# VOLATILIDADE
# Uma Introdução Quant

SÁVIO COELHO
Universidade de Fortaleza
saviocoelho@edu.unifor.br

Abril 2025

**Resumo**

A metrificação do risco é uma das áreas essenciais em finanças quantitativas. As diferenças entre modelos podem levar a diferentes medidas de risco e escolhas de alocação financeira diferentes. O presente artigo busca fornecer uma introdução a alguns desses modelos de engenharia financeira.

## Introdução

A natureza volátil do nosso mundo não é surpresa para ninguém. Vivemos cercados de incertezas e em uma condição fundamental de ignorância quase completa sobre as causas. O mundo está em processo de constante mudança e é difícil para a mente humana, a qual não evoluiu para esse tipo de tarefa, acompanhar tais fenômenos dinâmicos

Em finanças isso é particularmente sensível devido os fenômenos analisados, particularmente as flutuações nos retornos de ativos, serem notoriamente caracterizados por um comportamento aleatório. O risco acaba sendo a consequência dessa volatilidade.

A aleatoriedade que domina os fenômenos financeiros é tão grande que muitos autores argumentam que sequer é possível realizar uma análise completa do comportamento das flutuações e sua predição. O objeto de estudo das finanças seria assim tão aleatório quanto as partículas sub-atômicas analisadas pela física e a química. Não é incomum que pessoas que mergulhem no estudo desse tema em finanças tenham que aprender a arcana área da teoria da probabilidade para

lidar com essa característica do seu objeto de análise.

Todavia, os profissionais de finanças que mergulham profundamente nessa questão são poucos. Muitas vezes a massa do mercado os vê como místicos ou mesmo "magos"devido estudarem temas com nomes tão exotéricos como "*matrizes de Markov*", "*regressão quantílica*"e "*cálculo estocástico*".

O objetivo do presente texto é introduzir aqueles que desejam se aventurar em águas tão turvas de uma maneira simples. Será analisado as flutuações de retorno de um ativo financeiro e será revisado vários modelos que tentam determinar e prever a volatilidade dessa série histórica. O noviço que deseja entrar nessa jornada arcana irá ver tanto os modelos mais básicos como os mais...estranhos.

Apesar de abordar temas matemáticos complexos, não irei aqui me prender em demonstrações. Afinal, quem no mercado tem tempo para isso? O que irei tentar fazer como abordagem didática é introduzir os conceitos de uma maneira direta e prática. Assim, o presente texto tem uma visão mais de um *engenheiro financeiro* (ou quant, como são chamados) do que de um acadêmico matemático.

# 1 Escolhendo um Ativo

Primeiramente, é necessário escolher o ativo que iremos analisar. Dentre os ativos que podemos analisar volatilidade estão moedas, imóveis, criptoativos e ações. Os ativos de renda fixa, como letras do tesouro e dívida corporativa, não serão considerados, pois a medição de seu risco é diferente das outras classes de ativos[1].

Dentro do universo de ativos citados, irei analisar ações. A razão para essa escolha é bastante pragmática: é mais fácil. Os dados de ações possuem séries históricas longas, são de fácil acesso e mais fáceis de analisar do que, por exemplo, câmbio.

A ação que irei analisar será a NVIDIA Corporation (NVDA) listada na Nasdaq para o período de 01/01/2019 a 01/01/2025. A escolha se dá mais pelo *hype* em torno das questões de inteligência artificial do que qualquer outra coisa, acreditem. Não irei me prender em narrativas de *valuation* sobre a importância de determinada companhia[2], mas talvez seja importante sabermos algumas informações básicas. A NVIDIA é uma empresa especializada em unidades de processamento gráfico (GPUs) e soluções como a A100 e a plataforma CUDA

[1]Isso se deve a seu retorno ao longo do tempo ser "conhecido". A medição da volatilidade de um ativo de renda fixa é feita pela variação entre dois pontos fixos e envolve classes de modelos brownianos que não será abordado aqui

para inteligência artificial. Ela também atua em data centers com sistemas como o DGX, no setor automotivo com a plataforma DRIVE para carros autônomos, e em visualização profissional com as GPUs Quadro, usadas em design e animação 3D, impactando desde entretenimento até inovações científicas. Ou seja, uma empresa *beemmmm* importante.

A NVIDIA apresentou um forte crescimento em suas ações no período recente. Abaixo é representado a série histórica da cotação da ação NVDA[3].

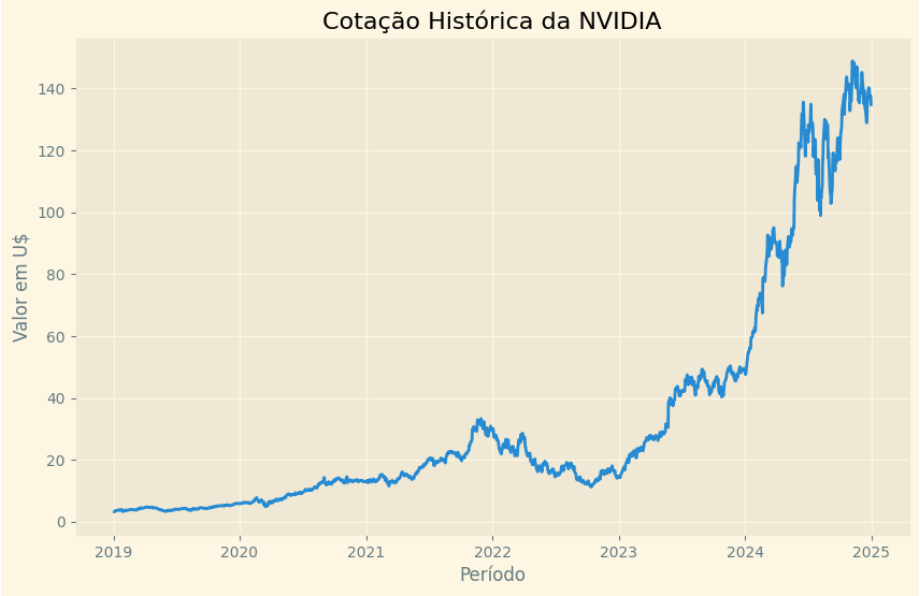


Figura 1: Cotação História da NVDA entre 01/01/2019 e 01/01/2025

## 2 Definindo Volatilidade

Para que possamos estudar a volatilidade é necessário antes definir o que ela é exatamente.

Em finanças a volatilidade pode ser definida como a medida de quão mutável são os preços de um ativo no mercado. Se os preços apresentam grandes movimentos de alta e queda seguidos, dizemos que a volatilidade é alta. Do contrário,

[2]Em minha visão a natureza das firmas é contingente. Alguns anos atrás a NVIDIA era uma companhia de esquisitões do Vale do Silício que competiam na sombra da Intel vendendo placas para gamers e mineradores de bitcoin. Ninguém jamais apostaria neles. Que técnica que valuation previu que ocorreria um salto de várias ordens de grandeza em seu retorno?

[3]Escolhi o estilo de cor salmão para os gráficos mais para imitar o Financial Times

quando eles são baixos, dizemos que a volatilidade é baixa. Imagine que você está acompanhando o preço de uma ação chamada "Macrohard". Em um dia tranquilo, o preço da Macrohard varia pouco, oscilando entre 10,00 e 10,20 — baixa volatilidade. Já em um dia agitado, com notícias sobre a empresa, o preço pula de 10,00 para 12,00, cai para 9,50 e sobe novamente para 11,80, tudo em poucas horas — alta volatilidade.

Os analistas financeiros não estão tão preocupados com o preço em si, mas com o *retorno*. O retorno é simplesmente a diferença entre o preço pago na compra de um ativo e seu preço de venda ou liquidação. Analiticamente é possível expressar como:

$$R_n = \frac{p_n - p_{n-1}}{p_{n-1}}$$

Onde $p_n$ equivale ao preço atual do ativo e $p_{n-1}$ é o preço pelo qual o ativo foi comprado. O numerador da fórmula indica assim a variação no preço do ativo do dia n a partir do dia n-1. Ao realizar a divisão pelo preço no dia n-1 temos o retorno em quatidade relativa. Se multiplicarmos esse retorno por 100 teremos o mesmo em percentual.

Contudo, uma maneira mais usual e correta de se calcular os retorno é por meio do uso de logaritmos naturais. Esses logaritmos são muito semelhantes aos que você viu na escola, porém, ao invés da base 10, é utilizado como base o número de Euler, $e$, que é aproximadamente 2,718...

O uso de logaritmos naturais no cálculo de retornos financeiros é fundamental devido às suas propriedades matemáticas. Uma das principais é a propriedade de aditividade. Enquanto os retornos aritméticos, como os definidos anteriormente, exigem o uso de fórmulas mais complexas para serem combinados ao longo de múltiplos períodos, os retornos logarítmicos podem ser somados diretamente, facilitando o cálculo do retorno total em intervalos mais longos. Se definirmos que o retorno financeiro logarítmico do n-nésimo dia, $U_n$, é definido como:

$$U_n = ln(1 + R_n)$$

Com $R_n$ sendo o retorno aritmético. Uma vantagem da nossa definição anterior é que podemos trabalhar o logaritmo natural de 1 + x por meio de uma série de Taylor em torno de x=0, dado que os retornos financeiros são geralmente dados em 0,01 por diante. Assim temos que:

$$ln(1 + R_n) = R_n - \frac{R_n^2}{2} + \frac{R_n^3}{3}...$$

Só que uma característica de finanças é que os retornos geralmente são muito pequenos, sobretudo para retornos diários. Geralmente os retornos são 2%, 1% ou as vezes até menos. Só que números como 0,01 elevados ao quadrado ou

ordens de grandeza maiores é um número *muito* pequeno. Ele é tão pequeno que podemos até dizer que é insignificante. Dessa forma, podemos considerar, para o caso dos retornos, o segundo termo da série de Taylor como equivalente a zero, gerando o resultado de que *aproximadamente*:

$$U_n = R_n$$

Essa conclusão é interessante, pois se tivermos que analisar, por exemplo, qual o retorno no período de três dias, sendo que cada dia teve retornos de 5%, 4% e 3%, teremos que o retorno total ao longo desses dias será:

$$ln(1 + 0,03) + ln(1 + 0,04) + ln(1 + 0,05)$$

Se fizermos as séries de Taylor completas para cada termo da soma teremos que $0,04875 + 0,0392 + 0,02955 = 0,1175$. Mas isso aproximado a fazermos o logaritmo natural em torno de zero tal que:

$$ln(1 + (0,05 + 0,03 + 0,04)) = 0,1248$$

Assim, a utilização do logarítmo natural garante que possamos ter uma aproximação bastante razoável de que ele vai nos dar o retorno cumulativo de vários dias ao longo de uma série histórica pela simples soma dos retornos individuais de cada dia.

Além disso, ao aplicar o logaritmo natural a dados que seguem uma distribuição log-normal, como é pressuposto dos preços de ativos financeiros, a transformação resulta em uma distribuição aproximadamente normal. Isso facilita e é extremamente importante para modelagem estatística como veremos em breve. Realizando o logaritmo natural da série histórica de preços da NVIDIA temos a série logarítmica abaixo:

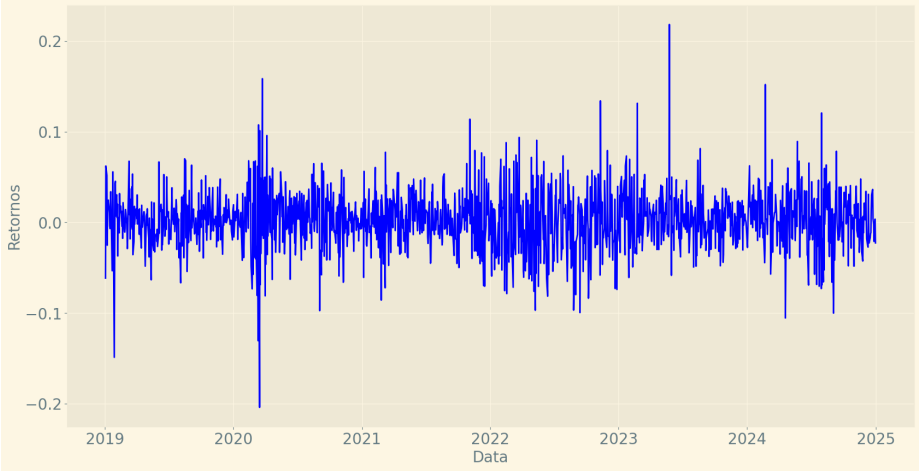


Figura 2: VaR Empírico da NVDA com Confiança de 95%

Como é possível ver a partir do gráfico os retornos da ação da NVIDIA são tudo, menos estáveis. Essa é uma característica que não é exclusiva da empresa, essa é a *volatilidade* dos ativos financeiros. Para medir essa volatilidade geralmente utilizamos a medida estatística do desvio-padrão, $\sigma$.

