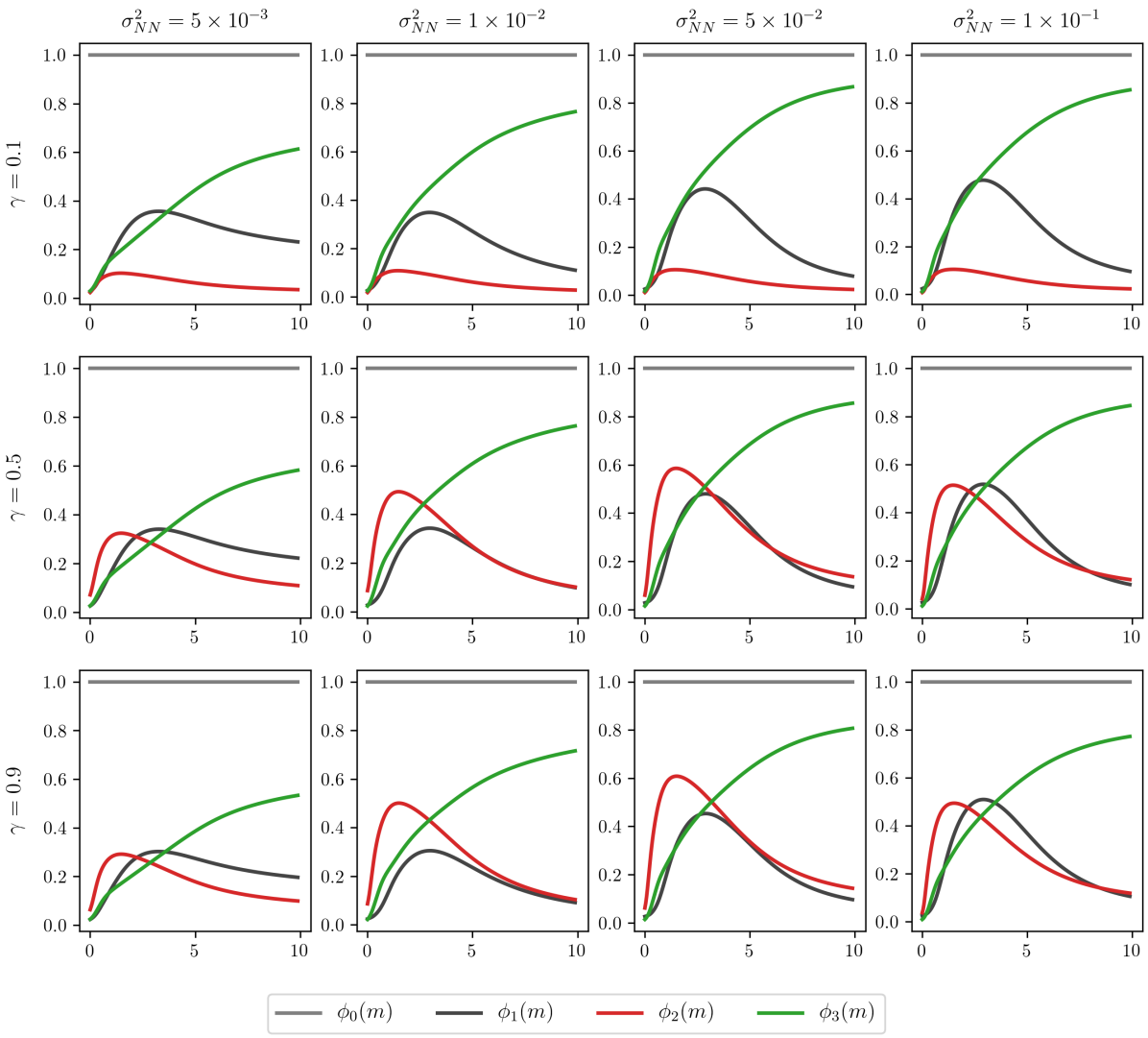

Figura 5.8: *Funções de carga estimadas do modelo 4-LNEE de cada combinação dos hiperparâmetros da distribuição priori* $\gamma \in [0.1, 0.5, 0.9]$ *e* $\sigma^2_{NN} \in [5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}]$*, fixando-se* $(a, b) = (0.1, 0.001)$ *e* $\lambda = 0.5$.

