UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Leandro Augusto Lacerda Campos

UMA ABORDAGEM GENERATIVA PROFUNDA PARA PREVISÃO DE DENSIDADE DE MÚLTIPLOS PASSOS

Belo Horizonte
2022

Leandro Augusto Lacerda Campos

UMA ABORDAGEM GENERATIVA PROFUNDA PARA PREVISÃO DE DENSIDADE DE MÚLTIPLOS PASSOS

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Fabrício Murai Ferreira
Coorientador: Cristiano Arbex Valle

Belo Horizonte
2022

Campos, Leandro Augusto Lacerda.

C198u Uma abordagem generativa profunda para previsão de densidade de múltiplos passos [manuscrito] / Leandro Augusto Lacerda Campos. — 2022.
94 f. il.; 29 cm.

Orientador: Fabrício Murai Ferreira.
Coorientador: Cristiano Arbex Valle.
Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Referências: f. 69-78.

1. Computação – Teses. 2. Redes neurais (Computação) – Teses. 3. Densidade - Previsão - Teses. 4.Testes de hipóteses estatísticas – Teses. I. Ferreira, Fabrício Murai. II. Valle, Cristiano Arbex. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. IV. Título.

CDU 519.6*85 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma abordagem generativa profunda para previsão de densidade de múltiplos passos

LEANDRO AUGUSTO LACERDA CAMPOS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. FABRÍCIO MURAI FERREIRA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. CRISTIANO ARBEX VALLE - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ROBERTO HIRATA JUNIOR
Departamento de Ciência da Computação - USP

PROF. BRUNO RIBEIRO
Departamento de Ciência da Computação - Purdue University

Belo Horizonte, 18 de julho de 2022.

Agradecimentos

Aos familiares e amigos, eu agradeço por me incentivarem a perseguir meus sonhos e por compreenderem a minha ausência enquanto eu me dedicava a esse empreendimento. E aos meus orientadores, eu sou grato por me ensinarem, por meio de conversas e de exemplos, o que é ser um pesquisador, e por me darem liberdade para exercer minha criatividade e dar vida às minhas ideias.

"*If the market, in effect, does not predict its fluctuations, it does assess them as being more or less likely, and this likelihood can be evaluated mathematically.*"

(Louis Bachelier)

Resumo

Nesta dissertação, nós introduzimos um modelo generativo profundo para previsão de densidade de múltiplos passos em sequências univariadas de retornos de ações. Ele é baseado em uma rede neural recorrente Bayesiana que opera sob a premissa de que a distribuição condicional dos retornos pertence a uma família de distribuições t de Student assimétrica. Também apresentamos medidas de acurácia preditiva baseadas na Transformação Integral da Probabilidade e na Discrepância Média Máxima (MMD) para avaliar, separadamente, a normalidade incondicional e a independência serial dos pseudo-resíduos normais gerados por modelos de distribuição condicional univariada. Usando essas medidas, propomos então um procedimento Bayesiano para comparar, conjuntamente, o desempenho de dois ou mais modelos concorrentes em múltiplas sequências univariadas. Esse procedimento controla automaticamente a probabilidade de um modelo ser declarado mais preciso do que outro por mera sorte e não por de fato possuir habilidade preditiva superior. Experimentos com dados reais mostram que introduzir caudas pesadas e assimetria na densidade condicional do modelo que nós propomos fornece melhorias significativas de precisão. Eles também demonstram que nosso modelo é melhor do que as alternativas convencionais em capturar as principais características distribucionais e da estrutura de dependência observadas em sequências de retornos.

Palavras-chave: previsão de densidade, redes neurais recorrentes bayesianas, avaliação de previsões de densidade, pseudo-resíduos, discrepância média máxima, teste de hipótese bayesiano, múltiplas comparações.

Abstract

In this dissertation, we introduce a deep generative model for multi-step density forecasting over univariate sequences of stock returns. It is based on a Bayesian recurrent neural network that operates under the premise that the conditional distribution of returns belongs to a family of skew Student's t-distributions. We also present measures of predictive accuracy based on Probability Integral Transformation and Maximum Mean Discrepancy (MMD) to separately evaluate the unconditional normality and the serial independence of the normal pseudo-residuals generated by univariate conditional distribution models. Using these measures, we then propose a Bayesian procedure to jointly compare the performance of two or more competing models over multiple univariate sequences. This procedure automatically controls the probability that one model will be declared more accurate than another because of luck, not because it has superior predictive ability. Experiments on real datasets show that introducing fat tails and skewness into the conditional density of our model we propose provides significant accuracy improvements. They also demonstrate that our model is better than mainstream alternatives in capturing the main distributional and dependency structure characteristics observed in sequences of returns.

Keywords: density forecasting, bayesian recurrent neural networks, density forecast evaluation, pseudo-residuals, maximum mean discrepancy, bayesian hypothesis testing, multiple comparisons.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

Nos últimos anos, aprendizado profundo tem ocupado o centro das atenções por conta de avanços significativos em problemas que eram considerados de difícil solução por métodos tradicionais de aprendizado de máquina [108], por exemplo reconhecimento de fala e de objetos, tradução de textos, geração de imagens e direção autônoma de veículos. Porém, a despeito da existência de aplicações bem-sucedidas na área, ainda não é consenso que modelos de aprendizado profundo alcançarão, em finanças, o mesmo sucesso que tem sido obtido em outros domínios [49, 28]. De fato, trata-se de um domínio bastante diferente desses. Em particular, cabe destacar duas das principais diferenças: a baixa razão sinal-ruído e o tamanho relativamente pequeno dos conjuntos de dados.

A razão sinal-ruído é uma medida da previsibilidade de um sistema. Na tarefa de previsão de retornos de uma ação de empresa, a baixa razão sinal-ruído pode ser explicada como consequência da hipótese do mercado eficiente. Com efeito, a forma fraca (semi-forte) dessa hipótese afirma que toda a informação disponível publicamente sobre uma determinada ação já está refletida em seu preço (e que esse preço muda instantaneamente para refletir novas informações tornadas públicas). Segue-se daí que não é possível prever o seu preço futuro a partir de dados públicos, uma vez que eles não contêm sinais que indiquem que a ação está precificada incorretamente.

É razoável questionar a validade da hipótese do mercado eficiente. Suponha, então, que ela é inválida e que, portanto, existe algum sinal, de acesso não restrito, que indique que o preço de uma ação é diferente do seu valor intrínseco. Uma vez que esse sinal é detectado, investidores atuarão de modo a explorar essa oportunidade de retorno garantido até exauri-la, isto é, até o preço convergir para o valor intrínseco. E considerando os recursos tecnológicos e financeiros à disposição de certos grupos de investidores, essa convergência ocorrerá em um curto intervalo de tempo. Logo, para efeitos práticos, podemos assumir que a dita hipótese é válida pelo menos na sua forma fraca [63], especialmente para ações de empresas que apresentam grande volumes de negociação e que são constantemente monitoradas por analistas e investidores, tais como aquelas que compõem o índice S&P 500.

Sob a hipótese do mercado eficiente na sua forma fraca, é aceitável supor que as oscilações diárias no preço de uma ação são respostas a eventos imprevisíveis. Por isso

modelos de aprendizado profundo geralmente produzem resultados insatisfatórios quando usados para prever a esperança condicional do próximo retorno diário de uma ação a partir de dados históricos. De fato, em tais circunstâncias, a melhor previsão para o próximo retorno diário com base nas informações até então disponíveis é que ele será igual a zero [91]. Esse fato parece excluir qualquer chance de previsão da dinâmica dos preços futuros de uma ação. Porém, análises empíricas mostram que os retornos diários não são independentes e que, portanto, existem padrões não-triviais nessa dinâmica que podem ser aprendidos. A boa notícia é que podemos empregar modelos generativos profundos para aprender a distribuição conjunta desses retornos e assim capturar outras características além da esperança condicional. Com uma boa aproximação dessa distribuição, é possível usar técnicas de finanças quantitativas, tais como diversificação e *hedging*, para reduzir a aleatoriedade dos mercados financeiros e gerenciar riscos associados a investimentos.

Outro desafio relacionado à aplicação de aprendizado profundo em finanças diz respeito ao tamanho das bases de dados. Parte do progresso recente em aprendizado profundo se deve à abundância de dados [108, 40], o que não ocorre em finanças da mesma forma que em outros domínios nos quais aprendizado profundo tem alcançado resultados estado-da-arte. Por exemplo, na tarefa de previsão de retornos diários de uma determinada ação, qualquer modelo está limitado ao número de observações da variável de interesse: retornos dessa ação dia após dia. E novas observações só são geradas com o passar do tempo. Se o conjunto de treinamento é pequeno, modelos de aprendizado profundo tendem ao sobre-ajuste, especialmente quando os dados são ruidosos [109, 38]. Nesse cenário, existem infinitas combinações de arquiteturas e parâmetros de redes neurais que se ajustam bem ao conjunto de treinamento, porém poucas, quando existem, conseguem generalizar de forma satisfatória.

Por causa dessas diferenças entre finanças e os domínios nos quais aprendizado profundo tem alcançado bons resultados, nós propomos combinar modelos generativos profundos e a abordagem Bayesiana para redes neurais com o objetivo de capturar tanto a natureza estocástica dos mercados financeiros quanto a incerteza decorrente de estados de mercado considerados inéditos ou raros com respeito a conjuntos de dados relativamente pequenos. O nosso modelo, que nós chamamos de DeepRisk, baseia-se em uma rede neural recorrente Bayesiana especializada no problema de previsão de densidade de múltiplos passos da distribuição condicional dos retornos diários de uma ação de empresa. Ele opera sob a premissa de que essa distribuição pertence a uma família de distribuições t de Student assimétrica. E a distribuição de probabilidade sobre seus pesos é aprendida de forma global a partir dos dados históricos de múltiplas sequências univariadas[1] de retornos, o que nos permite aumentar a sua complexidade e, potencialmente, a sua acurácia

[1]Nesse trabalho, nós dizemos que uma sequência de retornos $\{r_t\}$ é univariada quando o seu t-ésimo termo representa o retorno de uma única ação no passo de tempo t. Por sua vez, dizemos que ela é multivariada quando o seu t-ésimo termo é um vetor cuja i-ésima coordenada representa o retorno da ação $i = 1, \ldots, N$ no passo de tempo t, onde $N > 1$ é o número de ações associadas à sequência.

preditiva, sem sobre-ajuste [89].

É do nosso interesse comparar a calibragem condicional do DeepRisk com a de alternativas convencionais para avaliar se a adoção de um modelo mais complexo é compensada pelo ganho em acurácia preditiva. Além disso, consideramos importante comparar diferentes configurações do nosso modelo, cada uma delas adotando uma família de distribuição distinta para modelar a densidade condicional dos retornos, de modo a estimar o impacto da introdução de caudas pesadas e de assimetria nessa densidade. Para tanto, propomos utilizar a Transformação Integral da Probabilidade para calcular os pseudo-resíduos normais gerados pelos diferentes modelos para cada sequência univariada de retornos diários observados. Também apresentamos métricas baseadas na Discrepância Média Máxima (MMD) para medir os desvios dos pseudo-resíduos em relação às condições ideais de normalidade incondicional e de independência serial. Para realizar a comparação de modelos em múltiplas sequências univariadas de retornos, introduzimos um procedimento Bayesiano que admite a possibilidade de desempenhos correlacionados entre modelos rivais e que controla automaticamente a ocorrência de falsas descobertas.

Resumimos, a seguir, as principais contribuições do nosso trabalho:

- Mostramos a importância de utilizar uma família de distribuições assimétricas e de caudas pesadas para modelar a densidade condicional dos retornos diários de uma ação de empresa.

- Propomos uma abordagem baseada na Transformação Integral da Probabilidade e na Discrepância Média Máxima (MMD) para avaliar a calibragem condicional de previsões de densidade de múltiplos passos.

- Introduzimos um procedimento Bayesiano para comparar múltiplos modelos de previsões de densidade em várias sequências univariadas de retornos.

- Demonstramos que o nosso modelo é, em geral, melhor do que seus concorrentes tradicionais em capturar as principais propriedades distribucionais e da estrutura de dependência que nós observamos em sequências de retornos diários; nos piores cenários, ele é pelo menos tão bom quanto.

- Analisamos as previsões do DeepRisk para uma ação de empresa específica e mostramos como o nosso modelo responde adequadamente a diferentes situações de mercado.

Capítulo 2

Modelo Generativo Profundo

Esse capítulo tem por objetivo propor um modelo generativo profundo para previsão de densidade de múltiplos passos da distribuição condicional dos retornos diários de uma ação de empresa negociada em alguma bolsa de valores. O modelo proposto, chamado DeepRisk, baseia-se em uma rede neural recorrente Bayesiana implementada para capturar as principais propriedades distribucionais e da estrutura de dependência que nós observamos em sequências de retornos diários: caudas pesadas, assimetria entre perdas e ganhos, agrupamento de volatilidade, dependência de longo prazo e efeito de alavancagem.

2.1 Revisão de Literatura

Para introduzir conceitos, notações e terminologias utilizadas na especificação do nosso modelo, o DeepRisk, nós revisamos nessa seção algumas arquiteturas e técnicas de aprendizado profundo. Também apresentamos um levantamento de aplicações recentes de modelos generativos profundos em finanças.

2.1.1 Redes Neurais Recorrentes

Redes neurais recorrentes (RNNs, do inglês *Recurrent Neural Networks*) são uma família de redes neurais caracterizadas pela existência de laços em seus grafos computacionais [44]. Esses laços permitem modelar, qualquer que seja $t \in \mathbb{N}$, a relação entre uma sequência finita $x_1, \ldots, x_t$ de variáveis em $\mathbb{R}^m$ e uma variável y_t em $\mathbb{R}^n$. Com essa arquitetura, uma rede neural recorrente pode aproximar, tão bem quanto se queira, qualquer função mensurável $f : (\mathbb{R}^m)^* \to \mathbb{R}^n$ definida no conjunto $(\mathbb{R}^m)^* \equiv \bigcup_{i=1}^{\infty} (\mathbb{R}^m)^i$ de todas as sequências finitas de vetores em $\mathbb{R}^m$ [50]. Em particular, as RNNs são adequadas para

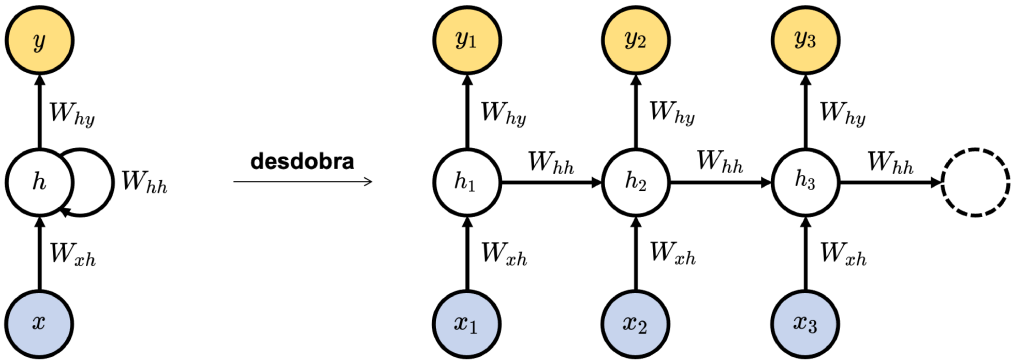


Figura 2.1: Rede neural recorrente simples. *(Esquerda)* Grafo computacional com laço. *(Direita)* Grafo computacional desdobrado no tempo, no qual a mesma coleção de pesos é utilizada em todos os passos de tempo. Fonte: Elaborado pelo autor, com inspiração em [40].

tarefas relacionadas a modelagem de séries temporais [89, 92, 99, 88].

Considere, por exemplo, a arquitetura de uma rede neural recorrente simples, representada na Figura 2.1. Essa RNN pode ser descrita pelo seguinte par de equações:

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h)\,, \tag{2.1}$$

$$y_t = \psi(W_{hy}h_t + b_y)\,, \tag{2.2}$$

onde $x_t \in \mathbb{R}^m$, $h_t \in \mathbb{R}^q$ e $y_t \in \mathbb{R}^n$ são a entrada, o estado latente exposto e a saída da rede no passo de tempo $t \in \mathbb{N}$, respectivamente; $W_{xh} \in \mathbb{R}^{q\times m}$, $W_{hh} \in \mathbb{R}^{q\times q}$ e $b_h \in \mathbb{R}^q$ são os pesos da camada oculta; $W_{hy} \in \mathbb{R}^{n\times q}$ e $b_y \in \mathbb{R}^n$ são os pesos da camada de saída; e $\phi(\cdot)$ e $\psi(\cdot)$ são as funções de ativação das camadas oculta e de saída, nessa ordem. Geralmente, definimos $h_0 = 0 \in \mathbb{R}^q$. Para tornar essa arquitetura profunda, basta empilhar múltiplas camadas ocultas, de modo que a saída de uma seja a entrada da subsequente, conforme indicado por [42].

É comum abordar uma rede neural recorrente como um modelo probabilístico. Dadas uma sequência $x_1, \ldots, x_t$ do espaço de entrada e um valor w para a sua coleção de pesos, uma RNN define uma probabilidade $p(y_t|x_1, \ldots, x_t, w)$ sobre o conjunto dos possíveis valores y_t do espaço de saída. Dessa forma, podemos utilizar o método de estimação de máxima verossimilhança para ajustar seus pesos:

$$\begin{aligned} w^{\text{MLE}} &= \arg\max_w \log p(\mathcal{D}|w) \\ &= \arg\max_w \sum_{i=1}^{N}\sum_{t=1}^{T(i)} \log p\left(y_t^{(i)}|x_1^{(i)}, \ldots, x_t^{(i)}, w\right), \end{aligned} \tag{2.3}$$

onde $\mathcal{D}$ é um conjunto de dados formado por pares de sequências de entrada e de saída, N é o número de pares nesse conjunto, e $T(i)$ é o tamanho das sequências do i-ésimo par. Se a função $\log p(\mathcal{D}|w)$ é derivável com respeito aos pesos, então esse ajuste pode ser feito aplicando o algoritmo de retropropagação no tempo. Esse algoritmo é uma extensão do

algoritmo de retropropagação padrão e opera sob a premissa de que uma RNN pode ser desdobrada no tempo em uma rede alimentada adiante.

Em sequências longas, a aplicação do algoritmo de retropropagação no tempo pode resultar no conhecido problema de explosão e extinção de gradiente [54], ao qual se atribui a dificuldade das redes neurais recorrentes em aprender dependências de longo prazo. Esse problema é próprio das RNNs [40], uma vez que ele ocorre por causa da profundidade que um grafo computacional desdobrado no tempo pode alcançar e da utilização da mesma coleção de pesos em todos os passos de tempo. A seguir, apresentamos três soluções parciais para esse problema: a arquitetura LSTM (do inglês *Long Short Term-Memory*) de [55], as conexões de salto dilatadas de [12] e as conexões residuais de [103].

LSTM Para todo passo de tempo t, a arquitetura LSTM introduz um estado latente interno $c_t \in \mathbb{R}^q$ para armazenar informações complementares àquelas do estado latente exposto h_t. Portas de entrada i_t, de esquecimento f_t e de saída o_t, todas em $\mathbb{R}^q$, definem a dinâmica do par (c_t, h_t):

$$i_t = \sigma(W_i[x_t, h_{t-1}]^\top + b_i)\,, \tag{2.4}$$
$$f_t = \sigma(W_f[x_t, h_{t-1}]^\top + b_f)\,, \tag{2.5}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[x_t, h_{t-1}]^\top + b_c)\,, \tag{2.6}$$
$$o_t = \sigma(W_o[x_t, h_{t-1}]^\top + b_o)\,, \tag{2.7}$$
$$h_t = o_t \odot \tanh(c_t)\,, \tag{2.8}$$

onde as funções $\sigma(\cdot)$ e $\tanh(\cdot)$ são a sigmoide e a tangente hiperbólica, respectivamente; o operador $\odot$ é o produto de Hadamard; e as matrizes W_i, W_f, W_c, W_o em $\mathbb{R}^{q\times(m+q)}$ e os vetores b_i, b_f, b_c, b_o em $\mathbb{R}^q$ são os pesos da rede. Tipicamente, definimos c_0 e h_0 como o vetor nulo em $\mathbb{R}^q$. Apesar dos pesos da LSTM serem constantes no tempo, as suas portas tornam a atualização e a exposição do seu estado latente dependentes do contexto, o que evita a ocorrência do problema de explosão e extinção de gradiente.

Conexões de salto dilatadas Suponha uma rede neural recorrente simples com L camadas ocultas, $L \geqslant 1$. Para $l = 1, \ldots, L$, dizemos que a l-ésima camada oculta é dotada de conexão de salto com dilatação $s^{[l]}$ quando a sua evolução é descrita por:

$$h_t^{[l]} = \phi\left(W_{xh}^{[l]} x_t^{[l-1]} + W_{hh}^{[l]} h_{t-s^{[l]}}^{[l]} + b_h^{[l]}\right)\,, \tag{2.9}$$

onde $x_t^{[l-1]}$ é a entrada da camada no passo de tempo t, com $x_t^{[0]} \equiv x_t$. Analogamente, podemos introduzir uma conexão de salto dilatada em uma camada oculta com arquitetura LSTM: basta alterar o atraso dos estados latentes interno e exposto de $t-1$ para $t - s^{[l]}$ no lado direito das Equações (2.4)-(2.8). Em geral, definimos a dilatação de modo que ela cresça exponencialmente em função da posição da camada:

$$s^{[l]} = B^{l-1}\,, l = 1, \ldots, L\,, \tag{2.10}$$

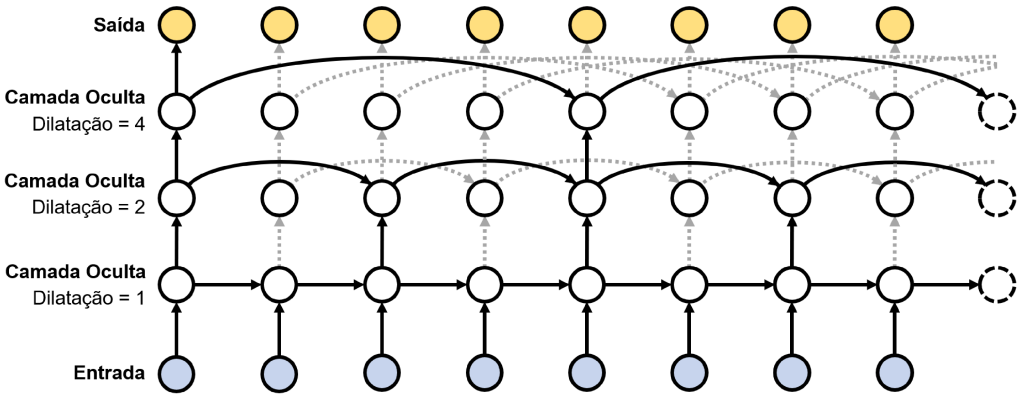


Figura 2.2: Rede neural recorrente simples composta por três camadas ocultas dotadas de conexões de salto com dilatações 1, 2 e 4. Fonte: Elaborado pelo autor, com inspiração em [12].

onde $B \in \{2, 3, 4, \dots\}$ é chamado fator de dilatação. A Figura 2.2 ilustra uma RNN exponencialmente dilatada. Para [12], definir a dilatação dessa forma reduz a distância média entre estados latentes de diferentes passos de tempo, o que previne a ocorrência de explosão e extinção de gradiente.

Conexões residuais Mesmo utilizando conexões de salto exponencialmente dilatadas, não é verdade que o desempenho de uma rede neural recorrente pode sempre melhorar conforme empilhamos mais camadas ocultas. Segundo [103], o desempenho melhora até uma determinada altura, para além da qual a rede se torna lenta e difícil de treinar. Para evitar esse problema, comumente associado à explosão e extinção de gradiente, esses autores propuseram as conexões residuais ilustradas na Figura 2.3. Para apresentá-las formalmente, considere uma rede neural recorrente simples com $l - 1$ camadas ocultas, $l \geqslant 2$. Suponha que nós desejamos adicionar-lhe mais uma camada oculta sem incorrer no risco de piorar seu desempenho. Nesse caso, nós podemos usar os estados latentes expostos $h_t^{[l]}$ para modelar diretamente a contribuição dessa camada para a rede, definindo:

$$x_t^{[l]} = h_t^{[l]} + x_t^{[l-1]}. \tag{2.11}$$

onde $x_t^{[l-1]}$ e $x_t^{[l]}$ são, respectivamente, as entradas e as saídas da camada adicional. Note que, sem conexão residual, teríamos $x_t^{[l]} = h_t^{[l]}$, pela definição da arquitetura profunda dessa rede que foi mencionada na Seção 2.1.1. De modo geral, dizemos que a l-ésima camada oculta de uma RNN simples tem conexão residual quando a sua saída é definida pela Equação (2.11). Munida de tal conexão, essa camada toma a forma da função identidade nos contextos em que ela não contribui para melhorar o desempenho da rede, o que nos dá ideia de como a conexão residual evita que a adição de camadas ocultas dificulte o fluxo de retropropagação de gradiente.

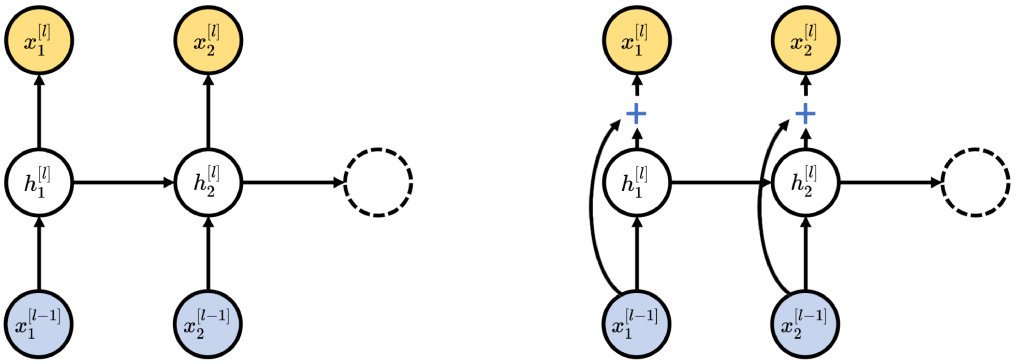


Figura 2.3: Comparação entre camadas ocultas de uma rede neural recorrente com respeito ao emprego de conexão residual. *(Esquerda)* Sem conexão residual. *(Direita)* Com conexão residual, representada pela aresta conectando a entrada da camada ao operador de adição em todo passo de tempo t. Fonte: Elaborado pelo autor.