O desvio-padrão dos retornos financeiros é dado pela fórmula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(U_i - \mu)^2}{N-1}}$$

Onde $N$ é o período ou tamanho da amostra de dados (anos, dias), $U_i$ é o retorno logarítmico para o dia $i$ e $\mu$ é a média dos retornos. Um valor alto do desvio padrão significa que a distribuição dos dados em torno de sua média é ampla, enquanto que um valor baixo significa que os dados estão mais próximos da média. Valor de desvio-padrão alto indicam retornos mais voláteis e valor baixo indica retornos mais estáveis.

Calculando o desvio-padrão para nossa série de retornos teríamos o valor de 0.03, o que é bem baixo.

# 3 O Value-at-Risk (VaR)

Ok, quantificamos a volatilidade, mas com que finalidade? Para que vou utilizar esse número de maneira prática? A volatilidade pode ser utilizada de diversas maneiras em finanças, porém uma das mais populares é para calcular o Value-at-Risk (VaR).

O VaR é uma métrica estatística utilizada no setor financeiro para quantificar o risco de perda de um investimento em um determinado período sob condições normais de mercado e com um nível de confiança específico.

Imagine que você é um investidor com uma carteira de ações no valor de R$ 100.000. Você quer saber o quanto pode perder com esse investimento, com 95% de confiança. Usando o VaR, você analisa os dados históricos e calcula que, em 95% dos dias, a perda máxima seria de até R$ 5.000. Isso significa que, em um dia típico (dentro desse intervalo de confiança), você não perderia mais que R$ 5.000, mas em 5% dos casos (os piores dias), a perda poderia ser maior. O VaR, nesse caso, é R$ 5.000.

Os quants gostam também de utilizar uma medida chamada de Condicional Value-at-Risk (CVaR). Ele é uma medida de risco que quantifica o risco de cauda (os "outliers") de um investimento. Enquanto o VaR estima a perda máxima esperada dentro de um intervalo de confiança específico, o CVaR vai além, calculando a média das perdas que *excedem* o VaR. Para fins didáticos,

contudo, iremos calcular apenas o VaR e ver como diferentes medidas de volatilidade alteram o VaR de um mesmo ativo.

Uma forma muito comum de se calcular o VaR é por meio da chamada abordagem empírica ou quantílica. Ela é uma metodologia não-paramétrica que utiliza dados históricos de retornos financeiros para determinar diretamente os percentis desejados. Dessa forma, evita-se a necessidade de assumir que a distribuição dos retornos obedeça a ou de estimar parâmetros como a volatilidade. O VaR empírico é dado por:

$$VaR_{\alpha} = -r((1-\alpha) \cdot n)$$

Onde $\alpha$ é o intervalo de confiança, $r$ é o retorno e $n$ é o número de observações. O VaR empírico então toma os últimos percentis negativos da distribuição dos retornos e determina que aquele deve ser o máximo valor a ser perdido dado o intervalo de confiança. Tomando um intervalo de 95%, um investimento inicial de R$ 100.000 e os dados de retorno da NVIDIA, o VaR empírico seria representado da seguinte forma na distribuição e teria um valor de R$ -4838,59 ou 0,0483859 para qualquer valor investido:

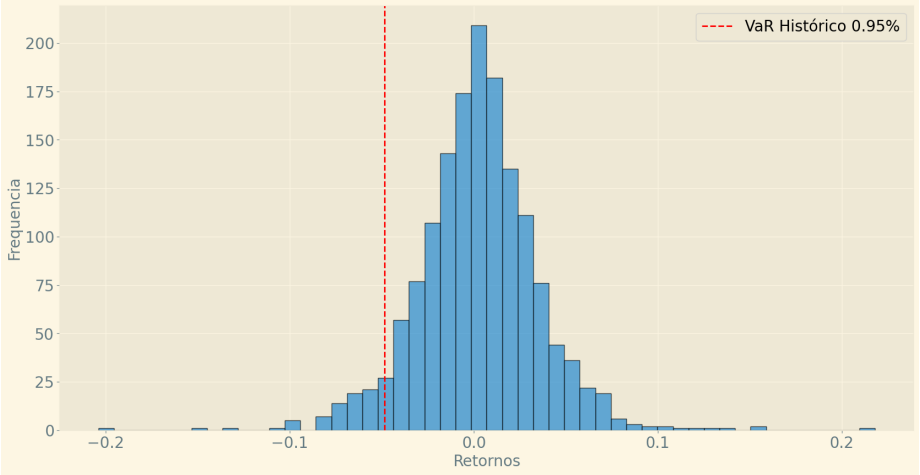


Figura 3: VaR Paramétrico com Desvio-Padrão Simples e Confiança de 95%

Apesar de suas vantagens, o VaR empírico apresenta algumas limitações. A mais notável é sua dependência dos dados históricos. O método é altamente dependente dos dados utilizados. Se os dados históricos não forem representativos, ou se o período de análise não capturar eventos extremos, o VaR histórico pode não ser preciso e o VaR gerado será viesado. Além disso, como ele é baseado apenas em observações passadas, o VaR empírico não consegue capturar mudanças estruturais que não ocorreram no passado, como inovações ou mudanças de política econômica.

A outra maneira de calcular o VaR é por meio da arbodagem paramétrica. Essa abordagem baseia-se na suposição de que os retornos dos ativos seguem uma distribuição normal. Essa abordagem utiliza as estatísticas de média e volatilidade dos retornos para determinar o VaR. Sua fórmula é dada por:

$$VaR_{\alpha} = -(C \cdot z \cdot \sigma)$$

Onde $C$ é o capital investido, $z$ é o *valor crítico*[4]para o nível de confiança dado uma distribuição normal e $\sigma$ é a medida de volatilidade. Como é possível observar, apesar de ser mais simples e não ser dependente de dados histórico, o VaR paramétrico depende de pressupostos de normalidade, sobretudo na ausência de valores extremos, e depender da estimação prévia de parâmetros, como a volatilidade.

Se assumirmos que a volatilidade dos retornos da NVIDIA é dado simplesmente pelo desvio-padrão da série histórica, ou seja 0,03291, e um valor crítico para distribuição normal de 1,645, o VaR paramétrico da NVDA será R$ -5346,39 ou 0,05446,39 de perda para o capital investido. Logo, já temos um valor diferente daquele calculado pelo VaR empírico. Abaixo temos a representação gráfica:

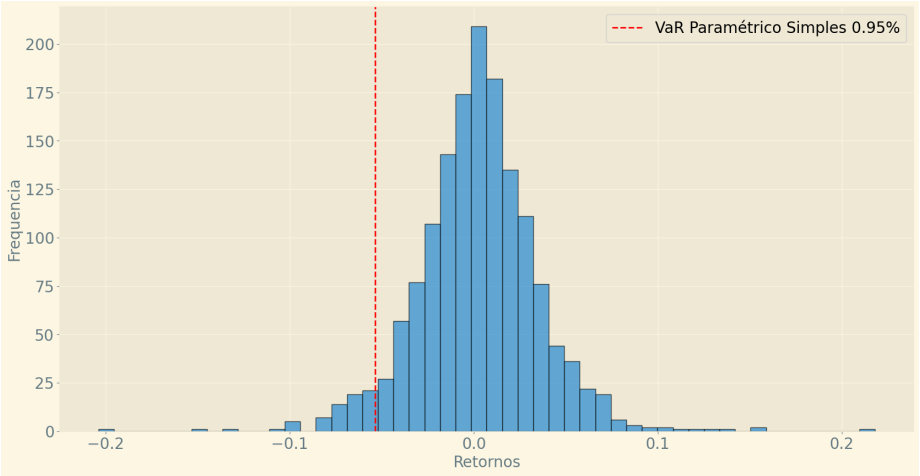


Figura 4: VaR Paramétrico com desvio-padrão simples e Confiança de 95%

Mas qual abordagem de VaR escolher? Segundo (DOBRÁNSZKY, 2009), ambos os métodos são bons desde que você tenha técnicas de normalização dos dados adequadas. Essas técnicas envolvem uma manipulação de conceitos estatísticos que não vem ao caso no presente artigo. Devido permitir realizar um

[4]O valor crítico é um conceito bastante importante dentro da estatística. Não cabe aqui discutir longamente sobre ele, mas recomendo leitura adicional para entender o conceito.

número maior de simulações, não ser dependente de dados históricos e funcionar mesmo com poucos dados, vamos seguir em nossos exemplos com o VaR paramétrico.

Uma questão se iremos trabalhar com um VaR paramétrico é: os dados seguem uma distribuição normal? Se olharmos para o que ocorre caso tentemos colocar nossa distribuição histogramática em uma distribuição normal, teremos a representação abaixo. Os dados claramente não se encaixam em uma distribuição normal. Eles apresentam leptocurtose (esse nome doentil simplesmente significa que o topo da distribuição dos nossos dados é "pontudo") e valores extremos formando caudas longas (a distribuição possui mais valores extremos que uma distribuição normal).

Se fizermos um Teste de Anderson-Darling para verificar a normalidade dessa distribuição encontraremos que o valor da estatística é 6.01. A um nível de significância $\alpha = 0.01$ teremos um valor crítico de apenas 1.08, o que nos permite rejeitar a hipótese nula de que a distribuição é uma distribuição normal.

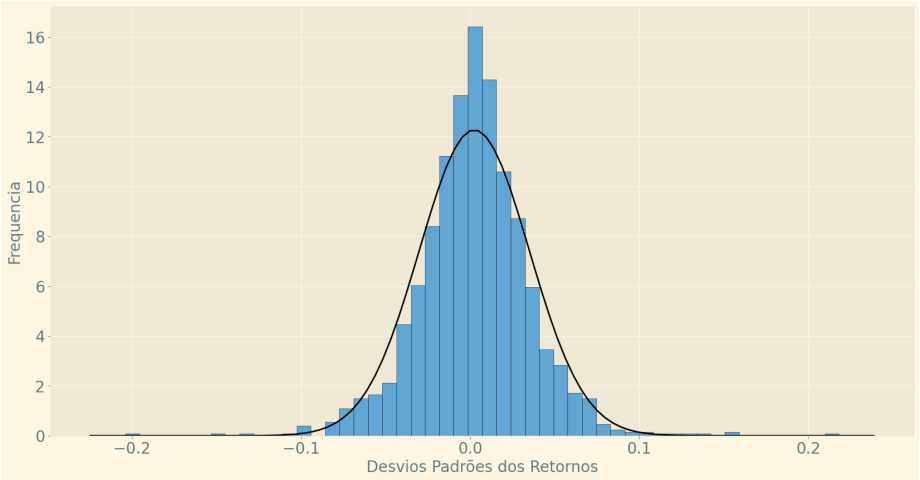


Figura 5: Distribuição Real vs Distribuição Normal

A ampla maioria dos dados financeiros apresenta essas características. Apesar de que séries históricas de longo prazo tenderem a se aproximar mais da distribuição normal do que séries de curto prazo, os dados sempre vão apresentar ao menos caudas longas em suas distribuições. Em seu estudo clássico sobre distribuições financeiras, (FAMA, 1965) demonstrou que as distribuições de dados financeiros tendem a apresentar caudas cada vez mais longas ao longo da série de tempo.