Capítulo 6

Conclusões

Neste trabalho foi proposto um modelo para previsão de curvas de taxas de juros que faz aprendizado de novas decomposições a partir de dados com uma rede neural. No Capítulo 4.2, o modelo proposto foi apresentado, assim como distribuições Priori com efeitos de regularização para os parâmetros do modelo e um procedimento de estimação em uma etapa.

No Capítulo 5.2, o modelo proposto foi avaliado em um experimento com dados artificiais, que mostrou que o modelo foi capaz de estimar com precisão o processo gerador original dos dados e de reproduzir boas propriedades nas decomposições de fatores obtidas.

No Capítulo 5.3, um experimento com 14 anos de dados extraídos das curvas de taxas de juros brasileira mostrou que o modelo proposto obteve resultados superiores a abordagens tradicionais da literatura de previsão de curvas de taxas de juros. Ao analisar as funções de carga estimadas pelo modelo neste experimento, foi possível identificar propriedades vantajosas nas decomposições obitdas, como suavidade, estabilidade assintótica e interpretabilidade dos fatores do modelo. Além disso, uma análise das séries temporais dos fatores filtrados mostrou que os fatores do modelo apresentaram comportamento mais estável e previsível que os competidores testados.

Por último, um experimento para analisar o impacto da distribuição Priori definida sob os parâmetros do modelo foi conduzido na Seção 5.3.3. Do experimento, foi possível verificar que a escolha de hiperparâmetros da distribuição Priori do modelo pode impactar positivamente a qualidade das previsões obtidas. Particularmente, verificou-se que uma penalidade mais forte para os pesos da rede neural ($\sigma^2_{NN} = 5 \times 10^{-3}$) e penalidades intermediárias para os parâmetros da matriz de transição ($\gamma = 0.5$) produziram previsões de melhor qualidade, especialmente para horizontes de tempo maiores.

6.1 Sugestões para Estudos Futuros

As análises abordadas neste trabalho podem ser estendidas para garantir maior robustez de resultados empíricos através de um procedimento validação-cruzada com janelas móveis para estimação e avaliação do modelo. Apesar do custo computacional elevado, este tipo

de análise pode isolar melhor o efeito dos ciclos econômicos nos resultados, que podem ser sensíveis à escolha de período utilizado nas séries temporais de interesse.

Uma outra extensão possível para estudar com maior profundidade os resultados obtidos de modelos baseados em decomposição seria estudar isoladamente o comportamento da série de fatores filtrados da decomposição de cada modelo. Ajustando-se um modelo para prever os fatores filtrado de um dos modelos de decomposição, seria possível quantificar a "previsibilidade"induzida pela escolha do modelo de decomposição, sem recorrer à análise visual do gráfico dos fatores filtrados, como feito neste trabalho.

Dadas as semelhanças diretas, uma ampla gama de extensões desenvolvidas para a classe de modelos Nelson e Siegel dinâmico também podem ser adaptadas para o modelo proposto neste trabalho, como por exemplo a inclusão de variáveis macroeconômicas na equação de evolução de estados (Diebold *et al.* (2006)) e a modelagem simultânea de curvas de múltiplos países (Diebold *et al.* (2008)).

Outra linha de estudo possível refere-se a ampliar as aplicações do modelo proposto a outros tipos de problema. Por exemplo, problemas de elaboração de estratégias de negociação para carteira de produtos de renda fixa, ou problemas de gestão de risco de carteiras de renda fixa. Em particular este último, que foi apontado como uma limitação do modelo proposto devido a dificuldade de realizar inferência completa nos parâmetros do modelo, poderia ser contornado com análises empíricas das distribuições dos erros de previsão em novas amostras de dados, por exemplo.