2.1.2 Redes Neurais Bayesianas

O objetivo do treinamento clássico é encontrar o melhor valor para os pesos de uma rede neural, tipicamente por meio da maximização da sua função de log-verossimilhança. Na inferência Bayesiana, por outro lado, o que se quer é inferir a distribuição a posteriori $p(w|\mathcal{D})$ sobre os pesos w após observar os dados $\mathcal{D}$. A vantagem em atribuir uma probabilidade aos pesos é melhorar a generalização da rede ao expressar incerteza em regiões do espaço de entrada para os quais há poucos ou nenhum exemplo de treinamento.

Chamamos de rede neural Bayesiana aquela na qual os pesos são modelados como variáveis aleatórias ao invés de valores pontuais. As predições dessa rede para um novo exemplo de teste x são, então, dadas pelo modelo médio Bayesiano:

$$\begin{aligned} p(y|x) &= \mathbb{E}_{p(w|\mathcal{D})}[p(y|x,w)] \\ &= \int p(y|x,w)p(w|\mathcal{D})dw\,, \end{aligned} \tag{2.12}$$

onde $p(y|x,w)$ é a distribuição preditiva para um dado valor de pesos w. Como o próprio nome induz, podemos interpretar essa equação como a média das predições geradas por um conjunto não-enumerável de redes neurais, ponderadas por uma medida de adequação entre os valores de pesos e os exemplos de treinamento.

Devido à forma funcional de uma rede neural e à dimensionalidade de seus pesos, a integral da Equação (2.12) não pode ser solucionada analiticamente. Segue daí a necessidade de recorrer a um método escalável de inferência aproximada. Nessa seção, apresentamos um procedimento de inferência variacional chamado BBB (do inglês *Bayes by Backprop*), que foi proposto por [9] a partir dos trabalhos de [43] e [52]. Outros métodos de inferência aproximada consistem, por exemplo, em multiplicar os pesos por variáveis aleatórias independentes com distribuição de Bernoulli [35], em considerar múl-

tiplos valores de pesos obtidos ao longo de trajetórias de descida estocástica de gradiente [102, 72, 61], e em aplicar um método Hamiltoniano de Monte Carlo (HMC) baseado em decomposições simétricas [17].

O BBB é um procedimento compatível com o algoritmo de retropropagação padrão que tem por objetivo encontrar o parâmetro η que minimiza a divergência de Kullback-Leibler (KL) entre uma distribuição variacional $q(w|\eta)$ e a verdadeira distribuição a posteriori sobre os pesos:

$$\begin{aligned} \eta^\star &= \arg\min_\eta D_{\mathrm{KL}}[q(w|\eta)\|p(w|\mathcal{D})] \\ &= \arg\min_\eta \int q(w|\eta)\log\frac{q(w|\eta)}{p(w)p(\mathcal{D}|w)}dw \\ &= \arg\min_\eta D_{\mathrm{KL}}[q(w|\eta)\|p(w)] - \mathbb{E}_{q(w|\eta)}[\log p(\mathcal{D}|w)]\,, \end{aligned} \tag{2.13}$$

onde $p(w)$ é a distribuição a priori sobre os pesos antes de observar os dados $\mathcal{D}$. A função de perda implícita é denominada energia livre variacional ou limite inferior esperado, e minimizá-la é equivalente a maximizar a função de log-verossimilhança de uma rede neural sujeito a um termo de complexidade sobre os seus pesos.

2.1.3 Redes Neurais Generativas e Aplicações em Finanças

Na seção anterior, nós descrevemos uma forma de atribuir probabilidade aos pesos de uma rede neural para expressar incerteza sobre qual valor de pesos é apropriado. Nessa seção, por sua vez, apresentamos como modelar a incerteza sobre a qual saída associar cada exemplo do espaço de entrada. Para ilustrar a diferença entre as duas incertezas, considere o experimento de lançar uma determinada moeda e observar a face superior. A primeira incerteza é sobre a probabilidade de obter cara. Ela está relacionada ao fato de termos pouca ou nenhuma informação sobre a moeda utilizada; sendo assim, essa incerteza pode diminuir a cada repetição do experimento. A segunda é sobre a face que vamos observar. Essa incerteza é inerente à natureza estocástica do experimento.

Uma rede neural é chamada generativa quando ela define, explícita ou implicitamente, uma função de probabilidade no espaço de saída. Considere a tarefa de calcular, para uma determinada ação, a probabilidade $p(r_{t+1}|r_1,\ldots,r_t)$ de observar o retorno r_{t+1} dados os retornos passados $r_1,\ldots,r_t$, com $t \geqslant 1$. Se assumirmos que essa probabilidade pertence a uma família paramétrica conhecida, a tarefa em questão se resume a encontrar o valor do parâmetro θ_{t+1} tal que $p(r_{t+1}|r_1,\ldots,r_t) = p(r_{t+1}|\theta_{t+1})$. Nesse caso, a função de probabilidade pode ser avaliada explicitamente. Por outro lado, podemos optar por não fazer qualquer suposição quanto à distribuição condicional dos retornos e amostrar de

uma distribuição aprendida, definida implicitamente, que possui propriedades estatísticas semelhantes às da verdadeira, desconhecida.

De acordo com [2], a abordagem mais comumente empregada para modelar a distribuição condicional dos retornos consiste em assumir que ela é Gaussiana com média zero. Sob essa hipótese, a evolução dos retornos é univocamente determinada pela dinâmica do desvio padrão condicional, também chamado volatilidade. Para prever a volatilidade futura a partir dos retornos passados, [106], [71] e [13] propõem modelos de variáveis latentes baseados em redes neurais recorrentes e em inferência variacional. Adotando um horizonte de previsão de um dia e métricas baseadas na log-verossimilhança, experimentos com múltiplas séries de retornos diários sugerem a superioridade das redes neurais em relação às alternativas tradicionais: modelos da família GARCH (do inglês *Generalized Autoregressive Conditional Heteroskedasticity*) [10] e de volatilidade estocástica [97]. Em particular, resultados experimentais de [110] indicam a adequação da RNN dilatada apresentada na Seção 2.1.1.

Ainda sobre previsão de volatilidade, alguns trabalhos investigam a utilização de dados alternativos, complementares aos retornos passados, como entrada de modelos baseados em redes neurais. [64] testam empregar o valor estimado dos parâmetros de um ou mais modelos da família GARCH; cada um desses modelos foi projetado para capturar diferentes características da estrutura de dependência das séries de retornos. [90] e [107], por sua vez, usam áudio e texto de eventos de divulgação de resultados para prever a volatilidade durante períodos de curto e longo prazos após a realização de tais eventos. Resultados experimentais indicam que a inclusão de variáveis exógenas pode proporcionar ganhos em acurácia preditiva, medida em termos da diferença entre a volatilidade predita e o desvio padrão dos retornos diários observados durante um determinado período.

Embora conveniente, a hipótese de normalidade dos retornos não condiz com os dados observados. De fato, a densidade incondicional empírica do retorno diário de uma ação apresenta um pico mais acentuado em torno da média e suas caudas decrescem a uma velocidade muito menor do que ocorreria caso a distribuição condicional verdadeira fosse Gaussiana [86, 19, 91]. Em particular, nós observamos mais valores extremos do que o esperado quando fazemos tal suposição. Nesse contexto, é mais apropriado trabalhar com distribuições paramétricas com caudas mais pesadas, tais como t de Student, normal inversa Gaussiana, hiperbólica generalizada, Johnson SU e lambda generalizada. Outra opção é evitar assumir essa e outras hipóteses sobre a distribuição dos retornos, empregando modelos que são totalmente orientados a dados como aqueles baseados em redes adversárias generativas (GANs, do inglês *Generative Adversarial Networks*).

As GANs foram propostas por [41] e se baseiam em um jogo hipotético no qual uma rede neural generativa precisa competir contra uma rede adversária. A aplicação delas em finanças é recente e um caso de uso típico consiste em gerar simulações para enriquecer os conjuntos de dados que são utilizados para ajustar, testar e agregar modelos

ou estratégias de investimento [68, 81]. Resultados empíricos obtidos por [101] e [96] sugerem que modelos baseados em GANs são capazes de aprender as principais características distribucionais e da estrutura de dependência de séries de retornos. Em particular, [101] utilizam a função W de Lambert para facilitar o aprendizado de distribuições com caudas pesadas, casos em que as GANs geralmente não apresentam bons resultados [59].

2.2 Definição do Problema

Seja S_t o preço de mercado de um ativo financeiro no dia t. Assuma que esse ativo é uma ação de empresa negociada em alguma bolsa de valores. Passado um intervalo k de um ou mais dias, é provável que o preço S_{t+k} seja diferente de S_t e que essa diferença tenha sido causada por eventos de alguma forma relacionados com a empresa subjacente. Ao invés de tentar prever a ocorrência de todo evento que pode afetar o preço da ação, faz mais sentido modelar $S_t = S_t(\omega)$ como uma variável aleatória definida em algum espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$, encapsulando em $\omega \in \Omega$ todas as possíveis situações de mercado. É comum munir esse espaço de uma filtragem $\{\mathcal{F}_t, t \geqslant 0\}$ de σ-álgebras tal que $\mathcal{F}_t \subseteq \mathcal{F}_u \subseteq \mathcal{F}$ para $t \leqslant u$. Nós definimos $\mathcal{F}_0 = \{\varnothing, \Omega\}$ e interpretamos $\mathcal{F}_t$, $t \geqslant 1$, como a informação sobre o mercado que está disponível a um observador no dia t. Portanto, é natural assumir que S_t é $\mathcal{F}_t$-mensurável.

Uma forma conveniente de descrever a dinâmica dos preços $\{S_t, t \geqslant 0\}$ é por meio de seus retornos logarítmicos $\{r_t, t \geqslant 1\}$, onde $r_t = r_t(\omega)$ é uma variável aleatória $\mathcal{F}_t$-mensurável definida por:

$$r_t = \log(S_t) - \log(S_{t-1})\,. \tag{2.14}$$

Dado S_0, existe uma correspondência biunívoca entre a sequência de preços e a de retornos. De fato, segue da Equação (2.14) que

$$S_t = S_0 \exp\left(r_1 + \cdots + r_t\right), \quad t \geqslant 1\,. \tag{2.15}$$

Sendo assim, podemos escolher trabalhar com $\{r_t\}$ no lugar de $\{S_t\}$. A principal razão para fazer tal escolha é o fato de que os retornos são o resultado da aplicação, nos preços, de um procedimento comum em análise de séries temporais [84]. Esse procedimento tem por objetivo transformar uma sequência não estacionária de termos positivos em outra que pode ser modelada como fracamente estacionária.

Dado $t \geqslant 0$, seja $F_{t+1}(r) \equiv \mathbb{P}(r_{t+1} \leqslant r|\mathcal{F}_t)$ a função de distribuição condicional do retorno r_{t+1} dada a informação $\mathcal{F}_t$. Chamamos a sequência $\{F_{t+1}(r)\}$ de processo gerador de dados. Tendo em vista que o processo verdadeiro é desconhecido qualquer que

seja a ação de empresa subjacente, o nosso objetivo nesse capítulo é propor um modelo generativo profundo que induza uma boa aproximação de $\{F_{t+1}(r)\}$. Em particular, nós queremos que esse modelo capture propriedades não-triviais observadas em realizações desse processo. Essas propriedades são conhecidas como fatos estilizados e as principais delas são apresentadas na Figura 2.4 e descritas a seguir:

Caudas pesadas. A densidade incondicional empírica dos retornos r_{t+1} apresenta caudas que decrescem a uma velocidade muito menor do que ocorreria caso a distribuição condicional verdadeira fosse Gaussiana.

Assimetria entre perdas e ganhos. Grandes oscilações negativas no preço de uma ação são mais frequentes do que grandes variações positivas, o que sugere a possibilidade de assimetria na distribuição condicional dos retornos.

Agrupamento de volatilidade. Retornos de grande (pequena) magnitude tendem a ser seguidos por retornos de grande (pequena) magnitude. Em outras palavras, eventos de alta (baixa) volatilidade tendem a se agrupar no tempo, o que indica que os desvios padrões condicionais $\mathrm{SD}(r_{t+1}|\mathcal{F}_t) \equiv \sqrt{\mathrm{Var}(r_{t+1}|\mathcal{F}_t)}$ apresentam valores positivos de autocorrelação ao longo de vários dias.

Dependência de longo prazo. A correlação empírica entre as variáveis $|r_t|$ e $|r_{t+k}|$ é significativa quando $k \in \mathbb{N}$ é pequeno e ela se aproxima lentamente de zero, a uma taxa hiperbólica, conforme k aumenta. Uma sequência fracamente estacionária que tem essa característica é dita ter memória longa [91].

Efeito de alavancagem. A correlação empírica entre os retornos r_t e os desvios padrões condicionais $\mathrm{SD}(r_{t+1}|\mathcal{F}_t)$ tende a ser negativa. Esse fenômeno explica a tendência de crescimento (queda) da volatilidade após a ocorrência de um dia de negociação com retorno negativo (positivo).

2.3 Especificação do Modelo

Suponha que $\mathcal{G}_t$ é a σ-álgebra gerada pela sequência de retornos de uma ação de empresa até o dia t. Como $\mathcal{G}_t \subseteq \mathcal{F}_t$ e $\mathcal{G}_t \subseteq \mathcal{G}_{t+1}$, temos que $\{\mathcal{G}_t\}$ é uma sub-filtragem de $\{\mathcal{F}_t\}$. Para $t \geqslant 1$, o nosso objetivo é modelar uma densidade condicional $p(r_{t+1}|\mathcal{G}_t)$ para o retorno r_{t+1} tal que a função de distribuição condicional induzida,

$$\hat{F}_{t+1}(r) \equiv \int_{-\infty}^{r} p(x|\mathcal{G}_t)dx\,, \tag{2.16}$$

aproxime a verdadeira, porém desconhecida, função de distribuição condicional $F_{t+1}(r)$. Em nosso modelo, o qual chamamos de DeepRisk, nós adotamos uma forma paramétrica para a densidade condicional e a representamos genericamente por:

$$p(r_{t+1}|\theta(r_1,\ldots,r_t;w))\,, \tag{2.17}$$

onde w é uma coleção finita de pesos e $\theta_{t+1} \equiv \theta(r_1,\ldots,r_t;w)$ é um vetor de parâmetros dinâmico no tempo. Dado um valor de pesos w, esse modelo opera sob a premissa de que θ_{t+1} é observável no dia t; mais formalmente, que θ_{t+1} é $\mathcal{G}_t$-mensurável.

Para dar forma à densidade condicional $p(r_{t+1}|\theta_{t+1})$, nós utilizamos uma família de locação-escala com parâmetros que são capazes de capturar características distribucionais não-triviais comuns em sequências de retornos, a saber, caudas pesadas e assimetria entre perdas e ganhos. Em particular, nós escolhemos a distribuição t de Student assimétrica de [30] com a seguinte parametrização:

$$\mu_{t+1} = (\theta_{t+1})_1\,, \tag{2.18}$$
$$\sigma^2_{t+1} = \log\{1+\exp[(\theta_{t+1})_2]\}^2\,, \tag{2.19}$$
$$\nu_{t+1} = \log\{1+\exp[(\theta_{t+1})_3]\}+2\,, \tag{2.20}$$
$$\gamma_{t+1} = \log\{1+\exp[(\theta_{t+1})_4]\}\,, \tag{2.21}$$

onde $(\theta_{t+1})_i$, $i=1,\ldots,4$, é a i-ésima coordenada do vetor de parâmetros do dia $t+1$, μ_{t+1} é a esperança condicional $\mathbb{E}(r_{t+1}|\mathcal{G}_t)$, σ^2_{t+1} é a variância condicional $\mathrm{Var}(r_{t+1}|\mathcal{G}_t)$, ν_{t+1} é o parâmetro que ajusta o peso das caudas, e γ_{t+1} é o parâmetro que controla o grau de assimetria. Como $\nu_{t+1} > 2$ por definição, μ_{t+1} e σ^2_{t+1} existem e estão bem definidas. É importante destacar que a esperança e a variância condicionais de r_{t+1} com respeito a $\mathcal{G}_t$ são apenas aproximações dos respectivos momentos condicionais verdadeiros. Com efeito, tendo em vista a nossa definição de $\mathcal{G}_t$, é razoável supor que $\mathcal{G}_t \subsetneq \mathcal{F}_t$ para $t \geqslant 1$; em outras palavras, o nosso modelo explora somente um subconjunto próprio da informação relevante sobre o mercado.

Para descrever a dinâmica dos vetores de parâmetros θ_{t+1}, nós implementamos uma rede neural recorrente com L camadas ocultas seguida de uma camada densa (função afim) sem ativação:

$$\theta_{t+1} = W_{x\theta}x^{[L]}_{t+1} + b_\theta\,, \tag{2.22}$$

onde $x^{[L]}_{t+1}$ são as saídas da RNN, e $W_{x\theta}$ e b_θ são os pesos específicos dessa camada. Em outras palavras, dado um valor para a coleção de pesos w, essa rede define θ_{t+1} como uma função dos retornos passados $r_1,\ldots,r_t$. Para a hipótese de t assumir um valor grande, nós aplicamos na rede a arquitetura LSTM. Também a munimos de conexões de salto exponencialmente dilatadas entre os passos de tempo, usando fator de dilatação igual a B, e de conexões residuais a partir da segunda camada oculta quando $L \geqslant 2$. Acreditamos que a não-linearidade e a capacidade de memória dessa rede são capazes de capturar várias

propriedades da estrutura de dependência das sequências de retornos diários, tais como agrupamento de volatilidade, dependência de longo prazo e efeito de alavancagem.

Finalmente, nós atribuímos à coleção de pesos w uma distribuição de probabilidade para expressar incerteza quanto a estados de mercado considerados inéditos ou raros com respeito ao conjunto de dados $\mathcal{D}$. Dessa forma, a densidade condicional de r_{t+1} é dada pelo seguinte modelo médio Bayesiano:

$$\begin{aligned} p(r_{t+1}|\mathcal{G}_t) &\equiv \mathbb{E}_{p(w|\mathcal{D})}[p(r_{t+1}|\theta(r_1,\ldots,r_t;w))] \\ &= \int p(r_{t+1}|\theta(r_1,\ldots,r_t;w))p(w|\mathcal{D})dw\,. \end{aligned} \tag{2.23}$$

Nós aproximamos a verdadeira distribuição a posteriori $p(w|\mathcal{D})$ sobre os pesos por meio de uma distribuição variacional $q(w|\eta)$. Para especificar essa distribuição, nós definimos $\eta = (\mu, \rho)$ e adotamos uma normal multivariada diagonal com vetor de médias μ e vetor de variâncias $\sigma^2 = \log(1 + \exp(\rho))^2$. Seguindo [9], nós escolhemos uma mistura de duas Gaussianas com médias iguais a zero para a priori dos pesos:

$$p(w) = \prod_j \pi\mathcal{N}(w_j|0,\sigma_1^2) + (1-\pi)\mathcal{N}(w_j|0,\sigma_2^2)\,, \tag{2.24}$$

onde w_j é o j-ésimo peso da rede, $\pi \in (0,1)$ é o peso da primeira densidade componente, e σ_1^2 e σ_2^2 são as variâncias da primeira e da segunda componente, respectivamente. Definindo $\sigma_1 > \sigma_2$ e $\sigma_2 \ll 1$, nós damos a essa mistura um formato leptocúrtico: pico mais acentuado em torno de zero e caudas mais pesadas do que uma curva normal. Uma priori definida dessa forma mantém muitos pesos concentrados em torno de zero; ao mesmo tempo, ela torna provável que alguns pesos assumam valores relativamente grandes, com sinal positivo ou negativo.

2.3.1 Inferência Variacional e Aprendizagem

Suponha disponível, para treinamento, um conjunto de dados formado pela sequência de retornos diários de N ações de empresas. Para cada uma dessas ações, nós geramos múltiplos exemplos de treinamento escolhendo diferentes datas da sequência original para o passo de tempo $t = 1$. Cada exemplo é um par ordenado composto por uma sequência de entrada $r_1,\ldots,r_T$ e outra de saída $r_2,\ldots,r_{T+1}$, ambas com tamanho fixo de T dias, onde $T \gg 1$ é um hiperparâmetro. O conjunto $\mathcal{D}$ contendo os exemplos de treinamento é então aleatoriamente particionado em uma coleção finita de mini-lotes: $\mathcal{D}^{(1)},\ldots,\mathcal{D}^{(C)}$. A essa coleção aplicamos o procedimento de inferência aproximada chamado BBB, apresentado na Seção 2.1.2.

O BBB tem por objetivo encontrar o par de parâmetros $\eta = (\mu, \rho)$ que minimiza a energia livre variacional $\mathcal{L}(\mathcal{D}, \eta)$ do modelo, a função objetivo da Equação (2.13). Para trabalhar com mini-lotes de exemplos de treinamento, [43] propôs a seguinte adaptação dessa função de perda:

$$\mathcal{L}^{(i)}\left(\mathcal{D}^{(i)}, \eta\right) = \frac{1}{C} D_{\mathrm{KL}}[q(w|\eta) \| p(w)] - \mathbb{E}_{q(w|\eta)}[\log p(\mathcal{D}^{(i)}|w)]\,, \tag{2.25}$$

onde $\mathcal{D}^{(i)}$ é o i-ésimo mini-lote, com $i = 1, \ldots, C$. Como $\sum_i \mathcal{L}^{(i)}\left(\mathcal{D}^{(i)}, \eta\right) = \mathcal{L}(\mathcal{D}, \eta)$, minimizar a Equação (2.25) para todos os mini-lotes é equivalente a minimizar a energia livre variacional do modelo [9]. Para evitar cálculos envolvendo integrais, nós invocamos a definição da divergência KL e utilizamos simulação de Monte Carlo para obtermos o valor aproximado dessa equação:

$$\mathcal{L}^{(i)}\left(\mathcal{D}^{(i)}, \eta\right) \approx \frac{1}{M_w} \sum_{m=1}^{M_w} \left\{ \frac{1}{C} \log \frac{q\left(w^{(m)}|\eta\right)}{p\left(w^{(m)}\right)} - \log p\left(\mathcal{D}^{(i)}|w^{(m)}\right) \right\}, \tag{2.26}$$

onde $w^{(1)}, \ldots, w^{(M_w)}$, $M_w \geqslant 1$, são valores de peso amostrados de $q(w|\eta)$. O Algoritmo 2.1 descreve o procedimento de treinamento apresentado nesse parágrafo.

Algoritmo 2.1: Procedimento de treinamento

Defina $\sigma = \log(1 + \exp(\rho))$.
for $i = 1$ *até* C **do**
  for $m = 1$ *até* M_w **do**
    Amostre $\epsilon^{(m)} \sim \mathcal{N}(0, I)$.
    Defina $w^{(m)} = \mu + \sigma \odot \epsilon^{(m)}$.
    Calcule $q\left(w^{(m)}|\eta\right) \equiv \mathcal{N}(w^{(m)}|\mu, \sigma^2)$.
    Calcule $p\left(w^{(m)}\right)$ via Equação (2.24).
    Calcule $\log p\left(\mathcal{D}^{(i)}|w^{(m)}\right)$ a partir da Equação (2.27).
  end
  Aproxime a perda $\mathcal{L}^{(i)}\left(\mathcal{D}^{(i)}, \eta\right)$ usando a Equação (2.26).
  Calcule o gradiente com respeito aos parâmetros μ e ρ.
  Atualize os parâmetros μ e ρ.
end

Com relação à função de verossimilhança do modelo, nós nos baseamos na reparametrização proposta por [105] para a função de densidade da distribuição t de Student assimétrica de [30]:

$$p(r_t|\theta_t) = \frac{2}{\gamma_t + \frac{1}{\gamma_t}} f(z_t \gamma_t^{-\operatorname{sgn} z_t}|\nu_t) \frac{\tilde{\sigma}_t}{\sigma_t}\,, \quad t \geqslant 2\,, \tag{2.27}$$

onde

$$f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \tag{2.28}$$

$$z_t = \tilde{\mu}_t + \frac{r_t - \mu_t}{\sigma_t}\,\tilde{\sigma}_t\,, \tag{2.29}$$

$$\tilde{\mu}_t = M_{1,t}\left(\gamma_t + \frac{1}{\gamma_t}\right), \tag{2.30}$$

$$\tilde{\sigma}_t^2 = (1 - M_{1,t}^2)\left(\gamma_t^2 + \frac{1}{\gamma_t^2}\right) + 2M_{1,t}^2 + 1\,, \tag{2.31}$$

$$M_{1,t} = \frac{2}{\nu_t - 1}\,\frac{\sqrt{\pi(\nu_t-2)}\Gamma\left(\frac{\nu_t}{2}\right)}{\Gamma\left(\frac{\nu_t+1}{2}\right)}\,. \tag{2.32}$$

Note que $f(\cdot|\nu)$ é a função de densidade da distribuição t de Student padronizada com ν graus de liberdade, $\operatorname{sgn}(\cdot)$ é a função sinal e $(\mu_t, \sigma_t, \nu_t, \gamma_t)$ é o vetor de parâmetros distribucionais definido pelas Equações (2.18)-(2.21) em função do vetor θ_t.