Todavia, a distribuição normal ainda é a maneira mais simples de se realizar os cálculos. Dessa forma, assumir que os dados possuem uma distribuição

normal serve como pressuposto simplificador da análise em primeiro momento. Não tomemos aqui o termo "simplificador"de maneira negativa. A simplificação é necessária em primeiro momento para que seja mais fácil abstrair e absorver os conceitos. Iremos realizar isso ao longo do texto e posteriormente, provavelmente em outro momento, trabalhar o relaxamento do pressuposto de normalidade nos modelos de risco.

A questão após essa simplificação é: o desvio-padrão simples é realmente adequado para se estimar o VaR paramétrico? A resposta é não. Calcular o desvio-padrão dos retornos de séries históricas longas, como o caso analisado da NVIDIA, é inadequado devido ele calcular a média de períodos com altas e quedas fortes. Assim, ele tenderia a ignorar eventos extremos nos dados de retorno e acabaria sendo uma medida viesada.

Por causa disso, seguiremos o estudo da volatilidade olhando como a volatilidade pode ser calculada para se considerar essas variações nos dados.

# 4 Médias Móveis

Uma das maneiras mais populares de se contornar o problema dos eventos extremos no cálculo de volatilidade é realizar a suavização da série temporal por meio de média móvel da mesma. A suavização é um conjunto de técnicas estatísticas utilizadas para reduzir o ruído em dados temporais, permitindo a identificação mais clara de tendências e padrões. A média móvel pode ser calculada por:

$$MM = \frac{1}{n}\sum_{j=1}^{n} p_{i=j}$$

Onde $n$ é o número de termos na série histórica e $p$ é o termo em determinado momento. A fórmula é bastante simples se olharmos de outra forma, ela só diz que:

$$p_i = \frac{p_{i+1} + p_{i+2} + ...p_{i+n}}{n}$$

Se fizermos o cálculo da média móvel simples (SMA) da série de retornos da NVDA teremos que ela gera a representação abaixo. Uma vez que vamos tratar dados diários de retorno, a média móvel ocorrerá com uma janela de 30 dias. Podemos então ver que ocorre realmente uma suavização em relação aos picos da série.

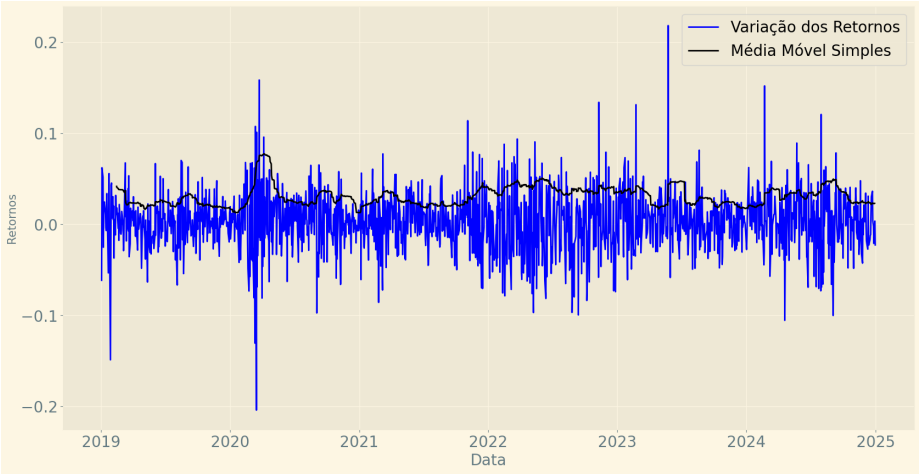


Figura 6: Média móvel dos retornos do NVDA

O valor da volatilidade da série passa então a ser 0.022520. Se fizermos um cálculo do VaR paramétrico com a nova estimativa para volatilidade da série teremos que a perda máxima com 95% de confiança será R$ -3704.13 ou -0,0370413 para cada investimento inicial. A representação do VaR ficará da seguinte forma:

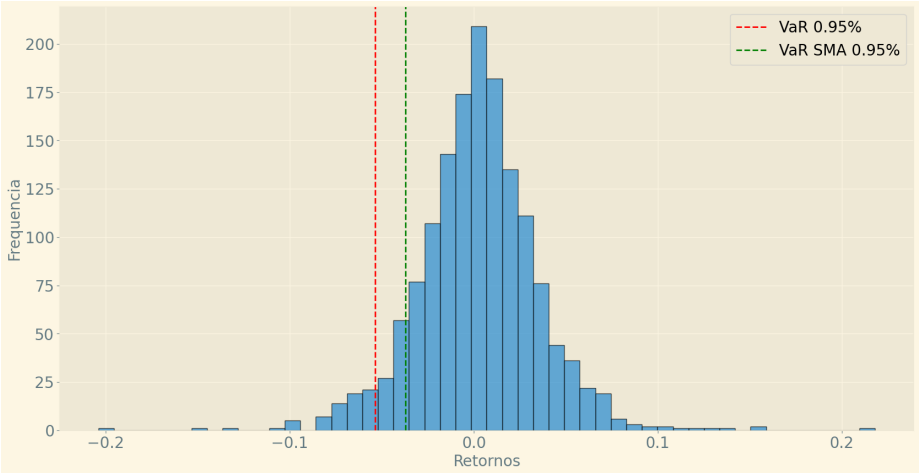


Figura 7: VaR Paratmétrico para $\sigma$ por Média Móvel (SMA)

Contudo, a média móvel realmente é uma boa maneira de estimar a volatilidade? Apesar de realizar a suavização das séries temporais, as médias móveis atribuem igual importância a dados ao longo de toda a série. Logo, dados de vários anos atrás terão o mesmo peso que dados de ontem ou segundos atrás.

Entretanto, o mercado muitas vezes atribui mais importância a dados recentes do que a dados mais antigos. Uma crise financeira em 2008 não tem o mesmo impacto hoje em dia (apesar de que perfeitamente ainda podemos ter algum) do que uma guerra dois anos atrás. Dessa forma, a média móvel pode ser uma medida não responsiva para mudanças estruturais no mercado e, portanto, uma medida inadequada.

# 5 Médias Móveis Ponderadas Exponenciais (EWMA)

Uma vez que as médias móveis simples apresentam a limitação de ponderarem igualmente os dados, temos que encontrar uma abordagem que supere esse problema. Uma forma de fazer isso é utilizando uma variação da média móvel chamada de ponderação exponencial ou em inglês *exponentially weighted moving average* (EWMA).

Em uma média ponderada, você não trata todos os dados de maneira igual. Alguns são considerados mais importantes do que outros para a análise e devido isso recebem um peso maior. No caso do EWMA os dados recentes recebem maior peso do que dados de períodos passados, uma vez que consideramos que dados recentes chamam mais atenção dos agentes do mercado.

No EWMA a volatilidade será dada pela raiz da média ponderada da volatilidade dos dias anteriores e dos retornos em determinado dia. A ponderação é feita por meio de um parametro de suavização da série temporal $\lambda$. O EWMA é dado por:

$$\sigma_e = \sqrt{\lambda\sigma_{n-1}^2 + (1-\lambda)R^2}$$

Todavia, o valor da volatilidade no período $n-1$ ainda não é conhecido. Você poderia solucionar isso escolhendo arbitrariamente o período como o anterior, $n-1$, contudo novamente existira o problema de não é dado a volatilidade no período $n-2$ e assim por diante. Todavia, voltando período a período dessa forma, por meio de um processo de iteração, podemos reescrever a expressão como:

$$\sigma_e = \sqrt{(1-\lambda)\sum_{i=1}^{\infty}\lambda^{i-1}R_{n-i+1}^2}$$

Essa reformulação é interessante, pois mostra a propriedade recursiva do EWMA, que permite expressar o valor atual como uma média ponderada de todas as observações anteriores, com pesos que diminuem exponencialmente à medida que os dados vão cada vez mais ao passado, exatamente como queremos. Uma vez que o parametro $0 \leq \lambda \leq 1$, se ele for elevado ao expoente $i-1$ o seu valor será cada vez menor tendendo a 0. Dessa forma, não é necessário regredir

ao infinito para fins práticos no EWMA financeiro e a análise pode ser realizada considerando que o primeiro valor *é* 0.

O parametro $\lambda$ é definido arbitrariamente pelo analista. É uma prática comum dos agentes do mercado calcular o EWMA seguindo a prática do relatório Riskmetrics do J.P Morgan Chase e considerar que $\lambda = 0.95$. Se fizemos isso para nossa série de retornos teremos uma nova estimativa de volatilidade de 0.031378. O EWMA realiza um processo de suavização da série temporal muito mais reativo do que a média móvel simples:

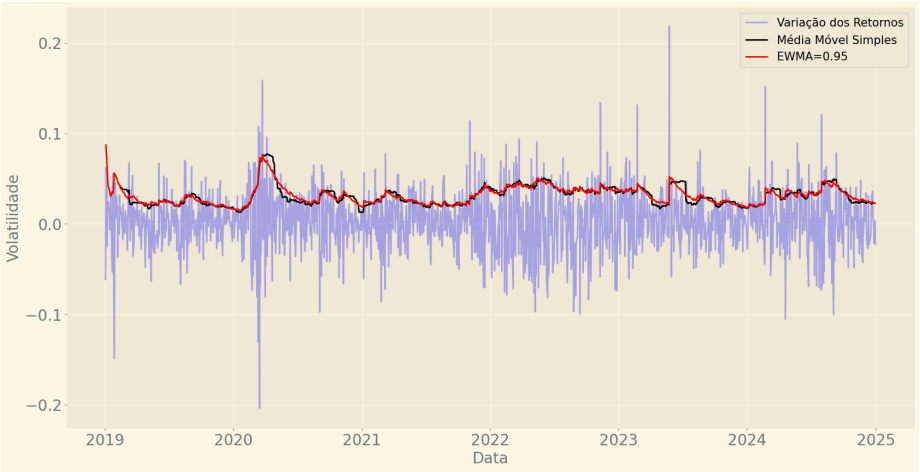


Figura 8: EWMA para $\lambda = 0.95$

Se substituirmos essa nova medida de volatilidade no nosso cálculo de VaR teremos um valor máximo de perda a um nível de confiança de 95% de R$ -5161.34 ou 0,0516134 para cada valor de capital investido inicialmente.

Esse valor é bem próximo da estimativa do VaR simples, mas existe diferenças quanto ao valor; mostrando como o VaR pode ser sensível a mínimas variações na estimativa de volatilidade. A representação na curva normal é dada por:

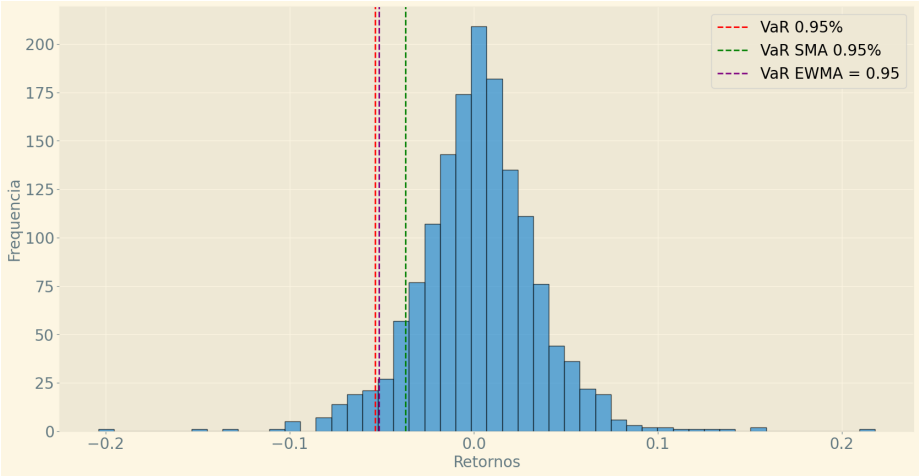


Figura 9: VaR com $\sigma$ dado por EWMA com $\lambda = 0.95$

A grande questão do EWMA é que os valores do parametro $\lambda$ são selecionados de maneira arbitrária. Valores de lambda mais altos vão gerar maiores suavizações, mais próximo da média móvel, e valores menores vão gerar estimações mais reativas a variações de mercado. Qual o valor correto de $\lambda$?

