

Leonardo Santos Jacintho Blois Lopes

Previsão do retorno de portfólios financeiros a partir da combinação entre o
Modelo dos Fatores e HMM

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientador: Davi Michel Valladão
Co-Orientador: Mateus Cavalcante Waga

Departamento de Engenharia Industrial
Rio de Janeiro, 3 de julho de 2023

AGRADECIMENTOS

Aos meus avós Ana Luiza e Alberto, pelo amor, por tanto terem me apoiado e incentivado na trajetória acadêmica e profissional e por me inspirarem diariamente.

À minha mãe Ana Cristina e meus irmãos Sergio e Gustavo, pelo amor, companheirismo e apoio incondicional.

Aos meus amigos, pelas risadas e momentos de descontração, mas também pelos papos sérios e conselhos.

Ao Colégio de São Bento, por ter me formado como pessoa, aluno e cidadão e ter me apresentado à uma segunda família.

À PUC-Rio, o CTC e o Departamento de Engenharia Industrial e todo seu corpo docente, direção, administração e funcionários em geral, pela oportunidade e por terem disponibilizado a melhor estrutura e organização possíveis.

Ao professor e amigo Bruno Fanzeres, pelas aulas, pesquisas e conversas que tanto me motivaram a seguir buscando a excelência.

Aos meus orientadores, Davi e Mateus, que me guiaram e auxiliaram no desenvolvimento desse projeto.

RESUMO

Este trabalho de conclusão de curso propõe o uso de *Hidden Markov Models* (HMM) para prever os fatores de risco do Modelo dos Fatores e, em seguida, utiliza esses fatores para prever os retornos de portfólios no mercado financeiro. Além disso, realiza uma avaliação do risco associado a esses investimentos. O estudo visa explorar a aplicação do HMM nesse contexto e fornecer *insights* sobre a relação entre os fatores de risco identificados e os retornos do mercado financeiro. Os resultados obtidos podem contribuir para uma melhor compreensão dos investimentos e auxiliar na gestão de risco de portfólios.

Palavras-chave: HMM, Modelo dos Fatores, Retornos, Medidas de Risco

ABSTRACT

This undergraduate thesis proposes the use of the Hidden Markov Model (HMM) to predict the risk factors of the Five-Factor Model and subsequently employs these factors to forecast portfolio returns in the financial market. Additionally, it conducts an evaluation of the associated risk of these investments. The study aims to explore the application of HMM in this context and provide insights into the relationship between the identified risk factors and market returns. The obtained results can contribute to a better understanding of investments and assist in portfolio risk management.

Key-words: HMM, Five-Factor Model, Returns, Risk Measures

SUMÁRIO

1	INTRODUÇÃO	1
2	REFERENCIAL TEÓRICO	3
2.1	Modelo dos Fatores	3
2.2	<i>Hidden Markov Models</i>	5
3	METODOLOGIA	9
3.1	Dados	9
3.2	Modelo	11
4	RESULTADOS	14
5	CONCLUSÃO	22

LISTA DE FIGURAS

1	Ilustração do funcionamento do algoritmo <i>Forward-Backward</i> (ZHANG, 2004)	7
2	Ilustração do funcionamento do algoritmo de Viterbi (ZHANG, 2004)	8
3	Gráfico da função densidade de probabilidade estimada dos fatores de risco . .	11
4	Fluxograma do processo de cálculo da previsão de retorno dos portfólios financeiros	13
5	Gráfico dos retornos previstos para o setor de energia	17
6	Gráfico dos retornos previstos para o setor de energia com as diferentes metodologias	20

LISTA DE TABELAS

1	Equações de Fama-French	14
2	Erro Absoluto Médio e Erro Quadrático Médio do modelo	18
3	Comparação da previsão utilizando diferentes metodologias	19
4	Probabilidade do modelo acertar valorização ou desvalorização	21

1 INTRODUÇÃO

A previsão de retornos no mercado financeiro é um desafio complexo enfrentado por investidores. Através de previsões precisas, os investidores podem tomar decisões informadas sobre quando comprar, vender ou manter determinados ativos em sua carteira, buscando otimizar sua alocação de recursos e alcançar seus objetivos financeiros.

Existem várias fontes de incerteza e complexidade que dificultam a tarefa de prever com precisão os futuros retornos de um ativo. Fatores econômicos, políticos e eventos inesperados podem afetar significativamente o mercado financeiro, tornando a previsão de retornos ainda mais desafiadora. Dessa maneira, compreender e prever os fatores que impulsionam esses retornos é essencial para a tomada de decisões e a gestão de riscos.

Com isso, ao obter previsões precisas, os gestores podem avaliar e quantificar os riscos associados aos investimentos de forma mais efetiva. Isso permite uma alocação de recursos mais informada e uma melhor compreensão dos potenciais ganhos e perdas. A previsão precisa de retornos também auxilia na identificação de cenários de risco, permitindo a implementação das estratégias de mitigação adequadas. Além disso, facilita a avaliação do desempenho do portfólio e o acompanhamento contínuo dos riscos envolvidos.

O Modelo dos Fatores tem sido amplamente utilizado para descrever os retornos de ativos por meio da consideração de fatores específicos, como tamanho da empresa, valor, lucratividade e investimento (FAMA; FRENCH, 2015). Já os *Hidden Markov Models*, introduzidos primeiramente por Baum e Petrie (1966), pode ser utilizado para descrever o retornos de ativos financeiros considerando não apenas a dependência das observações passadas (indiretamente), mas também a influência de estados ocultos que afetam a geração desses retornos.

Nesse contexto, uma união entre o Modelo dos Fatores e o HMM surge como uma abordagem promissora, devido à sua capacidade de capturar a dependência temporal dos dados e identificar e considerar fatores subjacentes. Portanto, essa união permite uma interação complexa entre os fatores de risco e os estados ocultos, fornecendo uma visão mais completa da dinâmica dos retornos de ativos.

No entanto, a escolha do modelo mais adequado para a previsão de retornos depende das características dos dados, da disponibilidade de informações relevantes e dos objetivos e necessidades específicas dos investidores. É necessário considerar as limitações e os desafios inerentes à aplicação do Modelo dos Fatores com o HMM e avaliar sua relevância em rela-

ção a outras abordagens disponíveis. É fundamental realizar análises criteriosas, considerar as limitações do modelo, ponderar o custo-benefício da complexidade e compará-lo com outras abordagens disponíveis a fim de determinar sua eficácia em diferentes cenários e contextos de mercado.