2.3.2 Previsão de Múltiplos Passos

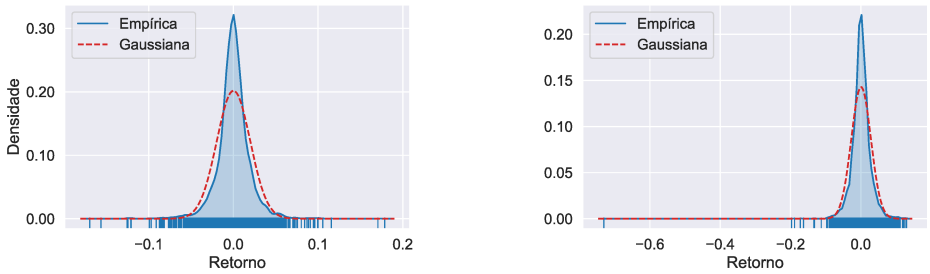
Dados os retornos diários de uma ação de empresa até o dia t, o DeepRisk define a seguinte distribuição preditiva para a sequência de retornos dos próximos $K \geqslant 1$ dias, onde K é o maior horizonte de previsão que nos interessa:

$$p(r_{t+1}, \ldots, r_{t+K}|r_1, \ldots, r_t) = \mathbb{E}_{q(w|\eta)}\left[\prod_{k=1}^{K} p(r_{t+k}|\theta(r_1, \ldots, r_{t+k-1}; w))\right]. \tag{2.33}$$

A esperança da Equação (2.33) não pode ser solucionada analiticamente. Porém, nós podemos estimá-la usando simulação de Monte Carlo, seguindo o procedimento que está descrito a seguir. Amostre um valor de pesos w da distribuição variacional $q(w|\eta)$ e então gere um valor para o retorno r_{t+1} da distribuição $p(r_{t+1}|\theta(r_1, \ldots, r_t; w))$. Se $K = 1$, o procedimento termina aqui. Do contrário, suponha simulada uma sequência de valores para os retornos $r_{t+1}, \ldots, r_{t+k-1}$, onde $1 < k \leqslant K$. Para amostrar um valor para o retorno r_{t+k}, tome um exemplo da distribuição $p(r_{t+k}|\theta(r_1, \ldots, r_{t+k-1}; w))$. Especificamos, dessa forma, um procedimento recursivo para obter uma realização da distribuição preditiva. Repita essa procedimento um número M_r grande de vezes.

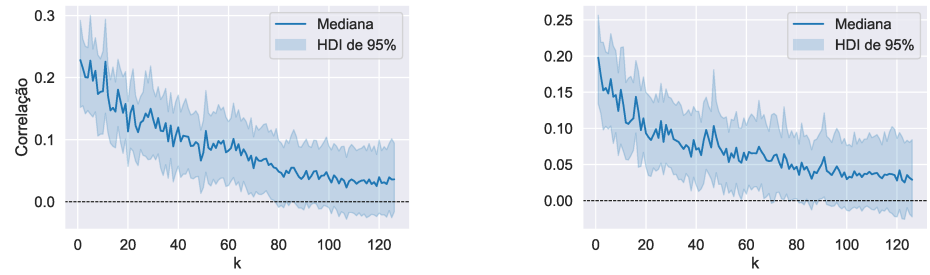
Uma desvantagem do procedimento descrito no parágrafo anterior é o fato dele não tirar proveito do poder de computação paralela em massa oferecido pelas GPUs (unidades de processamento gráfico). Para contornar essa deficiência, podemos aplicá-lo em um mini-lote de dados composto por M_c cópias dos retornos diários observados. Dessa forma,

cada uma das M_r repetições do procedimento produz, paralelamente, M_c previsões para os próximos K dias. Note, porém, que a escolha de $M_c > 1$ quando o tamanho da amostra é mantido constante está relacionada a um custo de oportunidade: um valor grande aumenta a paralelização mas pode comprometer a exploração da distribuição variacional $q(w|\eta)$ sobre os pesos e fazer com que a distribuição simulada não aproxime bem a distribuição preditiva do modelo. Para se ter uma ideia de como a paralelização pode comprometer a exploração da distribuição variacional, pense no caso mais extremo, onde $M_c = M_r$: nesse caso, nós amostramos uma única coleção de pesos w da distribuição variacional $q(w|\eta)$ para simular todas as M_r sequências de retornos para os próximos K dias.

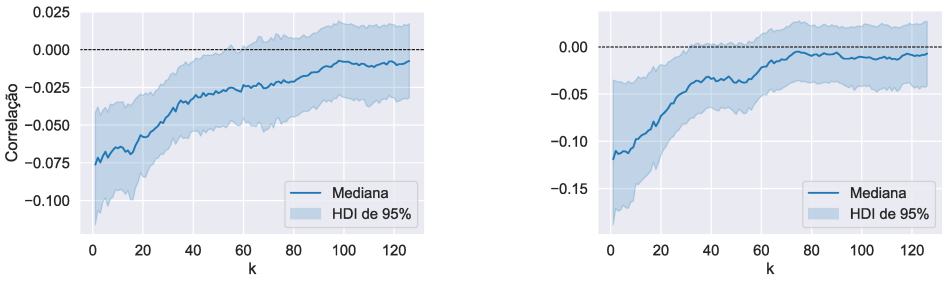


(a) Caudas pesadas e assimetria entre perdas e ganhos

(b) Agrupamento de volatilidade



(c) Dependência de longo prazo



(d) Efeito de alavancagem

Figura 2.4: Fatos estilizados com dados diários da Microsoft (esquerda) e da Apple (direita) no período de 03/01/2000 a 30/12/2016. *(a)* Densidade incondicional empírica dos retornos. *(b)* Sequência de retornos. *(c)* Função de autocorrelação para a sequência de retornos absolutos. *(d)* Correlação empírica entre retorno e volatilidade k dias adiante. Medianas e intervalos de mais alta densidade (HDIs) foram calculados por *bootstrap* estacionário. Volatilidades foram estimadas pelo modelo EGARCH(1,1) com distribuição t de Student. Fonte: Elaborado pelo autor.

Capítulo 3

Comparação de Modelos em Múltiplas Sequências

O objetivo do presente capítulo é propor um procedimento Bayesiano para comparação conjunta da habilidade preditiva fora da amostra de dois ou mais modelos de previsão de densidade em múltiplas sequências univariadas de retornos de ações. Na Seção 3.1, nós revisamos alguns dos principais testes na literatura de finanças para comparação de distribuições. Na Seção 3.2, propomos duas medidas baseadas na Transformação Integral da Probabilidade e na Discrepância Média Máxima (MMD) para avaliar a acurácia preditiva de modelos de densidade condicional. E na Seção 3.3, apresentamos e analisamos um teste de habilidade preditiva equivalente baseado em uma abordagem Bayesiana hierárquica.

3.1 Comparação de Modelos de Previsão de Densidade em Finanças

Previsão de densidade é uma atividade central nas finanças modernas. Por exemplo, para precificar ativos e selecionar carteiras de investimento, usamos distribuições para modelar a variabilidade observada em variáveis econômicas e para expressar incerteza quanto à situação corrente dos mercados. E para estimar perdas associadas a eventos adversos de alto impacto mas baixa chance de ocorrência, nós empregamos modelos probabilísticos para calcular medidas de risco tais como *Value at Risk* (VAR) e *Expected Shortfall* (ES). Por conta dessas e outras aplicações, a avaliação de modelos de previsão de densidade tem ocupado posição de destaque na agenda de pesquisas em finanças e economia. Nessa linha, podemos citar as contribuições de [85], [67], [1], [8], [26] e [25]. Nos próximos parágrafos, dedicamos nossa atenção a alguns testes fora da amostra que nos permitem comparar, dada uma medida de acurácia, a habilidade preditiva de dois ou mais modelos de previsão de densidade concorrentes.

Dados um grupo de $M \geqslant 2$ modelos de previsão de densidade possivelmente mal especificados e uma medida de acurácia preditiva, [4] e [21] propõem testes estatísticos que nos permitem identificar o modelo que fornece a melhor aproximação fora da amostra da densidade condicional de uma sequência de retornos de ação. Se $M = 2$, então a comparação do par de modelos é direta: os autores testam a hipótese nula de habilidades preditivas equivalentes contra as alternativas de que um dos rivais é superior ao outro. Quando $M > 2$, eles seguem [51] e [100] em espírito, apontando um dos concorrentes como referência e testando a hipótese nula de que nenhum dos outros modelos fornece uma aproximação mais precisa da densidade condicional verdadeira contra a alternativa de que pelo menos um dos rivais supera a referência. A comparação em pares, na qual nenhum modelo precisa ser destacado como referência, segue como um caso particular.

Ao executarmos muitas comparações em pares usando uma mesma coleção de conjuntos de dados, nós enfrentamos um problema conhecido como mineração de dados ou espionagem de dados (em inglês, *data mining* ou *data snooping*, respectivamente). Esse problema diz respeito à possibilidade de um modelo ser declarado mais preciso do que outro por mera sorte e não por de fato possuir habilidade preditiva superior. Quanto maior essa possibilidade, maior a probabilidade de rejeitarmos incorretamente a hipótese nula de acurácias equivalentes; chamamos esse tipo de erro de falsa descoberta. Para ilustrar, suponha que um procedimento de avaliação relativa de modelos envolve C comparações independentes de pares de modelos. E assuma que o nível de significância de cada teste de hipótese é α. Então a probabilidade de ocorrer pelo menos uma falsa descoberta é dada por $\alpha_C = 1 - (1 - \alpha)^C$. Conforme C cresce, α_C tende rapidamente para 1. Por exemplo, quando $C = 10$ e $\alpha = 0{,}05$, temos $\alpha_C \approx 0{,}40$. Sendo assim, precisamos adotar alguma técnica para controlar α_C, como a correção de Bonferroni e outras que foram propostas por [53] e [57], ou algum método para controlar a proporção esperada de falsas descobertas, como aqueles apresentados por [7] e [6].

A medida de acurácia preditiva empregada por [4] é o critério de informação de Kullback–Leibler (KLIC). Segundo os próprios autores, o emprego do KLIC equivale ao uso do negativo da função de log-verossimilhança (NLL), medida de acurácia que conduz à escolha do modelo que, na média, atribui maior probabilidade a eventos que de fato já ocorreram. Por sua vez, [21] introduzem uma medida de acurácia preditiva que é o análogo distribucional do erro quadrático médio, denotada aqui por dMSE. Na fórmula dessa medida, substituímos a distribuição condicional verdadeira do retorno r_{t+k}, que é desconhecida, pelo seu estimador dado pela função indicadora do evento $[r_{t+k} \leqslant r]$, onde $k \geqslant 1$ e $r \in \mathbb{R}$. De acordo com [21], não há vantagens claras de uma medida sobre a outra. Porém, esses autores destacam que o dMSE facilita a implementação de procedimentos nos quais o interesse é medir a acurácia preditiva de intervalos de confiança ou de previsões de densidade mas em regiões específicas do suporte. Para [4], o KLIC é conceitualmente mais simples e computacionalmente mais eficiente.

Nenhum dos procedimentos anteriores foi especificado para comparar múltiplos modelos em várias sequências de retornos de ação. Para esse problema, podemos usar um dos seguintes testes que foram sugeridos por [24]: o teste ANOVA de medidas repetidas [31] e o teste de Friedman [33]. Como o primeiro teste é baseado em suposições bem restritivas sobre os dados, tais como normalidade e esfericidade, nós o preterimos em benefício do segundo. No teste de Friedman, nós classificamos os modelos para cada conjunto de dados, sendo a primeira posição ocupada pelo rival com a maior acurácia preditiva e assim sucessivamente. Nos casos de empate, atribuímos a posição média. Em seguida, para cada modelo avaliado, nós tomamos a média das posições sobre os conjuntos de dados. E então testamos a hipótese nula de habilidades preditivas equivalentes contra a alternativa de que pelo menos um rival tem acurácia diferente dos demais. Quando a hipótese nula é rejeitada, nós podemos empregar o teste de Nemenyi [79] e comparar todos os pares de modelos. Sob esse teste, as acurácias preditivas de dois modelos são significativamente diferentes se suas posições médias distam por pelo menos um valor crítico devidamente ajustado para controlar o erro de falsas descobertas ao se realizar $M(M-1)/2$ comparações, onde M é o número de modelos avaliados. [104] usam o NLL para medir acurácia preditiva e os testes de Friedman e de Nemenyi para comparar modelos de previsão de densidade em sequências de retornos de ação.

3.2 Medidas de Acurácia Preditiva

O objetivo dessa seção é propor duas medidas de acurácia preditiva complementares para comparação de modelos de densidade condicional. Para introduzi-las, considere um modelo de previsão de densidade de $k \geqslant 1$ passos adiante e a sequência de retornos diários realizados $\{r_t\}$ de uma ação de empresa. Seja $\hat{F}_{t+k}$ a função de distribuição condicional do retorno r_{t+k} que é induzida por esse modelo com respeito a $\mathcal{G}_t$, a sigma-álgebra gerada pelos retornos passados $r_1, \ldots, r_t$. Ao final de cada dia, nós podemos calcular qual é a probabilidade $u_{t+k} \equiv \hat{F}_{t+k}(r_{t+k})$, sob esse modelo, de observarmos um retorno menor do que o retorno r_{t+k} realizado. Chamamos as probabilidades u_{t+k}, $t \geqslant 1$, de pseudo-resíduos e a transformação que associa, a cada retorno realizado, o pseudo-resíduo correspondente, de Transformação Integral da Probabilidade.

Para simplificar, suponha que o horizonte de previsão é $k = 1$. Se o modelo em análise é o correto[1] e, para $t \geqslant 1$, $\mathcal{G}_t$ contém toda informação relevante para a formação de r_{t+1}, então os pseudo-resíduos são condicionalmente imprevisíveis. Do contrário, nós po-

[1] Nesse trabalho, dizemos que um modelo é o correto quando não podemos rejeitar a hipótese de que ele coincide com a verdadeira, porém desconhecida, função de distribuição condicional dos dados.

deríamos usar $\mathcal{G}_t$ para prever u_{t+1} e, então, construir um modelo melhor do que o correto, o que seria uma contradição. Logo, sob as hipóteses de corretude do modelo e de completude dos $\mathcal{G}_t$'s, os pseudo-resíduos $\{u_{t+1}\}$ devem ser variáveis aleatórias independentes e identicamente distribuídas de $\mathcal{U}(0,1)$ [25]. Verificar se os pseudo-resíduos satisfazem a essas condições pode ser problemático devido ao suporte limitado da distribuição uniforme [16, 2]. Por conta disso, tornou-se comum transformá-los nos pseudo-resíduos normais:

$$z_{t+1} \equiv \Phi^{-1}(u_{t+1}) = \Phi^{-1}\left[\hat{F}_{t+1}(r_{t+1})\right], \tag{3.1}$$

onde Φ^{-1} é a inversa da função de distribuição da normal padrão $\mathcal{N}(0,1)$. Segundo [8], a sequência $\{u_{t+1}\}$ é independente e identicamente distribuída de $\mathcal{U}(0,1)$ se, e somente se, a sequência $\{z_{t+1}\}$ é independente e identicamente distribuída de $\mathcal{N}(0,1)$. Por conta dessa equivalência, nós podemos optar por focar nossa atenção nos pseudo-resíduos normais.

Quando o horizonte de previsão é $k > 1$, a sequência $\{z_{t+k}\}$ de pseudo-resíduos normais apresenta dependência serial mesmo sob as hipóteses de corretude do modelo e de completude dos $\mathcal{G}_t$'s. Para se ter uma ideia sobre essa dependência serial, considere as previsões de horizonte $k > 1$ geradas em dois dias consecutivos: r_{t+k}, gerada no dia t, e $r_{(t+1)+k}$, gerada no dia $t+1$. Existe uma interseção de $k-1$ dias entre os horizontes das duas previsões; essa interseção sugere a dependência entre os respectivos pseudo-resíduos z_{t+k} e $z_{(t+1)+k}$. De modo geral, dado $k \geqslant 1$, se o modelo em análise é o correto e, para $t \geqslant 1$, a sigma-álgebra $\mathcal{G}_t$ contém toda informação relevante que está disponível a um observador no dia t, então $\{z_{t+k}\}$ é identicamente distribuída de $\mathcal{N}(0,1)$ e todo par (z_{s+k}, z_{t+k}) de termos da sequência é independente quaisquer que sejam $s, t \geqslant 1$ tais que $|s-t| \geqslant k$ [67]. Chamamos essas duas condições necessárias de normalidade incondicional e independência serial, respectivamente.

Um modelo satisfaz a condição de normalidade incondicional com respeito a uma sequência de retornos quando as funções de distribuição condicional induzidas estão, na média, corretamente calibradas. A seguir, descrevemos uma situação que ilustra como a violação da normalidade incondicional pode levar um investidor a subestimar riscos. Dada uma ação de empresa, suponha que a distribuição incondicional verdadeira de seus retornos diários é uma t de Student com média, desvio padrão e graus de liberdade iguais a 0, 2 e 4, respectivamente. Suponha também que um investidor usa uma normal com iguais média e desvio padrão para modelar esses retornos. Nesse cenário, o investidor observa, em particular, uma quantidade maior de retornos negativos de grande magnitude do que aquela esperada sob o modelo por ele escolhido. Por exemplo, definindo $u = 0{,}01$, a proporção de dias nos quais o pseudo-resíduo normal observado é menor do que $\Phi^{-1}(u) \approx -2{,}6486$ supera a fração prometida u; o valor esperado dessa proporção é aproximadamente igual a 0,015. Caso o investidor possua ações dessa empresa, ele corre mais risco de falir do que ele estima correr por meio de seu modelo. Se, por outro lado, o modelo escolhido fosse o verdadeiro, essa proporção seria igual à fração prometida u para

todo $u \in (0,1)$.

E um modelo satisfaz a condição de independência serial com respeito a uma sequência de retornos se, para calibrar a função de distribuição condicional $\hat{F}_{t+k}$, $t \geqslant 1$, ele usa corretamente toda informação relevante que está disponível para um observador no dia t. A situação hipotética descrita a seguir mostra que a normalidade incondicional não é suficiente e que a independência serial também é uma característica desejável em um modelo de previsão de densidade. Considere uma ação de empresa e assuma que o modelo verdadeiro de seus retornos diários é desconhecido e que o modelo escolhido para representar esses retornos induz uma distribuição incondicional que coincide com a verdadeira. Para simplificar, defina o horizonte de previsão $k = 1$. Então, para todo $u \in (0,1)$, a proporção de dias nos quais o pseudo-resíduo normal z_{t+1} do modelo é menor do que $\Phi^{-1}(u)$ coincide com a fração prometida u. Ainda assim, o investidor não estaria satisfeito com o seu modelo se os eventos $[z_{t+1} < \Phi^{-1}(u)]$, $t \geqslant 1$ e u fixado, ocorressem de forma agrupada no tempo. De fato, para fixar ideias, defina $u = 0{,}01$. Mesmo que a proporção de dias nos quais os eventos ocorressem fosse igual a u, se a ocorrência desses eventos se concentrasse em pequenos intervalos de tempo, o risco de falência do investidor seria maior do que se essas ocorrências estivessem distribuídas de forma independente ao longo do tempo.

Nesse trabalho, nós não estamos interessados em testar se um dado modelo de previsão de densidade está, de acordo com definições propostas por [74], correta ou completamente calibrado. Queremos, na verdade, comparar múltiplos modelos e distinguir aquele que apresenta a melhor calibragem condicional. O melhor dentre os modelos concorrentes não é outro senão aquele que produz os pseudo-resíduos normais que mais se aproximam de uma sequência identicamente distribuída de $\mathcal{N}(0,1)$ e cujos termos são dois-a-dois independentes caso eles estejam a pelo menos k passos de distância, onde $k \geqslant 1$ é o horizonte de previsão. Para medir essa proximidade, nós utilizamos duas funções de divergência estatística que medem, separadamente, desvios em relação à normalidade incondicional e à independência serial. Acreditamos que um procedimento de comparação de modelos baseado nessas duas funções tem mais valor em aplicações práticas do que um procedimento baseado em uma das duas medidas de acurácia preditiva introduzidas na seção anterior. De fato, o primeiro procedimento é, como diria [25], construtivo: quando a habilidade preditiva de um modelo é declarada superior, o procedimento fornece indícios sobre se a superioridade detectada está relacionada a como as densidades condicionais são modeladas, a como as estruturas de dependência temporal são especificadas, ou a ambas alternativas.

3.2.1 Normalidade Incondicional

Sejam $\xi > 0$ uma constante real e $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ um núcleo Gaussiano definido por:

$$\forall x, x' \in \mathbb{R}\,, \quad K(x, x') = e^{-\frac{|x-x'|^2}{2\xi^2}}\,. \tag{3.2}$$

Chamamos ξ de largura do núcleo Gaussiano. Como K é um núcleo simétrico positivo definido, existe um espaço $\mathcal{H}$ de funções em $\mathbb{R}$ que é munido de produto interno $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, é completo (toda sequência de Cauchy é convergente) e possui a chamada propriedade de reprodução:

$$\forall f \in \mathcal{H}\,, \forall x \in \mathbb{R}\,, \quad f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}}\,.$$

Dizemos que $\mathcal{H}$ é um espaço de Hilbert de reprodução (EHR) associado ao núcleo K e denotamos por $\|\cdot\|_{\mathcal{H}}$ a norma que é induzida pelo produto interno nesse espaço. Para mais informações sobre núcleos simétricos positivos definidos e os EHRs a eles associados, consulte [75].

De modo geral, estamos interessados em mapear distribuições de probabilidade no espaço $\mathcal{H}$ e então usar uma função de distância definida nesse espaço para comparar as distribuições mapeadas. Para essa tarefa, definimos a seguinte representação média para uma dada distribuição $\mathbb{P}_x$ em $\mathcal{B}$, a σ-álgebra de Borel na reta:

$$\mu[\mathbb{P}_x] = \mathbb{E}_x[K(x, \cdot)]\,, \tag{3.3}$$

onde $\mathbb{P}_x(B) = \mathbb{P}[x \in B]$ para todo $B \in \mathcal{B}$ e o operador de esperança é interpretado como uma integral de Bochner [76]. Sobre a definição dessa integral, consulte [27]. Como o núcleo K é limitado, uma vez que $|K(x, x')| \leqslant 1$ para todo $x, x' \in \mathbb{R}$, a Proposição 2 de [94] então garante que $\mu[\mathbb{P}_x] \in \mathcal{H}$ para toda distribuição $\mathbb{P}_x$ na classe $\mathcal{P}(\mathcal{B})$ de probabilidades em $\mathcal{B}$. Usamos essa representação para definir uma pseudo-métrica em $\mathcal{P}(\mathcal{B})$ chamada Discrepância Média Máxima (MMD, do inglês *Maximum Mean Discrepancy*):

$$\mathrm{MMD}(\mathbb{P}_x, \mathbb{P}_y) = \|\mu[\mathbb{P}_x] - \mu[\mathbb{P}_y]\|_{\mathcal{H}}\,. \tag{3.4}$$

Aplicando a propriedade de reprodução, [47] mostram que nós podemos obter uma expressão mais simples para o quadrado da MMD:

$$\mathrm{MMD}^2(\mathbb{P}_x, \mathbb{P}_y) = \mathbb{E}_{x,x'}[K(x, x')] - 2\mathbb{E}_{x,y}[K(x, y)] + \mathbb{E}_{y,y'}[K(y, y')]\,, \tag{3.5}$$

onde x' é uma cópia independente de x e y' é uma cópia independente de y.

De acordo com [34], o núcleo K dado pela Equação (3.2) e empregado ao longo dessa seção é um núcleo característico, o que implica que o mapeamento $\mu : \mathbb{P}_x \mapsto \mu[\mathbb{P}_x]$ é injetivo. Segue-se daí que a MMD definida em $\mathcal{P}(\mathcal{B})$ é uma métrica e não apenas

uma pseudo-métrica[2], donde $\text{MMD}(\mathbb{P}_x, \mathbb{P}_y) = 0$ se, e somente se, $\mathbb{P}_x = \mathbb{P}_y$. Nesse caso, a representação $\mu[\mathbb{P}_x]$ captura todas as características da distribuição $\mathbb{P}_x$. E a MMD detecta diferenças entre duas distribuições distintas mesmo quando elas coincidem em um número infinito de momentos.

Seja $Z = \{z_1, \ldots, z_T\}$ uma sequência de pseudo-resíduos normais. Assuma que Z é estacionária e que a sua distribuição incondicional é $\mathbb{P}_z \in \mathcal{P}(\mathcal{B})$. Como essa distribuição é desconhecida, nós não podemos computar a sua representação média verdadeira $\mu[\mathbb{P}_z]$. Para atuar como proxy dessa representação, nós definimos a seguinte representação média empírica:

$$\mu[Z] = \frac{1}{T}\sum_{t=1}^{T} K(z_t, \cdot)\,. \tag{3.6}$$

Uma vez que $\mathcal{H}$ é um espaço vetorial e, portanto, é fechado para as operações de soma e de multiplicação por escalar, temos que $\mu[Z] \in \mathcal{H}$. Pela linearidade da esperança e por Z ser identicamente distribuída, temos que a representação empírica é um estimador não-enviesado da representação verdadeira. E se ainda assumirmos a independência dos pseudo-resíduos, então a lei forte dos grandes números de Kolmogorov garante que $\mu[Z]$ converge quase certamente para $\mu[\mathbb{P}_z]$.

Para medir a normalidade incondicional de uma sequência $Z = \{z_1, \ldots, z_T\}$ de pseudo-resíduos normais, nós propomos a seguinte função de divergência:

$$D(\mathbb{P}_z) = \text{MMD}^2(\mathcal{N}, \mathbb{P}_z)\,, \tag{3.7}$$

onde $\mathcal{N}$ é a distribuição normal padrão. [87] apresenta uma expressão analítica para estimar essa divergência:

$$D_k(Z) = \left(\frac{\xi^2}{2+\xi^2}\right)^{1/2} - \frac{2}{T}\left(\frac{\xi^2}{1+\xi^2}\right)^{1/2}\sum_{t=1}^{T} e^{-\frac{|z_t|^2}{2(1+\xi^2)}} + \frac{1}{|I_k|}\sum_{(s,t)\in I_k} K(z_s, z_t)\,, \tag{3.8}$$

onde $k \geqslant 1$ é um inteiro, $I_k = \{(s,t) : 1 \leqslant s,t \leqslant T, |s-t| \geqslant k\}$ e ξ é a largura do núcleo Gaussiano K. Na expressão original de [87], temos $k = 1$ e, consequentemente, $|I_k| = T(T-1)$. A versão generalizada $D_k(Z)$ é importante no contexto de previsões de múltiplos períodos, onde k é o horizonte da previsão. Com efeito, mesmo quando o modelo de previsão de densidade é o correto, $|s-t| < k$ implica na dependência entre os pseudo-resíduos z_s e z_t da sequência Z correspondente [67].