## 5.1 Otimização do $\lambda$ em um EWMA

Uma das dificuldades de se estimar a volatilidade de uma série temporal por médias móveis ponderadas é que o parâmetro é dado de forma arbitrária. É o analista que escolhe qual será o valor de $\lambda$, com valores ou altos demais ou baixos demais. Valores altos diminuem a sensibilidade do EWMA a mudanças recentes, tornando-o lento para reagir a alterações significativas nos dados e aproximando ele de uma média móvel simples. Valores baixos aumentam a sensibilidade do modelo a flutuações recentes, atribuindo maior peso a essas observações e diminuindo a capacidade do analista de ver tendências na série temporal.

Uma forma de ilustrar o caos que pode surgir dessa indeterminação do parâmetro é ver que diferentes valores de $\lambda$ vão produzir diferentes estimações de volatilidade. O simular valores de $\lambda = 0.95$, $\lambda = 0.92$ e $\lambda = 0.80$ temos as seguintes séries:

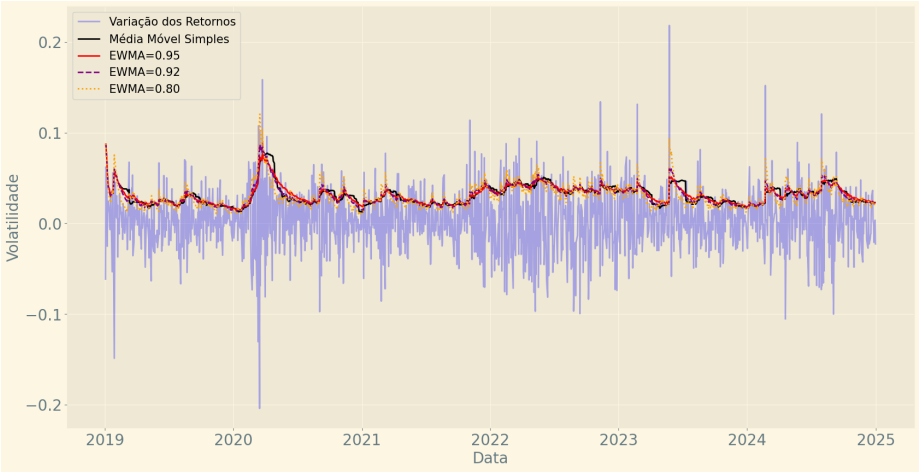


Figura 10: Simulações de EWMA para vários valores de $\lambda$

O valor mais baixo, $\lambda = 0.80$, produziu uma série extremamente reativa enquanto que o valor mais alto, $\lambda$, uma série bem mais suave e próxima da série da média móvel simples. Mas qual desses valores é o correto? Qual o analista deveria escolher? A escolha errada de um valor do parâmetro de suavização levará a uma estimação errada da volatilidade e do VaR, o que gerará perdas financeiras.

Uma forma de realizar isso é por meio de otimização. Esse processo consiste em ajustar variáveis, como nosso parâmetro, de forma a melhorar o máximo possível uma determinada métrica. No presente texto iremos seguir o processo de otimização indicado por (BOLLEN, 2015). A otimização consiste em avaliar o quanto a série estimada com o EWMA se aproxima de uma série tomada como *benchmark* para análise.

Idealmente, em uma utopia, o *benchmark* escolhido seria a volatilidade "real"do ativo. Todavia, a volatilidade "real"é uma variável não-observável e não temos como ter acesso direto a ela. A razão para isso é que ela é uma variável latente. Variáveis latentes são características ou fatores subjacentes que não podem ser observados ou medidos diretamente, mas que influenciam as variáveis observáveis.Uma forma de contornar essa limitação é desenvolvendo "aproximações"ou *proxys*. Em matemática, uma *proxy* é uma variável ou um objeto que serve como substituto de algo mais complexo ou cuja observação concreta não é possível.Para o caso da volatilidade uma boa aproximação de seu valor 'real' é a volatilidade realizada de cada período de operação. A volatilidade realizada é feita por meio da raiz da soma dos quadrados dos retornos diários do ativo:

$$\sigma_r = \sqrt{\sum_{i=1}^{n} U^2}$$

Mas por qual razão elevar os retornos ao quadrado daria uma ideia da volatilidade? A soma dos quadrados dos retornos, sem considerar a média, mostra a magnitude total das variações dos retornos. Esse é um dos componentes do cálculo da variância, que é o quadrado do desvio-padrão, e também do próprio desvio-padrão. Assim o quadrado dos retornos da série analisada daria uma aproximação, ainda que deficiente, da volatilidade 'real' da mesma.

A otimização proposta busca verificar quanto a volatilidade estimada difere da volatilidade realizada para diferentes parâmetros. Para quantificar essa diferença entre o modelo EWMA e a 'realidade' utilizaremos duas métricas estatística chamada raiz do erro quadrático médio (*RMSE* em inglês) e o erro médio absoluto (*MAE* em inglês).

O RMSE mede a magnitude da diferença entre os valores previstos pelo modelo e o benchmark que tomamos pela 'realidade'. Ele faz isso por meio de uma média da soma do quadrado dos erros de tal forma que pode ser dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\sigma_r - \sigma_e)^2}$$

A lógica do RMSE é que a exponenciação quadrática dos erros garantirá que todas as diferenças entre os valores estimados e do *benchmark* sejam positivas e que os valores extremos sejam mais penalizados. A raiz quadrada garante que a média desses valores seja fornecida na mesma unidade que a dos dados originais. Todavia, a exponenciação também faz com que a métrica do RMSE seja sensível a outliers, tipo de dados que são comuns em distribuições financeiras. Por essa razão que ele deve ser avaliado com outras métricas como o MAE.

O MAE é muito semelhante ao RMSE, porém ao invés da exponenciação temos os valores absolutos dos erros. Uma vez que não existe o quadrado no cálculo, a raiz se faz desnecessária para retornar à unidade original. A lógica do MAE é que o valor absoluto de cada erro contribui proporcionalmente para o valor final da métrica. Isso o torna menos sensível a outliers, porém ao mesmo tempo o faz indiferente a erros grandes. Ele pode ser dado pela fórmula:

$$\frac{1}{n}\sum_{i=1}^{n}|\sigma_r - \sigma_e|$$

Devido suas vantagens e desvantagens, o recomendado é que o analista utilize *ambas* as métricas na hora de avaliar um modelo.

Contudo, um problema surge ao tentar aplicar essas métricas a dados financeiros. A razão é que esses dados possuem a característica de apresentarem *heterocedasticidade*.

Esse nome que horrível que parece uma doença na realidade é somente uma palavra muito grande que os estatísticos criaram para se mostrarem. A palavra vem da união de dois termos gregos: 'hetero' que significa diferente e 'skedasis' que significa dispersão. Ela apenas quer dizer que a variância dos retornos (o quadrado do desvio-padrão) não é *constante*.

Quando construímos um modelo baseado no pressuposto de normalidade, uma das suposições mais importantes é que os erros tenham variância constante ao longo de todas as observações. Essa suposição é conhecida como *homocedasticidade*. quando isso não ocorre, quando os erros variam a depender do valor das variáveis ao longo da série de tempo, dizemos que existe heterocedasticidade.

Essa variabilidade desigual nos erros pode ter consequências importantes na análise estatística. Especialmente, a heterocedasticidade afeta a precisão das estimações (apesar de não necessariamente tornar um determinado modelo viesado) e a qualidade dos testes estatísticos, como p-valor.

Na prática, a heterocedasticidade significa que, ao longo do conjunto de dados, os erros do modelo não se dispersam de forma uniforme. Isso pode ser visualizado de forma gráfica. Se a dispersão dos erros pelos valores previstos não forem distribuídos aleatoriamente em torno de zero, mas sim apresentarem uma forma cônica, dizemos que os dados apresentam heterocedasticidade. Dados financeiros apresentam bastante essa característica. Abaixo vemos a dispersão de erros da NVDA e identificamos a forma cônica característica da heterocedasticidade.

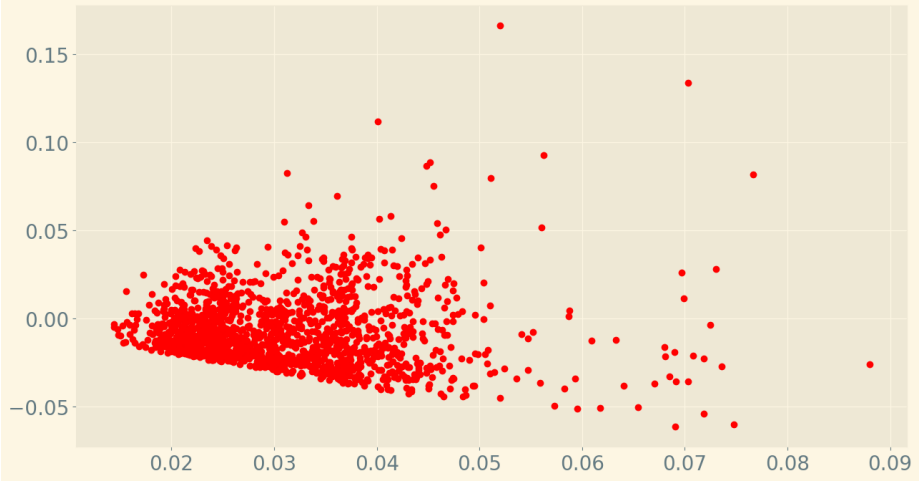


Figura 11: Dispersão dos Erros pelos Valores Previstos da NVDA

A heterocedasticidade é particularmente sensível para o que foi exposto aqui, pois quando os erros apresentam variância não constante, tanto o RMSE quanto

o MAE podem ser afetados de maneira desigual. Segundo (ANDERSEN; BOLLERSLEV; LANGE, 1999), o RMSE e o MAE tendem a aumentar na presença de dispersão não-aleatória (homocedástica), refletindo uma piora na precisão geral das previsões do modelo. Para controlar a heterocedasticidade dos erros, os autores recomendar modificar tanto o RMSE e o MAE para controlar para essa característica por meio das estatísticas ajustadas para heterocedasticidade:

$$HRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(1 - \frac{\sigma_r}{\sigma_e})^2}$$

$$HMAE = \frac{1}{n}\sum_{i=1}^{n}|1 - \frac{\sigma_r}{\sigma_e}|$$

Onde $\sigma_r$ é o valor 'real' da volatilidade da série temporal, $\sigma_e$ é o valor estimado, $\frac{\sigma_r}{\sigma_e}$ é o erro relativo e $1 - \frac{\sigma_r}{\sigma_e}$ é o mesmo erro relativo em termos reais, variando em torno de 1 (com 1 sendo o grau de acerto da previsão perfeita).

A ideia do processo de otimização é realizar um processo de minimização onde seja procurado o parámetro, no caso $\lambda$, que forneça o menor HRMSE e o menor HMAE possível. Vamos considerar que ambas as métricas tem igual importância em nossa análise. Assim, a nossa função objetivo pode ser expressa por:

$$f(\lambda) = 0.5 \cdot HRMSE + 0.5 \cdot HMAE$$

$$0 \leq \lambda \leq 1$$

Se realizarmos essa otimização por L-BFGS [5]para os valores calculados de $\sigma_e$ e $\sigma_r$ teremos que o $\lambda$ otimizado heterocedástico será de 0.96. Esse valor parece muito próximo do valor padrão utilizado pelo mercado de $\lambda = 0.95$ e a diferença insignificante. Porém, será isso mesmo?.