Uma corrente de pesquisa possível diz respeito as melhorias do método de otimização utilizado para obter estimativas pontuais dos parâmetros do modelo, uma vez que métodos de otimização baseados em gradiente em modelos de espaço de estados podem sofrer de problemas de instabilidade e exigir uma escolha cuidadosa de pontos iniciais. Uma alternativa popular em modelos de fatores dinâmicos é o algoritmo *Expectation-Maximization* (EM) de estimação de parâmetros, visto que este exibe comportamento robusto e usualmente se aproxima rapidamente de uma vizinhança de um ponto ótimo local, apesar da convergência para um ótimo local ser lenta. Como observado por Watson e Engle (1983), o algoritmo EM pode ser utilizado para se obter bons pontos iniciais de um algoritmo de otimização baseado em gradiente, combinando assim os benefícios de ambas abordagens.

Para o modelo proposto neste trabalho, uma solução de estimação híbrida com o algoritmo EM ainda exigiria a utilização de métodos iterativos na etapa de maximização do algoritmo, devido a ausência de soluções analíticas para a estimação de parâmetros de uma rede neural. Apesar disso, uma abordagem híbrida de otimização baseada no algoritmo EM e em métodos baseados em gradiente ainda desfrutaria de ganhos de performance e robustez, dado que o número de execuções do Filtro de Kalman, que é a operação de maior custo computacional do cálculo da função objetivo, seria reduzido significativamente.

Referências

[Allen-Zhu *et al.* 2019a] Zeyuan Allen-Zhu, Yuanzhi Li e Zhao Song. "A convergence theory for deep learning via over-parameterization". Em: *International Conference on Machine Learning*. PMLR. 2019, pgs. 242–252 (citado na pg. 29).

[Allen-Zhu *et al.* 2019b] Zeyuan Allen-Zhu, Yuanzhi Li e Yingyu Liang. "Learning and generalization in overparameterized neural networks, going beyond two layers". Em: *Advances in neural information processing systems*. 2019, pgs. 6158–6169 (citado na pg. 29).

[Ang e Piazzesi 2003] Andrew Ang e Monika Piazzesi. "A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables". Em: *Journal of Monetary economics* 50.4 (2003), pgs. 745–787 (citado na pg. 14).

[Anisimov 2020] Maksim Anisimov. "Forecasting the term structure in emerging markets using extensions of the dynamic nelson-siegel model". Em: *Higher School of Economics Research Paper No. WP BRP* 228 (2020) (citado nas pgs. 21, 38).

[Athans 1974] Michael Athans. "The importance of kalman filtering methods for economic systems". Em: *Annals of Economic and Social Measurement, Volume 3, number 1*. NBER, 1974, pgs. 49–64 (citado na pg. 32).

[Barron 1991] Andrew R Barron. "Complexity regularization with application to artificial neural networks". Em: *Nonparametric functional estimation and related topics*. Springer, 1991, pgs. 561–576 (citado na pg. 38).

[Bingham *et al.* 2019] Eli Bingham *et al.* "Pyro: deep universal probabilistic programming". Em: *The Journal of Machine Learning Research* 20.1 (2019), pgs. 973–978 (citado na pg. 44).

[Bowsher e Meeks 2008] Clive G Bowsher e Roland Meeks. "The dynamics of economic functions: modeling and forecasting the yield curve". Em: *Journal of the American Statistical Association* 103.484 (2008), pgs. 1419–1437 (citado nas pgs. 2, 36).

[Buntine e Weigend 1991] Wray L. Buntine e Andreas S. Weigend. "Bayesian backpropagation". Em: *Complex Systems* 5.6 (1991) (citado na pg. 30).

