Inferência sobre modelos de mistura no contexto financeiro

Emanuel Victor da Silva Favorato, 12558151
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo

02 de Julho, 2024

Resumo

O presente relatório visa investigar a aplicação de Modelos de Mistura Gaussiana (GMM), Hidden Markov Model (HMM) e Modelos de Mistura Gaussiana Bayesianos (BGMM) em dados financeiros, a fim de interpretar as dinâmicas dos regimes de mercado, valendo-se do uso dos índices S&P 500 e Ibovespa para tais ajustes, comparamos a eficácia dos métodos clássicos via Expectation-Maximization (EM) e Bayesianos por Inferência Variacional Bayesiana na modelagem de retornos. Os resultados obtidos apontam para a melhor adequabilidade dos modelos gaussianos GMM e BGMM, com destaque para o segundo, que se mostrou bastante eficiente em capturar períodos de maiores retornos e prejuízos. Para a análise comparativa, foram observados os retornos médios ajustados, análise gráfica, bem como o custo computacional.

Palavras-chave: GMM, HMM, BGMM, EM, Inferência Variacional Bayesiana, retornos, regimes, modelagem.

1 Introdução

Com o avanço da Computação Financeira e dos algoritmos Bayesianos, muito vem sendo explorado acerca da utilização de métodos estatísticos sofisticados e de Aprindizado de Máquina para a identificação de regimes financeiros que possam trazer resultados robustos, de modo a otimizar ganhos e minimizar riscos dentro do contexto financeiro, bem como antecipar tendências de crescimento, estagnação e queda, oferecendo importantes informações acerca do melhor momento para a compra e venda de ações.

Neste relatório, serão exploradas a aplicação e comparação de três abordagens bastante poderosas para identificar regimes de mercado: o Modelo de Mistura Gaussiana (GMM), utilizando o algoritmo de Expectation-Maximization (EM), o Hidden Markov Model (HMM), também fazendo uso do algoritmo EM, e o Modelo de Mistura Gaussiana Bayesiano (BGMM), valendo-se do algoritmo de Inferência Variacional Bayesiana (Variational Bayesian Inference). Tais métodos serão aplicados nas bases de dados temporais dos índices S&P500 e Ibovespa, principais medidores das bolsas estado-unidense e brasileira, respectivamente, a fim de observar aqueles que melhor capturam as nuances de mercado existentes nos dois países.

2 Metodologia

Para a aplicação dos Modelos de Mistura Gaussiana e Hiden Marcov, foram empregadas duas abordagens para a estimação dos parâmetros, a abordagem clássica, através do algoritmo EM, e a abordagem Bayesiana, pelo Variational Bayesian Inference, este último aplicado apenas para Mistura Gaussiana. A aplicabilidade, características e utilização das metodologias empregadas será discorrida a seguir:

2.1 Modelo de Mistura Gaussiana(GMM)

O Modelo de Mistura Gaussiana(GMM) (Przyborowski and Ślezak, 2022) trata-se de um modelo probabilístico que assume que os dados são gerados a partir de uma mistura de várias distribuições normais, cujos parâmetros são desconhecidos. No problema apresentado, cada componente Gaussiano representa um regime de mercado

distinto, sendo o algoritmo EM utilizado na estimação iterativa dos parâmetros dessas, de modo a obter uma aproximação para as estimativas de máxima-verossimilhança.

- Vantagens: flexibilidade na modelagem de dados multimodais, implementação relativamente simples e facilidade de interpretação.
- Desvantagens: os retornos podem não necessáriamente ser gaussianos, qualidade de ajuste dependente das condições iniciais da iteração, necessita de um número fixo de componentes a serem definidos previamente, o que pode, ou não, corresponder a realidade.

  A seguir, é possível observar algumas características do GMM (Przyborowski and Ślezak, 2022):

  1. **Função de densidade para o modelo de mistura**:

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \tag{1}$$

  em que K é o número de componentes, π_k são os pesos das componentes (de tal forma que $\sum_{k=1}^{K} \pi_k = 1$), μ_k são as médias e Σ_k as covariâncias.

  2. **Distribuição das componentes gaussianas**

  A distribuição de cada componente Gaussiana pode ser observada através da expresão (Kayabol, 2015):

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{2}$$

  onde $\mathbf{x}$ é o ponto de dados d-dimensional, $\boldsymbol{\mu}$ é o vetor de médias d-dimensional, $\boldsymbol{\Sigma}$ é a matriz de covariâncias $d \times d$, $|\boldsymbol{\Sigma}|$ é o determinante da matriz de covariâncias e $\boldsymbol{\Sigma}^{-1}$ é a inversa da matriz de covariâncias.

  3. **Probabilidade a posteriori**:

  A probabilidade a posteriori representa a probabilidade de uma determinada observação pertencer a uma componente, cuja fórmula está expressa a seguir:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)} \tag{3}$$

  Observação: quanto mais próximo de 1, maiores as taxas de aceitação, os gráficos com os valores obtidos para essas probabilidades podem ser encontrados no código em anexo como "Responsabilidades", porém sua análise não será discorrida, devido a complexidade associada à interpretação.

2.2 Hidden Markov Model(HMM)

O modelo HMM (Chandrika et al., 2020) (Rabiner and Juang, 1986), ou Modelo Oculto de Markov em português, é um procedimento estatístico que assume que os dados são provenientes de um processo de Markov oculto, sendo bastante útil na modeladem de séries temporais financeiras, onde os estados de mercado são, muitas vezes, não observáveis de forma clara. Assim como no caso anterior, o algoritmo EM (Przyborowski and Ślezak, 2022) é utilizado para estimar os parâmetros do modelo.

- Vantagens: bastante eficaz na captura de dependências temporais nos dados financeiros, o que facilita modelar transições entre regimes de mercado, bem como compreender seu dinamismo e complexidade, podendo também realizar inferência sobre estados não observáveis, o que pode fornecer mais informações acerca desses regimes.
- Desvantagens: custo computacional, especialmente para grandes bases de dados, pressupostos sobre a independência dentro dos estados ocultos podem não ser satisfeitos, número de estados deve ser especificado previamente, o que, na prática, pode não coincidir com a realidade.

Extendendo as informações retiradas de (Rabiner and Juang, 1986), temos as fórmulas a seguir:

1. **Distribuição Inicial** (π):

$$\pi_i = P(z_1 = i) \tag{4}$$

2. **Matriz de Transição** (A):

$$a_{ij} = P(z_{t+1} = j \mid z_t = i) \tag{5}$$

sendo cada z_t um estado oculto no tempo t.

3. **Distribuição de Emissão** (B):

$$b_j(o_t) = P(o_t \mid z_t = j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(o_t - \mu_j)^2}{2\sigma_j^2}\right) \tag{6}$$

4. **Algoritmo Forward**:

É o algoritmo usado no cálculo da probabilidade de um vetor de observações O, dados π, A e B.

$$\alpha_t(j) = P(o_1, o_2, \ldots, o_t, z_t = j \mid \pi, A, B) \ = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t) \tag{7}$$

sendo N o número total de estados.

5. **Algoritmo Backward**:

Utilizado no cálculo da probabilidade para observações futuras dado o estado atual.

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \ldots, o_T \mid z_t = i, \pi, A, B) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{8}$$

6. **Reestimativa dos parâmetros da distribuição Inicial** (π):

$$\pi_i = \gamma_1(i) \tag{9}$$

- **Matriz de Transição** (A):

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{10}$$

- **Distribuição de Emissão** (B):

$$\mu_j = \frac{\sum_{t=1}^{T} \gamma_t(j) o_t}{\sum_{t=1}^{T} \gamma_t(j)}, \quad \sigma_j^2 = \frac{\sum_{t=1}^{T} \gamma_t(j)(o_t - \mu_j)^2}{\sum_{t=1}^{T} \gamma_t(j)} \tag{11}$$

7. **Probabilidade de estar no estado i no tempo t** (γ):

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{12}$$

8. **Probabilidade de estar no estado i no tempo t e no estado j no tempo $t+1$** (ξ):

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{13}$$

2.3 Modelo de Mistura Gaussiana Bayesiano (BGMM)

O Modelo de Mistura Gaussiana Bayesiano (BGMM) (Kayabol, 2015) corresponde a uma extensão do GMM, onde são incorporados métodos Bayesianos. A utilização da inferência variacional (David M. Blei and McAuliffe, 2017) no ajuste do modelo oferece certa robustez, em comparação com os métodos clássicos.