O valor da volatilidade estimada com o EWMA otimizado, $\sigma_{e_o}$, é de 0.0314945. Abaixo temos um gráfico da série estimada pelo modelo e as variações:

[5]A otimização L-BFGS é uma variação do método BFGS, uma solução hessiana para problemas de otimização. Essa é a abordagem padrão utilizado tanto pelo *solver* do Excel como pelo Scikit-Learn.

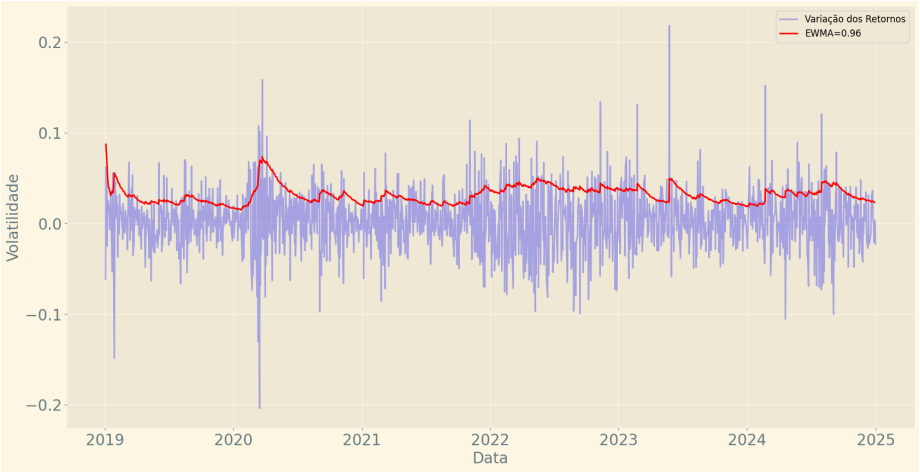


Figura 12: EWMA para $\lambda = 0.96$

O VaR calculado para a otimização fica R$ -5180.39. Isso significa que existe uma diferença de R$ 19,06 entre a estimação padrão e a estimação otimizada. A diferença é insignificante e os VaR são próximos.

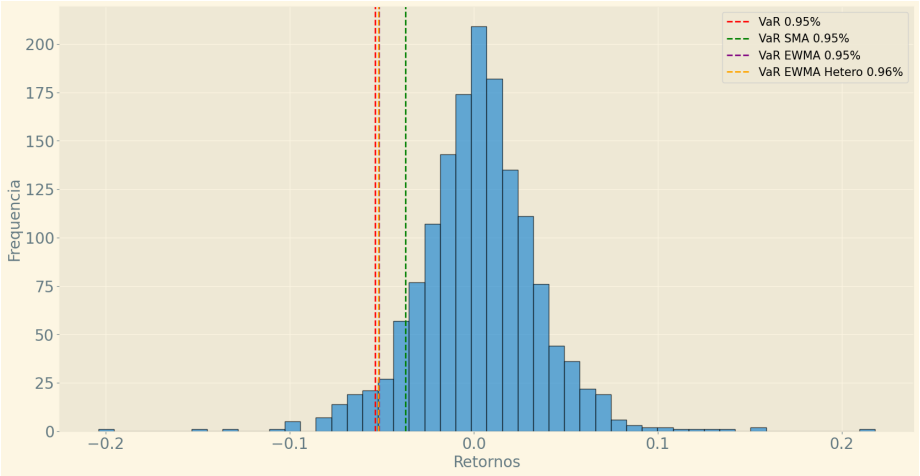


Figura 13: VaR para o EWMA otimizado

Podemos concluir a partir disso que a prática de mercado está correta? Não. Perceba que o valor de $\lambda$ é dado por uma otimização com base nos dados históricos da série de retornos. Se utilizarmos outra série temporal ou outro ativo, pode ser que o valor otimizado de $\lambda$ mude e fique divergente do padrão de 0.95. Além disso, observe que a função objetivo também determinará o valor da otimização. Dependendo de como a função objetivo é especificada, os

pesos atribuídos a cada métrica, o valor será diferente. Deveria o HRMSE ter o mesmo peso que o HMAE ou seria o HRMSE mais importante? Ou seria o HMAE? Essas questões servem para mostrar que questões métricas não são meramente um problema de cálculo, mas um processo crítico que exige que o analista conheça bem a lógica matemática por trás.

Assim, o recomendado é que, ao rodar um modelo EWMA, seja utilizado um algorítimo que ajuste para mudanças estruturais na série temporal.

# 6 Artes Arcanas 101: Os Modelos GARCH

No exemplo anterior aprendemos basicamente a realizar um controle da heterocedasticidade presente nos dados financeiros. O método de otimização consistiu basicamente em determinar a melhor estimação da volatilidade em razão da presença de variação não-aleatória dos erros.

No geral, em modelagem estatística, o desejável é que a heterocedasticidade seja controlada. Em modelos de regressão, por exemplo, por mais que a heterocedasticidade não gere resultados viesados, ela gera intervalos de confiança extremamente curtos devido uma 'compressão' do desvio-padrão. Assim, os resultados do modelo estimado com a regressão possuiriam baixa precisão e pouco valor analítico.

Contudo, muitas vezes a heterocedasticidade é uma característica fundamental dos dados e não pode ser facilmente controlada. Nos dados financeiros, alguns períodos são tipicamente mais arriscados que outros, e esse risco não ocorre de forma aleatória ao longo do tempo. Os momentos de alta volatilidade tendem a ocorrer próximos uns dos outros — um fenômeno que os quants chamam de *volatility cluster*. Nos dados de cotação da NVIDIA é possível observar a formação de ao menos dois desses clusters durante a Pandemia de 2020 e no Boom da IA começando em 2022.

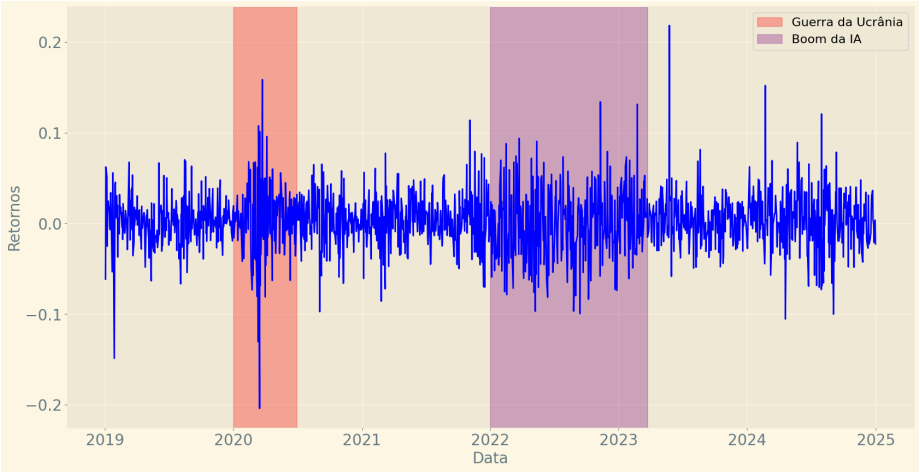


Figura 14: Clusters de Volatilidade na Série da NVDA

O primeiro a tentar lidar com esse problema foi (ENGLE, 1982). Em seu estudo das séries de tempo da inflação inglesa, Engle propôs que, em vez de assumir uma variância constante, a variância ao longo do tempo é uma função dos retornos quadrados passados, permitindo que a volatilidade seja modelada como uma média ponderada dos mesmos retornos. Ao fazer isso, a volatilidade estimada não se torna constante ao longo do tempo (homocedástica), mas será influenciada pela volatilidade passada. O nome dessa técnica é modelagem ARCH (*Autoregressive Conditional Heteroscedastic*). Esse modelo é uma média móvel ponderada onde os pesos são parámetros a serem estimados a partir dos dados.

Apesar de serem boas abordagens, os modelos ARCH geralmente não se mostraram bons o suficiente do ponto de vista empírico. Uma das razões para isso é que os modelos incorporavam os retornos passados, mas não a volatilidade passada, o que os tornava não muito eficientes em capturar dependência temporal. Por exemplo, se a volatilidade foi alta ontem, é provável que seja alta hoje também. Essa incapacidade de capturar a dependência temporal da volatilidade tornava os modelos ARCH simples inadequados.

Para solucionar esse problema, (BOLLERSLEV, 1986) desenvolveu o modelo GARCH (*Generalised Autoregressive Conditional Heteroscedastic*). Mas o que exatamente é um modelo GARCH?

Um modelo ARCH de ordem $n$ assume que a volatilidade no tempo, $\sigma_t$, depende linearmente dos quadrados dos retornos passados. Podemos expressar isso como:

$$\sigma_t = \sqrt{\omega + \alpha U_{t-i}^2}$$

Onde $\omega$ é uma constante e $\alpha$ é um parâmetro de sensibilidade para a mudança nos retornos. Para entender essa último parâmetro, basta entender que os agentes de mercado reagem à informação de que ocorreu mudança nos retornos de determinado ativo, de forma que a volatilidade não depende meramente de como esses retornos variam, mas de como esses sentimentos afetam os retornos.

Como dito anteriormente, esse modelo ARCH possui a deficiência de não levar em conta que a volatilidade presente depende da volatilidade passada. Para superar essa limitação, o modelo GARCH de ordem $p,q$ estima a volatilidade tanto com base nos retornos passados como também a volatilidade passada. Podemos expressão o modelo como:

$$\sigma_t = \sqrt{\omega + \alpha U_{t-i}^2 + \beta \sigma_{t-i}^2}$$

Onde $\beta$ é um parâmetro que mede a sensibilidade da volatilidade presente em relação a volatilidade passada. Novamente, a lógica é a mesma da sensibilidade aos retornos. Como colocado por (ENGLE, 2001), o modelo GARCH afirma que o melhor preditor da variância no próximo período é uma média ponderada da volatilidade de longo prazo, da volatilidade prevista para este período e da nova informação deste período, que é o retorno quadrado mais recente. Isso faz com que o modelo GARCH seja, em termos práticos, um algoritmo de atualização bayesiano.

## 6.1 Grimório Econométrico: Estimação Paramétrica e Máxima Verossimilhança

O modelo GARCH mais simples que pode existir é o GARCH(1,1). Os '1' dentro do parêntese se referem à quantidade de parâmetros do modelo. Geralmente, os modelos tem dois tipos de parâmetros: autoregressivos e médias móveis. Os econometristas irão se arvorar senhores da sabedoria universal dizendo que o termo autoregressivo se refere a parte GARCH do modelo e as médias móveis aos termos ARCH. Não perguntem a eles o que isso significa, eles não vão saber responder. Em termos práticos isso significa apenas que existirá um conjunto de parâmetros que darão conta da volatilidade passada (parte autoregressiva) e um para os retornos passados (parte da média móvel).

O modelo GARCH mais simples tem apenas um lag, o que significa que $i = 1$ e ele prevê apenas um único período para frente. A grande dificuldade com esse modelo é estimar os parâmetros $\alpha$ e $\beta$. Esses dois parâmetros também são chamados de 'reação'($\alpha$), uma vez que vai determinar o quanto a volatilidade passada vai afetar a volatilidade presente, e de 'persistência'($\beta$)

Algo que deve ser observado, todavia, é que se definirmos que os valores são tais que:

$$\alpha + \beta = 1$$

Teremos então na verdade um modelo EWMA uma vez que $\beta = 1 - \alpha$. Isso é interessante, pois mostra que o modelo EWMA na realidade *é um tipo* de modelo GARCH. O que vai diferenciar os dois é que em um modelo GARCH os parâmetros sempre serão tais que $\alpha + \beta < 1$.