[Cajueiro *et al.* 2009] Daniel O Cajueiro, Jose A Divino, Benjamin M Tabak *et al.* "Forecasting the yield curve for brazil". Em: *Central Bank of Brazil Working Paper Series* 197 (2009) (citado nas pgs. 35, 49).

[Caldeira *et al.* 2010] Joao Caldeira, Guilherme V Moura e Marcelo Savino Portugal. "Efficient yield curve estimation and forecasting in brazil". Em: *Revista Economia, January/April* (2010) (citado nas pgs. 2, 35, 49).

[Caldeira *et al.* 2016a] João F Caldeira, Guilherme V Moura, André AP Santos e Fabricio Tourrucôo. "Forecasting the yield curve with the arbitrage-free dynamic nelson–siegel model: brazilian evidence". Em: *EconomiA* 17.2 (2016), pgs. 221–237 (citado nas pgs. 2, 21, 35, 38).

[Caldeira *et al.* 2016b] João F Caldeira, Guilherme V Moura e André AP Santos. "Predicting the yield curve using forecast combinations". Em: *Computational Statistics & Data Analysis* 100 (2016), pgs. 79–98 (citado nas pgs. 2, 35).

[Carvalho *et al.* 2009] Carlos M Carvalho, Nicholas G Polson e James G Scott. "Handling sparsity via the horseshoe". Em: *Artificial Intelligence and Statistics*. PMLR. 2009, pgs. 73–80 (citado na pg. 31).

[Chan *et al.* 1979] YT Chan, AGC Hu e JB Plant. "A kalman filter based tracking scheme with input estimation". Em: *IEEE transactions on Aerospace and Electronic Systems* 2 (1979), pgs. 237–244 (citado na pg. 32).

[Choudhry 2019] Moorad Choudhry. *Analysing and interpreting the yield curve*. John Wiley & Sons, 2019 (citado na pg. 8).

[Christensen *et al.* 2009] Jens HE Christensen, Francis X Diebold e Glenn D Rudebusch. *An arbitrage-free generalized Nelson–Siegel term structure model*. 2009 (citado na pg. 1).

[Cox *et al.* 2005] John C Cox, Jonathan E Ingersoll Jr e Stephen A Ross. "A theory of the term structure of interest rates". Em: *Theory of valuation*. World Scientific, 2005, pgs. 129–164 (citado na pg. 1).

[Cybenko 1989] George Cybenko. "Approximation by superpositions of a sigmoidal function". Em: *Mathematics of control, signals and systems* 2.4 (1989), pgs. 303–314 (citado na pg. 27).

[De Pooter 2007] Michiel De Pooter. "Examining the nelson-siegel class of term structure models: in-sample fit versus out-of-sample forecasting performance". Em: *Available at SSRN 992748* (2007) (citado na pg. 2).

[Delle Monache *et al.* 2011] Luca Delle Monache, Thomas Nipen, Yubao Liu, Gregory Roux e Roland Stull. "Kalman filter and analog schemes to postprocess numerical weather predictions". Em: *Monthly Weather Review* 139.11 (2011), pgs. 3554–3570 (citado na pg. 32).

[Diebold e Li 2006] Francis X Diebold e Canlin Li. "Forecasting the term structure of government bond yields". Em: *Journal of econometrics* 130.2 (2006), pgs. 337–364 (citado nas pgs. 1, 2, 20, 21, 35, 36).

[Diebold e Rudebusch 2013] Francis X Diebold e Glenn D Rudebusch. *Yield curve modeling and forecasting: the dynamic Nelson-Siegel approach*. Princeton University Press, 2013 (citado nas pgs. 2, 38).

[Diebold *et al.* 2006] Francis X Diebold, Glenn D Rudebusch e S Boragan Aruoba. "The macroeconomy and the yield curve: a dynamic latent factor approach". Em: *Journal of econometrics* 131.1-2 (2006), pgs. 309–338 (citado nas pgs. 1, 60).