Sob a hipótese de que a sequência Z de pseudo-resíduos normais é independente e identicamente distribuída de $\mathbb{P}_z$, o estimador $D_1(Z)$ é não-enviesado e converge em probabilidade para $D(\mathbb{P}_z)$ [87, 46]. É razoável esperar que Z falhe em satisfazer a suposição

[2]Nesse trabalho, o núcleo subjacente às representações e à MMD definidas no espaço $\mathcal{H}$ é o núcleo Gaussiano K introduzido pela Equação (3.2), a menos que seja dito o contrário. Logo, em todas as nossas considerações sobre tais representações e essa MMD, assuma, sempre que necessário, que o núcleo subjacente é K.

de independência, uma vez que os modelos de previsão de densidade são, na prática, apenas uma aproximação do modelo verdadeiro. No presente trabalho, nós não apresentamos resultados teóricos que garantem essa convergência de $D_1(Z)$ quando a condição de independência é violada. Mas resultados experimentais relacionados à normalidade incondicional sugerem que o teste proposto para comparação de modelos é consistente mesmo na presença de dependência serial nos pseudo-resíduos: conforme o tamanho da sequência Z aumenta, a probabilidade de rejeitar incorretamente uma hipótese nula converge para o valor tolerado, enquanto a probabilidade de não rejeitá-la quando ela é falsa converge para zero. Para mais informações sobre esse teste e os resultados experimentais, consulte a Seção 3.3.

Para trabalhos futuros, recomendamos estabelecer condições que a sequência Z de pseudo-resíduos normais deve satisfazer para garantir a convergência de $D_k(Z)$ para $D(\mathbb{P}_z)$ e, consequentemente, a consistência do teste proposto com base na normalidade incondicional. Um caminho que parece promissor consiste em assumir que o processo gerador de Z é α-misturado: em particular, que a dependência entre os pseudo-resíduos z_s e z_t tende para zero quando $|s-t|$ cresce. Essa é uma suposição razoável, uma vez que vários modelos GARCH e de volatilidade estocástica são, sob condições suaves, β-misturados e, por consequência, α-misturados [11]. Pelo Teorema 14.1 de [23], transformações de Z pelo núcleo K também seriam, nessa circunstância, α-misturadas, o que colocaria à nossa disposição algumas leis fracas dos grandes números para provar a desejada convergência [3]. Para mais informações sobre processos misturados, consulte [23].

3.2.2 Independência Serial

Seja $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ um núcleo simétrico positivo definido dado por:

$$\forall (x, y), (x', y') \in \mathbb{R}^2 , \quad L((x, y), (x', y')) = K(x, x')K(y, y') , \tag{3.9}$$

onde K é o núcleo Gaussiano introduzido pela Equação (3.2). O espaço de Hilbert de reprodução (EHR) associado ao núcleo L é o espaço produto tensorial $\mathcal{H} \otimes \mathcal{H}$ sobre $\mathbb{R}^2$, sendo $\mathcal{H}$ o EHR induzido pelo núcleo K [47]. Segundo [76], esse espaço produto tensorial é isomórfico ao espaço dos operadores de Hilbert-Schmidt que estão definidos e que tomam valor no espaço $\mathcal{H}$.

Considere um par (x, y) de variáveis aleatórias com distribuição $\mathbb{P}_{xy} \in \mathcal{P}(\mathcal{B}^2)$, onde $\mathcal{P}(\mathcal{B}^2)$ é a classe das probabilidades definidas em $\mathcal{B}^2$, a σ-álgebra de Borel em $\mathbb{R}^2$. Nessa seção, o nosso objetivo é determinar se x e y são independentes. Em outras palavras, nós estamos interessados em testar se a distribuição conjunta $\mathbb{P}_{xy}$ é igual à medida produto

$\mathbb{P}_x \times \mathbb{P}_y \in \mathcal{P}(\mathcal{B}^2)$ dada pelas marginais, condição necessária e suficiente para a independência entre as duas variáveis aleatórias. Com esse propósito, definimos as seguintes representações:

$$\mu[\mathbb{P}_{xy}] = \mathbb{E}_{x,y}[L((x,y),\cdot)] \quad = \mathbb{E}_{x,y}[K(x,\cdot)K(y,\cdot)]\,, \tag{3.10}$$

$$\mu[\mathbb{P}_x \times \mathbb{P}_y] = \mathbb{E}_x\mathbb{E}_y[L((x,y),\cdot)] = \mathbb{E}_x[K(x,\cdot)]\mathbb{E}_y[K(y,\cdot)]\,. \tag{3.11}$$

Como $L \equiv K \otimes K$ e K é limitado, essas representações estão bem definidas e, portanto, pertencem ao espaço $\mathcal{H} \otimes \mathcal{H}$ quaisquer que sejam as variáveis aleatórias x e y. Podemos então utilizar a Discrepância Média Máxima (MMD) definida nesse espaço para medir a distância entre as distribuições $\mathbb{P}_{xy}$ e $\mathbb{P}_x \times \mathbb{P}_y$:

$$\mathrm{MMD}(\mathbb{P}_{xy}, \mathbb{P}_x \times \mathbb{P}_y) = \|\mu[\mathbb{P}_{xy}] - \mu[\mathbb{P}_x \times \mathbb{P}_y]\|_{\mathcal{H}\otimes\mathcal{H}}\,. \tag{3.12}$$

Seguindo os passos de [95], nós invocamos o isomorfismo declarado no parágrafo anterior para apresentar a seguinte igualdade:

$$\mathrm{MMD}^2(\mathbb{P}_{xy}, \mathbb{P}_x \times \mathbb{P}_y) = \|\mathcal{C}_{xy}\|^2_{\mathrm{HS}} \equiv \mathrm{HSIC}(\mathbb{P}_{xy})\,, \tag{3.13}$$

onde $\|\cdot\|_{\mathrm{HS}}$ é a norma de Hilbert-Schmidt e $\mathcal{C}_{xy} : \mathcal{H} \to \mathcal{H}$ é um operador de covariância cruzada entre as variáveis aleatórias x e y. A medida de dependência HSIC é chamada Critério de Independência de Hilbert-Schmidt, foi proposta por [48] e pode ser definida em termos do núcleo Gaussiano K da seguinte forma:

$$\begin{aligned}\mathrm{HSIC}(\mathbb{P}_{xy}) = \mathbb{E}_{xy}\mathbb{E}_{x'y'}[K(x,x')K(y,y')] - 2\mathbb{E}_x\mathbb{E}_y\mathbb{E}_{x'y'}[K(x,x')K(y,y')] \\ + \mathbb{E}_x\mathbb{E}_y\mathbb{E}_{x'}\mathbb{E}_{y'}[K(x,x')K(y,y')]\,.\end{aligned} \tag{3.14}$$

Como o núcleo K é característico e invariante por translação [76], $\mathrm{HSIC}(\mathbb{P}_{xy}) = 0$ se, e somente se, as variáveis aleatórias x e y são independentes[3] [45]. Logo, o HSIC é capaz de detectar qualquer forma de dependência entre duas variáveis aleatórias, não apenas dependências lineares como ocorre com os coeficientes de correlação de Pearson, Kendall e Spearman. Em particular, o HSIC analisa, conjuntamente, as relações de dependência em cada um dos infinitos momentos amostrais das variáveis aleatórias, incluindo, mas não se limitando a, média, variância, assimetria e curtose amostrais.

Seja $Z = \{z_1, \ldots, z_T\}$ uma sequência de pseudo-resíduos normais. Assuma que Z é estacionária e que a sua distribuição incondicional é $\mathbb{P}_z \in \mathcal{P}(\mathcal{B})$. Dado $0 \leqslant s < T$, defina uma sequência $W_s = \{w_{s,1}, \ldots, w_{s,T-s}\}$ tal que $w_{s,t} = (z_t, z_{t+s})$, $1 \leqslant t \leqslant T - s$. Por ser construída a partir de Z, a sequência W_s é estacionária e, potencialmente, serialmente dependente. Então defina outra sequência, $W^*_s = \{w^*_{s,t}\}$, contendo $T - s$

[3] Nesse trabalho, o núcleo subjacente às representações e medidas definidas no espaço $\mathcal{H} \otimes \mathcal{H}$ é o núcleo L introduzido pela Equação (3.9), a menos que seja dito o contrário. Logo, em todas as nossas considerações sobre tais representações e medidas, assuma, sempre que necessário, que o núcleo subjacente é L.

cópias independentes de $w_{s,1}$. Cada termo $w^*_{s,t}$ retém a dependência entre as variáveis aleatórias z_t e z_{t+s}, mas a sequência W^*_s é, por construção, serialmente independente. Além disso, W^*_s também é estacionária e a sua distribuição incondicional coincide com a da sequência W_s, isto é, $\mathbb{P}_{w^*_s} = \mathbb{P}_{w_s}$, ambas em $\mathcal{P}(\mathcal{B}^2)$.

Para medir a independência serial de uma sequência $Z = \{z_1, \ldots, z_T\}$ de pseudo-resíduos normais, nós propomos a seguinte função de divergência:

$$\begin{aligned} R_k(\mathbb{P}_z) &= \sum_{s=k}^{S} \frac{\text{HSIC}(\mathbb{P}_{w_s})}{\text{HSIC}(\mathbb{P}_{w_0})} \\ &= \sum_{s=k}^{S} \frac{\|\mathbb{E}_{z_1,z_s}[K(z_1,\cdot)K(z_s,\cdot)] - \mathbb{E}_z[K(z_1,\cdot)]^2\|^2_{\mathcal{H}\otimes\mathcal{H}}}{\|\mathbb{E}_z[K(z_1,\cdot)^2] - \mathbb{E}_z[K(z_1,\cdot)]^2\|^2_{\mathcal{H}\otimes\mathcal{H}}}, \end{aligned} \tag{3.15}$$

onde $k \geqslant 1$ é o horizonte de previsão e $S \geqslant k$ é o horizonte máximo a ser considerado para mensuração da independência serial. Usamos o horizonte de previsão k para parametrizar essa função pois mesmo quando o modelo subjacente é o correto, $|s-t| < k$ implica na dependência entre os pseudo-resíduos z_s e z_t da sequência Z correspondente [67]; logo, não há motivo para checar independência sob essa condição. Outro ponto a se observar sobre a medida proposta é que cada termo $s \in \{k, \cdots, S\}$ do seu somatório é o análogo em $\mathcal{H} \otimes \mathcal{H}$ do quadrado da autocorrelação $\rho(s)$ em $\mathbb{R}$. Por fim, para estimar o valor dessa divergência, podemos empregar as sequências W^*_s's e o estimador não-enviesado e consistente do HSIC proposto por [93]:

$$\text{HSIC}(W^*_s) = \frac{1}{\tilde{T}(\tilde{T}-3)} \left[\text{tr}(\tilde{L}_s^2) + \frac{(\mathbf{1}'\tilde{L}_s\mathbf{1})^2}{(\tilde{T}-1)(\tilde{T}-2)} - \frac{2}{\tilde{T}-2}\mathbf{1}'\tilde{L}_s^2\mathbf{1}' \right], \tag{3.16}$$

onde $\text{tr}(\cdot)$ é a função traço, $\tilde{L}_s$ é uma matriz quadrada de dimensão $\tilde{T} \times \tilde{T}$, com $\tilde{T} = T - s$, tal que $(\tilde{L}_s)_{i,j}$ é igual a $L(w^*_{s,i}, w^*_{s,j})$ se $i \neq j$ ou 0 se $i = j$, e $\mathbf{1}$ é um vetor de dimensão apropriada cujos elementos são iguais a 1, sendo $\mathbf{1}'$ o respectivo vetor transposto.

Na prática, as sequências W^*_s's não estão disponíveis. Então precisamos substituí-las pelas sequências W_s's na Equação (3.16). Porém, no presente trabalho, nós não apresentamos resultados teóricos que garantem a convergência de $\text{HSIC}(W_s)$ para $\text{HSIC}(\mathbb{P}_{w_s})$, $s \geqslant 1$. Por outro lado, resultados experimentais sobre independência serial sugerem que o teste proposto para comparação de modelos é consistente mesmo na presença de dependência serial nas sequências W_s's. Sobre esse teste e os resultados experimentais, consulte a Seção 3.3. Para trabalhos futuros, recomendamos estabelecer condições que a sequência Z de pseudo-resíduos normais deve satisfazer para garantir a mencionada convergência. Conforme destacado na Seção 3.2.1, um caminho que parece promissor consiste em assumir que o processo gerador de Z é α-misturado. Nesse caso, o Teorema 14.1 de [23] garante que o processo gerador das sequências W_s's também seria α-misturado.

3.3 Teste Bayesiano para Habilidade Preditiva Equivalente

Propomos um novo teste para habilidade preditiva equivalente que é baseado em uma abordagem Bayesiana hierárquica. Na nossa abordagem, dois ou mais preditores de densidade[4] concorrentes são analisados conjuntamente e os resultados de desempenho, medidos por uma das medidas de acurácia definidas anteriormente nesse capítulo, são comparados dois-a-dois, de forma emparelhada, em uma mesma coleção de sequências univariadas. Para avaliar as taxas de rejeição do teste proposto, nós conduzimos uma análise de poder Bayesiana em sua forma prospectiva. Quando a coleção de sequências contém pelo menos 200 exemplos com no mínimo 756 observações cada, resultados sugerem que nosso teste é poderoso e está bem calibrado com respeito a ocorrências de falsas descobertas. Em particular, o teste demonstra boa capacidade para detectar diferenças de habilidade preditiva e também é robusto a especificação incorreta da estrutura de dependência e da distribuição condicional ao se avaliar, respectivamente, a normalidade incondicional e a independência serial.

3.3.1 Modelo Hierárquico e Decisão do Teste

Nessa seção, nós apresentamos um modelo hierárquico Bayesiano que analisa conjuntamente os resultados de desempenho de dois ou mais preditores de densidade em múltiplas sequências univariadas. Ele retorna a distribuição a posteriori da diferença média estritamente padronizada (SSMD, do inglês *Strictly Standardized Mean Difference*) entre desempenhos para cada par de preditores. Com base nessa distribuição, nós propomos um procedimento de tomada de decisão para rejeitar ou não a hipótese nula de habilidades preditivas equivalentes. O nosso modelo hierárquico se destaca por controlar automaticamente as falsas descobertas. Além disso, ele é flexível o suficiente para acomodar eventuais valores atípicos de desempenho e potenciais correlações entre os preditores concorrentes.

Assumimos que o desempenho de um preditor de densidade é medido por uma das medidas de acurácia preditiva propostas na Seção 3.2. Para descrever conjuntamente a distribuição sobre os desempenhos $x_1, \ldots, x_N$ de M preditores concorrentes em N sequências

[4]Para evitar ambiguidades nessa seção, chamamos os modelos de previsão de densidade de preditores de densidade.

univariadas, com $x_n = (x_{n,1}, \ldots, x_{n,M})$, escolhemos a versão padronizada da distribuição t de Student multivariada:

$$x_1, \ldots, x_N \sim \text{MVT}(\nu, \mu, \text{diag}(\sigma)\Psi\,\text{diag}(\sigma)) \tag{3.17}$$

onde $\mu = (\mu_1, \ldots, \mu_M)$ é o vetor de médias, $\text{diag}(\sigma)$ é a matriz diagonal com o vetor $\sigma = (\sigma_1, \ldots, \sigma_M)$ de desvios padrões, $\Psi = [\rho_{i,j}]$ é a matriz de correlação, e $\nu > 2$ é o número de graus de liberdade. Optamos por uma distribuição multivariada ao invés de escolher uma distribuição univariada para cada preditor porque precisamos dos coeficientes de correlação $\rho_{i,j}$'s para calcular a SSMD entre desempenhos para cada par de preditores. E a preferência pela distribuição t de Student em detrimento da distribuição normal se justifica pela flexibilidade proporcionada pelo parâmetro ν. De fato, quando o valor desse parâmetro é pequeno (por exemplo, quando $\nu < 30$), a distribuição t de Student tem caudas pesadas e pode ser usada para descrever dados que contêm valores extremos. Do contrário (quando $\nu \geqslant 30$), essa distribuição é praticamente uma Gaussiana.

Os parâmetros do modelo hierárquico são estimados por inferência Bayesiana. Para evitar suposições restritivas quanto aos preditores de densidade em análise, nós especificamos distribuições a priori pouco informativas, expressando assim uma grande incerteza prévia com relação aos valores dos parâmetros:

$$\mu_1, \ldots, \mu_M \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)\,, \tag{3.18}$$
$$\tilde{\sigma}_1, \ldots, \tilde{\sigma}_M \sim \text{Gama}(f(s, 5s), g(s, 5s))\,, \tag{3.19}$$
$$\tilde{\nu} \sim \text{Gama}(1, 1/28)\,, \tag{3.20}$$
$$\bar{\mu} \sim \mathcal{N}(m, 25s^2)\,, \tag{3.21}$$
$$\bar{\sigma} \sim \text{Gama}(f(s, 5s), g(s, 5s))\,, \tag{3.22}$$
$$\Psi \sim \text{LKJ}(\eta)\,, \tag{3.23}$$
$$\eta \sim \text{Gama}(f(2, 2), g(2, 2))\,, \tag{3.24}$$

onde m e s são, respectivamente, a média e o desvio padrão amostrais calculados a partir dos dados de todos os M preditores concorrentes em todas as N sequências univariadas, $\text{Gama}(\alpha, \beta)$ é a distribuição gama com parâmetros de forma α e de taxa β, e $\text{LKJ}(\eta)$ é a distribuição LKJ para matrizes de correlação com parâmetro de forma η. Além disso, definimos:

$$\tilde{\sigma}_i = \sigma_i \left[\frac{\tilde{\nu}}{\nu}\right]^{\frac{1}{2}}, \quad i = 1, \ldots, M\,, \tag{3.25}$$
$$\tilde{\nu} = \nu - 2\,, \tag{3.26}$$
$$f(x, y) = 1 + x g(x, y)\,, \tag{3.27}$$
$$g(x, y) = \frac{x + \sqrt{x^2 + 4y^2}}{2y^2}\,. \tag{3.28}$$

As Equações (3.18)-(3.22) são inspiradas nas prioris alternativas do pacote BEST[5]. E as Equações (3.23)-(3.24) são baseadas em sugestão de [73], segundo o qual a distribuição LKJ com η igual a 2 define uma priori vaga sobre Ψ que dá pouca credibilidade a correlações extremas.

O modelo hierárquico estima os parâmetros μ_i's aplicando encolhimento às estimativas $\bar{x}_i$'s, onde $\bar{x}_i$ é a média amostral dos resultados de desempenho do preditor i. De fato, como $\bar{\mu}$ é desconhecido, ele é estimado junto com os demais parâmetros de localização, o que faz com que a estimativa de cada μ_i seja influenciada pelos resultados de todos os outros preditores. Intuitivamente, todos os preditores contribuem para a estimativa de $\bar{\mu}$, que por sua vez restringe e torna mais precisa a estimativa relacionada a cada preditor: os μ_i's sofrem encolhimento em direção a $\bar{\mu}$ de modo inversamente proporcional à dispersão observada entre os preditores. Por causa desse encolhimento, um dos principais benefícios da estrutura hierárquica é controlar automaticamente as falsas descobertas quando se conduz múltiplas comparações [69, 36]. E como o estimador de encolhimento produz um erro quadrático médio menor do que o estimador de máxima verossimilhança [20, 77], o nosso modelo reduz o erro de estimação do desempenho esperado de cada preditor em comparação com a abordagem usual que toma a média aritmética dos resultados observados em todas as sequências univariadas.

Para cada combinação credível de médias, desvios padrões e coeficientes de correlação, nós calculamos a SSMD entre os resultados de desempenho para cada par (i, j), $i < j$, de preditores de densidade, denotada aqui por $d_{i,j}$. Essa medida de tamanho de efeito foi proposta por [111] e tem a seguinte definição:

$$d_{i,j} = \frac{(\mu_i - \mu_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{i,j}}}, \tag{3.29}$$

onde $\sigma_{i,j} = \rho_{i,j}\sigma_i\sigma_j$ é a covariância entre os preditores i e j. Com relação à d de Cohen [18], uma medida popular de tamanho de efeito também baseada na diferença padronizada entre médias, uma das vantagens da SSMD é que ela pode ser usada quando os preditores em análise têm desempenhos correlacionados. No nosso caso, é razoável esperar que preditores aninhados ou pertencentes à mesma família apresentem desempenhos correlacionados quando avaliados de forma emparelhada em uma dada coleção de sequências univariadas. Uma outra vantagem é ser facilmente interpretável em probabilidade. Por exemplo, se assumirmos que a distribuição sobre as diferenças é unimodal e tem variância finita, então $d_{i,j} < -3$ implica, pela desigualdade de Vysochanskij-Petunin, que a probabilidade do preditor i ser superior ao preditor j em uma dada sequência univariada é maior do que 0,95[6]. Com base nessa interpretação em probabilidade da SSMD, é possível

[5] https://CRAN.R-project.org/package=BEST

[6] Cabe recordar que o desempenho de um preditor é medido aqui por uma das funções de divergência apresentadas nesse capítulo: quanto menor a divergência, maior a habilidade preditiva.

propor critérios significativos para classificar a magnitude da diferença de desempenho entre dois preditores [112].

Por fim, para decidir sobre rejeitar ou não a hipótese nula de habilidades preditivas equivalentes para cada par de preditores em análise, utilizamos o intervalo de mais alta densidade (HDI) de 95% para resumir a distribuição a posteriori da SSMD entre resultados de desempenho. Esse intervalo delimita a região do suporte da distribuição que contém 95% dos valores com maior credibilidade. Por definição, cada ponto dentro do HDI tem densidade (ou credibilidade, nesse contexto) maior do que qualquer ponto fora desse intervalo. Dado um par (i, j), $i < j$, de preditores de densidade, se o HDI de 95% sobre $d_{i,j}$ contém o número real zero, então nós não rejeitamos a hipótese nula. Do contrário, dizemos que a diferença de desempenho tem credibilidade. Caso esse HDI esteja contido na semirreta dos reais negativos, aceitamos a hipótese alternativa que afirma que o preditor i é credivelmente superior ao preditor j. Uma decisão análoga é tomada quando o HDI contém apenas reais positivos.

3.3.2 Análise de Poder

Na Seção 3.3.1, nós propomos um teste Bayesiano para habilidade preditiva equivalente que compara o desempenho de múltiplos preditores de densidade em várias sequências univariadas. O objetivo da presente seção é analisar o poder desse teste para vários tamanhos de sequência. Seguindo [69], chamamos de poder a probabilidade de rejeitar a hipótese nula quando os dados amostrados são gerados por um modelo descritivo com parâmetros hipotéticos, conhecidos.

Para conduzir a análise de poder Bayesiana, nós empregamos a forma prospectiva apresentada por [69]. A primeira etapa da análise consiste em gerar dados sintéticos para representar os resultados de um experimento hipotético que refletem nossas suposições com relação às diferenças médias estritamente padronizadas entre os resultados de desempenho. Para tanto, nós definimos o processo gerador de dados (PGD) como sendo o seguinte modelo EGARCH(1,1):

$$r_t = \sigma_t \varepsilon_t \,, \tag{3.30}$$

$$\log \sigma_t^2 = \omega + \alpha \varepsilon_{t-1} + \gamma(|\varepsilon_{t-1}| - \mathbb{E}|\varepsilon_{t-1}|) + \beta \log \sigma_{t-1}^2 \,, \tag{3.31}$$

$$\varepsilon_t \sim F \,, \tag{3.32}$$

para $t \geqslant 1$, onde $\omega = -0{,}08$, $\alpha = -0{,}06$, $\gamma = 0{,}11$, $\beta = 0{,}99$ e F é a distribuição t de Student padronizada com $\nu = 8$ graus de liberdade. Então utilizamos o pacote *rugarch*[7]

[7] https://CRAN.R-project.org/package=rugarch

para gerar $N_{\text{ideal}} = 1500$ sequências univariadas de retornos diários de ações com $\mathcal{T} + T$ observações cada, $\mathcal{T} = 4032$ e $T \in \{126, 252, 504, 756, 1008\}$. Note que cada uma dessas sequências apresenta, em particular, caudas pesadas na sua distribuição condicional e efeito de assimetria em sua estrutura de dependência, devido à nossa escolha para a distribuição F e para a forma funcional de $\log \sigma_t^2$, respectivamente. Para mais informações sobre esses fatos estilizados, consulte a Seção 2.2.

Além do PGD, nós também especificamos $M = 3$ preditores concorrentes. O preditor 1 é o próprio PGD. O preditor 2 se diferencia do PGD por atribuir a F a distribuição normal padrão. E o preditor 3 é o modelo GARCH(1,1) com distribuição t de Student padronizada: ele se diferencia do primeiro preditor por não controlar o efeito que o sinal do retorno de um determinado dia tem sobre a volatilidade do dia seguinte. Para cada combinação de preditor e de sequência gerada, nós estimamos os parâmetros usando o pacote *rugarch* e as primeiras $\mathcal{T}$ observações; calculamos os pseudo-resíduos normais para o horizonte de previsão $k = 1$ com as demais T observações; e então aplicamos sobre os resíduos uma medida de acurácia: $D(\cdot)$ ou $R(\cdot)$. Por não capturar as caudas pesadas da distribuição condicional dos retornos simulados, esperamos que o preditor 2 apresente o pior desempenho se a medida for D, que mede a normalidade incondicional. E por não capturar o efeito de assimetria na estrutura de dependência das sequências amostradas, nós esperamos que o preditor 3 apresente o pior desempenho se a medida for R, que mede a independência serial.

A segunda etapa da nossa análise de poder consiste em usar os dados idealizados, formados pelas N acurácias de cada um dos M preditores, para inferir a distribuição a posteriori sobre os parâmetros do modelo hierárquico especificado na Equação (3.17). Além de expressar a incerteza que está implícita em nosso experimento hipotético, essa distribuição também captura as dependências entre os parâmetros. De acordo com [69], é mais fácil gerar os dados idealizados e, a partir deles, inferir a distribuição conjunta sobre os parâmetros do que especificá-la diretamente. Para a inferência, utilizamos a implementação do algoritmo NUTS [56] do pacote *NumPyro*[8] e obtemos uma amostra com 51000 combinações de parâmetros, dos quais descartamos os primeiros 1000 exemplos a título de *burn-in*.

Tendo a distribuição conjunta hipotética sobre os parâmetros com respeito ao experimento idealizado, passamos então para a análise de poder propriamente dita. Nessa etapa, nós executamos, repetidas vezes, a sequência de passos descrita a seguir. Cada repetição simula uma realização do teste planejado.

1. Escolhemos, de forma aleatória e com reposição, uma combinação de parâmetros da amostra gerada na segunda etapa da nossa análise de poder.