Neste trabalho, será testada a acurácia de um modelo que utiliza *Hidden Markov Models* para prever observações futuras de fatores de risco e o Modelo dos Fatores para, consequentemente, a partir desses fatores de risco, prever o retorno de ativos financeiros.

Utilizando a abordagem descrita acima, considerando uma análise absoluta de valorização ou desvalorização dos ativos, esse modelo obteve uma precisão de aproximadamente 55%. Contudo, de maneira geral, identificamos que, apesar de prever corretamente, de forma absoluta, para a maior parte dos casos, esse modelo obteve uma acurácia inferior à outras abordagens menos complexas, devendo ser analisado o custo benefício de sua aplicação.

2 REFERENCIAL TEÓRICO

2.1 Modelo dos Fatores

Em "The cross-section of expected stock returns (1992)", Eugene F. Fama e Kenneth R. French enxergam deficiências no modelo CAPM de Sharpe (1964) e Lintner (1965), principalmente no fato de o modelo considerar apenas uma variável explicativa (o excesso de retorno do mercado de ações) como responsável pelo retorno dos ativos.

A equação do modelo CAPM (*Capital Asset Pricing Model*) está representada abaixo:

$$R_{it} = R_{Ft} + \beta_i (R_{Mt} - R_{Ft})$$

onde,

- R_{it} é o retorno do ativo i para o período t
- R_{Ft} é o retorno livre de riscos
- R_{Mt} é o retorno do mercado de ações

Após isso, no ano seguinte ao questionamento, os autores propõem a existência de três fatores de risco principais que influenciam os retornos de ativos: o excesso de retorno do mercado de ações, a diferença entre retornos de empresas com alta e baixa capitalização de mercado (conhecido como fator de tamanho) e a diferença entre retornos de empresas com alto e baixo *book-to-market ratio* (conhecido como fator de valor) (FAMA; FRENCH, 1993).

O fator de tamanho reflete a percepção de que empresas menores podem ter um desempenho diferente das empresas maiores devido a diferentes características e dinâmicas do mercado, enquanto que o fator de valor visa capturar o efeito de empresas consideradas subvalorizadas (alto *book-to-market ratio*) e sobrevalorizadas (baixo *book-to-market ratio*) no mercado.

A equação do modelo de Três Fatores está representada abaixo:

$$R_{it} - R_{Ft} = a_i + b_i (R_{Mt} - R_{Ft}) + s_i \text{SMB}_t + h_i \text{HML}_t + e_{it}$$

onde,

- SMB_t é o fator de tamanho, que é dado pela diferença entre retornos em um portfólio diversificado de pequenas ações e um portfólio diversificado de grandes ações (*Small*

Minus Big)

- HML_t é o fator de valor, que é dado pela diferença entre retornos em um portfólio diversificado de ações com alto *book-to-market ratio* e um portfólio diversificado de ações com baixo *book-to-market ratio* (*High Minus Low*)

Posteriormente, em "A five-factor asset pricing model (2015)", Fama e French revisaram esse modelo, que já estava desatualizado. Estudos posteriores ao primeiro modelo destes autores, como Novy-Marx (2013) e Titman, Wei e Xie (2004), indicavam que o modelo era incompleto, pois não capturava a variação do retorno médio relacionada a rentabilidade e investimentos das companhias.

Com isso, motivados por esses estudos, Fama e French adicionaram mais duas variáveis explicativas ao *Three-Factor Model*: a diferença entre o retorno de empresas mais e menos rentáveis e a diferença entre retornos de empresas com baixo e elevado investimento, tornando-o *Five-Factor Model*.

Essas duas novas variáveis são adicionadas a partir da percepção de que empresas com diferentes estratégias de investimento e alocação de recursos podem gerar retornos discrepantes, o que torna o fator de investimento relevante na explicação dos retornos dos ativos, e de que empresas mais lucrativas tendem a apresentar um desempenho diferente das empresas menos lucrativas, refletindo a influência da lucratividade nos retornos dos ativos

A equação do modelo de Cinco Fatores está representada abaixo:

$$R_{it} - R_{Ft} = a_i + b_i (R_{Mt} - R_{Ft}) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{it}$$

onde,

- RMW_t é o fator de lucratividade, que é dado pela diferença entre retornos em um portfólio diversificado de ações com robusta lucratividade e um portfólio diversificado de ações com fraca lucratividade (*Robust Minus Weak*)
- CMA_t é o fator de investimento, que é dado pela diferença entre retornos em um portfólio diversificado de ações com baixo investimento, conservadoras, e um portfólio diversificado de ações com elevado investimento, agressivas (*Conservative Minus Agressive*)

Com isso, o Modelo dos Fatores é vantajoso pois permite, além de uma maior precisão na precificação de ativos, uma explicação mais detalhada dos retornos e uma melhor discriminação de risco.

Esses fatores de risco são disponibilizados mensalmente pelo próprio Kenneth French, e sua metodologia de cálculo pode ser verificada em Fama e French (2015).

2.2 *Hidden Markov Models*

Cadeias de Markov são processos estocásticos que descrevem a sequência de variáveis aleatórias ao longo do tempo, de maneira que satisfaça a propriedade de Markov. Ou seja, as probabilidades referentes a como o processo evolui no futuro dependem apenas do estado atual do processo e, portanto, são independentes de eventos no passado. Isso é, em processos markovianos, condicionar ao histórico é equivalente a condicionar somente para a última observação.

$$\mathbb{P}(X_t \mid X_1, X_2, \dots, X_{t-1}) = \mathbb{P}(X_t \mid X_{t-1})$$

HMMs são, na tradução literal, Modelos de Markov Ocultos. Isso se dá porque são processos em que, além de satisfazerem a propriedade de Markov, a observação vista no tempo t é gerada por um processo cujo estado no tempo t não é conhecido/visto diretamente pelo observador (oculto). Diferentes processos (ocultos) podem ter diferentes distribuições de probabilidade de observações.

