Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Otimização de portfólios utilizando métodos de agrupamento para redução do erro de estimação do método de Mínima Variância: Uma aplicação ao mercado de ações.

Lucas Calistro Lessa

31 de maio de 2021

Lucas Calistro Lessa

Otimização de portfólios utilizando métodos de agrupamento para redução do erro de estimação do método de Mínima Variância: Uma aplicação ao mercado de ações.

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Hudson da Silva Torrent

Porto Alegre
Maio de 2021

Lucas Calistro Lessa

Otimização de portfólios utilizando métodos de agrupamento para redução do erro de estimação do método de Mínima Variância: Uma aplicação ao mercado de ações.

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador e pela Banca Examinadora.

Orientador:_____
Prof. Dr. Hudson da Silva Torrent, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. João Frois Caldeira, UFSC
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Prof. Dr. Márcio Valk, UFRGS
Doutor pela Universidade Estadual de Campinas – Campinas, SP

Porto Alegre
Maio de 2021

“ Cerque-se de pessoas que te fazem feliz. Pessoas que te fazem rir, que te ajudam quando você precisa. Pessoas que genuinamente se importam. Eles são os únicos que valem a pena manter em sua vida. Os outros estão apenas de passagem.”

(Karl Marx)

Agradecimentos

Agradeço ao meu orientador Hudson e aos professores Márcio, Márcia e Gabriela que me guiaram neste trabalho.

Dedicatória

Dedico este trabalho ao meu pai.

Resumo

Os modelos econômicos para otimizações de portfolios são negativamente afetados pela imprecisão da estimação da matriz de covariância, que é usada nos algoritmos de otimização, para se obter os pesos ideais de acordo com as estratégias de mínima variância ou média-variância. Desta forma, este trabalho busca formas diferentes de estimar a matriz de covariância, afim de melhorar os resultados obtidos nos algoritmos de otimização de portfolios. Neste trabalho foi verificado o desempenho em termos de retorno médio, desvio padrão, índice de *Sharpe*, *turnover* e μ_{ct} de diferentes estratégias de otimização de portfólios (MIV, EW e modelo misto destas duas estratégias em duas etapas) em 3 diferentes conjuntos de dados balanceados (DJI, COMPSTAT e SP500). Utilizando dois métodos de agrupamento (K-means e Uhclust), aplicando certas tranformações nos dados e medidas de distância/dissimilaridade para aplicar estes métodos de agrupamento. Além de utilizar duas formas diferentes de estimar a matriz de covariância amostral (empírica e diagonal).

Palavras-Chave: Otimização de portfólio, Coeficiente de autocorrelação, Distância euclidiana, K-means, Uhclust, Matriz de covariância empírica, Matriz de covariância diagonal, 1/N, Mínima variância.

Abstract

The economic models for portfolio optimization are negatively affected by the inaccuracy of the estimation of the covariance matrix, which is used in the optimization algorithms, to obtain the ideal weights according to the minimum-variance or mean-variance strategies. In this way, this work seeks different ways to estimate the covariance matrix, in order to improve the results obtained in the portfolio optimization algorithms. In this work, was verified the performance in terms of average return, standard deviation, Sharpe ratio, turnover and μ_{ct} of different portfolio optimization strategies (MIV, EW and mixed model of this two strategies in two stages) in 3 different balanced data sets (DJI, COMPSTAT and SP500). Using two clustering methods (K-means and Uhclust), applying certain transformations to the data and distance/dissimilarity measures to apply these clustering methods. In addition to using two different ways to estimate the sample covariance matrix (empirical and diagonal).

Keywords: Portfolio optimization, Autocorrelation coefficient, Euclidian distance, K-means, Uhclust, Covariance matrix, Diagonal covariance matrix, 1/N, Minimum variance.

Sumário

Lista de Tabelas

1 Introdução

O mercado de ações é o ambiente no qual empresas de capital aberto negociam frações de seu patrimônio e nele existe uma incerteza intrínseca, pois é afetado por muitos fatores externos, como PIB, inflação, lei da oferta e demanda, projeções de mercado, até questões pessoais de pessoas chave de determinada empresa. Esses fatores são tantos que não existe uma maneira de prevê-los, nem antecipar em como esses fatores influenciam na procura pelos ativos (Martini, 2013). Muitos estudos propõem diferentes métodos envolvendo algoritmos de otimização para melhorar os rendimentos dos investimentos, por exemplo, (Santos e Tessari, 2012; Michaud, 1989; Best e Grauer, 1991; Mendes e Leal, 2005), entre outros. Este conhecimento impulsionaria os investidores a selecionar a carteira de investimentos de forma em que o ganho seja maximizado, evitando perdas de capital.