[8] https://num.pyro.ai/

2. Usamos a combinação de parâmetros escolhida no passo (1) para gerar $N_{\text{plan}} = 200$ observações do modelo hierárquico do nosso teste. Cada observação gerada é um vetor de tamanho M que contém um valor simulado para a acurácia de cada preditor. Essas observações representam a amostra do teste planejado.

3. Inferimos, novamente, a distribuição conjunta dos parâmetros do modelo. Mas, dessa vez, nós utilizamos os dados da simulação do passo (2) e produzimos uma amostra com 11000 combinações de parâmetros, descartando os primeiros 1000 exemplos a título de *burn-in*. Essa etapa e as próximas duas representam a execução do teste planejado.

4. Para cada combinação de parâmetros da amostra gerada no passo (3), calculamos a diferença média estritamente padronizada $d_{i,j}$ entre as acurácias do par (i, j) de preditores, onde $1 \leqslant i < j \leqslant M$.

5. Por fim, para cada par (i, j) de preditores, onde $1 \leqslant i < j \leqslant M$, nós calculamos o HDI de 95% sobre $d_{i,j}$ e verificamos se esse intervalo contém o número real zero ou, caso contrário, se é negativo ou positivo. Dizemos que o intervalo é negativo (positivo) se ele está contido na semirreta dos reais negativos (positivos).

Com base nos resultados de 1000 repetições desses passos para cada medida de acurácia e tamanho de sequência T, obtemos as probabilidades apresentadas nas Tabelas 3.1-3.2. Para cada uma dessas probabilidades, nós calculamos o HDI de 95% sobre a sua distribuição a posteriori, que nós assumimos ser uma distribuição Beta com priori uniforme não-informativa [69]. Esse intervalo serve para expressar a incerteza na estimativa de cada probabilidade: quanto maior o número de repetições da sequência de passos descrita acima, menor o seu comprimento. A análise prospectiva de poder revela que o teste Bayesiano proposto é poderoso, detectando corretamente diferenças entre os preditores quando $T \geqslant 252$. De fato, quando essa condição é verdadeira, os preditores 2 e 3 são declarados inferiores com respeito à normalidade incondicional e à independência serial, respectivamente: veja os HDIs de 95% sobre $d_{1,2}$ e $d_{2,3}$ quando a medida de acurácia é D, e sobre $d_{1,3}$ e $d_{2,3}$ quando a medida é R.

E com respeito às falsas descobertas, resultados da análise indicam que o teste proposto está bem calibrado quando $T \geqslant 756$. Com efeito, se $T \in \{756, 1008\}$ e a medida de acurácia é D, que mede a normalidade incondicional, então a probabilidade estimada de que o HDI de 95% sobre $d_{1,3}$ não contém o número real zero está entre 0,049 e 0,079. Logo, nós não podemos afirmar que essa probabilidade é diferente do valor nominal de 0,05, a nossa tolerância quanto à taxa de rejeição incorreta de uma hipótese nula. Analogamente, podemos chegar a uma conclusão similar com relação ao HDI de 95% sobre $d_{1,2}$ se $T \in \{756, 1008\}$ e a medida de acurácia é R, que mede a independência serial. Note que já era esperado que os pares de preditores $(1, 3)$ e $(1, 2)$ apresentassem,

Evento	Poder Bayesiano	Limite Inferior	Limite Superior
$T = 126$			
HDI de 95% sobre $d_{1,2}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{1,3}$ contém zero	0,806	0,781	0,830
HDI de 95% sobre $d_{2,3}$ é positivo	0,998	0,994	1,000
$T = 252$			
HDI de 95% sobre $d_{1,2}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{1,3}$ contém zero	0,890	0,870	0,908
HDI de 95% sobre $d_{2,3}$ é positivo	1,000	0,997	1,000
$T = 504$			
HDI de 95% sobre $d_{1,2}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{1,3}$ contém zero	0,925	0,908	0,940
HDI de 95% sobre $d_{2,3}$ é positivo	1,000	0,997	1,000
$T = 756$			
HDI de 95% sobre $d_{1,2}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{1,3}$ contém zero	0,937	0,921	0,951
HDI de 95% sobre $d_{2,3}$ é positivo	1,000	0,997	1,000
$T = 1008$			
HDI de 95% sobre $d_{1,2}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{1,3}$ contém zero	0,937	0,921	0,951
HDI de 95% sobre $d_{2,3}$ é positivo	1,000	0,997	1,000

Tabela 3.1: Análise de poder prospectiva Bayesiana usando a medida de normalidade incondicional. Se o HDI de 95% sobre $d_{i,j}$ é negativo, então dizemos que o preditor i é credivelmente superior ao preditor j. Por outro lado, se ele é positivo, então a conclusão é favorável ao preditor j. Mas se ele contém zero, então nós não rejeitamos a hipótese nula de habilidades preditivas equivalentes. Os limites inferior e superior são os limites do HDI de 95% sobre a distribuição a posteriori do poder Bayesiano, que nós assumimos ser uma distribuição Beta com priori uniforme não informativa.

respectivamente, habilidades equivalentes quanto à capacidade de capturar as caudas pesadas da distribuição condicional e o efeito de assimetria na estrutura de dependência das sequências hipotéticas amostradas.

Para trabalhos futuros, sugerimos repetir a análise de poder para vários valores de N_{plan}, tal como fazemos com T. Nós optamos por fixar $N_{\text{plan}} = 200$ em nosso estudo pois a base de dados disponível para os experimentos relatados no Capítulo 4 contém apenas 204 ações de empresas. Havendo mais ações disponíveis, ainda recomendamos introduzir, seguindo os passos de [69], uma região de equivalência prática (ROPE, do inglês *Region of Practical Equivalence*) sobre as SSMD's, as diferenças médias estritamente padronizadas. Com uma ROPE, nós podemos não apenas decidir pela rejeição ou não das hipóteses nulas de habilidades preditivas equivalentes entre preditores, mas também por aceitá-las.

Largura do núcleo Gaussiano nas medidas de acurácia preditiva Para ajustar a constante ξ de cada medida, utilizamos os mesmos PGD e preditores concorrentes

Evento	Poder Bayesiano	Limite Inferior	Limite Superior
$T = 126$			
HDI de 95% sobre $d_{1,2}$ contém zero	0,905	0,886	0,922
HDI de 95% sobre $d_{1,3}$ é negativo	0,396	0,366	0,427
HDI de 95% sobre $d_{2,3}$ é negativo	0,321	0,293	0,350
$T = 252$			
HDI de 95% sobre $d_{1,2}$ contém zero	0,920	0,902	0,936
HDI de 95% sobre $d_{1,3}$ é negativo	0,956	0,942	0,968
HDI de 95% sobre $d_{2,3}$ é negativo	0,941	0,925	0,954
$T = 504$			
HDI de 95% sobre $d_{1,2}$ contém zero	0,925	0,908	0,940
HDI de 95% sobre $d_{1,3}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{2,3}$ é negativo	1,000	0,997	1,000
$T = 756$			
HDI de 95% sobre $d_{1,2}$ contém zero	0,939	0,923	0,953
HDI de 95% sobre $d_{1,3}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{2,3}$ é negativo	1,000	0,997	1,000
$T = 1008$			
HDI de 95% sobre $d_{1,2}$ contém zero	0,944	0,929	0,957
HDI de 95% sobre $d_{1,3}$ é negativo	1,000	0,997	1,000
HDI de 95% sobre $d_{2,3}$ é negativo	1,000	0,997	1,000

Tabela 3.2: Análise de poder prospectiva Bayesiana usando a medida de independência serial.

especificados anteriormente nessa seção e seguimos os passos descritos a seguir. Nós geramos uma segunda coleção de 1500 sequências univariadas com $\mathcal{T} + T$ observações cada, onde $\mathcal{T} = 4032$ e $T = 1008$. Então, para cada combinação de preditor e de sequência simulada, nós estimamos os parâmetros usando as primeiras $\mathcal{T}$ observações e calculamos os pseudo-resíduos normais para o horizonte de previsão $k = 1$ com as demais T observações. Por fim, para a medida de normalidade incondicional D, nós escolhemos $\xi \in \{2^{-t} : t = 0, \ldots, 5\}$ que minimiza a magnitude da diferença entre as médias amostrais dos resultados de desempenho dos preditores 1 e 3; para a medida de independência serial R, dos preditores 1 e 2. Dessa forma, nós obtemos $\xi = 1/4$ para a primeira medida e $\xi = 1/16$ para a segunda. Nós empregamos esses valores na análise de poder apresentada nessa seção e nos experimentos descritos no Capítulo 4.

Horizonte máximo na medida de independência serial Para definir o parâmetro S da medida R de independência serial, que é dada pela Equação (3.15), nós seguimos a recomendação de [98] para a escolha do parâmetro m da estatística Ljung-Box, $Q(m)$. Levando em conta o horizonte mínimo da medida R, que é dado pelo horizonte de previsão $k \geqslant 1$, nós então definimos $S = k + \lfloor \ln T \rfloor$, onde T é o tamanho da sequência.

Capítulo 4

Experimentos

Nesse capítulo, nós apresentamos a metodologia e os resultados de três experimentos que nós conduzimos no âmbito do presente trabalho. O primeiro compara o DeepRisk com um conjunto de modelos concorrentes em relação à normalidade incondicional e à independência serial de seus pseudo-resíduos normais. O segundo analisa as previsões que o nosso modelo faz para os parâmetros da distribuição condicional dos retornos diários de uma ação de empresa. E o terceiro experimento avalia quão bem o DeepRisk replica as principais características estatísticas desses retornos em sequências longas.

4.1 Dados e Esquema de Previsão

Nós consideramos 204 ações de empresas que foram constituintes do índice S&P 500 durante todo o período de 31 de dezembro de 1999 a 03 de junho de 2020. Esse índice é composto por ações de 500 empresas americanas de grande capitalização, escolhidas a partir de critérios que incluem capitalização de mercado, liquidez, flutuação pública e classificação setorial [60]. Para verificar se uma ação foi constituinte desse índice no período supracitado, nós consultamos as seguintes fontes: registro de constituintes históricos mantido pela Siblis Research[1]; relatórios anuais do fundo de investimento SPDR S&P 500 ETF[2], disponíveis no sistema EDGAR[3] da Comissão de Valores Mobiliários dos Estados Unidos; e notícias publicadas sobre fusões, aquisições e mudanças em códigos de negociação e nomes de empresas. A Tabela A.1 apresenta código de negociação, nome e estatísticas descritivas das ações selecionadas.

Os preços de fechamento diários das ações selecionadas foram coletados em uma base de dados[4] fornecida pela QuoteMedia na plataforma Quandl[5]. Todas as sequências

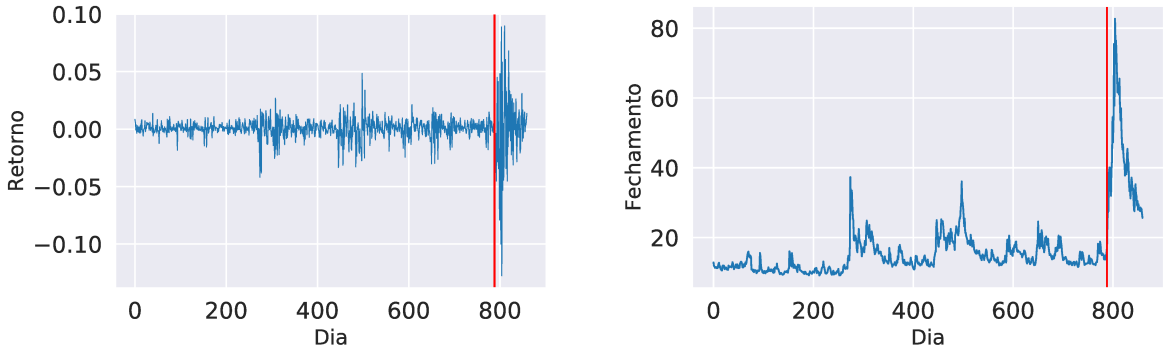
[1] https://www.siblisresearch.com
[2] https://www.ssga.com/us/en/individual/etfs/funds/spdr-sp-500-etf-trust-spy
[3] https://www.sec.gov/edgar/browse/?CIK=0000884394
[4] https://data.nasdaq.com/databases/EOD/data
[5] A plataforma Quandl foi transformada na Nasdaq Data Link.

coletadas incluem apenas os 5138 dias de negociação do período que se estende de 31 de dezembro de 1999 a 03 de junho de 2020. Essas sequências estão alinhadas no tempo, não contêm valores ausentes e já foram ajustadas pelo fornecedor de dados para pagamentos de dividendos, desdobramentos (*splits*, em inglês) e cisões (*spinoffs*, em inglês). Para transformá-las em sequências de retornos logarítmicos diários, nós aplicamos a Equação (2.14). O primeiro conjunto de dados de treinamento é formado pelas 4277 observações iniciais de cada sequência transformada, isto é, com os retornos diários de cada ação entre 03 de janeiro de 2000 e 30 de dezembro de 2016, inclusive. Os últimos 860 dias de negociação compõem a amostra na qual nós avaliamos as previsões de k dias adiante, com $k = 1, 5, 21$. Devido ao alto custo computacional e inspirados por [70], nós executamos os procedimentos de treinamento somente a cada mês (21 dias de negociação), utilizando uma janela deslizante de tamanho fixo contendo os 4277 dias de negociação antecedentes.

A Figura 4.1 ilustra a situação do mercado acionário americano no período de avaliação das previsões. Essa figura apresenta as séries diárias de retornos do índice S&P 500 e de valores de fechamento do índice VIX. Os valores de fechamento diários desse índice refletem a expectativa do mercado em relação à volatilidade do primeiro para os 30 dias subsequentes. Em função da drástica mudança na dinâmica das séries, nós podemos dividir o período de avaliação em duas partes. O primeiro subperíodo corresponde a um momento de estabilidade. Ele contém 789 observações, começando em 03 de janeiro de 2017 e terminando em 21 de fevereiro de 2020. Já o segundo subperíodo abrange uma pequena temporada de turbulência generalizada. Estendendo-se de 24 de fevereiro de 2020 a 03 de junho de 2020, esse subperíodo contém 71 dias de negociação e abrange o início e o pico da crise provocada pela pandemia do novo coronavírus (COVID-19). Por causa do tamanho do último subperíodo, nós decidimos avaliar o desempenho preditivo do DeepRisk apenas no período completo: de 03 de janeiro de 2017 a 03 de junho de 2020, totalizando 860 dias de negociação. De fato, de acordo com a análise de poder apresentada na Seção 4.4, o procedimento de comparação de modelos proposto no Capítulo 3 é poderoso e bem calibrado, dada uma coleção com aproximadamente 200 sequências de retornos, quando cada uma delas tem no mínimo 756 observações.

4.2 Implementação e Treinamento do DeepRisk

Essa seção descreve a implementação do DeepRisk e apresenta detalhes dos processos de ajuste do modelo e de previsão de múltiplos passos.



(a) S&P 500 (b) VIX

Figura 4.1: Comportamento dos índices S&P 500 e VIX no período de 03 de janeiro de 2019 a 03 de junho de 2020. A linha vertical em vermelho está posicionada no dia 21 de fevereiro de 2020. *(a)* Retornos logarítmicos diários do índice S&P 500. *(b)* Valores de fechamento diários do índice VIX. Fonte: Elaborado pelo autor.

Modelo Nós implementamos o modelo especificado na Seção 2.3 usando MXNet [14]. O número de camadas ocultas foi definido da seguinte forma: o maior número inteiro que é menor do que o $\log_B T$, onde B é o fator de dilatação das conexões de salto exponencialmente dilatadas e T é o tamanho das sequências truncadas utilizadas no treinamento do modelo. Essa fórmula parece explicar as escolhas para o número de camadas ocultas das redes neurais recorrentes dilatadas nos experimentos de [12]. Por fim, nós fixamos $\mu_{t+1} = 0$ para toda ação de empresa, onde μ_{t+1} é o retorno esperado da ação no dia $t+1$ dados seus retornos até o dia t. Tendo em vista as ações selecionadas para nossos experimentos, que estão posicionadas entre as mais negociadas e constantemente monitoradas do mercado acionário americano, é razoável supor que elas não oferecem oportunidades de lucro sem risco e que, portanto, essa é a melhor previsão que podemos oferecer para as esperanças condicionais. Mais formalmente, nós assumimos que as sequências de retornos dessas ações são diferenças martingais quadrado integráveis com respeito à filtragem $\{\mathcal{G}_t\}$ definida na Seção 2.3 [91]. Essa decisão também é aceitável sob o ponto de vista empírico. De fato, para toda ação na Tabela A.1, nós podemos observar que o módulo do retorno médio é consideravelmente menor do que o desvio padrão. Então não podemos rejeitar, estatisticamente, a hipótese de retorno médio igual a zero [16].

Inferência variacional e aprendizagem Inspirados pelos modelos de aprendizado profundo para séries temporais propostos por [80], [89] e [88], treinamos um único modelo univariado, descrito como global ou compartilhado, utilizando todas as sequências do conjunto de treinamento, na esperança dele aprender as características distribucionais e de estrutura de dependência que são comuns a essas sequências. Definimos o tamanho de cada mini-lote em 64 sequências truncadas e a quantidade de épocas em 512 iterações sobre o conjunto de treinamento. Os parâmetros variacionais μ e ρ foram inicializados seguindo os

mesmos procedimentos utilizados na implementação[6] da rede neural recorrente Bayesiana de [32]. Em cada iteração do algoritmo de treinamento sobre um mini-lote, nós tomamos apenas uma amostra da distribuição variacional sobre os pesos do modelo. Para configurar a taxa de aprendizagem, nós adotamos o método STLR (do inglês *Slanted Triangular Learning Rates*) proposto por [58], com os seguintes hiperparâmetros: cut_frac = 0,2, ratio = 32 e $\eta_{\max} = 0{,}01$. E para atualizar os parâmetros variacionais, nós empregamos o algoritmo Adam [66] com $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$ e $\epsilon = 10^{-8}$, limitando o gradiente médio por sequência ao intervalo $[-\text{clip\_gradient}, \text{clip\_gradient}]$, com clip_gradient dado pelo produto entre 0,25 e o tamanho T das sequências truncadas. O procedimento de treinamento está descrito na Seção 2.3.1.

Ajuste de hiperparâmetros O conjunto de dados utilizado foi o primeiro conjunto de treinamento definido na Seção 4.1. Desse conjunto, nós escolhemos, de forma aleatória, 25% das ações de empresa e as reservamos para validação. A ideia de separar os dois conjuntos por ação e não por data é evitar a possibilidade de ter um conjunto de validação com dados de apenas um regime de mercado. Nesse caso, nós correríamos o risco de escolher uma configuração de modelo que fosse adequada somente para períodos de estabilidade, por exemplo. Além disso, essa escolha condiz com o caráter global do DeepRisk, enfatizado no parágrafo anterior. Para a escolha da melhor configuração, nós adotamos como medida de comparação a função de log-verossimilhança do modelo avaliada nas sequências truncadas do conjunto de validação. O espaço de hiperparâmetros ajustáveis é apresentado na Tabela 4.1. Todas as configurações possíveis foram avaliadas.

Hiperparâmetro	Valores possíveis
Tamanho da sequência truncada (T)	63, 126, 252
Tamanho da camada oculta	16, 17, ..., 32
Fator de dilatação (B)	1, 2, 3
Peso da componente 1 da priori dos pesos (π)	1/4, 1/2, 3/4
Desvio padrão da componente 1 da priori dos pesos (σ_1)	$-\log \sigma_1 \in \{0, 1, 2\}$
Desvio padrão da componente 2 da priori dos pesos (σ_2)	$-\log \sigma_2 \in \{6, 7, 8\}$

Tabela 4.1: Espaço de hiperparâmetros ajustáveis.

Previsões de múltiplos passos Para gerar sequências de previsões de densidade, nós seguimos o procedimento recursivo descrito na Seção 2.3.2, definindo $K = 21$, $M_c = 100$ e $M_r = 50$. Sendo assim, para cada dia t do período de avaliação, nós geramos 5000 sequências de densidades preditivas para os dias $t+1, \ldots, t+21$. Cada uma dessas sequências está condicionada nos retornos observados entre os dias $(t-\tau+1)$ e t, onde τ é

[6] https://github.com/deepmind/sonnet/blob/v1/sonnet/examples/brnn_ptb.py

tal que $\tau + 21$ é igual ao tamanho T das sequências truncadas utilizadas no treinamento. Para calcular o pseudo-resíduo normal correspondente ao dia $t + k$, com $k = 1, 5, 21$, nós computamos essa estatística com respeito à densidade preditiva desse dia em cada uma das sequências preditas e, então, tomamos a média.

4.3 Comparação de Modelos

Para avaliar o desempenho do DeepRisk, nós o comparamos a dois grupos de modelos concorrentes. O primeiro é composto por diferentes configurações do modelo proposto. Cada uma dessas configurações é obtida adotando uma densidade condicional diferente para modelar os retornos logarítmicos diários: normal (NORM), normal assimétrica de [30] (SNORM) e t de Student (STD). Ao comparar o nosso modelo com os concorrentes desse grupo, nós o identificamos pela sigla SSTD, em referência à sua densidade condicional: t de Student assimétrica de [30]. Os dados e o esquema de previsão para treinamento e avaliação dos modelos desse grupo são aqueles apresentados na Seção 4.1. E a implementação deles faz uso das mesmas definições e estratégias de ajuste definidas na Seção 4.2.

O segundo grupo de modelos concorrentes é formado pelo modelo GARCH(1,1) [10] e algumas de suas variantes: GJR-GARCH(1,1), EGARCH(1,1) e CGARCH(1,1). As duas primeiras variantes, de [39] e [78], respectivamente, permitem capturar o efeito assimétrico do retorno observado em um dia na volatilidade dos dias seguintes. Ambos refletem o incremento na volatilidade provocado por uma queda no preço da ação. O segundo, porém, também modela o decremento na volatilidade provocado por uma alta no preço. Por sua vez, a variante CGARCH, proposta por [29], possibilita representar comportamentos de curto e de longo prazos do desvio padrão condicional dos retornos, fornecendo uma aproximação de processo de memória longa para a volatilidade [2]. Nesse grupo, a única densidade condicional usada para modelar os retornos logarítmicos diários foi a t de Student assimétrica de [30]. Sobre os parâmetros distribucionais e da estrutura de dependência de cada modelo, nós os estimamos para cada ação de empresa utilizando o pacote `rugarch` [37]. Os dados e o esquema de previsão também foram aqueles descritos na Seção 4.1. E para as previsões de múltiplos passos, nós seguimos o mesmo procedimento adotado para o DeepRisk: para cada dia t, nós geramos 5000 sequências de previsões para os dias $t+1, \ldots, t+21$, e então calculamos o valor médio dos pseudo-resíduos no dia $t+k$ para $k = 1, 5, 21$.

Uma alternativa aos modelos da família GARCH são os modelos de volatilidade estocástica [97]. Porém, como não há evidências de que esses modelos têm desempenho

preditivo melhor do que aqueles [83, 65], nós optamos por não incluí-los em nossos experimentos. E tendo em vista apenas a classe dos modelos generativos explícitos definida na Seção 2.1.3, podemos citar algumas alternativas mais recentes para modelar os retornos de uma ação de empresa: o modelo de [104], baseado em processo Gaussiano, e os modelos de [106], [71], [64] e [13], todos baseados em redes neurais recorrentes profundas. Esses modelos não foram comparados ao nosso pois eles operam sob a hipótese de que a densidade condicional dos retornos diários é Gaussiana. Logo, esses modelos não especificam parâmetros para controlar o peso das caudas e o grau de assimetria dessa densidade. Como o DeepRisk usa uma versão assimétrica da distribuição t de Student para dar forma a essa densidade, nós avaliamos que tal comparação não seria adequada.

4.3.1 Normalidade Incondicional

Com respeito à normalidade incondicional, o DeepRisk apresenta habilidade preditiva superior em todos os horizontes de previsão avaliados. Comparando-o com o primeiro grupo de modelos concorrentes, observamos pela Tabela 4.2 que o resultado do nosso modelo (SSTD) é, para $k = 1$, 10,83 vezes melhor do que aquele obtido pela configuração NORM e 2,49 vezes melhor do que aquele apurado para a configuração STD. E essas diferenças se tornam ainda maiores conforme o horizonte k aumenta. De modo geral, esses resultados indicam que introduzir caudas pesadas e assimetria no DeepRisk melhora significativamente a sua acurácia preditiva média ou, usando termos similares aos de [15], a sua calibragem incondicional.

Modelo	Horizonte		
	$k = 1$	$k = 5$	$k = 21$
NORM	8,600	9,212	10,781
SNORM	7,861	9,006	11,067
STD	1,975	1,845	1,703
SSTD	**0,794**	**0,696**	**0,553**

Tabela 4.2: Normalidade incondicional dos pseudo-resíduos normais do DeepRisk com diferentes distribuições de probabilidade. Quanto menor o resultado, menor a distância média entre a normal padrão teórica e a distribuição empírica dos pseudo-resíduos. Para cada horizonte de previsão, os melhores resultados estão em **negrito**. "Melhor" significa "a diferença estritamente padronizada entre esse resultado e o resultado com a menor distância média não é credível". As diferenças estritamente padronizadas estão reportadas no Anexo B. Todos os resultados nessa tabela foram multiplicados por 10^2.

Em relação aos modelos concorrentes da família GARCH, o DeepRisk também

apresenta habilidade preditiva superior em todos os horizontes de previsão avaliados, conforme aponta os resultados da Tabela 4.3. Esses resultados indicam que o nosso modelo é melhor do que as alternativas em capturar as características distribucionais observadas em sequências reais de retornos. Comparando-o com o melhor concorrente desse grupo, o EGARCH, observamos que o DeepRisk é 1,31 vezes melhor quando $k = 1$, 1,51 vezes melhor quando $k = 5$, e 1,96 vezes melhor quando $k = 21$. Essa diferença pode ser explicada, em partes, pelo fato do DeepRisk modelar a assimetria e o peso das caudas de forma dinâmica no tempo; tais parâmetros são constantes nos modelos da família GARCH. Além disso, novamente, as diferenças entre o DeepRisk e os concorrentes se tornam maiores conforme o horizonte k aumenta. De fato, essa é uma característica interessante do DeepRisk: dentre todos os modelos avaliados, apenas o DeepRisk, nas configurações STD e SSTD, e o modelo CGARCH não apresentam deterioração de desempenho com o aumento do horizonte de previsão. Essa característica pode indicar que, nesses casos, a distribuição condicional está convergindo adequadamente para a distribuição incondicional conforme k aumenta.