- Vantagens: incorporação de informações prévias, bastante útil na melhoria das estimativas dos parâmetros quando há menos dados disponíveis, robustez adicional através da quantificação incertezas, estimação do número de componentes, evitando a necessidade de pré-especificação.
- Desvantagens: a inferência variacional pode ser computacionalmente custosa, interpretação mais dificultosa dos resultados observados, implementação mais complexa.

Para a implementação Bayesiana, o pacote scikit-learn (Pedregosa et al., 2011) utiliza a priori de Dirichlet para a estimação dos pesos, a qual pode ser observada a seguir:

1. **Priori de Dirichlet**

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \tag{14}$$

onde $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ são os parâmetros da distribuição de Dirichlet.

2. **Priori para os parâmetros**

Para os parâmetros das componentes da mistura, podemos considerar a distribuição Normal-Wishart Inversa (Kayabol, 2015):

$$p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \mid \boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{W}_0) = \mathcal{N}(\boldsymbol{\mu}_i \mid \boldsymbol{\mu}_0, (\kappa_0 \boldsymbol{\Sigma}_i)^{-1}) \cdot \mathcal{W}^{-1}(\boldsymbol{\Sigma}_i \mid \boldsymbol{W}_0, \nu_0) \tag{15}$$

onde $\boldsymbol{\mu}_0$ é a média da priori gaussiana, κ_0 é o parâmetro de escala desta priori, ν_0 são os graus de liberdade da priori de Wishart e $\boldsymbol{W}_0$ é a matriz de escala da priori de Wishart.

3. **Posteriori para os parâmetros**

A distribuição a posteriori dos parâmetros é também uma distribuição Normal-Wishart Inversa:

$$p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \mid \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}_i \mid \boldsymbol{\mu}_N, (\kappa_N \boldsymbol{\Sigma}_i)^{-1}) \cdot \mathcal{W}^{-1}(\boldsymbol{\Sigma}_i \mid \mathbf{W}_N, \nu_N) \tag{16}$$

onde:

- N corresponde ao número de observações na amostra.
- $\boldsymbol{\mu}_N = \frac{\kappa_0 \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}}{\kappa_0 + N}$, média a posteriori das distribuições normais para os parâmetros $\boldsymbol{\mu}_i$.
- κ_0 é o parâmetro de escala.
- $\kappa_N = \kappa_0 + N$, atualização do parâmetro de escala para $\boldsymbol{\mu}_N$.
- $\nu_N = \nu_0 + N$, graus de liberdade atualizados da distribuição de Wishart para $\boldsymbol{\Sigma}_i$.
- $\mathbf{W}_N = \mathbf{W}_0 + S + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T$, matriz de escala atualizada da distribuição de Wishart para $\boldsymbol{\Sigma}_i$.
- $\mathbf{W}_0$ é a matriz de escala inicial.
- $\bar{\mathbf{x}}$ é a média amostral dos vetores de dados.

- $S = \sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$, corresponde a soma das matrizes externas das diferenças entre os vetores de dados e a média amostral $\bar{\mathbf{x}}$.

4. **Posteriori para os pesos ($\boldsymbol{\pi}$)**

A distribuição a posteriori é dada, também, por uma distribuição de Dirichlet atualizada pela informação dos dados:

$$p(\boldsymbol{\pi} \mid \mathbf{X}) = \text{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha} + \mathbf{n}) \tag{17}$$

onde $\mathbf{n} = (n_1, n_2, \ldots, n_K)$ é o vetor de contagem dos dados atribuídos a cada componente.

2.4 Algoritmos para a estimação dos parâmetros

2.4.1 Algoritimo Expectation-Maximization (EM)

Consiste em um método iterativo bastante útil para encontrar as estimativas de máxima verossimilhança aproximadas para os parâmetros de interesse, quando esses não podem ser encontrados pelo método tradicional.

- Vantagens: facilidade de implementação, convergência garantida, ampla aplicabilidade para a resolução de problemas em estatística.
- Desvantagens: convergência pode não ocorrer em um máximo global, em alguns problemas, pode vir a ser computacionalmente custoso, resultado final dependente do palpite inicial.

A seguir, tem-se um pseudo-código ilustrativo para o funcionamento do algoritmo, com base na visualização observada em (Przyborowski and Ślezak, 2022):

Algorithm 1 Algoritmo Expectation-Maximization (EM)

Require: conjunto de dados X, parâmetros do modelo θ, critério de parada ϵ
Ensure: parâmetros estimados θ^*

Vamos definir um "chute"inicial para o modelo $\theta^{(0)}$
inicialização das iterações $t \leftarrow 0$
Definindo uma flag de convergência convergiu $\leftarrow$ Falso
while não convergiu **do**
    // Passo E (Expectation)
    for $i = 1$ até n **do** ▷ Loop sobre cada dado iterado
        Calcular a log-verossimilhança completa esperada $Q(\theta \mid \theta^{(t)})$
        Estimação das variáveis latentes Z usando $\theta^{(t)}$
    end for
    // Passo M (Maximization)
    Atualizar $\theta^{(t+1)}$, maximizando $Q(\theta \mid \theta^{(t)})$ em relação a θ
    Incrementando a contagem de iterações $t \leftarrow t + 1$
    Verificando se a convergência foi satisfeita: ▷ Calcular mudanças nos parâmetros
    if $|\theta^{(t+1)} - \theta^{(t)}| < \epsilon$ **then**
        Observar se convergiu $\leftarrow$ Verdadeiro
    end if
end while
return $\theta^* = \theta^{(t+1)}$

2.4.2 Algoritmo de Inferência Variacional Bayesiana(Variational Bayesian Inference)

Trata-se de uma aproximação para as estimativas dos parâmetros desejados em modelos estatísticos complexos, assim como no caso anterior, a dificuldade em lidar com o problema proposto a torna bastante útil na estimação

paramétrica. Apesar do nome, e de apresentar certos princípios da abordagem Bayesiana, como a definição de distribuições a priori e a posteriori, seu foco principal está muito mais voltado à transformação de problemas inferenciais em problemas de otimização, do que na aplicação dos modelos Bayesianos pelos métodos tradicionais.

- Vantagens: convergência mais rápida que a de métodos usuais, ideal quando há grandes volumes de dados e modelos complexos, aplicação em vastos problemas na área de probabilidade e estatística.
- Desvantagens: escolha da priori interfere na qualidade da estimação, implementação mais complexa, dificuldades na aplicação do algoritmo.

Logo abaixo, é possível observar o respectivo pseudocódigo para a Inferência Variacional Bayeniana (VBI), com base em (David M. Blei and McAuliffe, 2017):

Algorithm 2 Inferência Variacional Bayesiana (VBI)

Require: vetor de dados X, verossimilhança para $p(X|\theta)$, priori $p(\theta)$, família variacional $q(\theta|\lambda)$, critério de parada ϵ

Ensure: aproximação para a distribuição a posteriori $q(\theta|\lambda)$

Vamos começar inicializando os parâmetros variacionais λ
repeat
  // Passo 1: atualização variacional
  for cada fator variacional $q_j(\theta_j)$ em $q(\theta|\lambda)$ **do**
    Atualizar $q_j(\theta_j) \propto \exp\left(E_{q_{-j}}[\log p(X, \theta)]\right)$ ▷ Atualização de cada fator variacional
  end for
  // Passo 2: maximização da ELBO (Evidência Inferior Limite)
  Vamos calcular ELBO(λ) ▷ Evidência Inferior Limite
  Verificando a convergência: Se $|\text{ELBO}(\lambda_{\text{novo}}) - \text{ELBO}(\lambda_{\text{antigo}})| < \epsilon$, pare
until convergência
return parâmetros variacionais otimizados λ e a posteriori aproximada $q(\theta|\lambda)$

A expressão para ELBO(λ) é definida a seguir:

$$\text{ELBO}(\lambda) = E_{q(\theta|\lambda)}[\log p(X, \theta)] - E_{q(\theta|\lambda)}[\log q(\theta \mid \lambda)] \tag{18}$$

sendo que:

- $p(X, \theta)$ é a função densidade conjunta de X e θ.
- $q(\theta \mid \lambda)$ é a densidade variacional aproximada para o vetor de parâmetros θ, parametrizada por λ.