[Diebold *et al.* 2008] Francis X Diebold, Canlin Li e Vivian Z Yue. "Global yield curve dynamics and interactions: a dynamic nelson–siegel approach". Em: *Journal of Econometrics* 146.2 (2008), pgs. 351–363 (citado na pg. 60).

[Duffee 2002] Gregory R Duffee. "Term premia and interest rate forecasts in affine models". Em: *The Journal of Finance* 57.1 (2002), pgs. 405–443 (citado na pg. 14).

[Duffie e Kan 1996] Darrell Duffie e Rui Kan. "A yield-factor model of interest rates". Em: *Mathematical finance* 6.4 (1996), pgs. 379–406 (citado na pg. 1).

[Durbin e Koopman 2012] James Durbin e Siem Jan Koopman. *Time series analysis by state space methods*. Oxford university press, 2012 (citado na pg. 31).

[Faria e Almeida 2018] Adriano Faria e Caio Almeida. "A hybrid spline-based parametric model for the yield curve". Em: *Journal of Economic Dynamics and Control* 86 (2018), pgs. 72–94 (citado nas pgs. 2, 35, 36).

[Filos *et al.* 2019] Angelos Filos *et al.* "A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks". Em: *arXiv preprint arXiv:1912.10481* (2019) (citado na pg. 41).

[Floc'h 2013] Le Floc'h *et al.* "Stable interpolation for the yield curve". Em: *Available at SSRN 2175002* (2013) (citado na pg. 11).

[Fortuna 2010] Eduardo Fortuna. "Mercado financeiro: produtos e serviços." Em: *Rio de Janeiro: Qualitymark* (2010) (citado nas pgs. 5, 7).

[Giannone *et al.* 2015] Domenico Giannone, Michele Lenza e Giorgio E Primiceri. "Prior selection for vector autoregressions". Em: *Review of Economics and Statistics* 97.2 (2015), pgs. 436–451 (citado na pg. 39).

[Hagan e West 2006] Patrick S Hagan e Graeme West. "Interpolation methods for curve construction". Em: *Applied Mathematical Finance* 13.2 (2006), pgs. 89–129 (citado na pg. 11).

[Hays *et al.* 2012] Spencer Hays, Haipeng Shen e Jianhua Z Huang. "Functional dynamic factor models with application to yield curve forecasting". Em: *The Annals of Applied Statistics* (2012), pgs. 870–894 (citado nas pgs. 2, 36).

[Heath *et al.* 1992] David Heath, Robert Jarrow e Andrew Morton. "Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation". Em: *Econometrica: Journal of the Econometric Society* (1992), pgs. 77–105 (citado na pg. 1).

[Hoerl e Kennard 1970] Arthur E Hoerl e Robert W Kennard. "Ridge regression: biased estimation for nonorthogonal problems". Em: *Technometrics* 12.1 (1970), pgs. 55–67 (citado na pg. 30).

[Hördahl *et al.* 2006] Peter Hördahl, Oreste Tristani e David Vestin. "A joint econometric model of macroeconomic and term-structure dynamics". Em: *Journal of Econometrics* 131.1-2 (2006), pgs. 405–444 (citado na pg. 14).

[Hull e White 1990] John Hull e Alan White. "Pricing interest-rate-derivative securities". Em: *The review of financial studies* 3.4 (1990), pgs. 573–592 (citado na pg. 1).

[Hull 2003] John C Hull. *Options futures and other derivatives*. Pearson Education India, 2003 (citado na pg. 10).

[Kauffmann *et al.* 2022] Piero C Kauffmann, Hellinton H Takada, Ana T Terada e Julio M Stern. "Learning forecast-efficient yield curve factor decompositions with neural networks". Em: *Econometrics* 10.2 (2022), pg. 15 (citado na pg. 3).