Modelo	Horizonte		
	$k = 1$	$k = 5$	$k = 21$
GARCH	1,484	1,525	1,518
GJR-GARCH	1,679	1,785	1,981
EGARCH	1,039	1,049	1,087
CGARCH	1,291	1,265	1,071
DeepRisk	**0,794**	**0,696**	**0,553**

Tabela 4.3: Normalidade incondicional dos pseudo-resíduos normais de diferentes modelos utilizando a distribuição t de Student assimétrica. As diferenças estritamente padronizadas entre os resultados estão reportadas no Anexo B. Todos os resultados nessa tabela foram multiplicados por 10^2.

4.3.2 Independência Serial

Com respeito à independência serial, o DeepRisk não apresenta habilidade preditiva superior em todos os horizontes de previsão analisados, tal como ocorre na avaliação da condição de normalidade incondicional. Por outro lado, nenhum outro modelo o supera em qualquer dos cenários estudados. Com efeito, de acordo com a Tabela 4.4, o nosso modelo (SSTD) tem desempenho equiparável ao da configuração STD em todos os horizontes de previsão, ao da configuração NORM quando $k = 5$, e ao da configuração SNORM quando $k = 21$. E com base na Tabela 4.5, nós podemos afirmar que o Dee-

pRisk é tão bom quanto os modelos concorrentes da família GARCH no horizonte $k = 21$, superando-os, porém, quando $k = 1$ e $k = 5$. Esses resultados nos permitem concluir que o DeepRisk é, de modo geral, melhor do que as alternativas convencionais em capturar os principais fatos estilizados relacionados à estrutura de dependência das séries de retornos diários: agrupamento de volatilidade, memória longa e efeito de alavancagem. Novamente, acreditamos que a superioridade do DeepRisk pode ser explicada pelo fato dele controlar dinamicamente os parâmetros distribucionais de assimetria e de curtose em função de acontecimentos recentes, o que lhe permite aproveitar melhor o conjunto de informações disponível no momento. Também pode ser explicada pela capacidade da rede LSTM em capturar dependências temporais que não são modeladas nas formas funcionais da volatilidade condicional dos modelos da família GARCH.

Modelo	Horizonte		
	$k = 1$	$k = 5$	$k = 21$
NORM	1,517	**0,902**	0,207
SNORM	1,718	1,114	**-0,097**
STD	**0,617**	**0,535**	**-0,324**
SSTD	**0,582**	**0,530**	**-0,193**

Tabela 4.4: Independência serial dos pseudo-resíduos normais do DeepRisk com diferentes distribuições de probabilidade. Quanto menor o resultado, menor a dependência serial média dos pseudo-resíduos. Para cada horizonte de previsão, os melhores resultados estão em **negrito**. "Melhor" significa "a diferença estritamente padronizada entre esse resultado e o resultado com a menor dependência serial média não é credível". As diferenças estritamente padronizadas estão reportadas no Anexo B. Todos os resultados nessa tabela foram multiplicados por 10^3.

Modelo	Horizonte		
	$k = 1$	$k = 5$	$k = 21$
GARCH	3,003	1,731	**0,219**
GJR-GARCH	2,317	1,195	**0,182**
EGARCH	1,615	0,894	**-0,110**
CGARCH	1,982	1,812	0,346
DeepRisk	**0,582**	**0,530**	**-0,193**

Tabela 4.5: Independência serial dos pseudo-resíduos normais de diferentes modelos utilizando a distribuição t de Student assimétrica. As diferenças estritamente padronizadas entre os resultados estão reportadas no Anexo B. Todos os resultados nessa tabela foram multiplicados por 10^3.

E sobre as tendências de melhora e de convergência das acurácias preditivas à medida que o horizonte de previsão k aumenta, temos a seguinte conjectura. Fixada uma ação de empresa, espera-se que o impacto que um fato ocorrido no dia t tem sobre o retorno realizado no dia $t + k$ diminui conforme k aumenta. Essa expectativa é suportada

pela hipótese do mercado eficiente em sua forma fraca, segundo a qual toda a informação disponível publicamente sobre uma ação já está refletida em seu preço. Ela também é corroborada empiricamente: basta observar o decaimento na função de autocorrelação empírica da sequência dada pelo valor absoluto dos retornos diários. Esse desaparecimento assintótico da dependência serial na sequência de retornos parece ser uma explicação razoável para a melhora de desempenho, medido aqui pela condição de independência serial, de todos os modelos conforme k aumenta. Portanto, dada essa medida de acurácia preditiva, não nos surpreende que as habilidades preditivas de quase todos os modelos analisados sejam equivalentes quando $k = 21$.

4.4 Análise das Previsões do DeepRisk

Para analisar as previsões do nosso modelo, nós conduzimos o seguinte experimento. Sejam $r_1, \ldots, r_{T'}$, com $T' = 5138$, os retornos logarítmicos diários da ação da Microsoft durante o período que se estende de 03 de janeiro de 2000 a 03 de junho de 2020. Essa sequência é apresentada na Figura 4.2. Para $m = 1, \ldots, 1000$, nós amostramos o conjunto de pesos $w^{(m)}$ da distribuição variacional $\mathcal{N}(w|\mu, \sigma^2)$ e então calculamos a previsão dos parâmetros distribucionais σ_{t+1}, ν_{t+1} e γ_{t+1} a partir de $\theta(r_1, \ldots, r_t; w^{(m)})$, com t indo de 1 até T'. Os parâmetros variacionais μ e σ são aqueles aprendidos a partir do primeiro conjunto de treinamento, definido na Seção 4.1. Além disso, nós fixamos $\mu_{t+1} = 0$ para $t \geqslant 1$. Para fins de comparação, nós adotamos um procedimento análogo para predizer os mesmos parâmetros utilizando o modelo EGARCH. Os resultados desse experimento são apresentados na Figura 4.3.

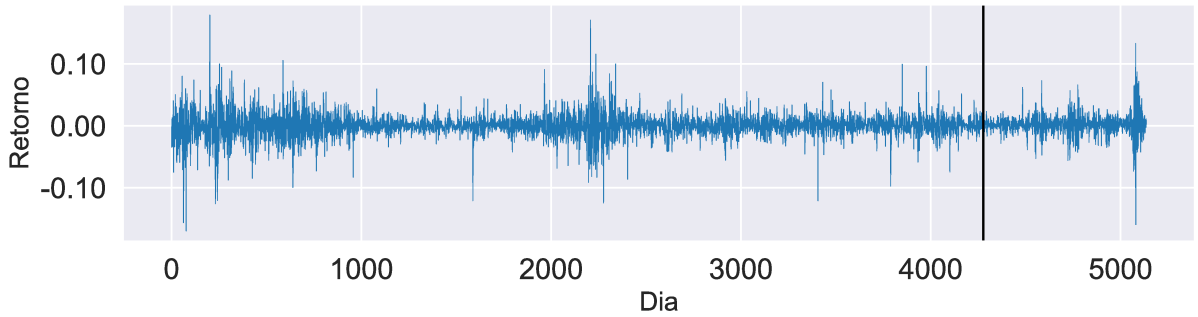
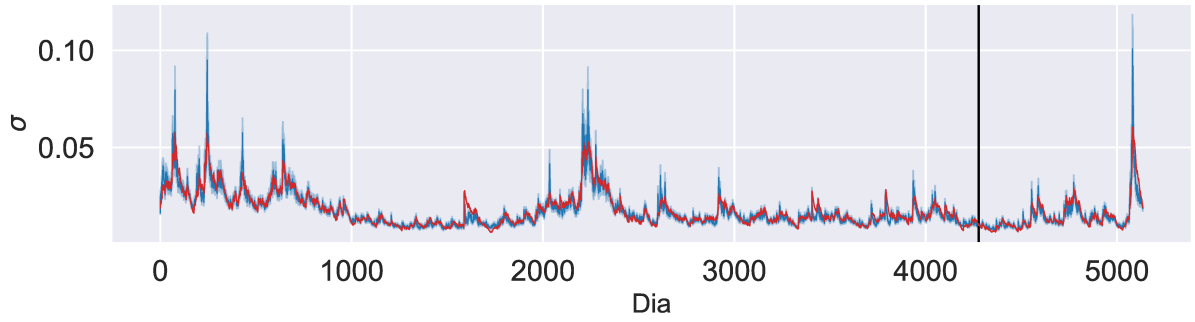
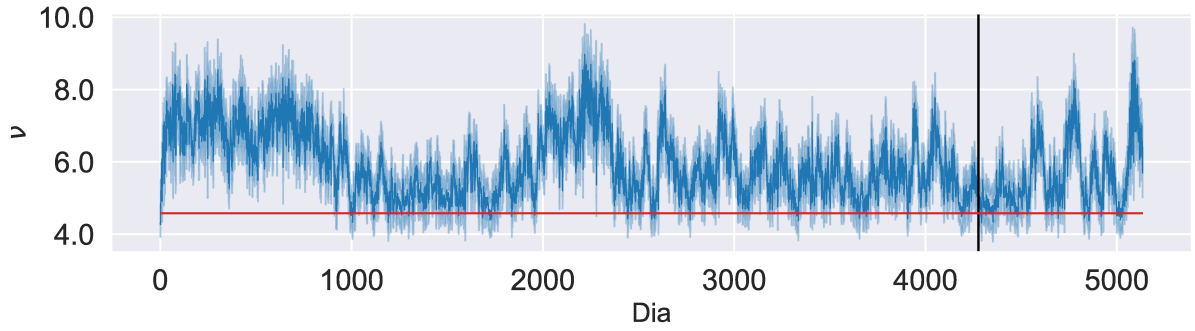


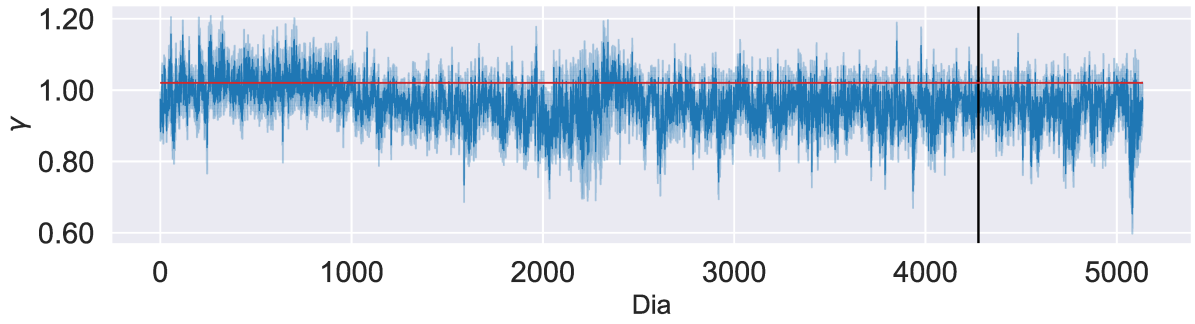
Figura 4.2: Sequência de retornos logarítmicos diários da ação da Microsoft no período de 03 de janeiro de 2000 a 03 de junho de 2020. A linha vertical em preto está posicionada no dia 03 de janeiro de 2017, início do período reservado exclusivamente para avaliação das previsões, isto é, o período para teste do DeepRisk e modelos concorrentes. Fonte: Elaborado pelo autor.



(a) Volatilidade



(b) Graus de Liberdade



(c) Assimetria

Figura 4.3: Previsão dos parâmetros da distribuição condicional dos retornos diários da ação da Microsoft. Em cada gráfico, a linha em azul escuro representa a mediana das previsões do nosso modelo, a área em azul claro identifica o respectivo intervalo de mais alta densidade (HDI) de 95%, e a linha em vermelho retrata as previsões do modelo EGARCH. A linha vertical em preto está posicionada no dia 03 de janeiro de 2017, início do período de avaliação das previsões. Fonte: Elaborado pelo autor.

As áreas em azul claro nos gráficos da Figura 4.3 representam os intervalos de mais alta densidade (HDIs) de 95% sobre as previsões do nosso modelo. O fato deles serem bastante estreitos sugere que o nosso modelo pode não quantificar bem a incerteza sobre os parâmetros distribucionais preditos. De fato, estudos recentes concluíram que procedimentos de inferência variacional, caso do método *Bayes by Backprop* (BBB) que nós empregamos, geralmente não produzem boas estimativas de incerteza [62, 102]. Sendo assim, ao invés de conduzir a nossa análise a partir desses intervalos, nós preferimos, por cautela, utilizar a mediana das previsões geradas.

Em relação às previsões de volatilidade, o nosso modelo parece se adequar melhor às diferentes situações de mercado do que o modelo EGARCH, o que é desejável quando se trata de prever uma medida de risco. Comparando os dois modelos, a volatilidade predita pelo DeepRisk tende a ser maior em períodos de turbulência e ligeiramente menor em períodos de estabilidade. Além disso, o nosso modelo se mostra mais sensível a mudanças drásticas no preço da ação subjacente. Por exemplo, no dia 16 de março de 2020 ($t = 5081$), quando o retorno da ação da Microsoft foi de -0,1595, a volatilidade predita pelo nosso modelo para o dia seguinte é de 0,1007. Essa previsão é cerca de 5,9 vezes o valor médio predito por esse modelo ao longo de todo período analisado e quase 1,7 vezes a volatilidade gerada pelo modelo alternativo para o mesmo dia.

Sobre os graus de liberdade, as previsões do nosso modelo variam entre 4,27 e 8,97, aproximadamente, enquanto aquelas geradas pelo modelo EGARCH são constantes ao longo do período analisado, com valor de 4,58. Em qualquer dos dois casos, nós podemos afastar a hipótese de normalidade da distribuição condicional dos retornos diários dessa ação. Segundo [69], as caudas da distribuição t de Student com graus de liberdade ν inferior a 30 são significativamente mais pesadas do que as caudas da distribuição normal; quando $\nu \geqslant 30$, as duas distribuições são praticamente equivalentes. Note que, sob o nosso modelo, ν_{t+1} tende a se adaptar ao mercado de forma similar à volatilidade σ_{t+1}. Durante os períodos de estabilidade, ν_{t+1} geralmente diminui de valor. Como consequência, os retornos se tornam mais concentrados em torno de zero, dentro do intervalo delimitado por $\pm\sigma_{t+1}$. Grandes variações no preço ainda são possíveis, por causa das caudas pesadas; porém, elas são menos prováveis. Por outro lado, em momentos de crise, ν_{t+1} tende a aumentar de valor, o que torna mais frequente a ocorrência de retornos com magnitude maior do que σ_{t+1}.

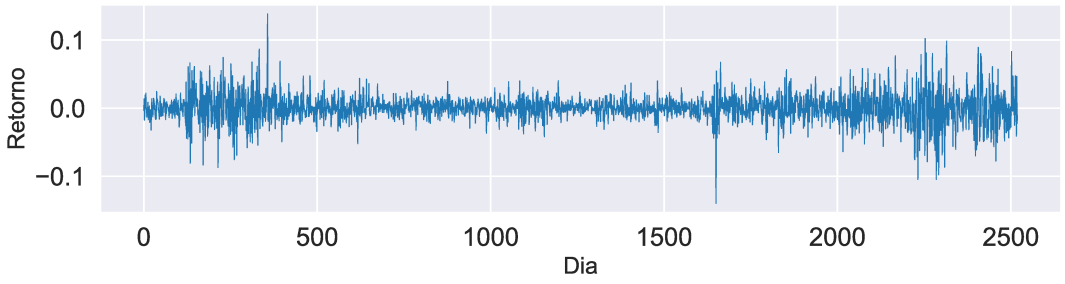
Por fim, a respeito da assimetria, os valores preditos pelo nosso modelo oscilam entre 0,65 e 1,16, ao passo que as previsões geradas pelo EGARCH são constantes e iguais a 1,02. Sob o nosso modelo, para $t \geqslant 1000$, γ_{t+1} parece flutuar em torno de um valor ligeiramente inferior a 1, ponto que indica simetria. De fato, o valor médio predito pelo nosso modelo nesse subperíodo é de aproximadamente 0,95. Ainda em relação a esse subperíodo, a mediana da correlação empírica entre γ_{t+1} e σ_{t+1} é de -0,3404 (fraca), e entre γ_{t+1} e ν_{t+1} é de -0,5767 (moderada). Logo, aparentemente, γ_{t+1} responde ao mercado de forma oposta aos outros dois parâmetros distribucionais. Sendo assim, em momentos de estabilidade, γ_{t+1} tende a crescer, tornando a distribuição condicional dos retornos diários da Microsoft mais próxima de simétrica. Por outro lado, em períodos de crise, γ_{t+1} tende a diminuir, tornando essa distribuição mais assimétrica à esquerda e, consequentemente, aumentando a probabilidade de ocorrência de retornos negativos de maior magnitude.

Embora o nosso modelo não tenha o objetivo de simular longas trajetórias de retornos, nós realizamos um segundo experimento para avaliar, visualmente, quão bem ele replica as principais características estatísticas dos retornos de uma ação de empresa

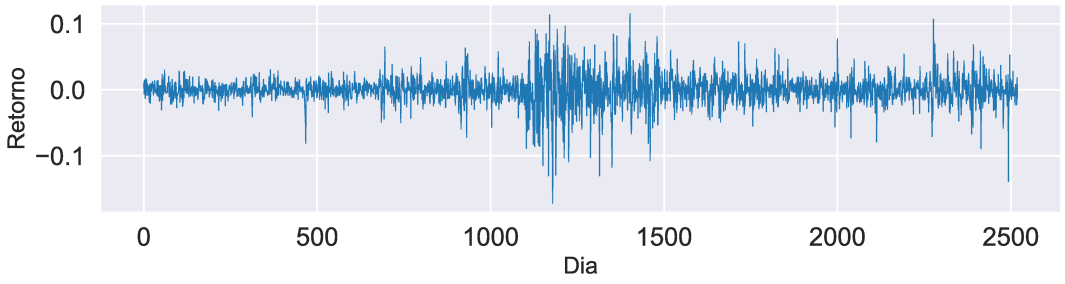
em um determinado período. Sejam $r_1, \ldots, r_{T'}$, com $T' = 2520$, os retornos logarítmicos diários da ação da Microsoft no período de 03 de janeiro de 2000 a 08 de janeiro de 2010. Para simular uma trajetória de tamanho T' condicionada nessa sequência, nós amostramos o conjunto de pesos w da distribuição variacional $\mathcal{N}(w|\mu, \sigma^2)$ e então, para $k = 1, ..., T'$, geramos $r_{T'+k}$ da distribuição $p(r_{T'+k}|\theta(r_1, \ldots, r_{T'}, r_{T'+1}, \ldots, r_{T'+k-1}; w))$. Os parâmetros variacionais μ e σ são aqueles aprendidos a partir do primeiro conjunto de treinamento, definido na Seção 4.1. Também fixamos $\mu_{t+1} = 0$ para todo $t \geqslant 1$. Repetimos esse procedimento até obter 1000 trajetórias. Finalmente, por inspeção visual, nós selecionamos 3 trajetórias que apresentam características similares às da sequência condicionante e outras 3 que falham em replicar pelo menos uma delas.

As Figuras 4.4 e 4.5 apresentam as trajetórias selecionadas. As Figuras 4.6, 4.7 e 4.8 comparam cada uma dessas trajetórias com a sequência condicionante em relação a: (i) caudas pesadas e assimetria entre perdas e ganhos, (ii) agrupamento de volatilidade e (iii) efeito de alavancagem. A Figura 4.6 sugere que a distribuição incondicional da Trajetória 4 é menos leptocúrtica do que a da sequência condicionante, o que indica que essa trajetória não representa adequadamente o fenômeno das caudas pesadas. Por sua vez, a Figura 4.7 aponta que as Trajetórias 5 e 6 falham em replicar o agrupamento de volatilidade. Em qualquer dos dois casos, a autocorrelação empírica do valor absoluto dos retornos é relativamente fraca quando o atraso é pequeno. E analisando especificamente a Trajetória 6, notamos que a autocorrelação converge para zero a uma taxa superior à da sequência condicionante. Para concluir, a Figura 4.8 deixa evidente que as Trajetórias 5 e 6 também não apresentam a característica conhecida como efeito de alavancagem. De fato, a correlação entre volatilidade e retorno passado na Trajetória 5 é fraca qualquer que seja o atraso. Na Trajetória 6, essa correlação converge para zero a uma taxa superior à da sequência condicionante.

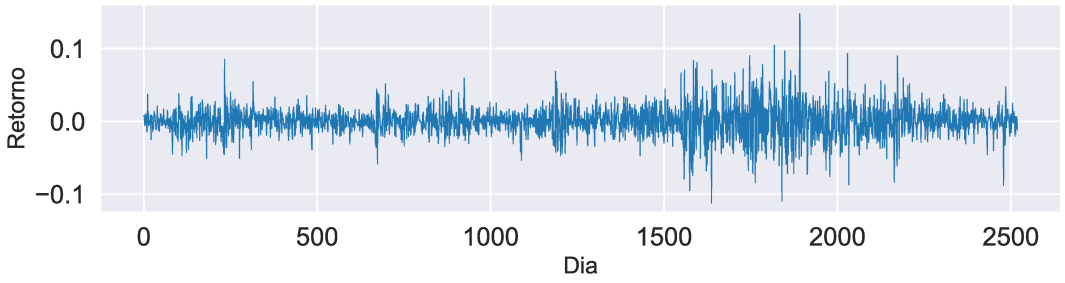
A principal hipótese que formulamos para explicar a geração de trajetórias que falham em replicar as principais características estatísticas dos retornos de uma ação é a adoção do método *teacher forcing* [41] para treinamento do modelo. Durante o treinamento, nós utilizamos sequências condicionantes formadas exclusivamente por retornos realizados. Por outro lado, para simular cada trajetória, nós amostramos suas observações condicionando o modelo a sequências parcialmente compostas por retornos realizados e por retornos simulados. Dessa forma, os espaços de entrada do modelo durante o treinamento e nesse experimento são, potencialmente, diferentes. Segundo [5], essa diferença pode resultar em erros que se acumulam rapidamente ao longo das sequências simuladas. Esses autores propuseram uma estratégia de aprendizado curricular para mitigar esse problema; ela consiste em escolher aleatoriamente, durante o treinamento do modelo, quando usar exemplos simulados no lugar de instâncias reais.



(a) Trajetória 1

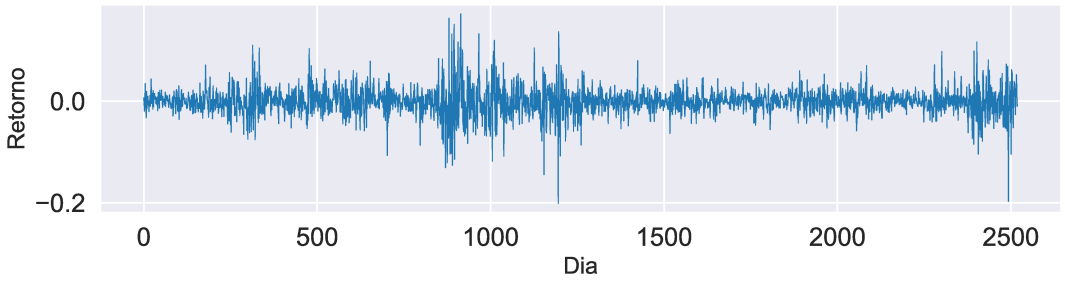


(b) Trajetória 2

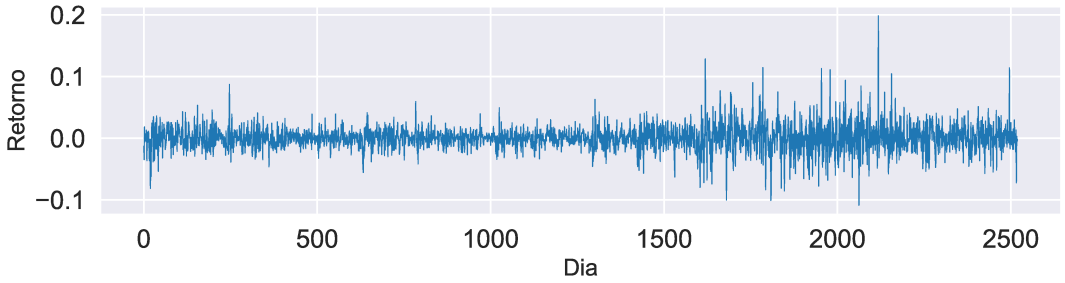


(c) Trajetória 3

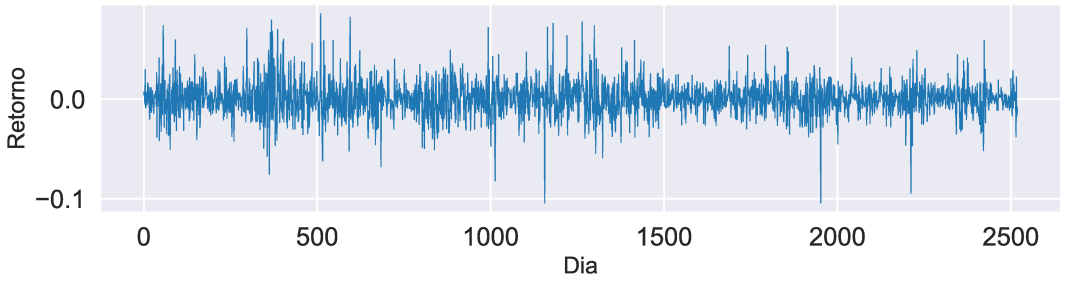
Figura 4.4: Trajetórias que apresentam características estatísticas similares às da sequência condicionante. Fonte: Elaborado pelo autor.