Os parâmetros serão estimados usualmente utilizando o método de Máxima Merossimilhança. Essa é uma técnica bastante refinada (e antiga) da estatística para otimização de parâmetros e apresenta um certo grau de dificuldade. Será exposto aqui uma versão simplificada.

A Máxima Verossimilhança é uma técnica estatística que busca determinar os parâmetros de um modelo probabilístico que tornam os dados observados mais prováveis. Seja $N = (N_1, N_2...N_n)$ uma amostra de $n$ observações independentes extraídas de uma distribuição com função densidade definida por $f(N|\theta)$, onde $\theta$ representa os parâmetros desconhecidos do modelo. Vamos dizer que a função de verossimilhança seja $L(\theta)$. Essa função diz basicamente: Qual a probabilidade de se observar os dados, N, dado os parâmetros, $\theta$.

Considerando que a distribuição seja normal, a probabilidade será dada pela distribuição normal que melhor se aproxima da distribuição dos dados observados. Podemos dizer que:

$$L(\theta) = \prod_{i=1}^{n} f(N|\theta)$$

No caso do modelo GARCH teremos uma função tal que:

$$L(\alpha, \beta, \omega) = \prod_{i=1}^{n} f(N|\alpha, \beta, \omega)$$

Então definimos uma regra de otimização para os melhores parâmetros tal que:

$$\theta_{max} = \max L(\alpha, \beta, \omega)$$

O que o método fará, e mais especificamente esse estranho símbolo $\prod$, será pegar várias distribuições normais e ir testando cada uma delas nos dados da amostra até encontrar a distribuição que melhor se encaixe, gerando assim os parâmetros mais prováveis de chegar perto dos dados reais.

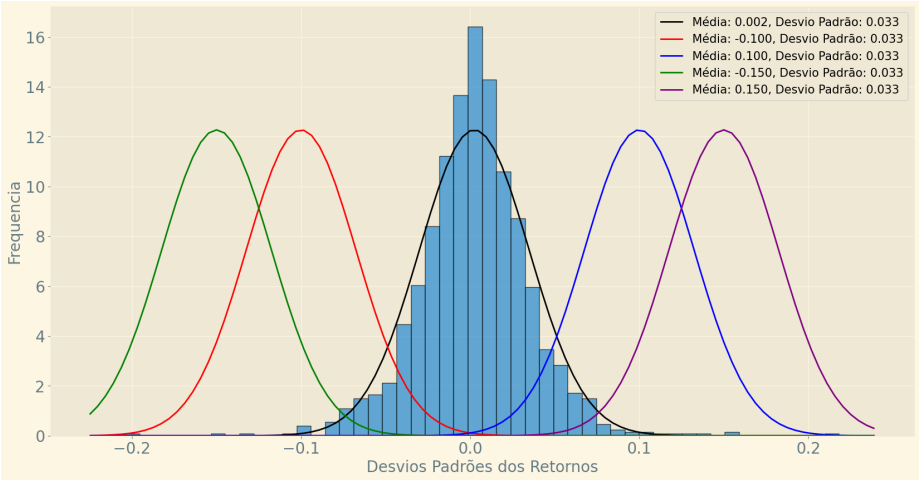


Figura 15: Ilustração das Múltiplas Curvas Normais geradas pela Máxima Verossimilhança

Se fizermos a estimação por verossimilhança com base em uma amostra dos dados da NVDA, teremos que:

- $\omega = 0.000106$
- $\alpha = 0.100001$
- $\beta = 0.800000$

Computando isso para um modelo GARCH dos dados da NVDA será gerado a seguinte estimativa da volatilidade do período:

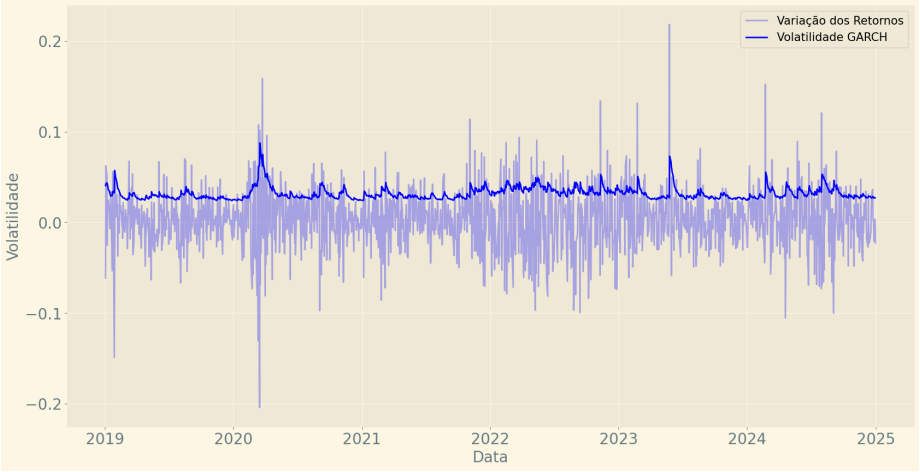


Figura 16: Volatilidade GARCH

A volatilidade do período estimada por esse modelo é de 0.03. Considerando apenas o último período temos uma volatilidade de 0.02. Se computarmos a primeira no nosso VaR paramétrico teremos uma perda máxima de -R$ 4934.56. Assim, a estimativa do modelo GARCH é em média semelhante a dos outros modelos.

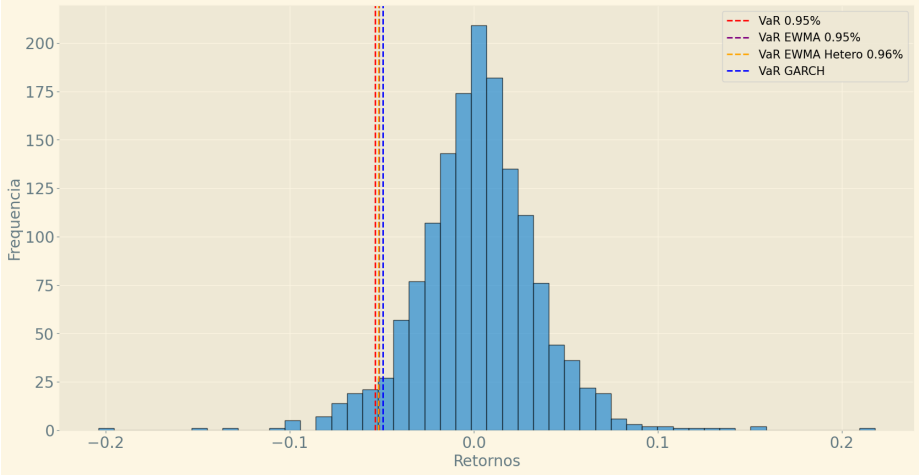


Figura 17: VaR estimado com volatilidade GARCH

## 6.2 Ilusões do Comportamento: O Efeito Alavancagem

Apesar de ser um modelo bastante elegante, o GARCH tem uma pequena limitação quando se trata de lidar com um comportamento específico dos mercados financeiros.

Considere a seguinte pergunta: *Toda volatilidade é igual?*

Se o IBOVESPA cair amanhã -5% você se sentiria igual a ele ter uma alta de +5%? A resposta provavelmente é não. Notícias ruins tendem a nos impactar mais do que notícias boas e tendemos a ser contaminados pelo pessimismo por mais tempo do que pelo otimismo.

Estudando a reação dos mercados a determinadas notícias, (ENGLE; NG, 1993) encontrou que após uma forte queda nas cotações do mercado, a volatilidade tendia a ficar mais alta e por mais tempo do que quando ocorria uma forte valorização. Os quants posteriormente chamariam esse fenômeno de volatilidade alavancada.

O grande problema com a alavancagem é que o modelo GARCH tradicional não conseghue capturar esse efeito. Para que o modelo consiga é necessário realizar uma modificação no mesmo e introduzir um parâmetro que dê conta

desse efeito. Se modificarmos o modelo GARCH original para introduzir esse parâmetro teremos:

$$\sigma_t = \sqrt{\omega + \alpha(U_{t-i}^2 - \delta) + \beta\sigma_{t-i}^2}$$

Onde $\delta$ é um coeficiente que captura o efeito de alavancagem, atribuindo maior peso a choques negativos. A estimação do modelo na presença desse parâmetro é algo *realmente* complicado e foge bastante do propósito do presente artigo. É recomendado fortemente ao leitor que leita integralmente o artigo de Engle e Ng para a demonstração.

Esse modelo, conhecido como GJR-GARCH, quando aplicado aos dados da NVDA gera uma estimação de volatilidade para o período analisado de 0.031.

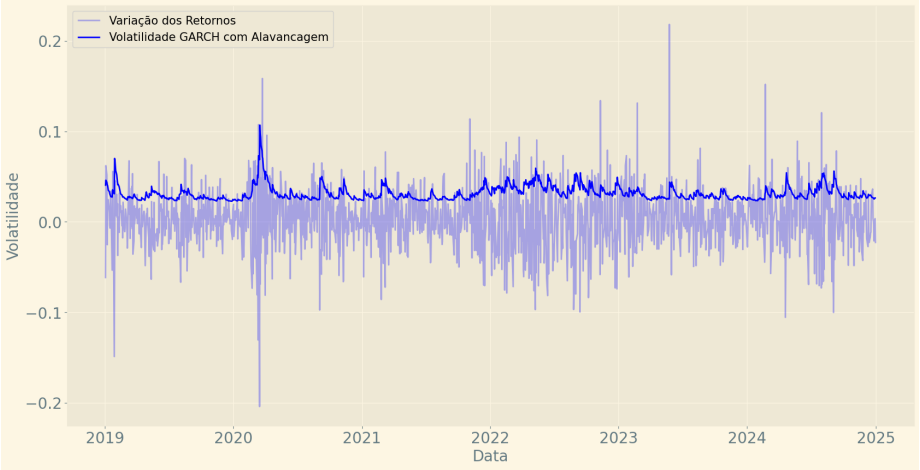


Figura 18: Volatilidade estimada com GJR-GARCH

Computando essa volatilidade para nosso VaR teremos uma perda máxima de -R$ 5099.04. Observe que esse é um valor bem próximo do estimado pelo modelo GARCH simples. Apesar de muito próximo, o modelo GJR-GARCH parece levamente mais ajustado aos dados dos retornos.

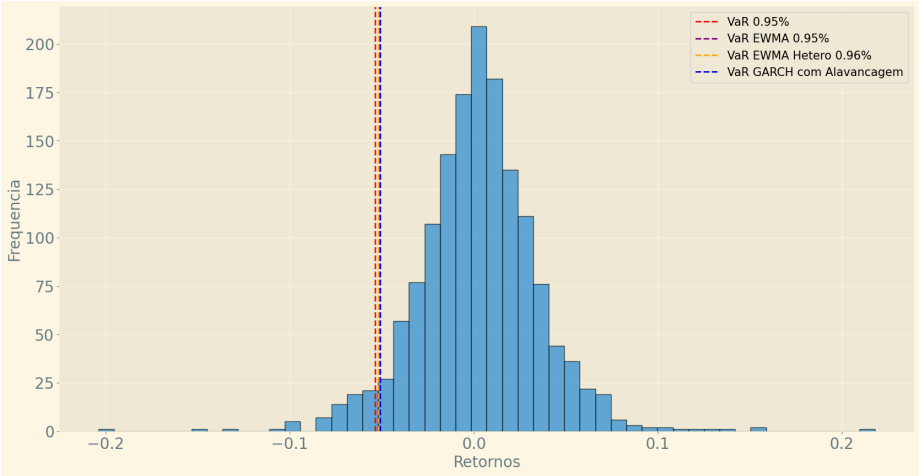


Figura 19: VaR calculado com volatilidade GJR-GARCH

# 7 Introdução às Artes das Trevas: Modelo Black-Scholes-Merton

Os assuntos tratados até aqui são mais do que suficientes. O leitor que tenha chegado a compreender minimamente um modelo GARCH já pode se considerar um mago financeiro treinado e pronto para a guerra. Porém, alguns podem querer estudar artes mais...obscuras. Os assuntos tratados a partir desse capítulo constituem estudos nas artes das trevas do estudo de volatilidade e exigirão uma mente mais refinida por parte do aluno.