Os modelos de otimização de portfólios de ativos de investimentos têm ganhado atenção desde o trabalho de média-variância proposto por Markowitz (1952), onde temos dois modelos: O modelo de Média Variância (MEV) e o modelo de Mínima Variância (MIV). Este método considera o *trade-off* entre o retorno esperado e o risco para determinar qual a alocação ótima dos ativos na carteira de investimento. Porém, a implementação destas estratégias na prática esbarra na dificuldade de se obter estimações mais precisas dos retornos esperados dos ativos e da matriz de covariância (Santos e Tessari, 2012).

Essas estimativas atualmente são obtidas via máxima verossimilhança, supondo retornos normalmente distribuídos. No entanto, seu desempenho é altamente sensível a desvios da distribuição empírica ou amostral de normalidade (DeMiguel e Nogales, 2009). Atualmente, na literatura financeira, existem diversas evidências de que esta suposição de normalidade dos retornos nem sempre é verdadeira, ver, por exemplo (Michaud, 1989; Best e Grauer, 1991; Mendes e Leal, 2005), entre outros. Assim, são esperadas defasagens na estimação da matriz de covariância via máxima verossimilhança que podem prejudicar a otimização das carteiras obtidas através destes estimadores. Considerando este problema da incerteza estatística na estimação da matriz de covariância na otimização de portfolios financeiros, os métodos de agrupamento são boas ferramentas para identificação de padrões e a sua aplicação na estimação da matriz de covariância pode resultar em um estimador mais preciso.

O trabalho de Massahi et al. (2020) utiliza o método de agrupamento K-means com o método de mínima variância em dois estágios para formular uma estratégia de seleção de portfólio. Uma desvantagem do método de agrupamento K-means é a necessidade de se definir previamente o número de grupos. Embora haja maneiras de se selecionar o número de grupos a partir dos dados, uma alternativa interessante

ao K-means é o método de agrupamento Uhclust, proposto por Valk e Cybis (2020). Nesse método, é possível testar a hipótese nula de que dois grupos são homogêneos contra a hipótese alternativa de que existe de fato dois grupos distintos. A aplicação sequencial dessa ideia determina a separação dos dados em grupos, de modo que o número de grupos é determinado por critérios estatísticos. Este trabalho busca investigar se a implementação do método de agrupamento Uhclust, proposto por Valk e Cybis (2020), com a metodologia de seleção de portfólio proposta por Massahi et al. (2020), é capaz de reduzir a variância na estimação da matriz de covariância. O que, por sua vez, poderá melhorar os resultados nos algoritmos de otimização de portfólios.

Além desta introdução, este trabalho está organizado da seguinte maneira: Na Seção 2 é apresentado o tratamento dos dados, os métodos de agrupamento, os métodos de estimação da matriz de covariância amostral, os métodos de otimização de carteiras e por último, as medidas de desempenho. A Seção 3 engloba a análise empírica das diferentes estratégias propostas neste trabalho. Por fim, a Seção 4 abrange a conclusão desde trabalho.

2 Metodologia

2.1 Pré-processamento dos dados

Muitas vezes negligenciado em aplicações financeiras, o pré-processamento de dados é uma etapa de grande importância em técnicas de mineiração de dados e que pode melhorar bastante o processo de agrupamento dos ativos (Massahi et al., 2020). Antes de performar o pré-processamento dos dados, o retorno diário das ações devem ser obtidas com base nos preços de fechamento ajustados como segue (Amenc e Le Sourd, 2003);

$$R_{it} = \log\left(\frac{P_{it}}{P_{i(t-1)}}\right), \tag{2.1}$$

onde R_{it} e P_{it} representam o log-retorno e o preço de fechamento ajustado do ativo i no dia t, respectivamente.

Para etapa de agrupamento, os dados precisam passar por algumas tranformaçoes de acordo com Massahi et al. (2020). As tranformações performadas nestes dados são *offset translation*, *amplitude scaling* e remoção da tendência linear (Soon e Lee, 2007).