Com isso, podemos caracterizar um HMM como $\text{HMM} = \{S, A, O, B, \pi\}$, onde:

- $S = \{X_1, X_2, \dots, X_n\}$ é um conjunto de n estados
- $A = \{a_{ij}\}$ é a matriz de transição de estados, em que a_{ij} representa a probabilidade de irmos do estado i para o estado j

$$a_{ij} = \mathbb{P}(S_t = X_j \mid S_{t-1} = X_i) \quad \text{e} \quad \sum_j^n a_{ij} = 1$$

- $O = \{o_1, o_2, \dots, o_T\}$ é a sequência de T observações
- $B = \{b_j(k)\}$ é a distribuição de probabilidade das observações no estado j

$$b_j = \mathbb{P}(o_t = k \mid S_t = X_j)$$

- $\pi = \{\pi_i\}$ é a distribuição de probabilidades iniciais dos estados

$$\pi_i = \mathbb{P}(S_1 = X_i) \quad \text{e} \quad \sum_{i=1}^n \pi_i = 1$$

Os parâmetros de um HMM são caracterizados por $\lambda = (A, B, \pi)$

De acordo com Rabiner (1989), com um HMM, temos 3 problemas fundamentais, são eles:

1. Qual a probabilidade de o modelo gerar uma sequência específica de observações?
2. Qual a sequência de estados que melhor explica uma sequência de observações?
3. Como ajustar os parâmetros do modelo de modo a maximizar a probabilidade de gerarmos uma sequência específica de observações?

Para resolver esses problemas fundamentais, temos 3 algoritmos, o algoritmo *Forward-Backward* (CHANG; HANCOCK, 1966), o algoritmo de Viterbi (VITERBI, 1967) e o algoritmo de Baum-Welch (BAUM; PETRIE *et al.*, 1970).

O algoritmo *Forward-Backward*, que resolve o primeiro problema, utiliza indução para criar dois *sets* de probabilidades, o *forward* e o *backward*. Depois, esses dois conjuntos são utilizados para formar um único grupo mais ajustado. Para esse algoritmo, vamos introduzir algumas variáveis auxiliares no decorrer do procedimento:

A parte *forward* do algoritmo pode ser descrita da forma abaixo:

$$\alpha_t(i) = \mathbb{P}(O_1, O_2, \dots, O_t, S_t = X_i \mid \lambda)$$

- Caso base: $\alpha_1(i) = \pi_i b_i(O_1)$
- Passo indutivo: $\alpha_{t+1}(j) = \left[\sum_{i=1}^n \alpha_t(i) a_{ij} \right] b_j(O_{t+1})$
- Passo final: $\mathbb{P}(O \mid \lambda) = \sum_{i=1}^n \alpha_T(i)$

Agora, ao contrário do *forward*, em que calculamos α avançando no tempo, para a parte *backward*, calcularemos β voltando no tempo:

$$\beta_t(i) = \mathbb{P}(O_{t+1}, O_{t+2}, \dots, O_T \mid S_t = X_i, \lambda)$$

- Caso base: $\beta_T(i) = 1$

- Passo indutivo: $\beta_t(i) = \sum_{j=1}^n a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$

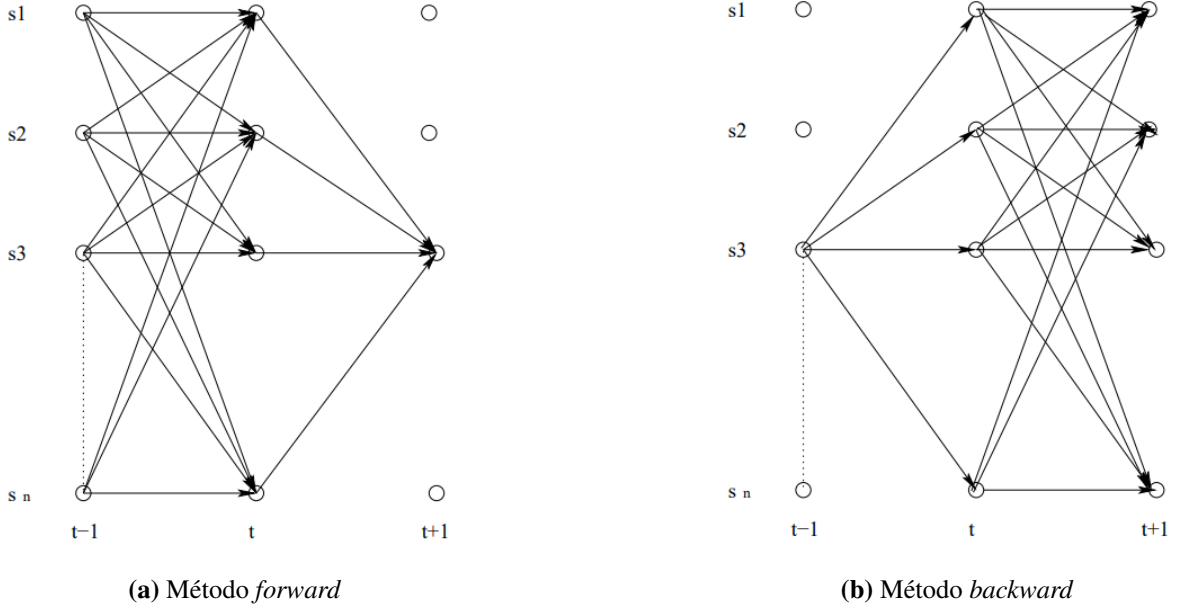


Figura 1: Ilustração do funcionamento do algoritmo *Forward-Backward* (ZHANG, 2004)

Como pode ser visto na figura 1, o funcionamento dos dois métodos do algoritmo *Forward-Backward* ocorre de maneira em que as variáveis *forward* e *backward* são calculadas indutivamente no tempo, sendo a *forward* avançando no tempo e a *backward* voltando.

Falando agora do algoritmo de Viterbi, que é utilizado para resolver o segundo problema, temos que:

$$\gamma_t(i) = \mathbb{P}(S_t = X_i \mid O, \lambda) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)}$$

Temos que a sequência de estados individualmente mais provável é

$$S' = \operatorname{argmax}_{1 \leq i \leq n} \gamma_t(i)$$

Entretanto, essa sequência pode não fazer sentido, uma vez que estando no estado X_i seja impossível se transferir para o estado X_j , pois pode haver um caso em que, $\gamma_t = X_i$ e $\gamma_{t+1} = X_j$ - o que representaria uma mudança de estado que não pode ter ocorrido.