(a) Trajetória 4

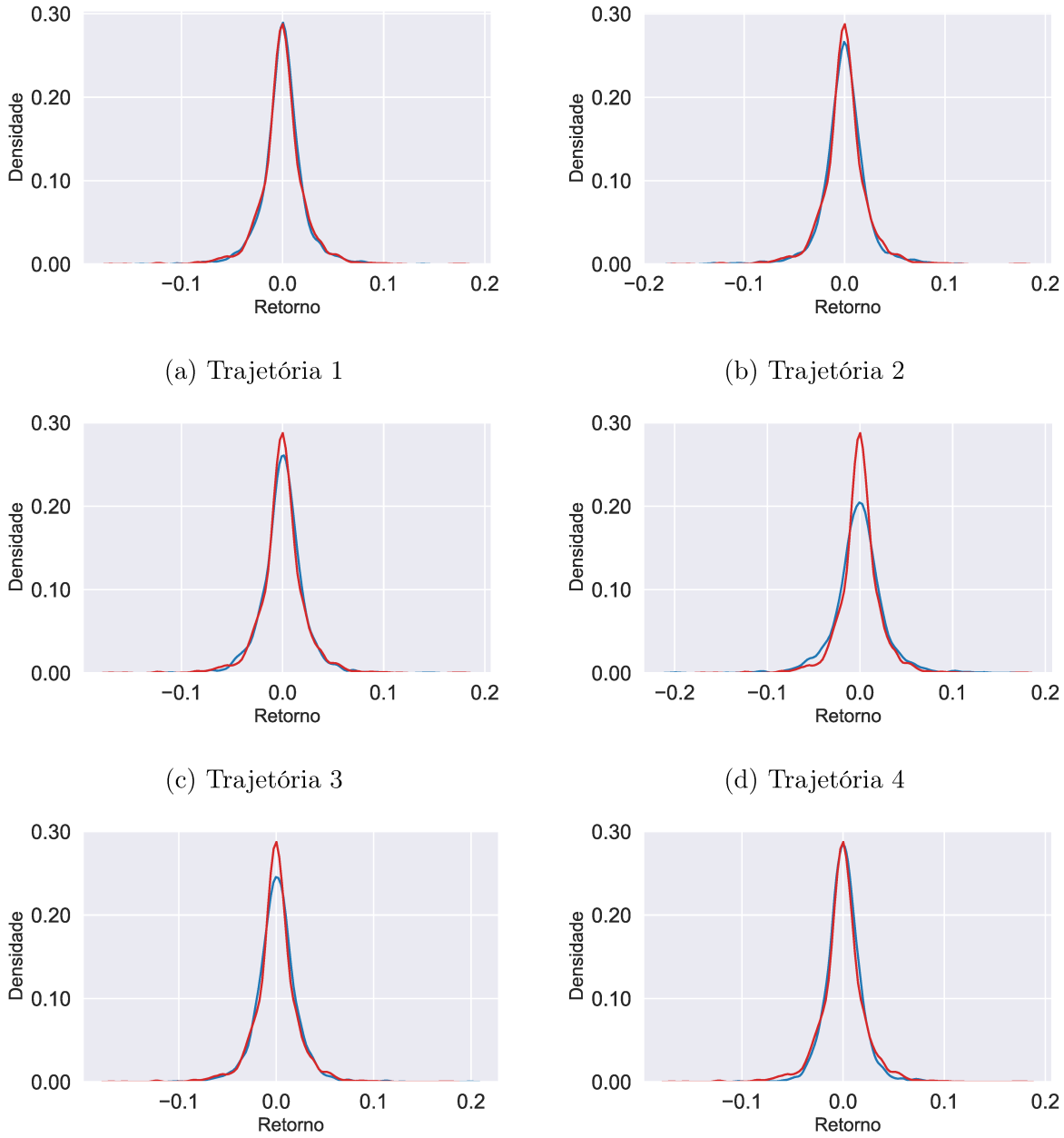


(b) Trajetória 5



(c) Trajetória 6

Figura 4.5: Trajetórias que falham em replicar alguma das principais características estatísticas da sequência condicionante. Fonte: Elaborado pelo autor.



(a) Trajetória 1

(b) Trajetória 2

(c) Trajetória 3

(d) Trajetória 4

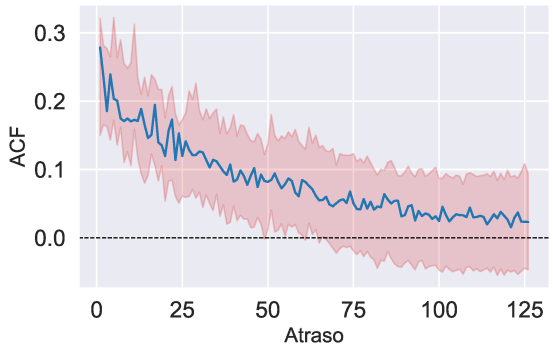
(e) Trajetória 5

(f) Trajetória 6

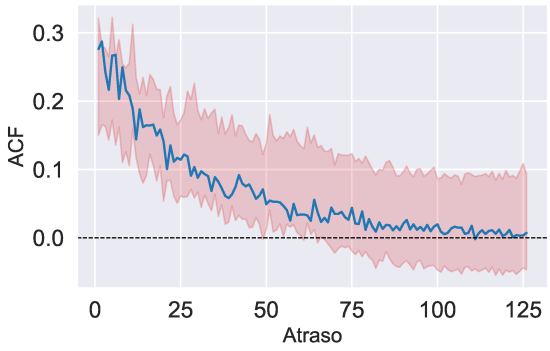
Figura 4.6: Comparação entre a densidade incondicional empírica de cada trajetória, em azul, e a da sequência condicionante, em vermelho. Fonte: Elaborado pelo autor.

(a) Trajetória 1

(b) Trajetória 2



(c) Trajetória 3



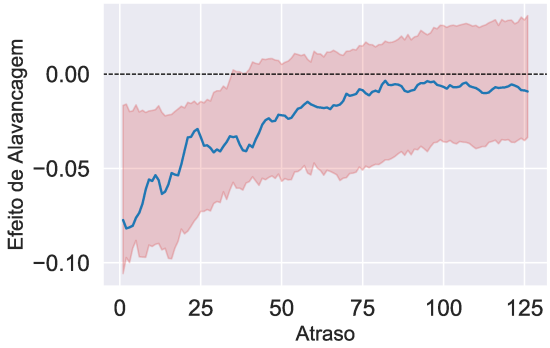
(d) Trajetória 4

(e) Trajetória 5

(f) Trajetória 6

Figura 4.7: Comparação entre a função de autocorrelação empírica do valor absoluto dos retornos de cada trajetória, em azul, e a da sequência condicionante, em vermelho. Utilizamos *bootstrapping* estacionário para estimar valores. As distribuições obtidas são resumidas da seguinte forma: mediana para as trajetórias e HDI de 95% para a sequência condicionante. Fonte: Elaborado pelo autor.

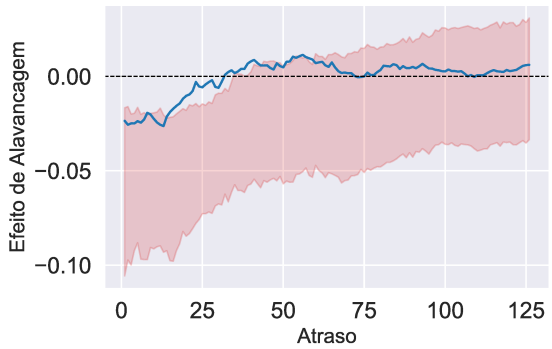
(a) Trajetória 1



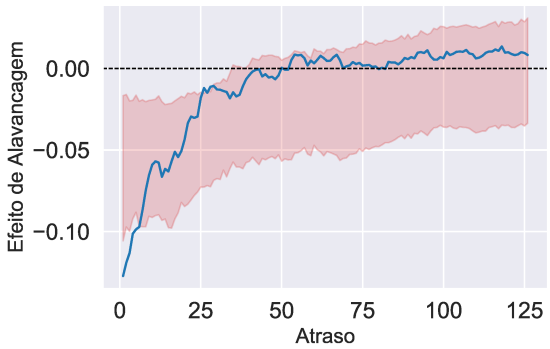
(b) Trajetória 2

(c) Trajetória 3

(d) Trajetória 4



(e) Trajetória 5



(f) Trajetória 6

Figura 4.8: Comparação entre o efeito de alavancagem estimado de cada trajetória, em azul, e o da sequência condicionante, em vermelho. Para estimar esse efeito, utilizamos o modelo EGARCH(1,1) para obter as volatilidades σ_t's e então calculamos, para o atraso $k = 1, \ldots, 126$, a correlação empírica entre σ_t e r_{t-k}. Utilizamos *bootstrapping* estacionário para estimar valores. As distribuições obtidas são resumidas da seguinte forma: mediana para as trajetórias e HDI de 95% para a sequência condicionante. Fonte: Elaborado pelo autor.

Capítulo 5

Conclusão

Nessa dissertação, nós propusemos um modelo generativo profundo para previsão de densidade de múltiplos passos em sequências univariadas de retornos diários de ações. Esse modelo, que nós chamamos de DeepRisk, é baseado em uma rede neural recorrente Bayesiana que foi especificada para capturar as principais propriedades distribucionais e da estrutura de dependência que nós observamos nessas sequências. Para avaliar a calibragem condicional das previsões do modelo, nós apresentamos duas funções de divergência baseadas na Transformação Integral da Probabilidade e na Discrepância Média Máxima (MMD). Aplicadas aos pseudo-resíduos normais calculados a partir das densidades preditas e dos respectivos retornos observados, essas funções nos permitem medir, separadamente, desvios em relação às condições ideais de normalidade incondicional e de independência serial. Usando esses desvios, introduzimos então um procedimento Bayesiano para comparar, conjuntamente, a acurácia preditiva de dois ou mais modelos concorrentes. Esse procedimento controla automaticamente a ocorrência de falsas descobertas; ele é, portanto, adequado para múltiplas comparações.

Experimentos com dados empíricos mostraram que introduzir caudas pesadas e assimetria na densidade condicional do DeepRisk fornece melhorias significativas de precisão e que esse modelo é melhor do que as alternativas convencionais em capturar as principais características distribucionais e da estrutura de dependência observadas em sequências reais de retornos. De fato, com respeito à normalidade incondicional, o DeepRisk apresentou desempenho credivelmente superior ao dos concorrentes em todos os horizontes de previsão avaliados. Já em relação à independência serial, o desempenho do nosso modelo foi, nos piores cenários, tão bom quanto o desempenho dos rivais mais competitivos. Análises qualitativas indicaram que o nosso modelo se ajusta bem a diferentes situações de mercado e que há espaço para aperfeiçoar a simulação de longas trajetórias de retornos, tarefa para qual o DeepRisk não foi projetado.

Para concluir, nós gostaríamos de esboçar direções promissoras para pesquisas futuras. A primeira delas diz respeito a avaliar outras famílias de distribuições para modelar a densidade condicional de sequências univariadas de retornos diários. Como exemplos, nós podemos citar as distribuições normal inversa Gaussiana, hiperbólica generalizada, Johnson SU e lambda generalizada. Para selecionar famílias candidatas, sugerimos a ado-

ção dos seguintes critérios: (i) existência de parâmetros para ajuste do grau de assimetria e do peso das caudas; e (ii) eficiência em tempo de execução e estabilidade numérica da implementação do procedimento de amostragem, da função de log-verossimilhança e sua derivada, e da função de distribuição acumulada (CDF). Dependendo do caso de uso, a inversa da CDF pode ser necessária. Também seria propício explorar a aplicação de redes adversárias generativas (GANs) para essa tarefa. Aqui vislumbramos duas possibilidades: (i) o desenvolvimento de técnicas e modelos que facilitem o aprendizado de distribuições com caudas pesadas, seguindo os passos de [59] e de [101]; e (ii) a adaptação do procedimento de comparação de modelos, proposto nessa dissertação, para situações nas quais existem concorrentes sem implementação disponível para a CDF, que é o caso das GANs.

A segunda direção consiste em estender o DeepRisk para o caso multivariado. Inspirados no trabalho de [99], um dos caminhos possíveis para essa extensão se baseia em modelos de fatores dinâmicos, considerados uma aplicação da análise de componentes principais para séries temporais. Um segundo caminho está relacionado à utilização de funções de cópula. Como exemplos, destacamos os modelos propostos por [88] e [22]. A vantagem das funções de cópula é a possibilidade de modelar, em uma primeira etapa, a distribuição de cada sequência univariada e, em uma etapa posterior, as dependências entre as sequências. Para selecionar cópulas para avaliação, recomendamos os mesmos critérios sugeridos no parágrafo anterior para a escolha de distribuições. Note que as características de assimetria e de peso nas caudas também são importantes nesse caso. De fato, elas permitem capturar o fato estilizado conhecido como dependência assimétrica nas caudas: grandes movimentos de preços, para baixo e para cima, são prováveis de ocorrer simultaneamente em duas ou mais ações de empresas. Essa extensão do DeepRisk levanta a necessidade de uma outra adaptação para a configuração multivariada: a do procedimento de comparação de modelos. Aqui a nossa principal sugestão é seguir os passos de [26], que propuseram um método de avaliação de previsões de densidade multivariada baseado na Transformação Integral da Probabilidade.

A terceira direção está relacionada ao processo de inferência e aprendizagem do DeepRisk. Para melhorar a calibragem da distribuição a posteriori sobre os pesos do modelo, nós sugerimos experimentar outros métodos de inferência aproximada, tendo em vista que resultados de nossos experimentos sugerem que o BBB (do inglês *Bayes by Backprop*) não produz boas estimativas de incerteza. Alguns dos procedimentos que nós consideramos promissores consistem em considerar múltiplos valores de pesos obtidos ao longo de trajetórias de descida estocástica de gradiente [102, 72, 61], e em aplicar um método Hamiltoniano de Monte Carlo (HMC) baseado em decomposições simétricas [17]. E para tornar as previsões de múltiplos passos mais precisas, sugerimos treinar o DeepRisk para mapear uma sequência de entrada em uma sequência de saída. Atualmente, usamos essa arquitetura sequência-para-sequência apenas para gerar as previsões. Para fixar o tamanho da sequência de saída, que pode ser diferente do da sequência de entrada,

recomendamos utilizar o maior horizonte de previsão do problema no qual o modelo será usado. Também aconselhamos avaliar a substituição, no treinamento, do método *teacher forcing* pela estratégia de aprendizado curricular proposta por [5]. O objetivo dessa estratégia é eliminar diferenças entre os espaços de entrada do modelo durante o treinamento e durante a geração das previsões.

A quarta e última direção está relacionada ao procedimento de comparação de modelos. Recomendamos estabelecer condições que as sequências de pseudo-resíduos normais devem satisfazer de modo a garantir a convergência das medidas de acurácia preditiva que nós propusemos. Um caminho que parece promissor consiste em assumir que o processo gerador dessas sequências é α-misturado [23]. Nós também sugerimos, conforme mencionado anteriormente nessa seção, estender o procedimento de comparação no sentido de suportar sequências multivaridas de retornos. Sobre o teste Bayesiano para habilidade preditiva equivalente, consideramos oportuno avaliar a introdução de uma região de equivalência prática (ROPE) sobre as diferenças médias estritamente padronizadas [69]. Com uma ROPE, podemos não apenas decidir pela rejeição ou não das hipóteses nulas, mas também por aceitá-las. A ROPE também ajuda a controlar falsas descobertas ao realizar múltiplas comparações. Por fim, indicamos conduzir uma nova análise de poder para esse teste, dessa vez variando tanto o tamanho das sequências quanto a quantidade delas. Nessa nova análise, seria interessante medir se a taxa de falsas descobertas ou a taxa de erro de família (FWER) está tão controlada quanto se espera, por construção, desse procedimento.

Referências Bibliográficas

[1] Gianni Amisano and Raffaella Giacomini. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2):177–190, 2007.

[2] Torben G. Andersen, Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold. Volatility and correlation forecasting. In G. Elliott, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, chapter 15, pages 777–878. Elsevier, 2006.

[3] Donald W.K. Andrews. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, 4(3):458–467, 1988.

[4] Yong Bao, Tae-Hwy Lee, and Burak Saltoğlu. Comparing density forecast models. *Journal of Forecasting*, 26(3):203–225, 2007.

[5] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[7] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[8] Jeremy Berkowitz. Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474, 2001.

[9] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

[10] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[11] Marine Carrasco and Xiaohong Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39, 2002.

[12] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark A Hasegawa-Johnson, and Thomas S Huang. Dilated recurrent neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] Sotirios P. Chatzis. Recurrent latent variable conditional heteroscedasticity. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2711–2715, 2017.

[14] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems, 2015.

[15] Peter F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, 1998.

[16] Peter F. Christoffersen. *Elements of Financial Risk Management*. Academic Press, San Diego, 2. edition, 2012.

[17] Adam D. Cobb and Brian Jalaian. Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting, 2020.

[18] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2. edition, 1988.

[19] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[20] Giorgio Corani, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon. Statistical comparison of classifiers through bayesian hierarchical modelling. *Machine Learning*, 106(11):1817–1837, November 2017.

[21] Valentina Corradi and Norman R. Swanson. Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, 135(1):187–228, 2006.

[22] Drew D. Creal and Ruey S. Tsay. High dimensional dynamic stochastic copula models. *Journal of Econometrics*, 189(2):335–345, 2015. Frontiers in Time Series and Financial Econometrics.

[23] James Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

[24] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, dec 2006.

[25] Francis X. Diebold, Todd A. Gunther, and Anthony S. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883, 1998.

[26] Francis X. Diebold, Jinyong Hahn, and Anthony S. Tay. Multivariate Density Forecast Evaluation and Calibration In Financial Risk Management: High-Frequency Returns on Foreign Exchange. *The Review of Economics and Statistics*, 81(4):661–673, 11 1999.

[27] Nicolae Dinculeanu. *Vector integration and stochastic integration in Banach spaces*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. John Wiley & Sons, Nashville, TN, January 2000.

[28] Matthew Francis Dixon and Igor Halperin. The four horsemen of machine learning in finance. *Available at SSRN 3453564*, 2019.

[29] R.F. Engle and G.G.J. Lee. A permanent and transitory component model of stock return volatility. In R.F. Engle and H. White, editors, *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, page 475–497. Oxford University Press, 1999.

[30] Carmen Fernandez and Mark F. J. Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[31] R.A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, 1959.

[32] Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian recurrent neural networks, 2019.

[33] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[34] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[35] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[36] Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

[37] Alexios Ghalanos. *rugarch: Univariate GARCH models*, 2014. R package version 1.4-0.

[38] C Lee Giles, Steve Lawrence, and Ah Chung Tsoi. Noisy time series prediction using recurrent neural networks and grammatical inference. *Machine Learning*, 44(1):161–183, jul 2001.

[39] Lawrence R Glosten, Ravi Jagannathan, and David E Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801, 1993.

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[41] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[42] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, May 2013.

[43] Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[44] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer, Berlin, 2012.

[45] Arthur Gretton. A simpler condition for consistency of a kernel independence test, 2015.

[46] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[47] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[48] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, page 63–77, Berlin, Heidelberg, 2005. Springer-Verlag.

[49] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273, 02 2020.

[50] Barbara Hammer. On the approximation capability of recurrent neural networks. *Neurocomputing*, 31(1):107–123, 2000.

[51] Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

[52] Geoffrey E. Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, page 5–13, New York, NY, USA, 1993. Association for Computing Machinery.

[53] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[54] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In John F. Kolen and Stefan C. Kremer, editors, *A Field Guide to Dynamical Recurrent Networks*, chapter 14, pages 237–243. Wiley-IEEE Press, 2001.

[55] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[56] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

[57] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[58] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.

[59] Todd Huster, Jeremy E. J. Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi Leslie, Cho-Yu Jason Chiang, and Vyas Sekar. Pareto gan: Extending the representational power of gans to heavy-tailed distributions, 2021.

[60] S&P Dow Jones Indices. S&p u.s. indices methodology, june 2021, 2021.

[61] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2019.

[62] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021.

[63] Mark Suresh Joshi. *The concepts and practice of mathematical finance*, volume 1. Cambridge University Press, 2003.

[64] Ha Young Kim and Chang Hyun Won. Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications*, 103:25–37, 2018.

[65] Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic volatility: Likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3):361–393, 1998.

[66] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[67] Malte Knüppel. Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33(2):270–281, 2015.

[68] Adriano Koshiyama, Nick Firoozye, and Philip Treleaven. Generative adversarial networks for financial trading strategies fine-tuning and combination. *Quantitative Finance*, 21(5):797–813, 2021.

[69] John Kruschke. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General*, 142, 07 2013.

[70] Sébastien Laurent, Jeroen V. K. Rombouts, and Francesco Violante. On the forecasting accuracy of multivariate garch models. *Journal of Applied Econometrics*, 27(6):934–955, 2012.

[71] Rui Luo, Weinan Zhang, Xiaojun Xu, and Jun Wang. A neural stochastic volatility model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[72] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[73] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, 2. edition, March 2020.

[74] James Mitchell and Kenneth F. Wallis. Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26(6):1023–1040, 2011.

[75] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2. edition, 2018.

[76] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[77] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[78] Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991.

[79] P. Nemenyi. *Distribution-free Multiple Comparisons*. Princeton University, 1963.

[80] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[81] Fernando De Meer Pardo and Rafael Cobo López. Mitigating overfitting on financial datasets with generative adversarial networks. *The Journal of Financial Data Science*, 2(1):76–85, 2020.

[82] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

[83] Ser-Huang Poon and Clive Granger. Practical issues in forecasting volatility. *Financial Analysts Journal*, 61(1):45–56, 2005.

[84] Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag New York, 2007.

[85] Barbara Rossi and Tatevik Sekhposyan. Alternative tests for correct specification of conditional predictive densities. *Journal of Econometrics*, 208(2):638–657, 2019.

[86] David Ruppert and David S. Matteson. *Statistics and Data Analysis for Financial Engineering*. Springer New York, 2015.

[87] Raif M. Rustamov. Closed-form expressions for maximum mean discrepancy with applications to wasserstein auto-encoders. *Stat*, 10(1):e329, 2021. e329 sta4.329.

[88] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank gaussian copula processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[89] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[90] Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 456–465, New York, NY, USA, 2020. Association for Computing Machinery.

[91] Albert N Shiryaev. *Essentials of Stochastic Finance*. World Scientific, January 1999.

[92] Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020. M4 Competition.

[93] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *J. Mach. Learn. Res.*, 13(1):1393–1434, may 2012.

[94] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, aug 2010.

[95] Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[96] Shuntaro Takahashi, Yu Chen, and Kumiko Tanaka-Ishii. Modeling financial time-series with generative adversarial networks. *Physica A: Statistical Mechanics and its Applications*, 527:121261, 2019.

[97] S. J. Taylor. Financial returns modelled by the product of two stochastic processes, a study of daily sugar prices 1961–79. In O. D. Anderson, editor, *Time Series Analysis: Theory and Practice 1*, pages 203–226, Amsterdam: North-Holland, 1982.

[98] R.S. Tsay. *Analysis of Financial Time Series*. CourseSmart. Wiley, 2010.

[99] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6607–6617. PMLR, 09–15 Jun 2019.

[100] Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

[101] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant gans: deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020.

[102] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc., 2020.

[103] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[104] Yue Wu, José Miguel Hernández-Lobato, and Zoubin Ghahramani. Gaussian process volatility model. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[105] Diethelm Wurtz, Yohan Chalabi, and Ladislav Luksan. Parameter estimation of arma models with garch/aparch errors. an r and splus software implementation. *Journal of Statistical Software*, 55(2):28–33, 2006.

[106] Xiuqin Xu and Ying Chen. Deep stochastic volatility model, 2021.

[107] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 441–451, New York, NY, USA, 2020. Association for Computing Machinery.

[108] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning, 2021.

[109] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021.

[110] Qiang Zhang, Rui Luo, Yaodong Yang, and Yuanyuan Liu. Benchmarking deep sequential models on volatility predictions for financial time series, 2018.

[111] Xiaohua Douglas Zhang. A pair of new statistical parameters for quality control in rna interference high-throughput screening assays. *Genomics*, 89(4):552–561, 2007.

[112] Xiaohua Douglas Zhang. *Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research*. Cambridge University Press, Cambridge, 2011.

Apêndice A

Ações Selecionadas

A Tabela A.1 apresenta as 204 ações de empresas selecionadas para nossos experimentos. As estatísticas descritivas consideram os retornos logarítmicos diários, em percentual, no período de 01 de janeiro de 2000 a 30 de dezembro de 2016.