Offset translation é feita removendo ou adicionando um certo *offset value*, de uma ou para uma sequência de valores. No caso deste trabalho, é utilizado a média como *offset value*, desta forma:

$$\mathit{offset\_l}_{o_i} = l_{o_i} - \bar{l}_{o_i}, \tag{2.2}$$

onde l_{o_i} são as observações do ativo i.

A *amplitude scaling* é calculada dividindo $\mathit{offset\_l}_{o_i}$ pelo desvio padrão de l_{o_i}

$$\mathit{amp\_l}_{o_i} = \frac{\mathit{offset\_l}_{o_i}}{DP(l_{o_i})}. \tag{2.3}$$

Para remover a tendência linear desta base de dados tranformada, a melhor reta ajustada aos dados é subtraída de $\mathit{amp\_l}_{o_i}$. Desta forma, é removida a influência do tempo desta série temporal. Estas 3 tranformações são utilizadas neste trabalho apenas para estimar as autocorrelações, após o agrupamento, as estratégias de investimento são aplicadas aos log-retornos.

2.2 Métodos de agrupamento

2.2.1 Uhclust

O método Uclust (Cybis et al., 2018) faz com que seja possivel testar a hipótese nula de que dois grupos são homogêneos contra a hipótese alternativa de que existe de fato dois grupos distintos, ou seja, que a separação dos dados em dois grupos é estatisticamente significativa. O método de agrupamento hierárquico Uhclust consiste em aplicar sequenciamente o método uclust (Valk e Cybis, 2020), seguindo uma construção hierárquica para divisão dos grupos. Essa divisão é baseada em uma estatística que maximiza a distância entre grupos e minimiza a distância dentro dos grupos. Uma vez que cada uma das subdivisões não podem mais ser significamente divididas, temos todos os grupos que são significamente distintos. Desde modo, este método começa com toda a amostra $G_0 = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ sendo particionada em 2 subgrupos, G_1 e G_2, se a amostra não for homogênea. Assim, para cada grupo G_i é aplicado novamente o Uclust que divide o grupo em 2 novos subgrupos G_l e G_k, sempre que G_i seja não-homogêneo. Repetindo assim para todo novo grupo G_i até que todos os grupos sejam considerados homogêneos pelo Uclust ou quando chega-se ao tamanho mínimo τ, que é o critério de parada do algoritmo. Sendo $\tau = 3$, como visto em Valk e Cybis (2020). Uma vantagem deste método sobre o K-means é não precisar fixar o número de grupos.

2.2.2 K-means

O algoritmo concebido por Hartigan e Wong (1979), visa particionar os dados em k grupos, de forma que a soma do quadrado da distância euclidiana dos dados agrupados até aos centros atribuídos aos agrupamentos seja minimizada. No mínimo, todos os centróides atribuídos aos agrupamentos estão na média do conjunto de dados que estão mais próximos ao centro do agrupamento. Suponha um conjunto de dados $X = \{x_1, \ldots, x_N\}, x_n \in \mathbb{R}^d$

$$E(m_1, \ldots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k)||x_i - m_k||^2, \tag{2.4}$$

onde m representa os centróides, x_i os pontos, C os grupos e $I(X) = 1$ se X é verdadeiro e 0 caso contrário.

2.2.3 Medidas de distância/dissimilaridade

Encontrar similaridade entre as observações é uma tarefa muito importante quando queremos agrupá-los. Algoritmos de aprendizagem não supervisionados como K-means se baseiam na teoria de que os pontos mais próximos são mais semelhantes de acordo com alguma medida de similaridade. Por isso, é necessário utilizar alguma medida de distância ou, pelo menos, de dissimilaridade entre as observações. Há várias medidas disponíveis na literatura, cada uma com suas vantagens e desvantagens, dependendo do problema de interesse. Neste trabalho, utilizamos uma das medidas utilizadas por Massahi et al. (2020), que se baseia na distância euclidiana entre as autocorrelações de cada ativo, conforme detalhado a seguir.