2.5 Bibliotecas e arquivos utilizados para os modelos

As análises foram geradas utilizando o software Python, as aplicações dos modelos GMM e BGMM, bem como os algoritmos utilizados para a estimação de seus respectivos parâmetros, foram implementados através do pacote scikit-learn (Pedregosa et al., 2011), já o HMM foi desenvolvido fazendo uso da ferramenta hmmlearn (Gramfort et al., 2014), os dados dos índices S&P500 e Ibovespa, por sua vez, foram extraídos por meio da biblioteca yfinance (Ran Aroussi et al., 2024).

3 Resultados

A fim de obter o modelo que melhor se adequasse aos regimes de mercado, os três modelos anteriormente descritos foram aplicados às séries temporais de S&P500 e Ibovespa no período de 01/01/2010 a 28/06/2024, conseiderando-se n = 3 componentes (regimes de mercado), para o Modelo de Mistura Gaussiana Bayesiano

(BGMM) também tolerou-se um número máximo de 10 componentes para analizar tais dinâmicas. As séries temporais de ambos os índices podem ser observadas a seguir:

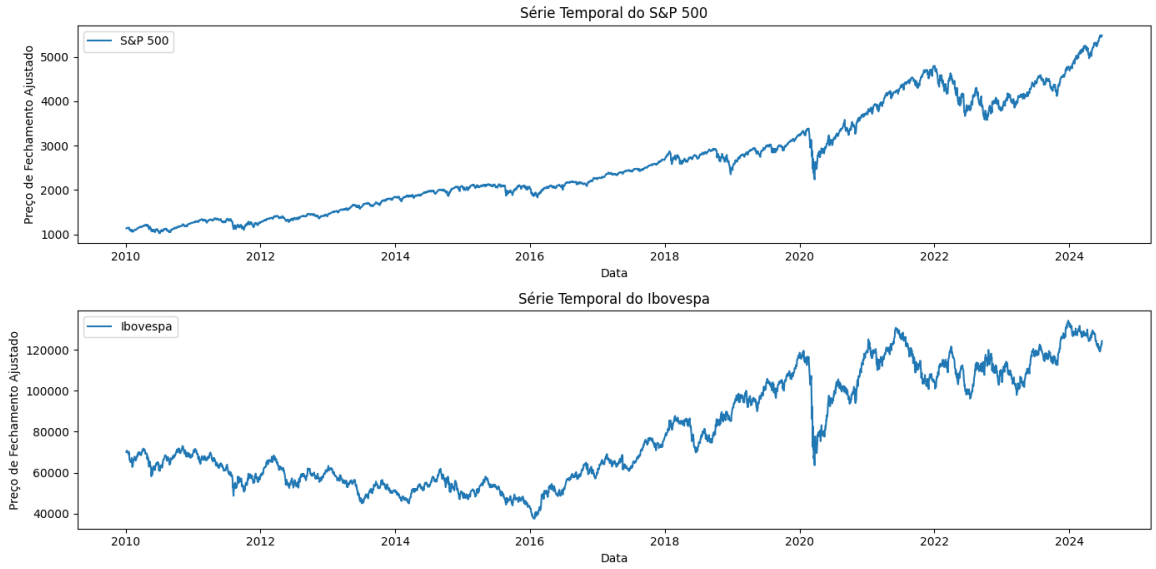

Figura 1: Série temporal dos índices S&P500 e Ibovespa

Como pode ser observado na figura 1, ambos os índices aparentam possuir mais de um regime de mercado, alternando entre sucessivos períodos de altas e baixas, tal relação é, aparentemente, mais expressa no índice da bolsa brasileira (Ibovespa), onde flutuações nas cotações se fazem mais visíveis, o que pode ser um indicativo de múltiplas componentes atuando dentro do mercado especulativo do país, algo que também pode ser visto de forma, possívelmente, menos acentuada no índice S&P500 dos Estados Unidos.

A fim de obter ganhos, e minimizar riscos associados à volatilidade de mercado, muitos investidores vêm recorrendo às ciências computacionais para enxergar tendências que apenas o conhecimento humano não seria capaz de reconhecer, algumas ferramentas ainda pouco exploradas, mas com grande aplicabilidade, são as de Estatística Computacional, a exemplo das observadas anteriormente.

3.1 Aplicação do GMM com Algoritmo EM

A partir das análises gráficas a seguir, é possível observar os resultados obtidos para S&P500 utilizando GMM com Algoritmo EM e os retornos proporcionais ajustados com base no modelo:

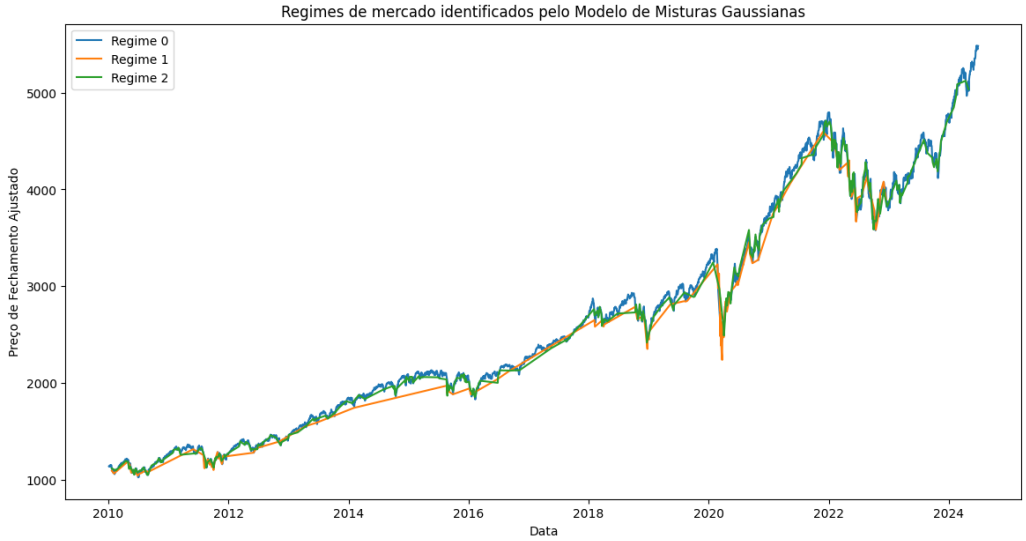

Figura 2: Componentes de mercado identificadas via GMM para S&P500

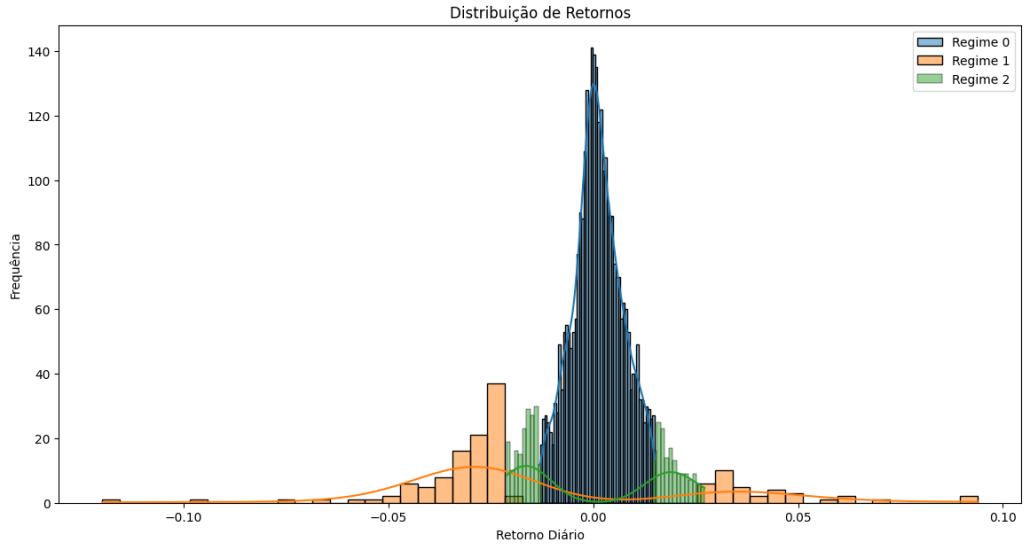

Figura 3: Retornos observados para cada regime via GMM para S&P500

Como pode ser observado, considerando-se n = 3 regimes de mercado, pode-se observar retornos aparentemente gaussianos centrados em zero para o regime 0, o que indica certa previsibilidade, bem como baixos retornos e volatilidade. Já para os regimes 1 e 2, observa-se bimodalidade, e simetria para o segundo, o que pode ser um indicativo de regimes distintos dentro das próprias componentes, alternando entre períodos de crescimento e queda, além disso, a maior dispersão dos dados para o caso 1 pode ser um indício de elevada volatilidade.