Por isso, o algoritmo de Viterbi calcula a sequência de estados mais provável pelo algoritmo abaixo, que leva em consideração as probabilidades de transição entre estados:

$$\delta_t(i) = \max_{S_1, S_2, \dots, S_{t-1}} \mathbb{P}(\{S_1, S_2, \dots, S_t = X_i\}, \{O_1, O_2, \dots, O_t\} \mid \lambda)$$

- Caso base: $\delta_1(i) = \pi_i b_i(O_1)$ $\psi_1(i) = 0$
- Passo indutivo: $\delta_t(j) = \max_{1 \leq i \leq n} [\delta_{t-1}(i) a_{ij}] b_j(O_t)$ $\psi_t(j) = \operatorname{argmax}_{1 \leq i \leq n} [\delta_{t-1}(i) a_{ij}]$
- Passo final: $P^* = \max_{1 \leq i \leq n} [\delta_T(i)]$ $q_T^* = \operatorname{argmax}_{1 \leq i \leq n} [\delta_T(i)]$

Na figura 2, está ilustrado o funcionamento do algoritmo de Viterbi para encontrar a melhor sequência de estados definida pela sequência de observações.

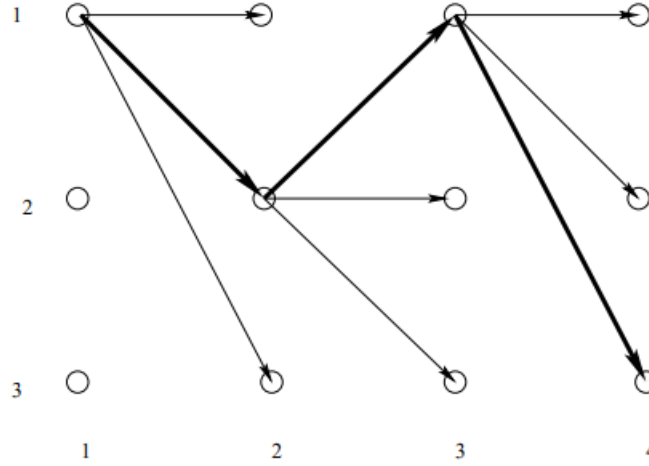


Figura 2: Ilustração do funcionamento do algoritmo de Viterbi (ZHANG, 2004)

Finalmente, o último problema dos HMMs não possui uma solução analítica conhecida. Todavia, podemos utilizar um algoritmo de maximização local, o algoritmo de Baum-Welch, que é um caso especial de algoritmo de Maximização da Expectativa. Para trabalhar iterativamente a fim de encontrar um modelo ótimo, temos:

$$\xi_t(i, j) = \mathbb{P}(S_t = X_i, S_{t+1} = X_j \mid O, \lambda) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^n \sum_{j=1}^n \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}$$

Sendo que, como $\gamma_t(i)$ é a probabilidade de estarmos no estado X_i em t e $\xi_t(i, j)$ é a probabilidade de sair do estado X_i para o estado X_j no tempo t (em ambos os casos, dada a sequência de observações e os parâmetros do modelo):

$$\gamma_t(i) = \sum_{j=1}^n \xi_t(i, j)$$

Com base na teoria exposta acima, aplicações práticas de HMMs permitem identificar diferentes regimes de mercado, como *Bull* e *Bear Markets*, tendências e volatilidades - o que é de grande utilidade para a gestão de uma carteira de investimentos.

3 METODOLOGIA

O método utilizado na pesquisa, visa unir o *Hidden Markov Model* ao *Five-Factor Model*. Dessa maneira, pode-se utilizar o HMM para explicar os fatores de risco de Fama e French.

O Modelo dos Fatores nada mais é do que a interpretação de retornos financeiros como uma regressão linear múltipla, onde a variável dependente, retorno dos portfólios, é explicado por variáveis independentes, os fatores de risco, e um erro aleatório.

Já um *Hidden Markov Model* considera uma observação como uma função probabilística do estado, logo visa inferir o estado oculto do sistema com base na sequência de observações a fim de, a partir desse estado inferido, descobrir as probabilidades referentes a uma próxima observação.

Contudo, neste trabalho, tenta-se explicar essas variáveis independentes do Modelo dos Fatores por meio do HMM. Dessa forma, ao invés de utilizar n HMMs para prever os retornos de n ativos financeiros, ao utilizar o Modelo dos Fatores, necessitamos somente de um HMM, para prever os 5 fatores de risco - que funcionam para n ativos financeiros.

Com isso, visa-se chegar a uma boa *proxy* para o retorno de portfólios para o próximo período t .

3.1 Dados

Para o cálculo proposto acima, foram utilizados dados históricos de retornos de 12 portfólios de diferentes setores da indústria e dados históricos dos 5 fatores de risco. Esses dados são fornecidos pelo próprio Kenneth French e disponibilizados em seu website (FRENCH, 2023a,b).

Os 12 setores da indústria estadunidense que analisaremos, e sua composição, estão listados abaixo:

- Bens de Consumo Não-Duráveis (*NoDur*)
 - Alimentos, tabaco, têxteis, vestuário, couro e brinquedos
- Bens de Consumo Duráveis (*Durbl*)
 - Carros, TVs, móveis e eletrodomésticos
- Indústria (*Manuf*)

- Maquinário, caminhões, aviões, mobília de escritório, papel e impressão comercial
- Energia (*Enrgy*)
 - Extração e produtos de petróleo, gás e carvão
- Químicos (*Chems*)
 - Produtos químicos e afins
- Equipamentos de Negócio (*BusEq*)
 - Computadores, *software* e equipamentos eletrônicos
- Telecomunicações (*Telcm*)
 - Transmissão telefônica e de televisão
- Utilities (*Utils*)
 - Utilidades públicas
- Comércio (*Shops*)
 - Atacado, varejo e alguns serviços (lavanderias, oficinas de reparo)
- Saúde (*Hlth*)
 - Planos de saúde, equipamentos médicos e medicamentos
- Financeiro (*Money*)
 - Bancos e financeiras
- Outros (*Others*)
 - Minas, construção, materiais de construção, transporte, hotéis, serviços empresariais, entretenimento

Seguindo as boas práticas de *machine learning*, esses dados serão divididos em 2 seções: treino, com 70% dos dados, e teste, com os 30% restante. O conjunto de treino é utilizado para formular e treinar o modelo, já o conjunto de teste é usado para realizar ajustes no modelo, evitar o *overfitting*, que ocorre quando o modelo se torna muito especializado nos dados de

treinamento e não generaliza bem para novos dados, e avaliar o desempenho final do modelo (GARETH *et al.*, 2013).