2.2.4 Coeficiente de autocorrelação

É a correlação entre todos os componentes de uma série temporal para t e $t-l$. O *lag* l indica como as mudanças no tempo $t-l$ podem afetar as mudanças futuras no tempo t. Uma das vantagens desta abordagem é a viabilidade de comparar séries temporais com tamanhos diferentes (Caiado et al., 2009). Um conjunto de séries temporais, uma para cada ativo, que supostamente segue um processo estocástico semelhante pode ser demonstrado como

$$X = x_{nt} : n = 1, \ldots, N; t = 1, \ldots, T = \begin{pmatrix} x_{11} & \cdots & x_{n1} & \cdots & x_{N1} \\ \vdots & & \vdots & & \vdots \\ x_{1t} & \cdots & x_{nt} & \cdots & x_{Nt} \\ \vdots & & \vdots & & \vdots \\ x_{1T} & \cdots & x_{nT} & \cdots & x_{NT} \end{pmatrix}, \tag{2.5}$$

onde x_{nt} representa a t-ésima ocorrência no ativo n.

A autocorrelação no *lag* $l (l = 1, \ldots, L = T-1)$ é escrita como

$$\rho_{nl} = \frac{\sum_{t=l+1}^{T}(x_{nt} - \bar{x}_n)(x_{n(t-l)} - \bar{x}_n)}{\sum_{t=1}^{T}(x_{nt} - \bar{x}_n)^2}, \tag{2.6}$$

onde $\bar{x}_n$ é a média da n-ésima série temporal. Formando assim, uma matriz $L \times N$. Ou seja, uma matriz onde cada coluna representa os *lags* de cada ativo.

No trabalho de Massahi et al. (2020), é utilizado o coeficiente de autocorrelação para gerar uma matriz dos dados após o pré-processamento dos mesmos. Assim, com esta matriz de autocorrelações dos dados após o pré-processamento e tranformações, aplica-se a distância euclidiana. Neste trabalho, será feito da mesma forma.

2.3 Métodos de estimação da matriz de covariância amostral

Para se obter a correlação dos log-retornos dos ativos devemos utilizar dados amostrais. Os métodos de otimização de portfólios dependem dessas estimações. Existem, porém, erros de estimação neste processo, afetando todo o processo de otimização de carteiras. A fim de melhorar os resultados nos processos de otimização, compararemos diferentes formas de estimar a matriz de covariância. Mais especificamente, a matriz de covariância amostral supõe que os log-retornos em t e $t-1$ não devem ter correlação, com média e desvio padrão constantes, isto é, independentes e igualmente distribuídos (i.i.d.), consideramos então que a matriz de covariância mantém-se constante ao longo do tempo. Infelizmente, sabe-se na literatura que a hipótese de log-retornos i.i.d. não se confirma na prática. Ledoit e Wolf (2003) apresentam que, por mais que a matriz de covariância amostral seja um estimador não-viesado, é um estimador bastante ruidoso da matriz de covariância populacional mesmo quando o tamanho da amostra é grande. Assim, utilizar toda a amostra é considerado ruim, permitindo pouca utilidade das informações mais recentes, pois no cálculo da matriz de covariância amostral, todas as observações têm o mesmo peso. A fim de contornar este problema, é utilizado uma janela móvel (*rolling window*).

Outra vantagem de utilizar a janela móvel é que desta maneira, a amostra fica mais focada nos acontecimentos recentes que impactam mais os preços de fechamento.

Uma janela móvel é simplesmente uma amostra dos dados de tamanho n, em que n, neste trabalho, é o número de dias ou meses. Desta maneira, é descartado as observações mais antigas.

2.3.1 Matriz de covariância empírica

O método mais simples é a matriz de covariância empírica. Assim, seja $\mathbf{R}$ uma matriz $N \times n$ com os log-retornos dos ativos

$$\mathbf{R} = \begin{pmatrix} R_{11} & \cdots & R_{21} & \cdots & R_{N1} \\ \vdots & & \vdots & & \vdots \\ R_{12} & \cdots & R_{22} & \cdots & R_{N2} \\ \vdots & & \vdots & & \vdots \\ R_{1n} & \cdots & R_{2n} & \cdots & R_{Nn} \end{pmatrix}, \tag{2.7}$$

A matriz de covariância amostral empírica, Σ, pode ser calculada deste modo:

$$\sum = \frac{1}{n-1} \sum_{t=1}^{n} (R_i - \mu)(R_i - \mu)^T, \tag{2.8}$$

com n sendo o tamanho do intervalo de estimação e $\mu = \frac{1}{n} \sum_{t=1}^{n} R_i$, com R_i sendo os log-retornos de cada ativo individualmente, assim μ é o vetor de médias amostrais dos log-retornos dos ativos no intervalo de tamanho n.