[Kingma *et al.* 2015] Durk P Kingma, Tim Salimans e Max Welling. "Variational dropout and the local reparameterization trick". Em: *Advances in neural information processing systems* 28 (2015), pgs. 2575–2583 (citado na pg. 31).

[Koopman *et al.* 2010] Siem Jan Koopman, Max IP Mallee e Michel Van der Wel. "Analyzing the term structure of interest rates using the dynamic nelson–siegel model with time-varying parameters". Em: *Journal of Business & Economic Statistics* 28.3 (2010), pgs. 329–343 (citado na pg. 1).

[LeCun *et al.* 2012] Yann A LeCun, Léon Bottou, Genevieve B Orr e Klaus-Robert Müller. "Efficient backprop". Em: *Neural networks: Tricks of the trade*. Springer, 2012, pgs. 9–48 (citado na pg. 37).

[Lee *et al.* 2017] Jaehoon Lee *et al.* "Deep neural networks as gaussian processes". Em: *arXiv preprint arXiv:1711.00165* (2017) (citado na pg. 31).

[Litterman e Scheinkman 1991] Robert Litterman e Jose Scheinkman. "Common factors affecting bond returns". Em: *Journal of fixed income* 1.1 (1991), pgs. 54–61 (citado nas pgs. 16, 17).

[Litterman 1986] Robert B Litterman. "Forecasting with bayesian vector autoregressions—five years of experience". Em: *Journal of Business & Economic Statistics* 4.1 (1986), pgs. 25–38 (citado na pg. 39).

[MacKay 1992] David JC MacKay. "A practical bayesian framework for backpropagation networks". Em: *Neural computation* 4.3 (1992), pgs. 448–472 (citado na pg. 30).

[Maltz 2002] A Maltz. "Estimation of zero coupon curves in datametrics". Em: *RiskMetrics Journal* 3.1 (2002), pgs. 27–39 (citado nas pgs. 11, 50).

[Matthews *et al.* 2018] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner e Zoubin Ghahramani. "Gaussian process behaviour in wide deep neural networks". Em: *arXiv preprint arXiv:1804.11271* (2018) (citado na pg. 31).

[Mineo *et al.* 2020] Eduardo Mineo, Airlane Pereira Alencar, Marcelo Moura e Antonio Elias Fabris. "Forecasting the term structure of interest rates with dynamic constrained smoothing b-splines". Em: *Journal of Risk and Financial Management* 13.4 (2020), pg. 65 (citado nas pgs. 2, 36).

[Neal 1992] Radford M Neal. *Bayesian training of backpropagation networks by the hybrid Monte Carlo method*. Rel. técn. Citeseer, 1992 (citado na pg. 30).

[Neal 1996] Radford M Neal. "Priors for infinite networks". Em: *Bayesian Learning for Neural Networks*. Springer, 1996, pgs. 29–53 (citado na pg. 31).

[Nelson e Siegel 1987] Charles R Nelson e Andrew F Siegel. "Parsimonious modeling of yield curves". Em: *Journal of business* (1987), pgs. 473–489 (citado nas pgs. 1, 17, 18, 20–22, 50).

[Oymak e Soltanolkotabi 2020] Samet Oymak e Mahdi Soltanolkotabi. "Towards moderate overparameterization: global convergence guarantees for training shallow neural networks". Em: *IEEE Journal on Selected Areas in Information Theory* (2020) (citado na pg. 29).

[Pasricha 2006] Gurnain Kaur Pasricha. "Kalman filter and its economic applications". Em: (2006) (citado na pg. 32).

[Paszke *et al.* 2019] Adam Paszke *et al.* "Pytorch: an imperative style, high-performance deep learning library". Em: *arXiv preprint arXiv:1912.01703* (2019) (citado na pg. 44).

[Plaut 1986] David C Plaut *et al.* "Experiments on learning by back propagation." Em: (1986) (citado na pg. 30).