Utilizando o mesmo modelo para o Ibovespa, temos:

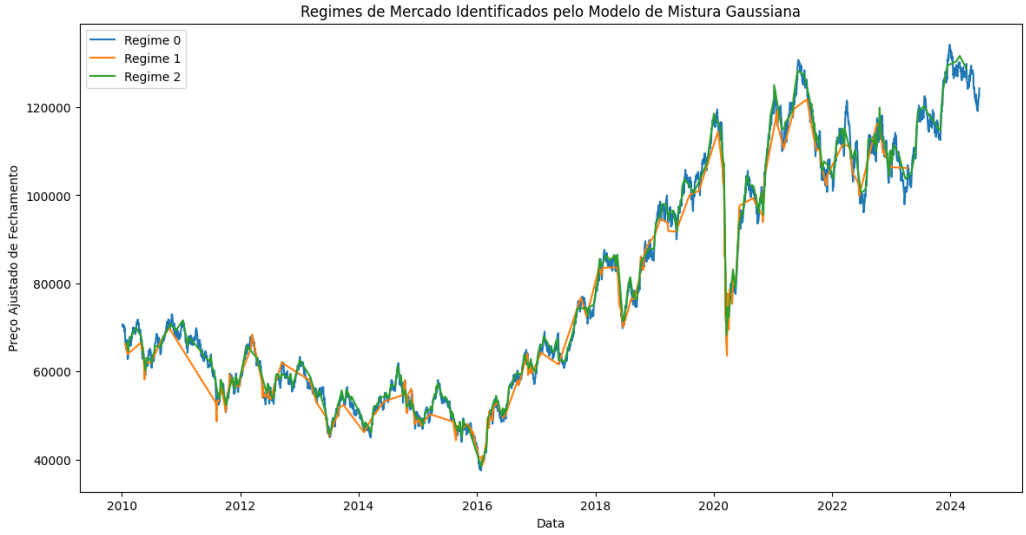

Figura 4: Componentes de mercado identificadas via GMM para Ibovespa

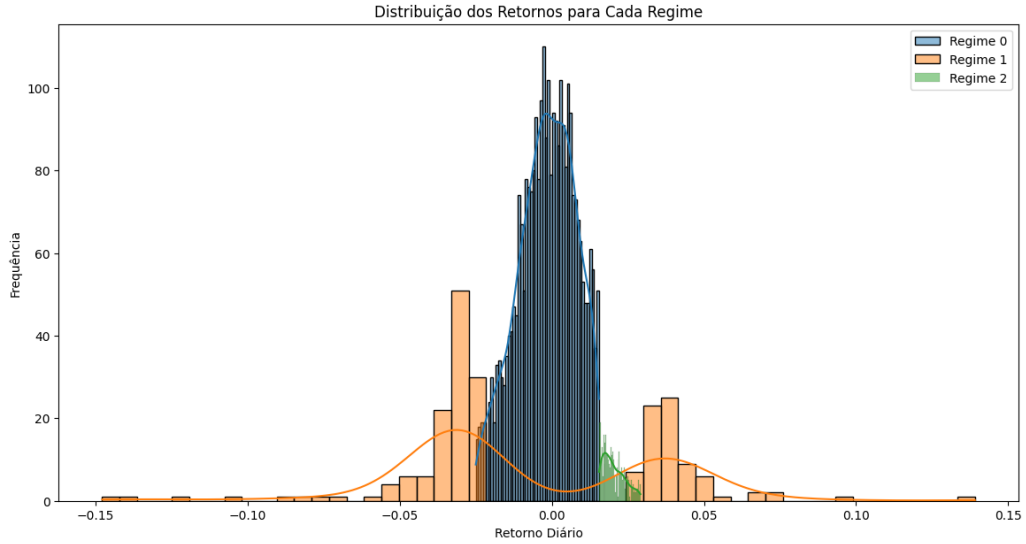

Figura 5: Retornos observados para cada regime via GMM para Ibovespa

É possível observar que os retornos apresentam razoável simetria levemente deslocada para o sentido negativo no regime 0, o que indica um mercado relativamente "calmo"ou em transição, sem grandes retornos ou perdas, para o regime 1, é possível observar a presença de bimodalidade, o que nos trás indícios de que existe mais de um regime de mercado dentro dessa componente, indicando a ocorrência de alternância entre períodos de ganho e prejuízo. O regime 2, por sua vez é assimétrico a direita, ou seja, está inserido no eixo positivo, o que sugere um indicativo para um período com consideráveis ganhos financeiros.

3.2 Aplicação do HMM com Algoritmo EM

Utilizando o Hidden Markov Model, temos os resultados obtidos para os retornos ajustados e os regimes observados para S&P500 apresentados a seguir:

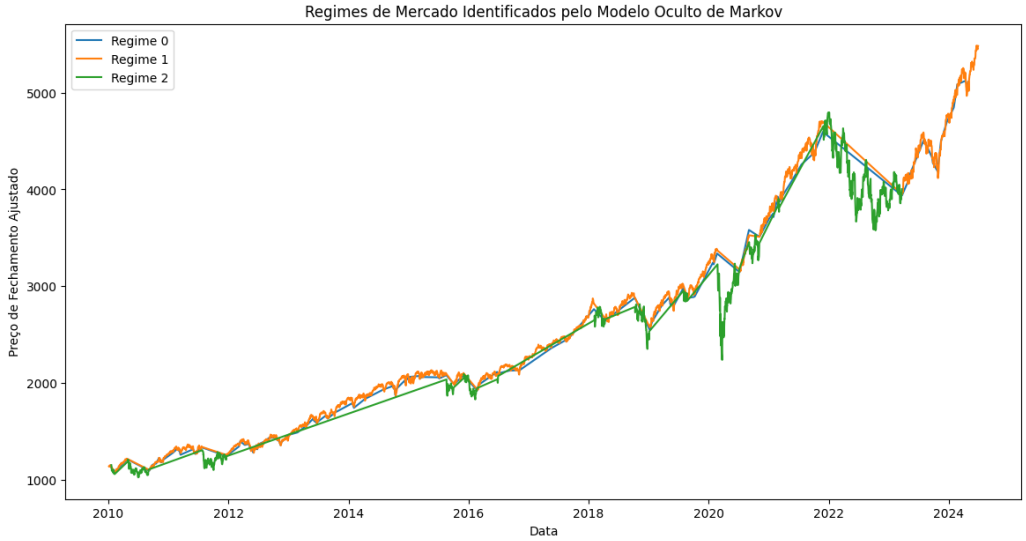

Figura 6: Componentes de mercado identificadas via HMM para S&P500

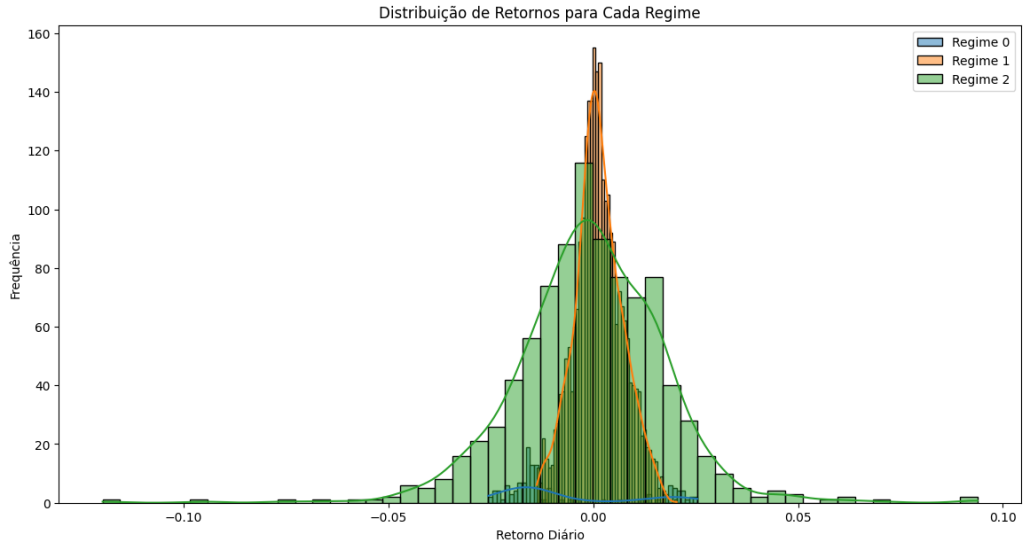

Figura 7: Retornos observados para cada regime via HMM para S&P500

Com base nas observações anteriores, é possível notar que, utilizando n = 3 regimes de mercado, os retornos ajustados para o regime zero apresentam bimodalidade, a qual indica a presença de múltiplos comportamentos, como já observado para casos anteriores, já o regime 1 apresenta distribuição gaussiana centrada em zero para os retornos com baixo desvio padrão, o que indica um período de poucos retornos e menor volatilidade, já para o regime 2, observamos uma distribuição gaussiana levemente deslocada para a esquerda, o que indica perdas pouco significativas, a elevada dispersão dos dados, entretanto, revela um regime com consideráveis incertezas acerca do mercado, resultando em uma volatilidade mais alta.

Já para o índice Ibovespa, os regimes identificados, bem como os retornos provenientes podem ser encontrados nas visualizações a segir:

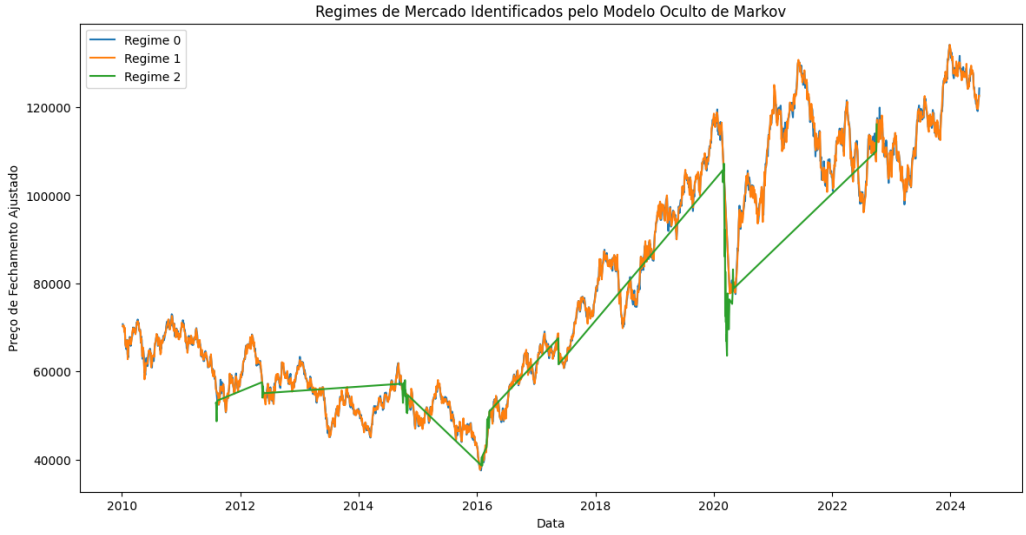

Figura 8: Componentes de mercado identificadas via HMM para Ibovespa

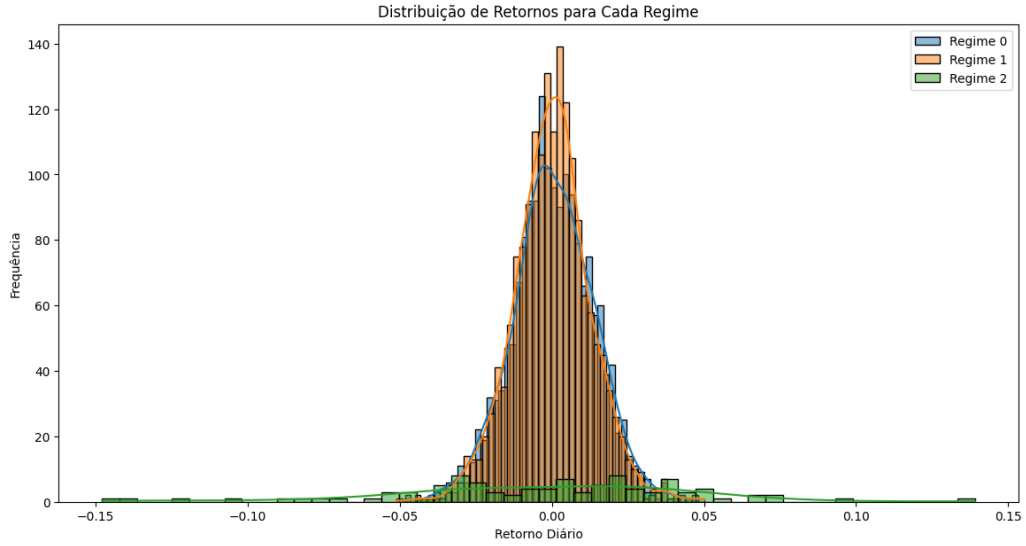

Figura 9: Retornos observados para cada regime via HMM para Ibovespa

A partir dos resultados observados e também dos ajustes das componentes,, temos que os retornos ajustados dos regimes 0 e 1 são bastante similares entre si, apresentando distribuições aproximadamente gaussianas centradas em zero e com dispersão moderada, indicando previsibilidade de seus respectivos regimes, baixos retornos e volatilidade moderada, já para a segunda componente, temos que o modelo, embora simétrico e centrado em zero, apresenta elevada dispersão, o que indica elevada volatilidade para esse tipo de regime, não sendo ideal para o investimento.

3.3 Aplicação do GMM com a abordagem Bayesiana

A implementação do modelo Bayesiano para GMM também pode nos trazer informações importantes acerca dos regimes de mercado, uma vez que não exige um número pré-fixado de componentes, no entanto, tendo em vista a complexidade do modelo foram testados dois cenários, um com tolerância máxima de n = 3, e outro com n = 10 componentes.