2.3.2 Matriz de covariâncial diagonal

A fim de tentar diminuir o ruído gerado na estimação da matriz de covariância amostral empírica, utilizaremos também a matriz de covariância amostral diagonal, que é, apenas, a variância de cada ativo, ignorando as covariâncias entre ativos. A matriz de covariância amostral diagonal, $\Sigma_{i,j}$, pode ser representada como:

$$\Sigma_{i,j} = \begin{cases} \sigma_i^2, & \text{se } i = j \\ 0, & \text{caso contrário.} \end{cases} \tag{2.9}$$

2.4 Métodos de otimização de carteiras

O conceito por trás de como a bolsa de valores funciona é bem simples. Operando como se fosse um leilão, as bolsas de valores permitem que compradores e vendedores negociem preços e façam negócios. Uma prática muito comum na bolsa de valores é a venda a descoberto (*short selling*) que consiste na venda de um ativo que o investidor não possui em carteira. Para fazer isso o investidor precisa realizar duas operações: Aluguel da ação que ele não tem e venda deste mesmo ativo. Por exemplo, se um investidor alugar mil ações de uma certa empresa por R$ 10 e colocar estas ações a venda. Por esta operação ele receberá R$ 10.000. No final do prazo do

aluguel, digamos que estas ações estão cotadas a R$ 6. Desta forma, o investidor vai recomprar estas mesmas ações que ele alugou para vender, pois elas foram alugadas e precisam ser devolvidas, porém, ele vai recompra-las por R$ 6.000, tendo assim, um lucro de R$ 4.000 (sem contar custos operacionais).

Investidores aplicam muitas técnicas para minimizar o risco e ao mesmo tempo aumentar o retorno. Entre estes técnicas, está a abordagem de otimização de carteiras, proposta inicialmente por Markowitz (1952), que deu origem a "Teoria Moderna do Portfólio", fazendo com que o processo de alocação de ativos seja abordada como um processo de otimização.

É importante também que exista uma frequência em que os investimentos nos ativos são reordenados, para isto, temos o rebalanceamento. Neste estudo usaremos o rebalanceamento com a maior frequência possível na amostra.

2.4.1 Modelo de carteira igualmente ponderada (EW)

Este é o modelo mais simples em que os pesos da carteira são todos iguais, pois, $w_i = 1/N$ e continuam nesta mesma forma em todos rebalanceamentos. Para monitorar os resultados, essa estratégia será utilizada como *benchmark*, pois é a estratégia mais simples que não depende de nenhuma forma de estimação e nenhuma técnica de otimização sendo amplamente usada como uma forma fácil de alocar recursos (Santos e Tessari, 2012). Encontra-se ainda, evidências empíricas de que o modelo de carteira igualmente ponderada têm uma performance superior a modelos de otimização de portfólios, como os modelos de média-variância e mínima variância (DeMiguel et al., 2007).

2.4.2 Modelo de mínima variância (MIV)

Este modelo é um caso especial do modelo de média-variância em que o parâmetro de aversão ao risco é infinito ($\gamma = \infty$). Grande parte da literatura acadêmica recente tem mostrado bastante interesse neste modelo pois sua estimação não leva em consideração o retorno esperado, eliminando uma estimativa, levando a menor erro de estimação, comparado ao modelo de média-variância (Ledoit e Wolf, 2003; DeMiguel e Nogales, 2009). Este modelo consiste em solucionar a seguinte expressão

$$\min_{w} \mathbf{w^T} \sum \mathbf{w}, \tag{2.10}$$

sujeito a $\mathbf{I}^T\mathbf{w} = 1$.

Onde $\mathbf{w} \in \mathbb{R}^N$ é o vetor de pesos da carteira e $\sum$ é a matriz de covariância amostral dos log-retornos da carteira. Podemos notar que este é um problema de otimização quadrática. A restrição $\mathbf{I}^T\mathbf{w} = 1$, onde $\mathbf{I} \in \mathbb{R}^N$ é um vetor de uns, que garante que a soma dos pesos da carteira seja um.

2.4.3 Modelo misto (EWMIV, MIVEW ou MIVMIV)

Este modelo incorpora tanto o modelo EW quanto o modelo MIV, de forma que a otimização ocorre em 2 etapas, cada etapa utilizando um dos modelos citados anteriormente. Os modelos mistos funcionam da seguinte forma:

- Primeiro estágio: Um método de agrupamento é aplicado aos dados pré-processados, dividindo os dados em grupos, cada grupo será visto como um ativo e um método de otimização de carteiras define os pesos de cada grupo.
- Segundo estágio: Dentro de cada grupo é aplicado um método de otimização de carteiras, definindo o peso de cada ativo dentro do grupo.