Os dados utilizados vão de julho de 1963 até março de 2023, e qualquer valor não computado (meses sem observação) é substituído pela média entre seu antecessor e seu sucessor, de modo que diminua seu impacto no modelo.

3.2 Modelo

Com o auxílio da linguagem de programação Julia, a partir dos dados mencionados acima, fazendo o uso da biblioteca `GLM.jl`, pode-se facilmente, com um modelo de regressão linear múltipla aplicado nos dados de treino, estimar os parâmetros do Modelo dos Fatores, obtendo-se assim os modelos e equações para o retorno de cada setor.

Sendo assim, conforme esperado, o grande desafio do modelo proposto passa a ser modelar o *Hidden Markov Model* para a série de dados histórica dos fatores de risco, obtendo-se um bom modelo de previsão para os fatores de risco.

Para o HMM, fazendo uso da biblioteca `HMMBase.jl`, utilizaremos uma normal multivariada para a distribuição das observações, o que pode ser justificado pela análise gráfica da função densidade de probabilidade, estimada por kernel, da base de treino (ZUCCHINI; BERZEL; NENADIC, 2003).

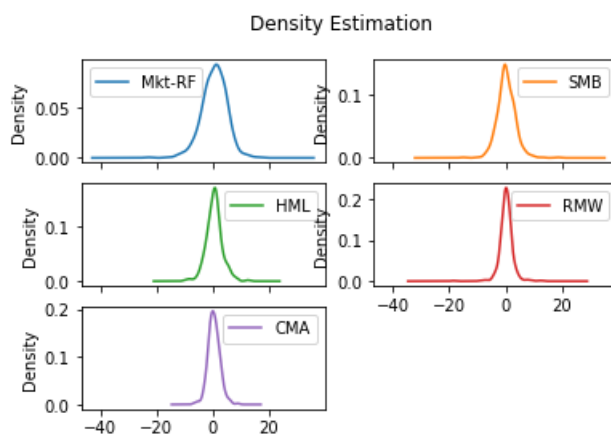


Figura 3: Gráfico da função densidade de probabilidade estimada dos fatores de risco

Além disso, estimaremos os parâmetros do modelo por meio do método de Baum-Welch, considerando três estado ocultos, uma matriz de covariância completa e um máximo de mil iterações.

Após o *fit* do HMM com os dados de treino do Modelo dos Fatores, validamos o modelo. Essa validação do modelo ocorre, e é necessária, para assertar que o modelo está performando da maneira planejada e para aferir se o modelo foi gerado corretamente, de modo que não esteja sobreajustado aos dados de teste - o que faria com que ele performasse insatisfatoriamente para um novo conjunto de dados.

Agora, para a previsão da próxima observação do HMM, utilizaremos a próxima observação esperada, que é calculado a partir da fórmula abaixo:

$$\mathbb{E}[O_t] = \sum_{j=1}^n \sum_{i=1}^n \mathbb{P}(S_{t-1} = X_j) \cdot \mathbb{P}(S_t = X_i \mid S_{t-1} = X_j) \cdot \mathbb{E}[O_t \mid S_t = X_i]$$

A probabilidade de estarmos em um estado j em $t - 1$, $\mathbb{P}(S_{t-1} = X_j)$, é calculada pelo método *forward*, o que nos permite, a partir da observação em $t - 1$, O_{t-1} , encontrar a probabilidade de estarmos, em $t - 1$, em cada estado do HMM. A probabilidade de irmos do estado j em $t - 1$ para um estado i em t , $\mathbb{P}(S_t = X_i \mid S_{t-1} = X_j)$, é calculada a partir da matriz de transição de estados. Já o valor esperado da observação no estado i em t , $\mathbb{E}[O_t \mid S_t = X_i]$, é calculado através da média das observações encontradas naquele estado do HMM.

É fácil perceber que, esse cálculo, para a primeira iteração, nada mais é do que:

$$\mathbb{E}[O_t] = \pi \cdot A \cdot B$$

onde:

- $\pi = \{\pi_i\}$, que é a matriz de probabilidades iniciais, ou seja, a probabilidade de iniciarmos em um estado i
- $A = \{a_{ij}\}$, que é a matriz de transição de estados, ou seja, a probabilidade de sair do estado i e ir para o estado j
- $B = \{b_j(k)\}$, que é um vetor com as distribuições de probabilidade das observações, dado que se está no estado j

Quando definimos o HMM, definimos que as observações seguiriam uma distribuição normal multivariada, logo $b_j(k) \sim \mathcal{N}(\mu_j, \Sigma_j)$. Dessa forma,

$$\mathbb{E}[b_j(k)] = \mu_j$$

Além disso, a matriz de transição de estados e a matriz de probabilidades iniciais são obtidas sem muito esforço após o *fit* do modelo.

A partir daí, para nosso caso, em que o primeiro passo é calcular a probabilidade de estarmos no estado j no tempo $t - 1$, vamos utilizar a variável auxiliar α , descrita anteriormente, que é obtida pelo método *forward* do algoritmo *forward-backward*.

Com isso, podemos utilizar a metodologia descrita acima para descobrir o valor esperado da próxima observação do HMM. Daí, basta inserirmos os valores "previstos" para os fatores de risco, via HMM, na regressão linear múltipla de Fama e French (estimada a partir dos dados de treino) para obter o valor esperado do retorno de um ativo A no próximo período de observação t .



Figura 4: Fluxograma do processo de cálculo da previsão de retorno dos portfólios financeiros

Na figura 4, podemos ver de forma resumida, o processo de cálculo da previsão dos retornos e seus subprocessos. A estimação dos fatores de risco será feita via HMM e, a fins de comparação, via EWMA e média móvel simples de três períodos.

Em seguida, afim de construir um intervalo de confiança para o retorno dos ativos no próximo período t , foram simulados 1000 cenários de retornos. Nessas simulações, o estado inicial (S_{t-1}) é escolhido a partir da distribuição das probabilidades *forward*; o estado seguinte (S_t) é escolhido a partir da distribuição de probabilidades da matriz de transição de estados e o valor da observação (O_t) é encontrado a partir da distribuição gaussiana multivariada citada anteriormente.