As visualizações gráficas a seguir nos permitem observar o ajuste para n = 3 regimes:

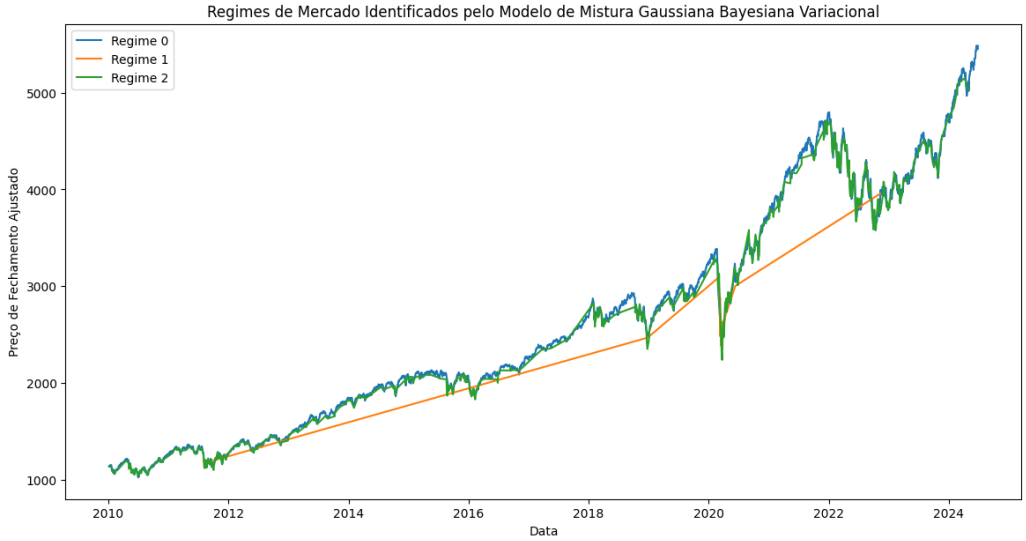

Figura 10: Componentes de mercado identificadas via GMM com abordagem Bayesiana para S&P500

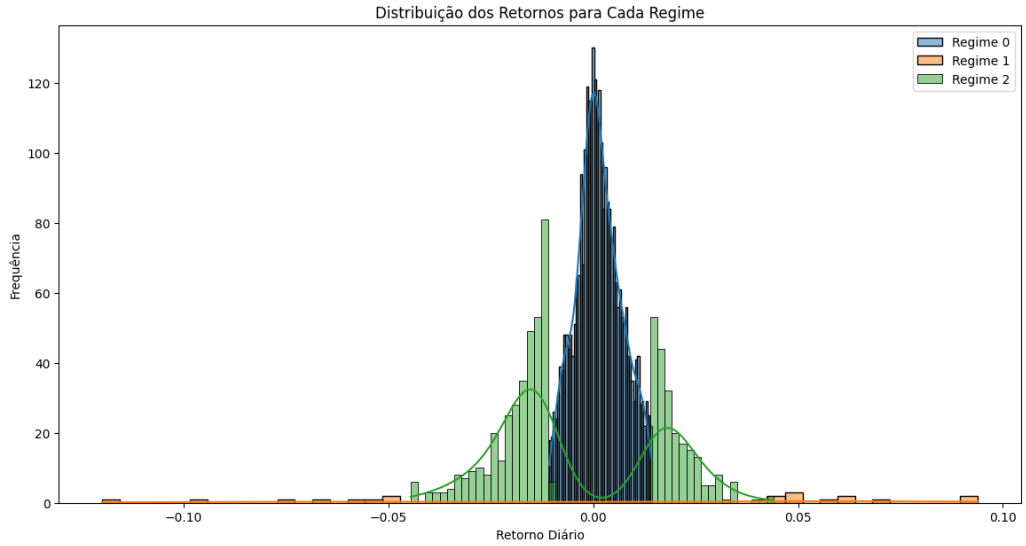

Figura 11: Retornos observados para cada regime via GMM com abordagem Bayesiana para S&P500

É possível observar que o regime 0 apresenta distribuição gaussiana centrada em zero e baixo desvio padrão, possuindo, assim, comportamento previsível e baixo retorno financeiro, para o regime 2, temos retornos bimodais, indicando a presença de alternância entre regimes de alta e baixa dentro de uma mesma componente, já para o regime 1, temos uma elevada dispersão dos dados de ganhos, além de possível bimodalidade, o que pode nos indicar elevada volatilidade de mercado, bem o como alternância de estado.

Já para o Ibovespa, tem-se:

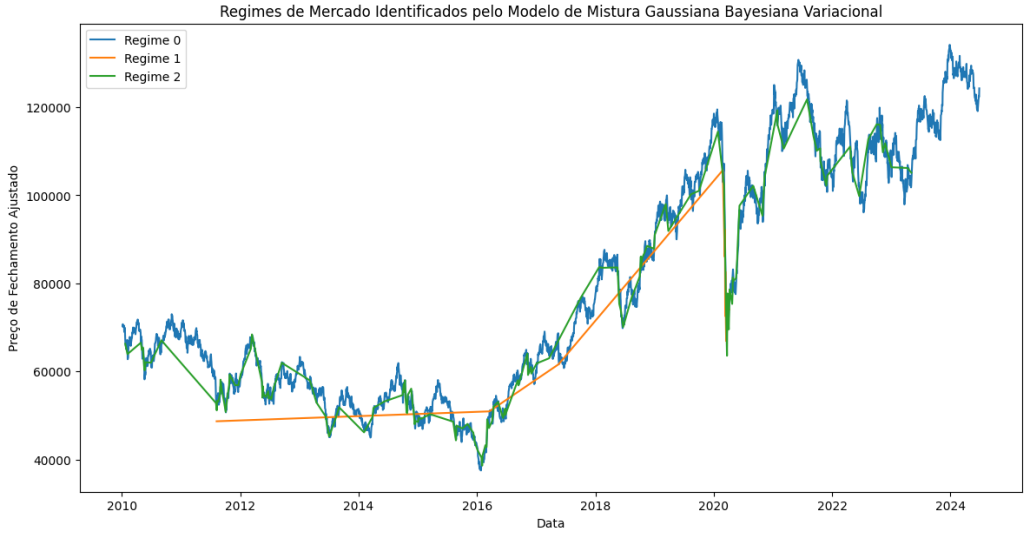

Figura 12: Componentes de mercado identificadas via GMM com abordagem Bayesiana para Ibovespa

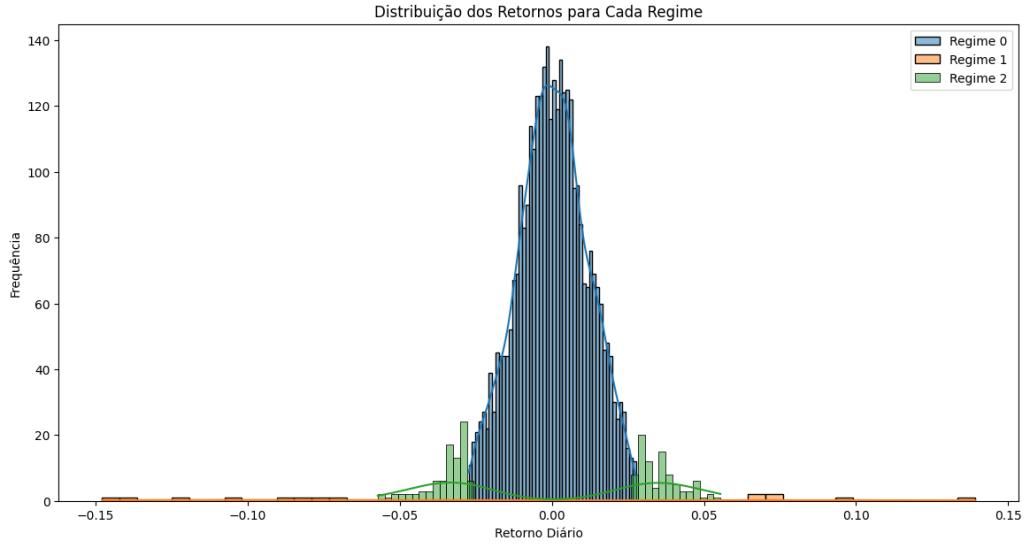

Figura 13: Retornos observados para cada regime via GMM com abordagem Bayesiana para Ibovespa

Assim como no caso anterior, o regime 0 apresenta distribuição gaussiana centrada em zero, mas com razoável desvio padrão, apresentando baixos retornos e moderada volatilidade, para o regime 1, a bimodalidade indica presença de alternância entre períodos de alta e baixa dentro da mesma componente, já para o caso 2, temos um elevado desvio padrão dos retornos, bem como aparente distribuição bimodal, o que aponta para uma elevada volatilidade de mercado e alternância de regime nesse período.