Assim, para obter o resultado final de otimização é necessário multiplicar o resultado do primeiro estágio com o resultado do segundo estágio. (Massahi et al., 2020).

2.5 Medidas de desempenho

É necessário avaliar o risco envolvido em cada uma das estratégias para se ter uma noção de desempenho dos modelos. As medidas utilizadas neste trabalho foram: Média, desvio padrão, índice de *Sharpe*, *turnover* e μ_{ct}. Seja T o número total de observações de cada conjunto de dados, $NT = T - n$, ou seja, $NT =$ numero de observações fora da janela móvel. Estes cálculos são aplicados nas T obervações que seguem a janela móvel. A média é simplesmente a média dos log-retornos dos ativos em todo o período estudado

$$\hat{\mu} = \frac{1}{T}\sum_{1}^{T} w_t R_{t+1}, \tag{2.11}$$

onde $w_t R_{t+1}$ é o retorno obtido da estratégia de investimento, que nada mais é que o produto entre os pesos designados à carteira no período e os respectivos log-retornos deste mesmo período.

O desvio padrão é dado por

$$\hat{\sigma} = \frac{1}{T}\sum_{1}^{T}(w_t R_{t+1} - \hat{\mu}). \tag{2.12}$$

O índice de *Sharpe* é dado por

$$IS = \frac{\hat{\mu}}{\hat{\sigma}}. \tag{2.13}$$

O *turnover* no tempo t é dado por

$$TO_t = \sum_{i=1}^{N} \left| w_{i,t+1|t} - w_{i,t|t-1}\frac{1 + y_{i,t}}{1 + \mathbf{w}_{t|t-1}^T \mathbf{y}_t} \right|, \tag{2.14}$$

onde $w_{i,j|t-1}$ é o i-éssimo elemento do vetor de pesos $\mathbf{w}_{t|t-1}$. O *turnover* indica o valor do portfólio que é vendido ou comprado quando ocorre o rebalanceamento no tempo t para $t+1$ É considerado a média dos retornos líquidos de custos de transação para quantificar o impacto dos custos de transação de cada estratégia diferente (DeMiguel e Nogales, 2009), desta forma:

$$\hat{\mu}_{ct} = \frac{1}{T}\sum_{t=1}^{T}\left[(1 + w_t^T R_{t+1})(1 - c\sum_{j=1}^{N}(|w_{j,t+1} - w_{j,t}|)) - 1\right], \tag{2.15}$$

onde c é a taxa paga a cada transação, neste trabalho, $c = 0,1$. Um μ_{ct} maior do que 0 significa que a estratégia teve retorno positivo levando em conta os custos de transação.

3 Análise Empírica

Neste trabalho foram utilizadas três bases de dados, de diferentes tamanhos e diferentes períodos, a fim de verificar os resultados em diferentes circunstâncias.

A primeira base de dados escolhida foi o Índice Dow Jones (DJI) foi selecionado no período de 02/03/2017 à 22/11/2019, através do banco de dados do Yahoo Finances. O segundo banco de dados compreende os retornos mensais de 16.855 companhias no nível da empresa do CRSP (COMPSTAT) de novembro de 1970 até dezembro de 2018 (http://www.crsp.org/resources/data). O último conjunto de dados selecionado contém retornos diários referentes a 500 ações negociadas bolsa de valores de Nova Iorque, no período de 1989 até 1996 (SP500), de acordo com estudos anteriores como Tola et al. (2008)

Todas estas bases de dados foram balanceadas, ficando com estas dimensões:

- DJI - 29 ativos e 684 preços de fechamento diários.
- COMPSTAT - 177 ativos e 537 preços de fechamento mensais.
- SP500 - 470 ativos e 1259 preços de fechamento diários.

A fim de diminuir ruídos na amostra, é utilizado uma janela móvel, neste caso, de tamanho n, proporcional a 1 ano no conjunto de dados DJI (253 dias), 21 anos na base COMPSTAT (253 meses) e 3 anos no conjunto de dados SP500 (756 dias). Neste trabalho, o parâmetro para o número de grupos foi fixado em 4 para o banco DJI, 7 para o banco COMPSTAT e 15 para o banco SP500, sem nenhum critério em especial.