Tabela A.1: Códigos de negociação, nomes e estatísticas descritivas das ações selecionadas

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
MMM	3M Company	0,040	1,47	-9,38	10,48	0,02	5,19
ABT	Abbott Laboratories	0,026	1,51	-17,60	11,75	-0,51	9,00
ADBE	Adobe Inc.	0,043	2,87	-35,32	21,50	-0,53	12,44
AES	AES Corporation	-0,022	3,58	-68,29	34,38	-1,82	48,44
AFL	Aflac Incorporated	0,032	2,44	-45,99	26,45	-1,38	44,11
APD	Air Products and Chemicals Inc.	0,045	1,79	-13,14	13,68	-0,17	5,42
ALL	Allstate Corporation	0,036	2,04	-23,80	19,63	-0,76	23,46
MO	Altria Group Inc	0,081	1,55	-14,90	15,17	-0,01	13,04
AEE	Ameren Corporation	0,032	1,35	-19,09	14,67	-0,59	17,65
AEP	American Electric Power Company Inc.	0,035	1,62	-25,86	18,10	-0,50	25,77

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
AXP	American Express Company	0,015	2,33	-19,35	18,77	-0,04	9,69
AIG	American International Group Inc.	-0,065	4,01	-93,63	50,68	-3,26	107,62
AMGN	Amgen Inc.	0,023	2,09	-14,41	14,06	0,24	5,64
ADI	Analog Devices Inc.	0,018	2,76	-17,68	16,48	0,24	5,01
AON	Aon Plc Class A	0,030	2,00	-36,14	19,96	-2,67	57,89
APA	Apache Corporation	0,032	2,41	-20,13	19,32	-0,10	5,12
AAPL	Apple Inc.	0,083	2,79	-73,12	13,02	-4,33	113,30
AMAT	Applied Materials Inc.	0,006	2,78	-15,10	22,83	0,31	4,69
ADM	Archer-Daniels-Midland Company	0,039	2,03	-18,43	15,99	-0,34	9,68
T	AT&T Inc.	0,016	1,66	-13,54	15,08	0,08	7,32
ADSK	Autodesk Inc.	0,052	2,70	-23,52	15,62	-0,37	6,91
ADP	Automatic Data Processing Inc.	0,029	1,58	-26,89	11,18	-1,13	24,63
AZO	AutoZone Inc.	0,075	1,77	-13,66	17,73	0,28	10,55
AVY	Avery Dennison Corporation	0,010	1,89	-15,87	11,45	-0,67	8,45
BKR	Baker Hughes Company Class A	0,030	2,58	-24,95	23,89	-0,23	7,27
BLL	Ball Corporation	0,068	1,71	-10,84	11,33	0,25	5,32
BAC	Bank of America Corp	0,008	3,04	-34,21	30,21	-0,34	25,54
BK	Bank of New York Mellon Corporation	0,013	2,47	-31,69	22,16	-0,11	17,64
BAX	Baxter International Inc.	0,030	1,69	-30,50	8,75	-2,42	37,82
BDX	Becton Dickinson and Company	0,049	1,53	-12,46	16,44	0,05	11,43
BBY	Best Buy Co. Inc.	0,022	3,06	-49,09	20,88	-2,07	34,81
BA	Boeing Company	0,039	1,92	-19,39	14,38	-0,30	5,81

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
BSX	Boston Scientific Corporation	0,016	2,40	-31,03	23,21	-0,30	13,52
BMY	Bristol-Myers Squibb Company	0,013	1,81	-25,44	13,67	-1,16	18,04
BF.B	Brown-Forman Corporation Class B	0,051	1,40	-10,41	12,04	0,14	5,35
CPB	Campbell Soup Company	0,022	1,40	-8,60	13,35	0,26	8,35
COF	Capital One Financial Corporation	0,018	3,12	-50,69	23,45	-1,27	26,04
CAH	Cardinal Health Inc.	0,032	1,81	-28,16	18,56	-1,17	22,49
CCL	Carnival Corporation	0,011	2,35	-38,38	15,26	-1,12	22,29
CAT	Caterpillar Inc.	0,043	2,06	-15,69	13,73	-0,07	4,41
CNP	CenterPoint Energy Inc.	0,020	2,37	-54,83	48,84	-2,22	131,67
CTL	CenturyLink Inc.	0,000	1,86	-25,61	16,29	-0,92	18,38
SCHW	Charles Schwab Corporation	0,015	2,82	-19,05	23,25	0,36	5,69
CVX	Chevron Corporation	0,037	1,64	-13,34	18,94	0,07	10,75
CI	Cigna Corporation	0,040	2,44	-47,91	21,14	-2,13	45,83
CINF	Cincinnati Financial Corporation	0,035	1,77	-22,40	16,77	-0,36	17,14
CSCO	Cisco Systems Inc.	-0,010	2,55	-17,69	21,82	0,17	8,61
C	Citigroup Inc.	-0,037	3,23	-49,47	45,63	-0,53	38,98
CTXS	Citrix Systems Inc.	0,009	3,35	-61,59	21,97	-1,80	33,18
CLX	Clorox Company	0,031	1,48	-12,36	12,41	-0,24	11,35
CMS	CMS Energy Corporation	0,019	2,01	-34,14	15,12	-2,27	35,00
KO	Coca-Cola Company	0,018	1,34	-10,60	13,00	0,04	8,66
CL	Colgate-Palmolive Company	0,024	1,40	-16,93	18,20	-0,07	18,01
CMA	Comerica Incorporated	0,021	2,48	-22,69	18,81	-0,22	11,45

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
CAG	Conagra Brands Inc.	0,033	1,42	-21,70	10,31	-0,90	18,21
COP	ConocoPhillips	0,037	1,91	-14,87	15,36	-0,34	5,64
ED	Consolidated Edison Inc.	0,039	1,17	-6,96	9,05	0,13	5,22
GLW	Corning Inc	-0,009	3,34	-43,42	19,59	-0,67	12,69
COST	Costco Wholesale Corporation	0,035	1,84	-24,25	13,98	-0,80	15,18
CSX	CSX Corporation	0,053	2,12	-17,64	14,10	-0,11	4,06
CMI	Cummins Inc.	0,065	2,61	-19,18	19,94	0,00	6,45
CVS	CVS Health Corporation	0,036	1,91	-26,13	13,04	-1,27	20,37
DHR	Danaher Corporation	0,051	1,62	-11,96	10,63	-0,06	4,97
DRI	Darden Restaurants Inc.	0,053	2,19	-23,95	18,20	-0,28	10,73
DE	Deere & Company	0,045	2,15	-15,33	14,87	-0,12	5,96
D	Dominion Energy Inc	0,048	1,36	-13,68	10,00	-0,56	10,11
DOV	Dover Corporation	0,024	1,89	-17,77	14,60	-0,18	5,34
DTE	DTE Energy Company	0,045	1,35	-9,28	12,20	0,11	7,03
DUK	Duke Energy Corporation	0,032	1,58	-16,14	14,98	-0,23	11,50
EMN	Eastman Chemical Company	0,040	2,08	-14,45	18,95	0,13	6,13
ETN	Eaton Corp. Plc	0,041	1,90	-12,51	17,37	0,01	5,18
ECL	Ecolab Inc.	0,047	1,53	-11,44	11,92	-0,08	6,44
EIX	Edison International	0,034	2,25	-42,67	30,28	-1,62	66,99
LLY	Eli Lilly and Company	0,015	1,69	-34,45	15,28	-1,84	45,72
EMR	Emerson Electric Co.	0,027	1,81	-16,27	14,32	-0,08	6,81
ETR	Entergy Corporation	0,040	1,55	-19,97	13,28	-0,50	13,05

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
EFX	Equifax Inc.	0,054	1,70	-13,89	11,49	-0,01	5,46
EXC	Exelon Corporation	0,032	1,67	-12,55	15,87	-0,04	7,73
XOM	Exxon Mobil Corporation	0,029	1,56	-15,03	15,86	0,03	10,35
FDX	FedEx Corporation	0,037	1,92	-15,65	11,18	-0,09	5,07
FITB	Fifth Third Bancorp	-0,004	3,37	-57,32	47,23	-0,34	56,48
FE	FirstEnergy Corp.	0,025	1,60	-10,77	16,29	0,03	10,44
F	Ford Motor Company	-0,009	2,76	-28,77	25,87	0,03	14,39
BEN	Franklin Resources Inc.	0,037	2,21	-16,58	14,94	-0,09	6,39
FCX	Freeport-McMoRan Inc.	0,016	3,34	-22,73	25,20	-0,19	5,19
GPS	Gap Inc.	-0,010	2,58	-23,64	24,09	-0,50	10,68
GD	General Dynamics Corporation	0,052	1,61	-13,21	11,09	-0,16	4,68
GE	General Electric Company	0,001	1,93	-13,68	17,98	0,05	8,70
GIS	General Mills Inc.	0,040	1,16	-11,92	9,01	-0,49	9,30
GPC	Genuine Parts Company	0,045	1,40	-9,48	9,65	0,19	4,09
GL	Globe Life Inc.	0,054	2,02	-15,23	40,89	1,67	50,10
HRB	H&R Block Inc.	0,029	2,13	-19,19	17,14	-0,64	10,08
HAL	Halliburton Company	0,029	2,84	-55,24	21,79	-1,72	38,19
HIG	Hartford Financial Services Group Inc.	0,008	3,71	-72,49	70,49	-0,52	84,38
HAS	Hasbro Inc.	0,042	1,98	-27,89	11,95	-1,02	18,03
HSY	Hershey Company	0,043	1,45	-12,71	22,54	1,17	23,74
HES	Hess Corporation	0,033	2,49	-21,27	15,44	-0,66	8,08
HD	Home Depot Inc.	0,023	2,04	-33,87	13,16	-0,97	22,17

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
HON	Honeywell International Inc.	0,026	2,05	-19,08	24,85	-0,23	13,03
HPQ	HP Inc.	-0,007	2,48	-22,35	15,95	-0,37	8,33
HUM	Humana Inc.	0,076	2,60	-28,02	18,49	-0,83	12,61
HBAN	Huntington Bancshares Incorporated	-0,001	3,46	-36,51	40,60	0,37	29,65
ITW	Illinois Tool Works Inc.	0,038	1,67	-10,13	12,26	0,10	5,00
INTC	Intel Corporation	0,006	2,42	-24,88	18,33	-0,49	8,85
IBM	International Business Machines Corporation	0,016	1,66	-16,89	11,35	-0,12	7,81
IFF	International Flavors & Fragrances Inc.	0,035	1,65	-30,08	14,94	-1,52	32,62
IP	International Paper Company	0,011	2,33	-20,48	19,79	0,01	8,73
IPG	Interpublic Group of Companies Inc.	-0,017	2,68	-35,40	30,10	-0,35	20,48
JNJ	Johnson & Johnson	0,031	1,22	-17,25	11,54	-0,47	15,86
JPM	JPMorgan Chase & Co.	0,024	2,56	-23,23	22,39	0,26	13,04
K	Kellogg Company	0,032	1,37	-9,69	10,29	0,13	7,30
KEY	KeyCorp	0,008	2,96	-40,55	43,34	-0,47	36,51
KMB	Kimberly-Clark Corporation	0,027	1,30	-11,54	9,02	-0,37	9,34
KLAC	KLA Corporation	0,020	2,96	-18,62	22,38	0,19	5,43
KSS	Kohl's Corporation	0,011	2,24	-20,82	14,99	-0,13	5,41
KR	Kroger Co.	0,034	1,75	-15,67	9,69	-0,41	6,19
LB	L Brands Inc.	0,046	2,36	-20,69	19,81	-0,09	5,92
LEG	Leggett & Platt Incorporated	0,033	1,92	-21,59	15,94	-0,48	11,33
LNC	Lincoln National Corporation	0,021	3,45	-50,89	36,23	-1,21	43,18
LIN	Linde plc	0,043	1,77	-13,47	13,86	0,08	6,82

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
LMT	Lockheed Martin Corporation	0,067	1,62	-11,55	13,71	0,14	7,01
L	Loews Corporation	0,039	1,85	-19,94	21,22	-0,47	18,52
LOW	Lowe's Companies Inc.	0,041	2,09	-11,57	16,92	0,32	4,79
MRO	Marathon Oil Corporation	0,030	2,47	-21,77	20,99	-0,19	7,83
MAR	Marriott International Inc. Class A	0,044	2,12	-23,64	14,12	-0,24	8,01
MMC	Marsh & McLennan Companies Inc.	0,018	1,82	-28,04	13,48	-0,90	21,89
MAS	Masco Corporation	0,018	2,50	-17,39	16,77	-0,17	5,80
MCD	McDonald's Corporation	0,036	1,49	-13,72	8,97	-0,16	6,47
MCK	McKesson Corporation	0,046	1,95	-25,71	14,87	-0,71	16,60
MDT	Medtronic Plc	0,021	1,64	-14,20	10,60	-0,53	7,65
MRK	Merck & Co. Inc.	0,012	1,78	-31,17	12,25	-1,44	28,73
MU	Micron Technology Inc.	-0,014	3,75	-26,19	21,06	-0,13	4,06
MSFT	Microsoft Corporation	0,010	1,98	-16,97	17,88	-0,12	9,54
TAP	Molson Coors Beverage Company Class B	0,038	1,70	-20,46	13,31	-0,43	12,26
MS	Morgan Stanley	-0,002	3,23	-29,97	62,59	1,31	45,48
MSI	Motorola Solutions Inc.	-0,025	2,83	-47,22	17,49	-1,53	26,89
NTAP	NetApp Inc.	-0,002	3,86	-23,45	34,30	0,26	8,46
NWL	Newell Brands Inc	0,021	2,09	-31,90	18,67	-0,87	21,63
NEM	Newmont Corporation	0,012	2,63	-15,19	22,45	0,19	4,42
NEE	NextEra Energy Inc.	0,054	1,44	-12,50	13,05	0,18	9,40
NKE	NIKE Inc. Class B	0,054	1,92	-21,65	13,34	-0,39	11,74
NSC	Norfolk Southern Corporation	0,048	2,13	-13,83	14,33	-0,02	3,84

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
NTRS	Northern Trust Corporation	0,019	2,24	-20,84	26,94	0,35	13,69
NOC	Northrop Grumman Corporation	0,062	1,59	-14,45	14,58	-0,17	7,66
NUE	Nucor Corporation	0,046	2,56	-22,10	24,51	-0,30	8,17
OXY	Occidental Petroleum Corporation	0,056	2,11	-20,45	16,64	-0,23	8,94
OMC	Omnicom Group Inc	0,019	1,84	-21,94	12,62	-0,37	10,35
ORCL	Oracle Corporation	0,009	2,57	-23,63	19,31	0,00	7,73
PCAR	PACCAR Inc	0,059	2,27	-13,14	17,26	0,02	4,37
PH	Parker-Hannifin Corporation	0,040	2,03	-12,42	12,81	-0,03	3,97
PAYX	Paychex Inc.	0,030	1,92	-14,14	13,35	0,04	5,68
PEP	PepsiCo Inc.	0,035	1,26	-12,71	13,86	0,04	11,65
PKI	PerkinElmer Inc.	0,027	2,52	-37,87	18,42	-1,27	24,70
PFE	Pfizer Inc.	0,013	1,61	-11,82	9,69	-0,22	5,31
PNW	Pinnacle West Capital Corporation	0,040	1,41	-10,37	9,45	-0,19	6,21
PNC	PNC Financial Services Group Inc.	0,034	2,45	-53,44	31,55	-1,46	67,96
PPG	PPG Industries Inc.	0,037	1,78	-12,21	13,83	0,01	5,19
PPL	PPL Corporation	0,043	1,60	-14,14	13,80	-0,53	9,10
PG	Procter & Gamble Company	0,020	1,39	-36,01	9,73	-4,27	110,41
PGR	Progressive Corporation	0,050	1,88	-21,36	21,49	0,23	16,79
PEG	Public Service Enterprise Group Inc	0,039	1,62	-10,99	15,81	0,03	8,15
PHM	PulteGroup Inc.	0,030	3,07	-20,45	20,73	0,15	4,10
QCOM	QUALCOMM Incorporated	-0,002	2,75	-18,44	17,12	-0,12	6,11
RTX	Raytheon Technologies Corporation	0,036	1,74	-33,20	12,79	-1,59	35,06

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
RF	Regions Financial Corporation	0,004	3,25	-52,88	39,48	-0,54	38,70
ROK	Rockwell Automation Inc.	0,034	2,67	-95,29	15,26	-10,87	381,57
SLB	Schlumberger NV	0,031	2,30	-20,34	13,90	-0,36	5,64
SEE	Sealed Air Corporation	0,018	2,44	-54,05	44,50	-2,34	108,19
SRE	Sempra Energy	0,054	1,58	-17,68	13,98	-0,46	11,39
SHW	Sherwin-Williams Company	0,067	1,82	-23,30	14,20	-0,56	14,82
SNA	Snap-on Incorporated	0,054	1,85	-17,57	14,14	0,01	8,66
SO	Southern Company	0,047	1,21	-8,85	10,50	0,24	6,70
LUV	Southwest Airlines Co.	0,037	2,24	-27,53	15,76	-0,65	9,35
SWK	Stanley Black & Decker Inc.	0,042	1,95	-15,38	11,84	0,01	5,43
STT	State Street Corporation	0,023	2,95	-89,25	27,27	-6,24	207,05
SYY	Sysco Corporation	0,034	1,49	-9,76	12,82	0,29	7,42
TROW	T. Rowe Price Group	0,042	2,41	-19,77	16,69	0,00	7,60
TGT	Target Corporation	0,022	2,05	-13,35	16,38	0,04	5,42
TXN	Texas Instruments Incorporated	0,015	2,58	-20,12	21,55	0,15	5,43
TXT	Textron Inc.	0,011	2,75	-38,06	39,78	-0,78	30,17
TMO	Thermo Fisher Scientific Inc.	0,057	1,87	-11,25	15,61	0,29	6,54
TJX	TJX Companies Inc	0,067	1,89	-14,59	16,25	0,24	5,50
TRV	Travelers Companies Inc.	0,041	1,88	-20,07	22,76	0,30	16,23
TFC	Truist Financial Corporation	0,027	2,18	-26,61	21,20	0,03	17,90
USB	U.S. Bancorp	0,036	2,22	-20,05	20,57	-0,12	14,48
UNP	Union Pacific Corporation	0,060	1,81	-15,09	8,67	-0,25	3,95

Continua na próxima página

Continuação da Tabela A.1

Código	Nome	Média	Desv. Pad.	Mínimo	Máximo	Assimetria	Curtose
UNH	UnitedHealth Group Incorporated	0,077	2,04	-20,62	29,83	0,26	19,81
UNM	Unum Group	0,015	2,84	-45,80	20,01	-2,75	50,13
VFC	V.F. Corporation	0,056	1,85	-14,64	13,60	0,15	6,08
VZ	Verizon Communications Inc.	0,017	1,60	-12,61	13,66	0,16	6,76
VMC	Vulcan Materials Company	0,033	2,18	-11,17	16,72	0,32	4,63
GWW	W.W. Grainger Inc.	0,044	1,74	-14,73	15,90	0,04	6,12
WBA	Walgreens Boots Alliance Inc	0,030	1,76	-16,24	15,39	-0,19	7,73
WMT	Walmart Inc.	0,007	1,52	-10,58	10,50	0,08	6,20
DIS	Walt Disney Company	0,035	1,96	-20,29	14,82	-0,09	9,41
WM	Waste Management Inc.	0,043	1,63	-13,70	23,32	0,78	16,77
WFC	Wells Fargo & Company	0,034	2,48	-27,21	28,34	0,88	26,71
WY	Weyerhaeuser Company	0,014	2,09	-18,83	13,14	-0,22	5,33
WHR	Whirlpool Corporation	0,034	2,42	-15,48	17,55	0,07	4,90
WMB	Williams Companies Inc.	0,018	3,92	-94,28	62,99	-3,33	112,77
XEL	Xcel Energy Inc.	0,036	1,81	-45,83	21,08	-4,72	133,41
XRX	Xerox Holdings Corporation	-0,017	2,89	-29,80	32,98	-0,26	16,81
XLNX	Xilinx Inc.	0,013	2,91	-23,69	16,60	-0,20	5,88
YUM	Yum! Brands Inc.	0,059	1,90	-23,29	14,30	-0,50	12,82

Apêndice B

Resultados de Comparação de Modelos

Para reportar credibilidade nos resultados de comparação de modelos da Seção 4.3, nós empregamos o teste Bayesiano para habilidade preditiva equivalente que é proposto na Seção 3.3. Para estimar as distribuições a posteriori sobre as diferenças médias estritamente padronizadas (SSMDs), nós utilizamos a implementação do algoritmo NUTS [56] do pacote *NumPyro* [82] e geramos amostras com 51000 diferenças, das quais descartamos os primeiros 1000 exemplos a título de *burn-in*. As Tabelas B.1-B.6 e as B.7-B.12 se referem aos resultados de normalidade incondicional e de independência serial, respectivamente.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 1$		
NORM	**SNORM**	0,560	0,877
NORM	**STD**	1,667	2,250
NORM	**SSTD**	1,834	2,470
SNORM	**STD**	1,472	2,000
SNORM	**SSTD**	1,735	2,334
STD	**SSTD**	1,102	1,547

Tabela B.1: HDIs de 95% sobre as SSMDs entre os resultados de normalidade incondicional do DeepRisk com diferentes distribuições de probabilidade e com horizonte de previsão $k = 1$. Para cada par de modelos comparados, o concorrente com a habilidade preditiva superior, quando há, está em **negrito**. "Superior" significa "o modelo com o melhor desempenho médio supera o modelo alternativo em todo ponto do HDI". Dado um ponto α do HDI, dizemos que o modelo A supera o modelo B quando $\alpha < 0$ ou que o modelo B supera o modelo A quando $\alpha > 0$. Cabe ressaltar que o desempenho de um modelo é medido aqui por uma função de divergência: quanto menor a divergência, maior a habilidade preditiva.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 5$		
NORM	**SNORM**	0,078	0,333
NORM	**STD**	1,689	2,272
NORM	**SSTD**	1,843	2,472
SNORM	**STD**	1,600	2,152
SNORM	**SSTD**	1,795	2,404
STD	**SSTD**	1,177	1,629

Tabela B.2: HDIs de 95% sobre as SSMDs entre os resultados de normalidade incondicional do DeepRisk com diferentes distribuições de probabilidade e com horizonte de previsão $k = 5$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 21$		
NORM	SNORM	-0,346	-0,096
NORM	**STD**	1,738	2,335
NORM	**SSTD**	1,847	2,476
SNORM	**STD**	1,786	2,392
SNORM	**SSTD**	1,923	2,564
STD	**SSTD**	1,246	1,718

Tabela B.3: HDIs de 95% sobre as SSMDs entre os resultados de normalidade incondicional do DeepRisk com diferentes distribuições de probabilidade e com horizonte de previsão $k = 21$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 1$		
GARCH	GJR-GARCH	-0,324	-0,100
GARCH	**EGARCH**	0,172	0,464
GARCH	**CGARCH**	0,123	0,375
GARCH	**DeepRisk**	0,120	0,377
GJR-GARCH	**EGARCH**	0,217	0,552
GJR-GARCH	**CGARCH**	0,152	0,424
GJR-GARCH	**DeepRisk**	0,146	0,420
EGARCH	CGARCH	-0,337	-0,102
EGARCH	**DeepRisk**	0,015	0,205
CGARCH	**DeepRisk**	0,077	0,303

Tabela B.4: HDIs de 95% sobre as SSMDs entre os resultados de normalidade incondicional de diferentes modelos utilizando a distribuição t de Student assimétrica e com horizonte de previsão $k = 1$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 5$		
GARCH	GJR-GARCH	-0,382	-0,125
GARCH	**EGARCH**	0,166	0,478
GARCH	**CGARCH**	0,142	0,420
GARCH	**DeepRisk**	0,140	0,432
GJR-GARCH	E**GARCH**	0,214	0,576
GJR-GARCH	**CGARCH**	0,170	0,476
GJR-GARCH	**DeepRisk**	0,168	0,481
EGARCH	CGARCH	-0,268	-0,058
EGARCH	**DeepRisk**	0,060	0,276
CGARCH	**DeepRisk**	0,096	0,345

Tabela B.5: HDIs de 95% sobre as SSMDs entre os resultados de normalidade incondicional de diferentes modelos utilizando a distribuição t de Student assimétrica e com horizonte de previsão $k = 5$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 21$		
GARCH	GJR-GARCH	-0,505	-0,210
GARCH	**EGARCH**	0,143	0,396
GARCH	**CGARCH**	0,179	0,444
GARCH	**DeepRisk**	0,175	0,454
GJR-GARCH	**EGARCH**	0,243	0,567
GJR-GARCH	**CGARCH**	0,235	0,544
GJR-GARCH	**DeepRisk**	0,213	0,518
EGARCH	CGARCH	-0,021	0,154
EGARCH	**DeepRisk**	0,121	0,362
CGARCH	**DeepRisk**	0,097	0,325

Tabela B.6: HDIs de 95% sobre as SSMDs entre os resultados de normalidade incondicional de diferentes modelos utilizando a distribuição t de Student assimétrica e com horizonte de previsão $k = 21$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 1$		
NORM	SNORM	-0,222	0,046
NORM	**STD**	0,185	0,469
NORM	**SSTD**	0,173	0,454
SNORM	**STD**	0,274	0,565
SNORM	**SSTD**	0,262	0,553
STD	SSTD	-0,118	0,154

Tabela B.7: HDIs de 95% sobre as SSMDs entre os resultados de independência serial do DeepRisk com diferentes distribuições de probabilidade e com horizonte de previsão $k = 1$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 5$		
NORM	SNORM	-0,246	0,019
NORM	STD	-0,007	0,264
NORM	SSTD	-0,014	0,258
SNORM	**STD**	0,085	0,384
SNORM	**SSTD**	0,069	0,367
STD	SSTD	-0,142	0,125

Tabela B.8: HDIs de 95% sobre as SSMDs entre os resultados de independência serial do DeepRisk com diferentes distribuições de probabilidade e com horizonte de previsão $k = 5$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 21$		
NORM	SNORM	-0,018	0,251
NORM	**STD**	0,033	0,326
NORM	SSTD	-0,007	0,270
SNORM	STD	-0,027	0,235
SNORM	SSTD	-0,093	0,163
STD	SSTD	-0,249	0,019

Tabela B.9: HDIs de 95% sobre as SSMDs entre os resultados de independência serial do DeepRisk com diferentes distribuições de probabilidade e com horizonte de previsão $k = 21$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 1$		
GARCH	**GJR-GARCH**	0,253	0,528
GARCH	**EGARCH**	0,383	0,679
GARCH	**CGARCH**	0,438	0,734
GARCH	**DeepRisk**	0,489	0,800
GJR-GARCH	**EGARCH**	0,148	0,422
GJR-GARCH	CGARCH	-0,047	0,216
GJR-GARCH	**DeepRisk**	0,312	0,599
EGARCH	CGARCH	-0,297	-0,031
EGARCH	**DeepRisk**	0,258	0,540
CGARCH	**DeepRisk**	0,307	0,599

Tabela B.10: HDIs de 95% sobre as SSMDs entre os resultados de independência serial de diferentes modelos utilizando a distribuição t de Student assimétrica e com horizonte de previsão $k = 1$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 5$		
GARCH	**GJR-GARCH**	0,203	0,475
GARCH	**EGARCH**	0,199	0,477
GARCH	CGARCH	-0,176	0,079
GARCH	**DeepRisk**	0,225	0,504
GJR-GARCH	EGARCH	-0,011	0,247
GJR-GARCH	CGARCH	-0,410	-0,147
GJR-GARCH	**DeepRisk**	0,056	0,316
EGARCH	CGARCH	-0,478	-0,202
EGARCH	**DeepRisk**	0,015	0,273
CGARCH	**DeepRisk**	0,295	0,587

Tabela B.11: HDIs de 95% sobre as SSMDs entre os resultados de independência serial de diferentes modelos utilizando a distribuição t de Student assimétrica e com horizonte de previsão $k = 5$.

Modelo A	Modelo B	Limite Inferior	Limite Superior
	$k = 21$		
GARCH	GJR-GARCH	-0,091	0,149
GARCH	EGARCH	-0,025	0,230
GARCH	CGARCH	-0,193	0,043
GARCH	DeepRisk	-0,014	0,249
GJR-GARCH	EGARCH	-0,030	0,221
GJR-GARCH	CGARCH	-0,187	0,047
GJR-GARCH	DeepRisk	-0,020	0,226
EGARCH	CGARCH	-0,256	0,006
EGARCH	DeepRisk	-0,083	0,143
CGARCH	**DeepRisk**	0,006	0,301

Tabela B.12: HDIs de 95% sobre as SSMDs entre os resultados de independência serial de diferentes modelos utilizando a distribuição t de Student assimétrica e com horizonte de previsão $k = 21$.