Considerando o caso em que um limite de 10 componentes foram permitidas, temos as seguintes observações para Ibovespa e S&P500:

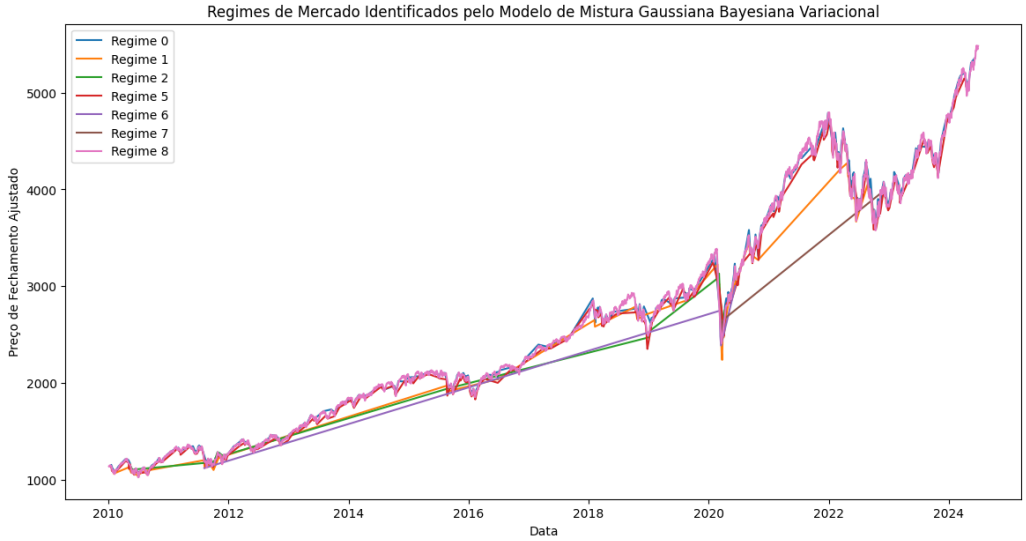

Figura 14: Regimes de mercado identificados via GMM Bayesiano com 10 componentes para S&P500

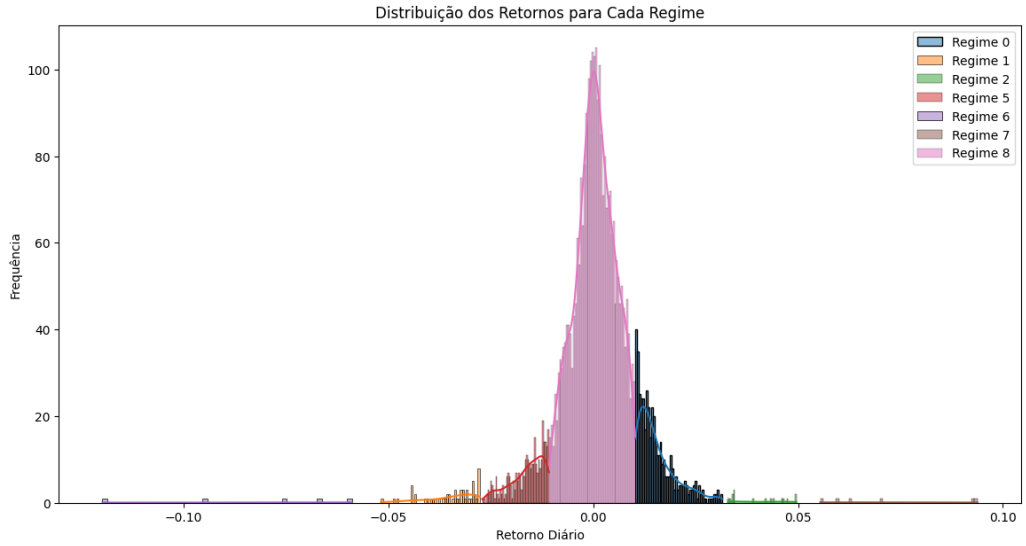

Figura 15: Retornos observados para cada regime via GMM com Bayesiano com 10 componentes para S&P500

Como é possível observar nos gráficos acima, foram encontrados pelos sete regimes significativos de mercado, sendo que o regime oito, centrado em zero, é o de maior ocorrência e indica um mercado estável e pouco volátil, enquanto as componentes zero, dois e sete encontram-se deslocadas para a direita, o que nos indica um período de crescimento econômico, propenso para bons ganhos financeiros, por sua vez um, cinco e seis encontram-se deslocadas para a esquerda, um indicativo de instabilidade econômica e perdas de ordem financeira.

As observações para a implementação Bayesiana do Ibovespa podem ser observadas logo abaixo:

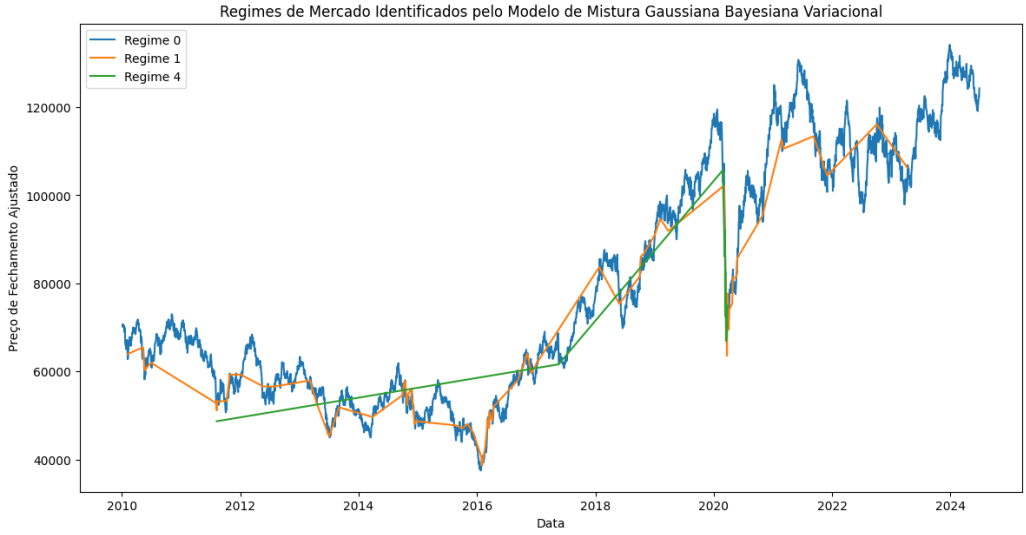

Figura 16: Regimes de mercado identificados via GMM Bayesiano com 10 componentes para Ibovespa

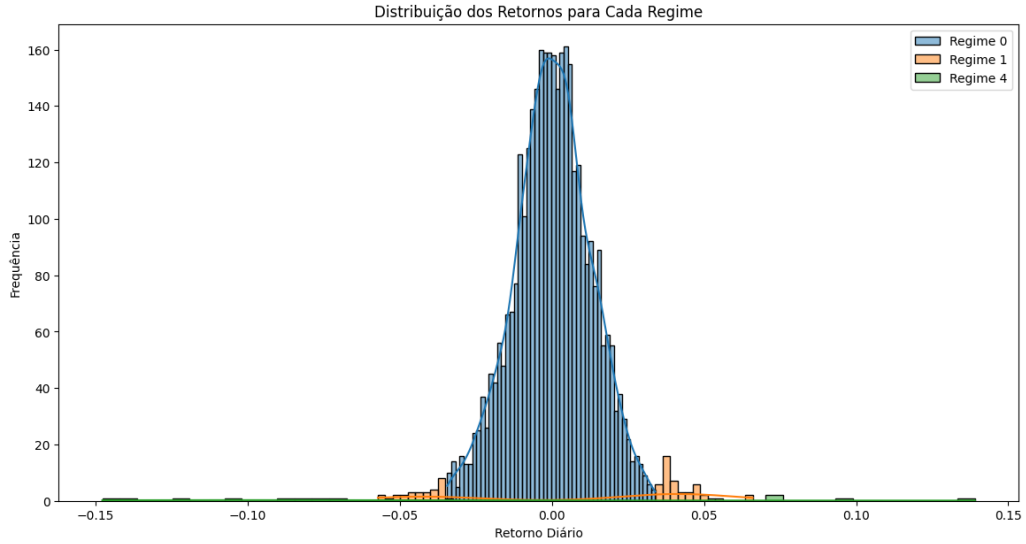

Figura 17: Retornos observados para cada regime via GMM com Bayesiano com 10 componentes para Ibovespa

No caso do Ibovespa, foram ajustados apenas três regimes de mercado, no regime 0, observamos um mercado estável e com mediana volatilidade, já para um, observamos a existência de mais de um período por componente, indicando alternância entre ganhos e perdas, em quatro, por sua vez, observa-se elevada dispersão dos dados e, portanto, elevada volatilidade.