As tabelas 3.1, 3.2 e 3.3 mostram os resultados anualizados das diferentes estratégias de otimização de portfólios em termos de retorno médio de investimento, desvio padrão, índice de *Sharpe* (IS), *turnover* e μ_{ct} nos 3 diferentes conjuntos de dados utilizados neste trabalho. As estratégias abordadas foram:

- Benchmark - Modelo de carteira igualmente ponderada.
- MIV - Modelo de mínima variância com matriz de covariância empírica e sem método de agrupamento.
- MIVEW-K - Modelo misto com matriz de covariância empírica e método de agrupamento K-means (primeira etapa MIV, segunda etapa EW).
- MIVEW-U - Modelo misto com matriz de covariância empírica e método de agrupamento Uhclust (primeira etapa MIV, segunda etapa EW).

- EWMIV-K - Modelo misto com matriz de covariância empírica e método de agrupamento K-means (primeira etapa EW, segunda etapa MIV).
- EWMIV-U - Modelo misto com matriz de covariância empírica e método de agrupamento Uhclust (primeira etapa EW, segunda etapa MIV).
- MIVMIVD-K - Modelo misto com matriz de covariância diagonal e método de agrupamento K-means (primeira etapa MIV, segunda etapa MIV).
- MIVMIVD-U - Modelo misto com matriz de covariância diagonal e método de agrupamento Uhclust (primeira etapa MIV, segunda etapa MIV).

Os resultados da Tabela 3.1 apresentam que em termos de retorno médio, o *benchmark* ficou atrás de todas estratégias, menos MIVEW-U. Neste sentido, nenhuma estratégia superou o MIV (18,91%). Vale destacar também que a estratégia com o melhor índice de *Sharpe* é EWMIV-U.

Observando a Tabela 3.2, nota-se que em termos de retorno médio, o *benchmark* ficou melhor do que todas as outras estratégias (11,79%). Vale destacar também que a estratégia com o melhor índice de *Sharpe* é também é o *benchmark*.

A Tabela 3.3 apresenta que em termos de retorno médio, o *benchmark* ficou melhor do que todas as outras estratégias (17,28%). Vale destacar também que a estratégia com o melhor índice de *Sharpe* também é o *benchmark*.

Os desvios padrões parecem estar todos nivelados, com apenas algumas estratégias com um valor mais diferente mas nada chamativo.

É fácil notar que o *benchmark*, em todas as bases deste estudo, superou as outras estratégias. O fato de este trabalho utilizar bases balanceadas, que não é o usual no mercado, beneficia a estratégia EW. Desta forma, é interessante comparar o MIV com as demais estratégias (exceto EW) para avaliar se o método de agrupamento Uhclust melhora o desempenho da estratégia MIV.

Olhando para o índice de *Sharpe*, todas as estratégias com agrupamento têm desempenho superior do que o MIV, menos no conjunto de dados DJI (Tabela 3.1). É possível que, por ser um conjunto de dados pequeno (29 ativos) comparado com as outras bases deste estudo, os métodos de agrupamento não sejam tão eficazes. Comparando os agrupamentos, o método Uhclust se mostrou superior em 7 das 9 comparações (2 envolvendo o modelo EWMIV).

Em questão de *turnover*, na base DJI (Tabela 3.1), nenhuma estratégia superou o EWMIV. Em geral, em termos de *turnover*, o método Uhclust ganhou em todas as comparações com o K-means e MIV. No conjunto de dados COMPSTAT (tabela 3.2), tirando o modelo EWMIV-U, os dois outros modelos que utilizam o método uhclust tiveram os melhores desempenhos, o mesmo foi verificado na base SP500 (Tabela 3.3).

Em termos de μ_{ct}, todos os modelos que utilizam o método Uhclust, superaram o K-means, exceto EWMIV-U. Todos os modelos que utilizam métodos de agrupamento também superaram o modelo MIV, exceto na base DJI (Tabela 3.1) e EWMIV-U, em todas as bases.

Fica evidente que os modelos que melhor performaram, sem levar em consideração o *Benchmark*, foram MIVEW-U e MIVMIVD-U.