Existem outras formas de se estimar e prever a volatilidade. Essas técnicas envolvem o estudo de uma área da matemática chamada de cálculo estocástico. Esseé um nome *fancy* e tenebroso para a área do cálculo que analisa processos aleatórios. Ao longo dos próximos capítulos será fornececido ao leitor uma compreensão simplificada dos modelos estocásticos.

Uma dessas formas alternativas de se estimar a volatilidade de um ativo é analisando um instrumento financeiro derivativo chamado de *opções*. De maneira simples, opções são instrumentos que garantem ao comprador o direito, *mas não a obrigação*, de comprar ou vender um ativo — geralmente ações — por um preço estabelecido (chamado de *"strike price"*) em uma data futura específica.

Para adquirir esse direito, o comprador paga um valor denominado "premium"ao vendedor da opção, também conhecido como lançador. Em contrapartida, o lançador assume a obrigação de vender ou comprar o ativo, caso o

comprador decida exercer a opção na data contratada. Se você tiver um direito de compra você tem uma *call option* e se tiver um direito de venda tem uma *put option*.

A lógica de se estimar a volatilidade pelo preço das opções é bastante simples. Uma vez que são instrumentos que garantem direito a vender ou comprar um ativo no futuro, as opções carregam na formação de seu preço as expectativas dos agentes econômicos sobre como o preço de um determinado ativo irá variar até essa data futura. Assim, um dos componentes na formação do preço de uma opção é a *volatilidade implícita* daquele ativo. É essa volatilidade implícita que será usada para calcular o VaR.

Porém, para acessar essa volatilidade implícita é necessário termos um modelo de formação do preço de uma opção. O modelo mais conhecido para esse fim é o infame Modelo Black-Scholes-Merton. Poucos são os modelos matemáticos tão criticados (e pouquíssimo compreendidos) quanto esse modelo financeiro. Sabendo disso, eu irei demonstrar esse modelo passo-a-passo. Compreender a intuição dele é necessária para a estimação da volatilidade implícita (e para pararmos com críticas tolas a ele).

Apesar de ser mais conhecido pelos trabalhos de Fischer Black e Myron Scholes, o modelo de precificação de opções fornecido por esses autores é em realidade não muito bom. Black e Scholes estimaram uma fórmula de precificação de opções com base no CAPM e essa solução não é muito elegante. O que será fornecido aqui (e o que é mais conhecido) será a fórmula de precificação de opções baseada em cálculo estocástico fornecida por (MERTON, 1973).

Vamos considerar que o valor de uma opção seja dado por uma função $V(S,t)$, onde $V$ é o valor da opção, $S$ é o preço do ativo base da opção (ações da NVIDIA, por exemplo) por exemplo e $t$ é o tempo corrente. O valor esperado dessa opção é determinado pelo lucro que se terá com o preço futuro do ativo *dado* seu preço corrente. Podemos expressar isso matematicamente como:

$$V(S,t) = E[K(S_T)|S_t = S]$$

Considerando agora que temos um portfólio de opções homogêneas, $\prod$, esse mesmo portfólio terá seu valor determinado por quanto você pagou pela opção menos a variação do preço do ativo base, pois afinal a opção lhe dará o direito de comprar ou vender o mesmo no final das contas. Assim:

$$\prod = V_t - \Delta S_t$$

Com $V_t$ sendo equivalente ao valor dado pela função preço da opção definida anteriormente. Dessa forma, podemos redefinir o valor do portfólio de opções como:

$$\prod = V(S_t,t) - \Delta S_t$$

Agora vamos dizer que o valor do ativo, e por consequência do portfólio, varie um pouco ao longo do tempo, com essa pequena variação sendo $dt$. O pouco definido aqui é pouco mesmo. Poderíamos falar de milissegundos e brincar com as artes da cronomancia, mas vamos focar na variação diária. Essa pequena variação pode ser definida por uma derivada de tal modo que:

$$d\prod = \prod_{t+dt} - \prod_{t}$$

Reescrevendo para a definição de $\prod$ feita anteriormente:

$$d\prod = (V(S_{t+dt}, t+dt) - V(S_t, t)) - \Delta(S_{t+dt} - S_t)$$

Todavia, a diferença no preço da opção dado pela função V em parênteses pode ser, em certa medida, aproximado por uma derivada parcial de V em relação ao tempo $t$. Podemos definir essa diferença como:

$$d\prod = \frac{\partial V}{\partial t}(S_t, t) \cdot dt + (\frac{\partial V}{\partial S}(S_t, t) - \Delta) \cdot dS + \frac{1}{2}\frac{\partial V^2}{\partial S^2}(S_t, t) \cdot dS^2$$

Aqueles com mais familiaridade com cálculo poderão dizer: *Mas isso está errado! Existe os termos de ordem secundária* $\frac{1}{2}\frac{\partial V^2}{\partial t^2}(S_t, t) \cdot dt^2$*... em uma derivada parcial que foram ignorados!*. Realmente, mas a razão para isso é a mesma lógica da série de Taylor tratada anteriormente. As variações de $dt$ são tão pequenas que a potência de seu número é basicamente insignificante e podemos considerar que, no limite, esses termos de segunda ordem são 0. Então o valor das variações de V pode ser dado aproximadamente pela derivada parcial.

Em seguida é necessário realizar alguns pressupostos sobre o comportamento do preço do ativo base. Opções são, no final das contas, instrumentos que dependem essencialmente de como o preço desses ativos irá variar. Mas como esses preços variam. Em seu paper original, (BLACK; SCHOLES, 1973) assumiram que o preço de um ativo de capital base para a opção seria determinado pelo CAPM. A genialidade de Merton foi observar que o preço dos ativos na realidade seguia um padrão mais próximo de um movimento browniano.

O movimento browniano é simplesmente um nome para o movimento aleatório de partículas. Um dificuldade de se utilizar diretamente o movimento browniano, tal qual ele é aplicado na física atômica, em finanças é que sua forma tradicional pressupõe que os valores possam ser negativos. Todavia, preços de ativos não podem ser negativos; seu valor mínimo é 0. Para solucionar isso, Merton utilizou uma versão modificada chamada de movimento browniano geométrico. O geométrico no nome é apenas para dizer que as variações são o produto de um número de euler, $e$, elevado a potência do movimento browniano.

Observando que os valores do preço do ativo base, S, estão mudando como um movimento browniano, podemos dizer que:

$$dS = \mu S_t \cdot dt + \sigma S_t dZ_t$$

Onde $\mu$ é a média dos valores de S e $dZ_t$ é uma variável que controla para a aleatoriedade do movimento browniano. A intuição dessa fórmula é que as variações extremamente pequenas nos valores de S serão determinadas pela média do preço de S em função das mudança mínimas no tempo, $\mu S_t \cdot dt$, mais um efeito aleatório dado pela volatilidade do preço, $\sigma S_t$, e a variável do movimento browniano, $dZ_t$. Um ponto importante é que a volatilidade do ativo é assumida como constante, não ocorrendo uma variação dela. De certa forma, é como se o ativo tivesse uma volatilidade intrínseca.

Agora vem a verdadeira mágica. Se for colocado o termo $dS_t$ na equação de $d\prod$, será necessário elevar ele ao quadrado em $dS_t^2$. Se fizermos isso teremos que:

$$(\mu S_t dt)^2 + 2(\mu S_t dt) \cdot (\sigma S_t dZ_t) + (\sigma S_t dZ_t)^2$$

Todavia, uma vez que $dt$ é uma variação de tempo tão pequena, podemos considerar qualquer termo composto por ele como sendo aproximadamente zero. Assim, $(\mu S_t dt)^2$ e $2(\mu S_t dt) \cdot (\sigma S_t dZ_t)$ podem ser considerados como equivalentes a zero. Assim, $dS_t^2$ pode ser simplificado em:

$$dS_t^2 = \sigma^2 S_t^2 dZ_t^2$$

Porém, graça a uma identidade matemática conhecida como Lema de Ito [6], é possível dizer que o quadrado de um movimento browniano, $dZ_t^2$ é equivalente a uma derivada do próprio tempo, de forma que $dZ_t = dt$. Assim, podemos reescrever que:

$$dS_t^2 = \sigma^2 S_t^2 dt$$

É possível então reescrever a equação do portfólio de opções para:

$$d\prod = \frac{\partial V}{\partial t}(S_t, t) \cdot dt + (\frac{\partial V}{\partial S}(S_t, t) - \Delta) \cdot dS + \frac{1}{2}\frac{\partial V^2}{\partial S^2}(S_t, t) \cdot \sigma^2 S_t^2 dt$$

Considerando que $\Delta$ pode ser tomado apenas como uma derivada parcial de V em relação a S tal que $\Delta = \frac{\partial V}{\partial S}(S_t, t)$, o termo $(\frac{\partial V}{\partial S}(S_t, t) - \Delta) = 0$. A equação então é reescrita como:

$$d\prod = \frac{\partial V}{\partial t}(S_t, t) \cdot dt + \frac{1}{2}\frac{\partial V^2}{\partial S^2}(S_t, t) \cdot \sigma^2 S_t^2 dt$$

Ou, isolando o termo da variação de tempo:

$$(\frac{\partial V}{\partial t}(S_t, t) + \frac{1}{2}\frac{\partial V^2}{\partial S^2}(S_t, t) \cdot \sigma^2 S_t^2) dt$$

Por fim, o modelo assume que os agentes são aversos ao risco e buscam neutralizar seus portfólios. Isso significa em termos práticos que ao longo de $dt$

[6] O Lema de Ito é um conceito matemático bastante avançado, o que daria quase um livro inteiro só para explicar ele. É recomendado ao leitor ler um livro de cálculo estocástico para entender.

eles buscam portfólios que sejam equivalentes ao ativo livre de risco da economia; usualmente títulos públicos do governo. Assim, o portfólio de opção deve ser igual ao portfólio de ativos cujos rendimentos são dados pela taxa livre de risco, $r_t$. A neutralização é dado pela equivalência:

$$(\frac{\partial V}{\partial t}(S_t, t) + \frac{1}{2}\frac{\partial V^2}{\partial S^2}(S_t, t) \cdot \sigma^2 S_t^2)dt = r_t \cdot \prod_t \cdot dt$$

Substituindo $\prod$ pelo portfólio $\prod_t$ dado pela diferença entre o preço do portfólio hoje e a estimativa de precificação, anulando o fator temporal $dt$:

$$\frac{\partial V}{\partial t}(S_t, t) + \frac{1}{2}\frac{\partial V^2}{\partial S^2}(S_t, t) \cdot \sigma^2 S_t^2 = r_t \cdot (V(S, t) - S_t \frac{\partial V}{\partial S}(S, t))$$

Esse é o modelo de Black-Scholes-Merton. A partir dele é possível pegar o preço de uma opção e isolar sua variável $\sigma^2$ para descobrir a volatilidade implícita.

## 7.1 O *Volatility Smile* com Black-Scholes-Merton

Um fato interessante sobre a volatilidade implícita das opções é que um mesmo tipo de opção pode ter diferentes datas de vencimento e cada uma com *strike prices* diferentes. Esse fato viola um dos pressupostos do modelo Black-Scholes-Merton que é o de que a volatilidade é constante ao longo da série temporal. Em realidade, ela apresenta variações na volatilidade a depender do strike da opção, refletindo o fato de que opções mais próximas do vencimento apresentam *strike prices* maiores.

Se computarmos a volatilidade de cada um dos contratos de opções vamos encontrar, caso seja observado a forma da série formada em gráfico, que a volatilidade implícita forma uma curva côncava em forma de sorriso. Os quants chamam isso de *volatility smile* (o sorriso da volatilidade).