4 RESULTADOS

Aplicando um modelo de regressão linear múltipla para os dados de treino de cada um dos portfólios analisados, temos as seguintes equações dos 5 fatores de Fama e French, sendo os valores na tabela os multiplicadores das variáveis independentes:

Portfólio	a_i	b_i	s_i	h_i	r_i	c_i
NoDur	-0.17306	0.96657	0.14594	0.03291	0.63093	0.40508
Durbl	-0.47088	1.14525	0.14624	0.42351	0.21815	0.06612
Manuf	-0.1980	1.10917	0.15230	0.11148	0.28225	0.09208
Enrgy	0.03446	0.91472	-0.17977	0.05612	0.15304	0.44448
Chems	-0.28830	1.02816	0.00259	-0.02457	0.51447	0.43347
BusEq	0.52517	1.00913	0.08621	-0.32087	-0.46904	-0.62546
Telcm	0.15068	0.79061	-0.30680	0.17912	-0.36654	-0.07691
Utils	-0.21887	0.71256	-0.12727	0.46087	0.08317	0.19503
Shops	-0.14736	1.06771	0.26003	0.10780	0.46473	-0.03442
Hlth	0.26091	0.89630	-0.14946	-0.53044	0.44474	0.41727
Money	-0.13972	1.15432	-0.03635	0.49862	0.15695	-0.09902
Other	-0.31278	1.12141	0.36849	0.12644	0.18473	-0.01301

Tabela 1: Equações de Fama-French

Com os modelos de Fama e French de cada portfólio em mãos, vamos montar o *Hidden Markov Model*. Com o HMM, torna-se possível descobrir o valor das variáveis independentes e, ao inserir os valores extraídos do HMM nas equações acima, temos uma *proxy* do retorno do portfólio para o período.

Ajustando um HMM de 3 estados para os dados de treino do Modelo dos Fatores, temos:

- Conjunto de Estados

$$S = \{1, 2, 3\}$$

- Distribuição de Probabilidades Iniciais dos Estados

$$\pi = \begin{bmatrix} 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

- Matriz de Transição de Estados

$$A = \begin{bmatrix} 86.9104\% & 0.0012\% & 13.0884\% \\ 3.2723\% & 92.2084\% & 4.5193\% \\ 9.8813\% & 2.6530\% & 87.4657\% \end{bmatrix}$$

- Distribuição de Probabilidade das Observações

- Estado 1

$$O_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$$

$$\mu_1 = \begin{bmatrix} 1.090405 & 0.483015 & 0.643454 & 0.366167 & 0.348872 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 19.1802 & 1.35257 & -7.0568 & 1.2558 & -5.22156 \\ 1.35257 & 9.66784 & -1.97758 & -1.03922 & -2.26284 \\ -7.0568 & -1.97758 & 8.35157 & -1.91221 & 4.93079 \\ 1.2558 & -1.03922 & -1.91221 & 2.61331 & -1.60341 \\ -5.22156 & -2.26284 & 4.93079 & -1.60341 & 4.93871 \end{bmatrix}$$

- Estado 2

$$O_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$$

$$\mu_2 = \begin{bmatrix} -0.756789 & 0.5637412 & 0.9821392 & 0.1634985 & 1.2444653 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 46.5648 & 7.78524 & -13.4779 & -13.9295 & -11.7708 \\ 7.78524 & 26.393 & -2.61474 & -12.0728 & -0.234102 \\ -13.4779 & -2.61474 & 22.4026 & 13.5419 & 10.7906 \\ -13.9295 & -12.0728 & 13.5419 & 24.2328 & 5.15683 \\ -11.7708 & -0.234102 & 10.7906 & 5.15683 & 10.5254 \end{bmatrix}$$

– Estado 3

$$O_3 \sim \mathcal{N}(\mu_3, \Sigma_3)$$

$$\mu_3 = \begin{bmatrix} 0.353900 & 0.026841 & 0.143240 & 0.209007 & 0.085053 \end{bmatrix}$$

$$\Sigma_3 = \begin{bmatrix} 10.7218 & 3.93357 & -0.292237 & -0.774159 & -0.623155 \\ 3.93357 & 5.04922 & -0.554063 & -0.763298 & -0.164115 \\ -0.292237 & -0.554063 & 2.92671 & -1.19392 & 1.59248 \\ -0.774159 & -0.763298 & -1.19392 & 1.70859 & -0.992239 \\ -0.623155 & -0.164115 & 1.59248 & -0.992239 & 1.75075 \end{bmatrix}$$

A partir desses modelos, vamos fazer a previsão dos fatores de Fama e French utilizando *Hidden Markov Models* e estimar o retorno de ativos financeiros a partir desses fatores previstos, conforme descrito anteriormente.

Abaixo, podemos ver o gráfico do comportamento do modelo testado. Em azul, temos os retornos obtidos através do valor esperado; em vermelho, podemos ver os limites superiores e inferiores do intervalo de confiança obtido através das simulações e; em verde, o retorno real da carteira.

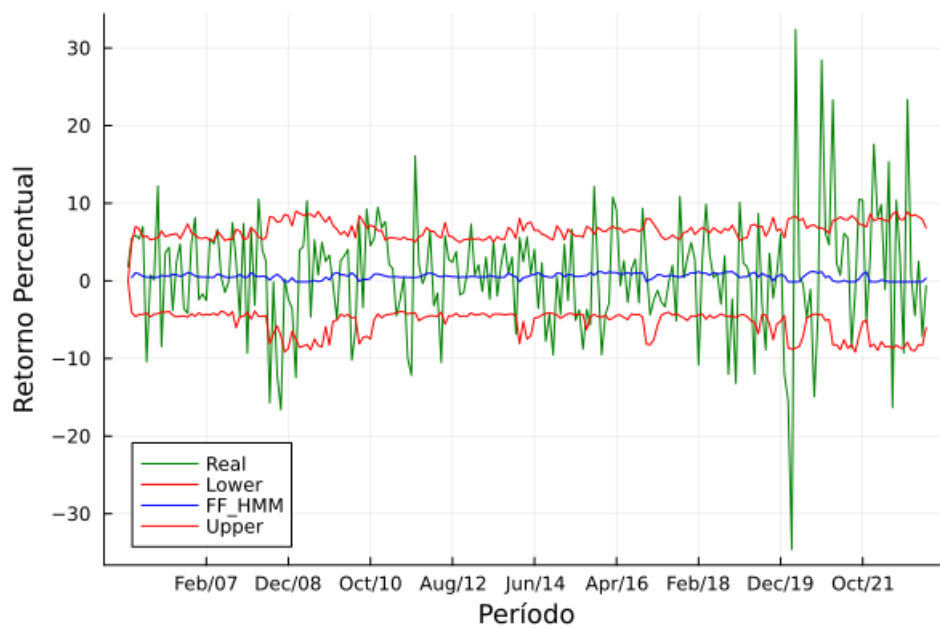


Figura 5: Gráfico dos retornos previstos para o setor de energia

Abaixo, para medir a acurácia do modelo, podemos ver uma tabela com os erros absoluto e quadrático médios do modelo para cada portfólio analisado:

Portfolio	MAE	MSE
NoDur	2.84026	13.8844
Durbl	6.36989	91.927
Manuf	4.42218	35.904
Enrgy	5.59908	59.8154
Chems	3.2647	19.6863
BusEq	4.42148	30.8876
Telcm	3.71598	23.3844
Utils	3.16833	16.4907
Shops	3.4547	21.4518
Hlth	3.2898	17.3861
Money	4.50434	36.8696
Other	4.02192	29.1357

Tabela 2: Erro Absoluto Médio e Erro Quadrático Médio do modelo

A partir da análise gráfica e da análise da tabela acima, é possível dizer que o resultado do modelo não é satisfatório, uma vez que ele aparenta não captar períodos de maior volatilidade.

Mesmo que em boa parte das observações o retorno real do ativo esteja dentro do intervalo de confiança, esse intervalo é muito grande. Ainda que dentro do intervalo de confiança, em boa parte dos casos, o valor esperado do próximo retorno do ativo ($\mathbb{E}[O_t]$) é muito distante do retorno real observado no período t - o que, no contexto de previsão para o mercado financeiro, é significativo.

Com isso, para entender se é ou não vantajosa a interpretação dos fatores como HMMs, vamos comparar a performance do modelo com modelos mais simples. Para isso, foram criados outros dois modelos, que calculam o valor esperado dos fatores com a média móvel exponencialmente ponderada (EWMA), com fator de decaimento padrão ($\lambda = 0.94$), e com a média móvel simples de 3 períodos (SMA).

A partir dessa previsão para os fatores, o retorno do ativo para o próximo período t é calcu-

lado através da mesma regressão múltipla utilizada para o HMM. Para essa análise, podemos verificar o gráfico e a tabela abaixo:

Portfolio	MAE			MSE		
	HMM	EWMA	SMA	HMM	EWMA	SMA
NoDur	2.84026	2.74629	2.50087	13.8844	12.6878	11.1324
Durbl	6.36989	6.25998	5.67685	91.927	89.5154	77.5282
Manuf	4.42218	4.18892	3.65146	35.904	32.7146	25.2677
Enrgy	5.59908	5.39076	5.16976	59.8154	56.4874	52.6606
Chems	3.2647	3.1018	2.86372	19.6863	18.0778	14.6784
BusEq	4.42148	4.14919	3.50993	30.8876	27.7648	20.7027
Telcm	3.71598	3.58536	3.28188	23.3844	22.1312	19.0591
Utils	3.16833	3.05903	3.17549	16.4907	15.3228	15.3562
Shops	3.4547	3.31109	3.02935	21.4518	20.1635	17.735
Hlth	3.2898	3.1432	2.89991	17.3861	16.2298	14.213
Money	4.50434	4.29372	3.71785	36.8696	33.5863	25.6253
Other	4.02192	3.81052	3.31975	29.1357	26.553	20.4402
Média	4.08939	3.91999	3.56640	33.0686	30.9362	26.1999

Tabela 3: Comparação da previsão utilizando diferentes metodologias

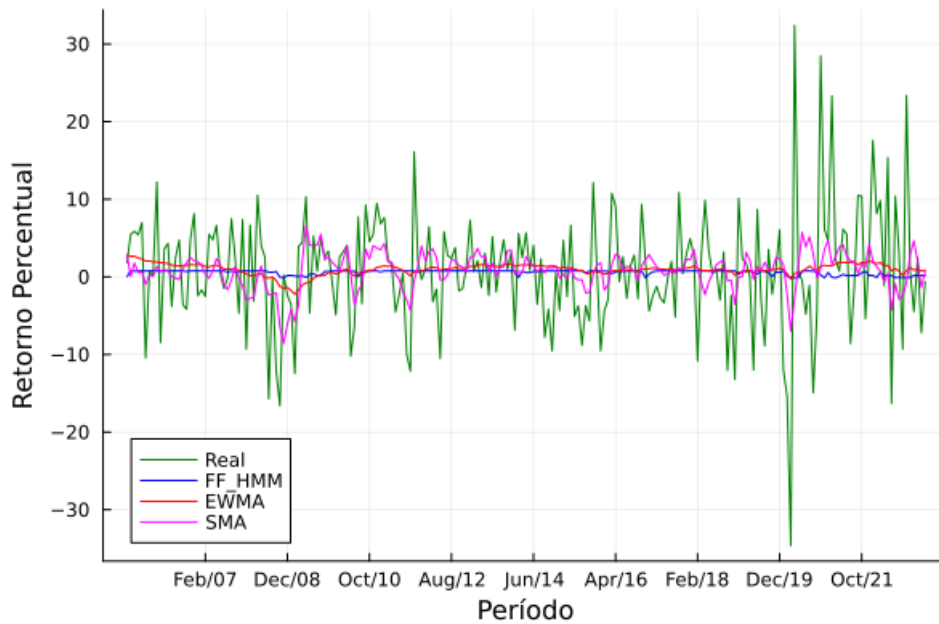


Figura 6: Gráfico dos retornos previstos para o setor de energia com as diferentes metodologias

Com isso, podemos ver que nenhum dos 3 métodos faz um bom papel em interpretar períodos de maior volatilidade. Um bom exemplo, que pode ser visto mais claramente no gráfico, é o período referente a pandemia da Covid-19, que impactou significativamente o mercado financeiro e o preço dos ativos - em que todos os métodos testados falham em identificar a grande volatilidade do mercado.

Agora, como o modelo se provou pouco eficiente para prever o próximo retorno de ativos financeiros, vamos tentar utilizar o modelo para fazer uma análise absoluta para o próximo período t . Isso é, utilizar o modelo visando somente prever a valorização ou desvalorização do ativo.