[Rall 1986] L. B. Rall. "The arithmetic of differentiation". Em: *Mathematics Magazine* 59.5 (1986), pgs. 275–282. issn: 0025570X, 19300980. url: http://www.jstor.org/stable/2689402 (citado na pg. 29).

[Safran e Shamir 2018] Itay Safran e Ohad Shamir. "Spurious local minima are common in two-layer relu neural networks". Em: *International Conference on Machine Learning*. PMLR. 2018, pgs. 4433–4441 (citado na pg. 29).

[Särkkä e García-Fernández 2019] Simo Särkkä e Ángel F García-Fernández. "Temporal parallelization of bayesian filters and smoothers". Em: *arXiv preprint arXiv:1905.13002* (2019) (citado nas pgs. 41, 44).

[Suimon *et al.* 2020] Yoshiyuki Suimon, Hiroki Sakaji, Kiyoshi Izumi e Hiroyasu Matsushima. "Autoencoder-based three-factor model for the yield curve of japanese government bonds and a trading strategy". Em: *Journal of Risk and Financial Management* 13.4 (2020), pg. 82 (citado nas pgs. 17, 36).

[Svensson 1994] Lars EO Svensson. *Estimating and interpreting forward interest rates: Sweden 1992-1994*. Rel. técn. National bureau of economic research, 1994 (citado nas pgs. 22, 35).

[Takada e Stern 2015] Hellinton H Takada e Julio M Stern. "Non-negative matrix factorization and term structure of interest rates". Em: *AIP Conference Proceedings*. Vol. 1641. 1. American Institute of Physics. 2015, pgs. 369–377 (citado nas pgs. 2, 17, 35, 36).

[Tikhonov 1963] Andrei Nikolajevits Tikhonov. "Solution of incorrectly formulated problems and the regularization method". Em: *Soviet Math.* 4 (1963), pgs. 1035–1038 (citado na pg. 30).

[Tikhonov 1943] Andrey Nikolayevich Tikhonov. "On the stability of inverse problems". Em: *Dokl. Akad. Nauk SSSR*. Vol. 39. 1943, pgs. 195–198 (citado na pg. 30).

[Vasicek 1977] Oldrich Vasicek. "An equilibrium characterization of the term structure". Em: *Journal of financial economics* 5.2 (1977), pgs. 177–188 (citado na pg. 1).

[Vicente e Tabak 2008] José Vicente e Benjamin M Tabak. "Forecasting bond yields in the brazilian fixed income market". Em: *International Journal of Forecasting* 24.3 (2008), pgs. 490–497 (citado na pg. 49).

[Watson e Engle 1983] Mark W Watson e Robert F Engle. "Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models". Em: *Journal of Econometrics* 23.3 (1983), pgs. 385–400 (citado na pg. 60).

[Weng *et al.* 2006] Shiuh-Ku Weng, Chung-Ming Kuo e Shu-Kang Tu. "Video object tracking using adaptive kalman filter". Em: *Journal of Visual Communication and Image Representation* 17.6 (2006), pgs. 1190–1208 (citado na pg. 32).

[Wengert 1964] R. E. Wengert. "A simple automatic derivative evaluation program". Em: *Commun. ACM* 7.8 (ago. de 1964), 463–464. ISSN: 0001-0782. DOI: 10.1145/355586.364791. URL: https://doi.org/10.1145/355586.364791 (citado na pg. 29).

[Williams 1995] Peter M Williams. “Bayesian regularization and pruning using a laplace prior”. Em: *Neural computation* 7.1 (1995), pgs. 117–143 (citado na pg. 30).

[Xiang e Zhu 2013] Ju Xiang e Xiaoneng Zhu. “A regime-switching nelson–siegel term structure model and interest rate forecasts”. Em: *Journal of Financial Econometrics* 11.3 (2013), pgs. 522–555 (citado na pg. 1).