É importante notar que, apesar de aparentar possuir mais regimes de mercado que o índice S&P500, como observado na figura 1, temos que a quantidade de componentes para o índice Ibovespa é, na verdade, muito inferior ao respectivo do mercado estado-unidense, como indicam as figuras 14 e 16, evidenciando que apenas a visualização de dados não seria suficiente para indicar as dinâmicas existentes nos mercados dos dois países.

3.4 Comparação dos modelos

3.4.1 Comparação dos retornos médios ajustados

Como o objetivo central do modelo é observar os modelos que nos trazem maior rentabilidade, uma boa análise seria a dos ganhos esperados ajustados. A seguir observamos essa análise para n = 3 componentes:

Regime (S&P500)	GMM(EM)	HMM(EM)	GMM(Bayesiano)
0	0.00101081	-0.00087408	0.00124033
1	-0.00386417	0.00118031	-0.00051437
2	0.00095746	-0.00108217	-0.00122026

Tabela 1: Tabela de comparação de retornos para S&P500 com base nos modelos investigados para 3 componentes

Regime (Ibovespa)	GMM (EM)	HMM (EM)	GMM (Bayesiano)
0	-0.00160705	0.00081646	0.00034332
1	-0.00224594	-0.00013976	-0.01440752
2	0.0075454	-0.00156363	0.00064632

Tabela 2: Tabela de comparação de retornos para Ibovespa com base nos modelos investigados para 3 componentes

Como podemos observar, para ambos os índices, não foi possível encontrar algum modelo que obtivesse ganhos ou perdas muito significativas quando comparado aos demais, entretanto, é importante salientar que, como já observado, os regimes possuem formas, comportamentos e volatilidade distintos, sendo a medida de posição μ dos retornos não muito eficaz para medir o quão bom é o desempenho de cada modelo.

A seguir, é possível observar a tabela para o BGMM quando permitimos um máximo de 10 componentes:

Regime (S&P500)	GMM (Baysiano) com as componentes significativas
0	0.0038524
1	-0.01332914
2	0.00942429
5	-0.00463777
6	-0.02700928
7	0.02324516
8	0.00079397

Tabela 3: Tabela de comparação de retornos para S&P500 com base nos modelos investigados para até 10 componentes

Regime (Ibovespa)	GMM (Baysiano) com as componentes significativas
0	0.00027275
1	0.00145974
4	-0.01729654

Tabela 4: Tabela de comparação de retornos para Ibovespa com base nos modelos investigados para até 10 componentes

Assim como para o caso com 3 componentes, temos que o caso com 10 componentes, embora possua médias aparentemente mais robustas, não esboçou uma performance significativa, no que tange a ganhos ou perdas, não nos permitindo extrair bastantes informações valiosas acerca do desempenho dos modelos explorados.

3.4.2 Comparação dos gráficos de retorno

Com base nos gráficos de retorno ajustado das figuras 3, 5, 7, 9, 11, 13, 15 e 17, temos que o modelo HMM via algoritmo EM é, possívelmente, o que menor interesse econômico apresenta aos investidores, uma vez que a volatilidade observada, bem como os baixos ganhos esperados não aparentam ser muito atrativos, a bimodalidade, bastante presente no GMM, também pode indicar certos desafios, uma vez que necessitaria de modelos mais complexos para tirar proveito do período de ganhos e evitar aqueles onde o mercado se encontra em maiores dificuldades.

Para o modelo BGMM, visualizamos que, com um máximo de 3 componentes, o modelo também pode ter seu desempenho afetado pela bimodalidade e pela volatilidade do mercado, entretanto, para o caso com até 10 regimes, temos que o impacto desses acaba por ser mitigado, sendo que, no caso do índice S&P500, existe a clara separação entre regimes onde haverá retorno financeiro, prejuízo e ganhos pouco significativos, algo bastante desejável para os investidores.

3.4.3 Comparação dos tempos de rodagem

Outra avaliação muito importante diz respeito à complexidade do algoritmo, que se traduz no seu tempo de execução. Um determinado modelo pode retornar ganhos robustos, no entanto, sua implementação pode ser computacionalmente inviável, tornando-o menos desejável e eficiente.

Na tabela a seguir é possível observar a comparação dos tempos de rodagem para os modelos apresentados com os índices S&P500 e Ibovespa:

Modelo	**Tempo(s) para S&P500**	**Tempo (s) para Ibovespa**
GMM(EM)	2.131124258041382	2.7216434478759766
HMM(EM)	4.38442063331604	1.625375509262085
BGMM(n=3)	1.677307367324829	2.801020622253418
BGMM(n=10)	3.247539520263672	8.80511212348938

Tabela 5: Tabela de comparação dos tempos de rodagem para cada um dos modelos apresentados

Como pode ser observado, temos que os tempos de rodagem para os três primeiros modelos são bastante similares, provavelmente devido ao fato de ambos receberem n = 3 componentes, já para o BGMM com até 10 regimes distintos, temos que o tempo até a execução é aproximadamente o dobro dos anteriores, o que indica que, apesar de mais preciso, o modelo com maior número de componentes também apresenta um custo computacional maior.

4 Conclusão

Podemos concluir que cada um dos três modelos nos retorna abordagens distintas para a identificação dos regimes de mercado, sendo bastante aplicáveis para as mais diversas situações, no entanto, é importante salientar que o modelo BGMM multiregimes, abordagem com tolerância de até 10 componentes, apresenta certa vantagem sobre o GMM e o HMM, sobretudo no que tange à capacidade de reduzir o impacto de cenários onde há existência de bimodalidade e elevada volatilidade de mercado, com notória performance no reconhecimento de períodos onde há crescimento, estagnação e queda nos retornos para o S&P500, algo que os demais algoritmos não foram capazes de identificar, em comparação ao HMM, o modelo GMM também se mostrou útil por mitigar o efeito da volatilidade dentro das componentes, no entanto, não foi capaz de identificar com clareza cenários de fortes ganhos financeiros, apenas tendo sido observados cenários com regimes mistos, ou com retornos reduzidos, salvo a componente 2 para o índice Ibovespa, demonstrando, por isso, ser menos eficiente que o cenário Bayesiano multiregimes, embora não tenha apresentado diferenças significativas quando comparado ao cenário Bayesiano com equivalente número de dinâmicas. O algoritmo HMM por sua vez, embora mais adaptado a esse tipo de problema, especialmente por levar em consideração a dependência temporal dentro das séries, nos trouxe cenários com retornos pouco significativos e elevada volatilidade, prejudicando seu desempenho para potencializar os ganhos.

Referências Bibliográficas

P. Chandrika, K. Visalakshmi, and K. Sakthi Srinivasan. Application of hidden markov models in stock trading. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1144–1147, 2020. doi: 10.1109/ICACCS48705.2020.9074387.

A. K. David M. Blei and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL `https://doi.org/10.1080/01621459.2017.1285773`.

A. Gramfort, G. Varoquaux, V. Dubourg, and et al. hmmlearn: Hidden markov models in python with scikit-learn. `https://hmmlearn.readthedocs.io/en/latest/`, 2014.

K. Kayabol. Bayesian gaussian mixture model for spatial-spectral classification of hyperspectral images. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1805–1809, 2015. doi: 10.1109/EUSIPCO.2015.7362695.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL `https://jmlr.org/papers/v12/pedregosa11a.html`.

M. Przyborowski and D. Ślezak. Approximation of the expectation-maximization algorithm for gaussian mixture models on big data. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6256–6260, 2022. doi: 10.1109/BigData55660.2022.10020450.

L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986. doi: 10.1109/MASSP.1986.1165342.

Ran Aroussi, Gabriel Gerrard, and Other Contributors. yfinance: Yahoo finance market data downloader. `https://github.com/ranaroussi/yfinance`, 2024.