Com esse novo objetivo, temos a seguinte performance do modelo:

Portfólio	HMM	EWMA	SMA
NoDur	56.2791%	60.4651%	65.5814%
Durbl	53.4884%	60.4651%	67.4419%
Manuf	53.4884%	63.2558%	70.6977%
Enrgy	51.6279%	59.5349%	66.5116%
Chems	59.0698%	64.186%	68.8372%
BusEq	53.9535%	61.8605%	66.5116%
Telcm	54.8837%	63.7209%	71.1628%
Utils	58.1395%	60.9302%	60.9302%
Shops	60.0%	66.0465%	71.6279%
Hlth	53.0233%	60.9302%	67.4419%
Money	53.4884%	63.2558%	69.3023%
Other	51.6279%	61.3953%	73.0233%
Média	54.9225%	62.1705%	68.2558%

Tabela 4: Probabilidade do modelo acertar valorização ou desvalorização

Dessa forma, percebe-se que, de forma absoluta, o modelo tem um bom desempenho, uma vez que prevê corretamente, para a aproximadamente 55% das vezes, se o portfólio irá valorizar ou desvalorizar. Entretanto, modelos com menor grau de complexidade, como o EWMA e como o de médias móveis de 3 períodos, possuem maior acurácia, de aproximadamente 62% e 68% respectivamente.

5 CONCLUSÃO

A tentativa de utilizar *Hidden Markov Models* para prever o modelo dos fatores e, consequentemente, obter o retorno de um ativo financeiro para o próximo período, foi explorada neste trabalho. No entanto, os resultados obtidos indicaram que esse modelo performou pior em comparação a abordagens de menor grau de complexidade, como a previsão via EWMA (*Exponentially Weighted Moving Average*) e SMA (*Simple Moving Average*).

Os HMMs são modelos estatísticos poderosos, amplamente utilizados em diversas áreas, incluindo finanças. Sua capacidade de modelar a dinâmica oculta por trás dos dados tornou-os uma escolha atrativa para previsão de séries temporais financeiras. No entanto, no contexto deste estudo, o desempenho do HMM foi insatisfatório.

Uma das razões possíveis para o desempenho inferior do HMM em relação aos modelos mais simples é sua própria complexidade. *Hidden Markov Models* são modelos paramétricos que requerem a estimação de vários parâmetros, incluindo os estados ocultos e as probabilidades de transição.

Por outro lado, as abordagens mais simples, como a previsão via EWMA e via média móvel, possuem menor número de parâmetros e são mais fáceis de serem implementadas. Nesse sentido, embora possam não capturar toda a complexidade do mercado financeiro, esses modelos mais simples geralmente apresentam desempenho razoável em previsões de curto prazo.

Portanto, com base nos resultados obtidos neste estudo, é recomendado que sejam exploradas outras abordagens de previsão que possam oferecer um equilíbrio entre complexidade e desempenho. Além disso, a seleção adequada do modelo de previsão deve levar em consideração não apenas a precisão das previsões, mas também a simplicidade, a interpretabilidade e a facilidade de implementação.

É importante ressaltar que os resultados específicos podem variar de acordo com o conjunto de dados utilizado, o horizonte de previsão, as características dos ativos financeiros e outros fatores. Portanto, estudos futuros podem explorar diferentes modelos, técnicas e conjuntos de dados para avaliar a eficácia do HMM em diferentes cenários financeiros.

Em suma, embora a tentativa de utilizar *Hidden Markov Models* para prever o modelo dos fatores e obter o retorno de um ativo financeiro tenha sido interessante, os resultados indicaram que modelos de menor grau de complexidade apresentaram um desempenho superior. Isso sugere que, para esse contexto específico, abordagens mais simples podem ser mais eficazes

na previsão de curto prazo. No entanto, é sempre válido explorar outras técnicas e abordagens para encontrar a melhor solução para cada problema de previsão financeira.

BIBLIOGRAFIA

BAUM, Leonard E; PETRIE, Ted. Statistical inference for probabilistic functions of finite state Markov chains. **The annals of mathematical statistics**, JSTOR, v. 37, n. 6, p. 1554–1563, 1966.

BAUM, Leonard E; PETRIE, Ted *et al.* A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. **The annals of mathematical statistics**, JSTOR, v. 41, n. 1, p. 164–171, 1970.

CHANG, R; HANCOCK, Je. On receiver structures for channels having memory. **IEEE Transactions on Information Theory**, IEEE, v. 12, n. 4, p. 463–468, 1966.

FAMA, Eugene F; FRENCH, Kenneth R. A five-factor asset pricing model. **Journal of financial economics**, Elsevier, v. 116, n. 1, p. 1–22, 2015.

FAMA, Eugene F; FRENCH, Kenneth R. Common risk factors in the returns on stocks and bonds. **Journal of financial economics**, Elsevier, v. 33, n. 1, p. 3–56, 1993.

FAMA, Eugene F; FRENCH, Kenneth R. The cross-section of expected stock returns. **the Journal of Finance**, Wiley Online Library, v. 47, n. 2, p. 427–465, 1992.

FRENCH, Kenneth R. **5 Industry Portfolios**. New Hampshire: Dartmouth College, 2023. Disponível em: <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/5_Industry_Portfolios_CSV.zip>.

FRENCH, Kenneth R. **Fama/French 5 Factors (2x3)**. New Hampshire: Dartmouth College, 2023. Disponível em: <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Research_Data_5_Factors_2x3_CSV.zip>.

GARETH, James *et al.* **An introduction to statistical learning: with applications in R**. New York: Springer, 2013.

LINTNER, John. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. **The review of economics and statistics**, JSTOR, p. 13–37, 1965.

NOVY-MARX, Robert. The other side of value: The gross profitability premium. **Journal of financial economics**, Elsevier, v. 108, n. 1, p. 1–28, 2013.

RABINER, Lawrence R. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, Ieee, v. 77, n. 2, p. 257–286, 1989.

SHARPE, William F. Capital asset prices: A theory of market equilibrium under conditions of risk. **The journal of finance**, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964.

TITMAN, Sheridan; WEI, KC John; XIE, Feixue. Capital investments and stock returns. **Journal of financial and Quantitative Analysis**, Cambridge University Press, v. 39, n. 4, p. 677–700, 2004.

VITERBI, Andrew. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. **IEEE transactions on Information Theory**, IEEE, v. 13, n. 2, p. 260–269, 1967.

ZHANG, Yingjian. Prediction of financial time series with Hidden Markov Models. Simon Fraser University, 2004.

ZUCCHINI, Walter; BERZEL, A; NENADIC, O. Applied smoothing techniques. **Part I: Kernel Density Estimation**, v. 15, p. 1–20, 2003.