Sim, a volatilidade sorri para o mercado. Você pode achar esse fato estranho, mas ele é apenas o Cheshire do mundo das finanças. Basta lembrar da passagem:

> - Eu não sabia que gatos de Cheshire sempre sorriam. Na verdade, eu não sabia que gatos podiam sorrir.
>
> - Todos eles podem - disse a Duquesa - E a maioria deles sorri.
>
> - Não conheço nenhum que sorria - Alice disse com muita educação, sentindo-se bastante contente por ter entrado em uma conversa.
>
> - Você não sabe muito - disse a Duquesa - e isso é um fato.

A análise do volatility smile é importante para verificar qual a volatilidade de opções cujos *strike prices* estão abaixo ou acima do preço do ativo base, conforme indicado por (DERMAN; MILLER, 2016). O ponto mínimo da curva

côncava indica as opções *in-the-money*, ou seja cujo *strike price* é equivalente ao preço do ativo base.

Abaixo temos o *volatility smile* calculado para a NVIDIA utilizando o modelo Black-Scholes-Merton. Uma das razões que escolhi trabalhar com uma ação americana e não brasileira é justamente devido o mercado de opções no Brasil ainda não ser bem desenvolvido. Não teria dados suficientes para realizar esse exercício.

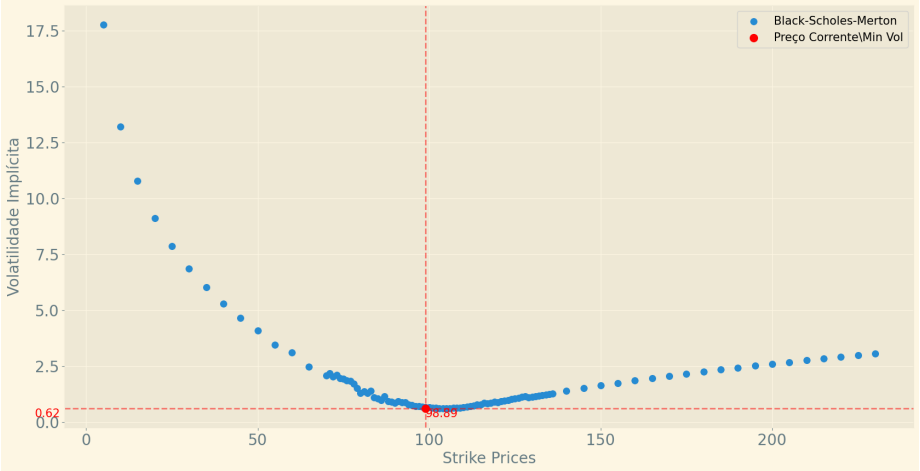


Figura 20: Volatility Smile do NVDA com Black-Scholes-Merton

O interessante é que é possível comparar os cálculos da volatilidade implícita do modelo Black-Scholes-Merton com os cálculos fornecidos pelo por plataformas de finanças como a Bloomberg. É perfeitamente possível ir diretamente nesses sites e conseguir diretamente os dados da volatilidade implícita.

Todavia, o cálculo dessas plataformas tem algumas suposições. A taxa de juros livre de risco considerada é equivalente a *FED Fund Rate* (taxa de juros básica dos EUA), sendo que poderia ser considerado uma série de outras taxas no lugar dela. O melhor é construir e rodar o próprio algoritmo. No que foi construído no exemplo anterior foi considerado que $r = 0.433$.

O outro ponto é que por vezes a volatilidade implícita dessas plataformas pode refletir eventos extremos não capturados pelo modelo Black-Scholes-Merton. Um exemplo disso são os contratos recentes de Abril/2025, pois eles foram severamente afetados pela incerteza das políticas do presidente Trump, chegando a bater 0.11 nos contratos de 17/04/2025.

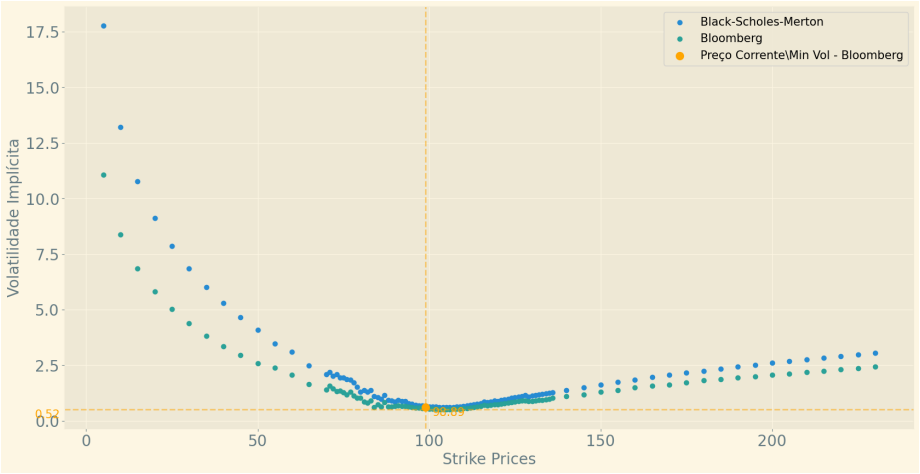


Figura 21: Volatility Smile Black Scholes Merton vs Bloomberg

Se considerarmos a volatilidade das opções precificadas exatamente no mesmo valor do preço do ativo base, 0.059 ,teremos um VaR de R$ -9704.63. Esse de longe é o valor mais discrepante de todos os já calculados até aqui. Um ponto que é necessário levantar é que o modelo Black-Scholes-Merton assume que a volatilidade seja constante, mas isso não é um pressuposto crível como é possível ver no *volatility smile.*Além disso, a volatilidade foi calculada utilizando os contratos correntes de opções do NVDA e não a volatilidade das opções no período delimitado.Como então lidar com esse problema?

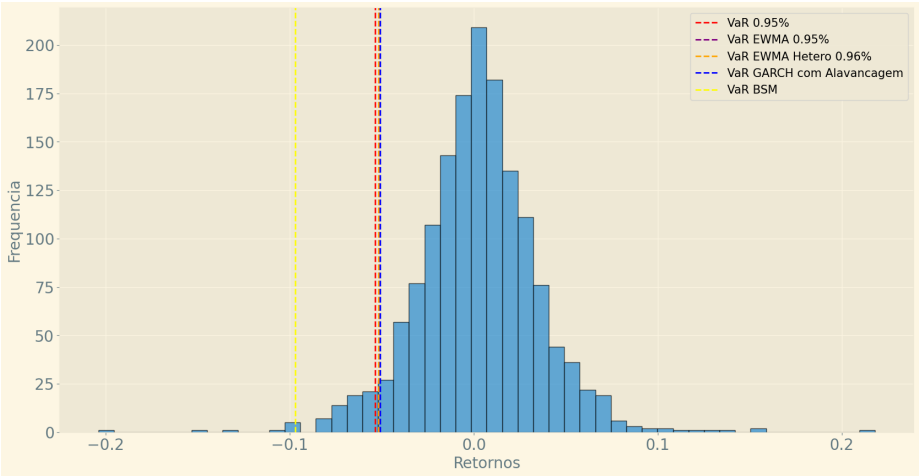


Figura 22: VaR com Black-Scholes-Merton(BSM)

# 8 Qual modelo gera o melhor VaR?

Uma pergunta que pode ser feita é: *Qual dos modelos gera o melhor VaR?*
Essa pergunta pode ser respondida de diversas formas, porém uma abordagem bastante popular é realizar o backtesting dos VaR estimados com as volatilidades dos modelos em relação aos dados dos retornos.

O teste utiliza os dados históricos dos retornos, contando o número de vezes que os valores excederam o intervalo de confiança estipulado para o VaR. O número de violações pode ser discriminado em: limites superiores, quando o retorno exceder o intervalo de confiança no lado direito da cauda, e limites inferiores, quando o retorno for mais negativo do que a perda determinada pelo VaR. O teste realizado aqui irá considerar os valores negativos, ou seja $-\sigma$.

Os modelos geraram as seguintes estimativas:

| **Modelo** | $\sigma$ | **VaR** |
|---|---|---|
| EWMA | 0.031 | 0.05 |
| EWMA Otimizado | 0.03 | 0.051 |
| GARCH | 0.029 | 0.049 |
| GJR-GARCH | 0.032 | 0.05 |
| Black-Scholes | 0.059 | 0.097 |

Se for analisado os desvios dos VaRs estimados em relação aos dados dos retornos por meio de backtesting, é possível o número de dias em que ocorreram violações do VaR estimado. Quanto menos violações, melhor. Realizando esse teste, é possível verificar que as percentagens, de um total de 1509 observações diárias, com violações são:

| **Modelo** | **% Violações** |
|---|---|
| EWMA | 4.63% |
| EWMA Otimizado | 4.63% |
| GARCH | 4.72% |
| GJR-GARCH | 4.63% |
| Black-Scholes | 0.46% |

É possível ver na tabela que os modelos EWMA, tanto normal quanto com otimização, possuem eficiências semelhantes para o conjunto de dados analisado. O interessante é que o modelo GARCH teve um desempenho *pior* do que os modelos EWMA e o modelo GJR-GARCH um desempenho apenas equivalente. Assim, por mais que seja mais elegante, o modelo GARCH não teve um desempenho melhor do que uma simples média móvel ponderada. Esse resultado está em linha com o que já tinham observado (GALDI; PEREIRA, 2007),

de que o modelo clássico do Riskmetrics de uma volatilidade calculada por suavização exponencial, além de mais simples, gera resultados tão bons quanto de um modelo GARCH mais complexo.

O fato de o modelo Black-Scholes-Merton ter gerado o melhor VaR é curioso. O modelo utilizou dados de vários contratos de opções e poderia ser considerado que não é uma estimativa crível devido não estar diretamente ligado aos dados analisados. Essa é uma crítica válida e ilustra como, mesmo aparentemente tendo validade quantitativa, um modelo deve ser analisado também pelo prisma teórico antes de qualquer afirmação mais causal. Entretanto, é importante observar que a volatilidade da série histórica dos contratos para mesmo período varia de 0.4 a 0.6, dentro do que foi calculado. Cabe ao analista avaliar se isso é válido.

# Referências


ANDERSEN, T. G.; BOLLERSLEV, T.; LANGE, S. Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon. *Journal of empirical finance*, Elsevier, v. 6, n. 5, p. 457–477, 1999.

BLACK, F.; SCHOLES, M. The pricing of options and corporate liabilities. *Journal of political economy*, The University of Chicago Press, v. 81, n. 3, p. 637–654, 1973.

BOLLEN, B. What should the value of lambda be in the exponentially weighted moving average volatility model? *Applied Economics*, Taylor & Francis, v. 47, n. 8, p. 853–860, 2015.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, Elsevier, v. 31, n. 3, p. 307–327, 1986.

DERMAN, E.; MILLER, M. B. *The volatility smile*. [S.l.]: John Wiley & Sons, 2016.

DOBRÁNSZKY, P. Comparison of historical and parametric value-at-risk methodologies. *Available at SSRN 1508041*, 2009.

ENGLE, R. Garch 101: The use of arch/garch models in applied econometrics. *Journal of economic perspectives*, American Economic Association, v. 15, n. 4, p. 157–168, 2001.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, JSTOR, p. 987–1007, 1982.

ENGLE, R. F.; NG, V. K. Measuring and testing the impact of news on volatility. *The journal of finance*, Wiley Online Library, v. 48, n. 5, p. 1749–1778, 1993.

FAMA, E. F. The behavior of stock-market prices. *The journal of Business*, JSTOR, v. 38, n. 1, p. 34–105, 1965.

GALDI, F. C.; PEREIRA, L. M. Valor em risco (var) utilizando modelos de previsão de volatilidade: Ewma, garch e volatilidade estocástica. *BBR-Brazilian Business Review*, FUCAPE Business School, v. 4, n. 1, p. 74–95, 2007.

MERTON, R. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, v. 4, p. 141–183, 1973.