Na Tabela 3.4, podemos notar a comparação entre o Uhclust e o K-means nos dois modelos que o Uhclust melhor performou. Nesta comparação, o índice de *Sharpe* aumentou em 13% e 24% no conjunto de dados DJI, em 5% e 18% na base

COMPSTAT e 11% na base SP500. Olhando para o *turnover*, em todas as bases ele diminuiu, 88% e 87% no conjunto de dados DJI, 84% e 71% na COMPSTAT e 63% e 62% na base SP500.

Na comparação entre Uhclust e MIV (Tabela 3.5), o método Uhclust foi superior em todos os aspectos (menos na base DJI). Aumentando o índice de *Sharpe* em 54% e 55% na base COMPSTAT e 195% e 196% na base SP500. Diminuindo o *turnover* em 64% e 65% no conjunto de dados DJI, em 86% e 87% na base COMPSTAT e também, 70% e 73% na base SP500.

	$\hat{\mu}$ (%)	$\hat{\sigma}$ (%)	IS	TO_t	μ_{ct}
Benchmark	0,1057155	0,1429381	0,7395892	0,0077564	0,1046140
MIV	0,1891397	0,1238582	1,5270665	0,0887114	0,1781753
MIVEW-K	0,1241037	0,1222188	1,0154221	0,2869327	0,0916358
MIVEW-U	0,1409424	0,1220109	1,1551623	0,0315683	0,1370212
EWMIV-K	0,1086085	0,1324361	0,8200822	0,4619216	0,0572660
EWMIV-U	0,0766963	0,1325256	0,5787278	0,0249482	0,0738380
MIVMIVD-K	0,1131957	0,1207257	0,9376272	0,2393369	0,0863515
MIVMIVD-U	0,1389747	0,1194369	1,1635825	0,0310255	0,1351260

Tabela 3.1: Medidas de desempenho - base DJI

	$\hat{\mu}$ (%)	$\hat{\sigma}$ (%)	IS	TO_t	μ_{ct}
Benchmark	0,1179971	0,1435329	0,8220908	0,0528538	0,1116933
MIV	0,0886102	0,1732321	0,5115118	1,3327332	-0,0562736
MIVEW-K	0,0845195	0,1115204	0,7578838	1,1512539	-0,0401942
MIVEW-U	0,0956505	0,1197623	0,7986696	0,1807911	0,0755751
EWMIV-K	0,0771515	0,1081998	0,7130461	1,1365105	-0,0451085
EWMIV-U	0,0843527	0,1091277	0,7729728	1,2184664	-0,0476038
MIVMIVD-K	0,0726705	0,1087918	0,6679778	0,5654734	0,0119715
MIVMIVD-U	0,0900753	0,1131853	0,7958215	0,1634314	0,0720056

Tabela 3.2: Medidas de desempenho - base COMPSTAT

	$\hat{\mu}$ (%)	$\hat{\sigma}$ (%)	IS	TO_t	μ_{ct}
Benchmark	0,1728363	0,1085642	1,5920196	0,0089291	0,1714468
MIV	0,0565596	0,1301678	0,4345132	0,8916669	-0,0375753
MIVEW-K	0,0880562	0,0764753	1,1514327	0,7097564	0,0108091
MIVEW-U	0,1096408	0,0853985	1,2838724	0,2608947	0,0805301
EWMIV-K	0,0475920	0,0907999	0,5241415	2,6397886	-0,2284941
EWMIV-U	0,0595588	0,0912479	0,6527137	3,0762078	-0,2658529
MIVMIVD-K	0,0887445	0,0765678	1,1590322	0,6139501	0,0218574
MIVMIVD-U	0,1088848	0,0844752	1,2889568	0,2333865	0,0828396

Tabela 3.3: Medidas de desempenho - base SP500

	IS	%	TO_t	%
DJI				
MIVEW-K	1,0154221	-	0,2869327	-
MIVEW-U	1,1551623	0,13	0,0315683	(0,88)
MIVMIVD-K	0,9376272	-	0,2393369	-
MIVMIVD-U	1,1635825	0,24	0,0310255	(0,87)
COMPSTAT				
MIVEW-K	0,7578838	-	1,1512539	-
MIVEW-U	0,7986696	0,05	0,1807911	(0,84)
MIVMIVD-K	0,6679778	-	0,5654734	-
MIVMIVD-U	0,7958215	0,19	0,1634314	(0,71)
SP500				
MIVEW-K	1,1514327	-	0,7097564	-
MIVEW-U	1,2838724	0,11	0,2608947	(0,63)
MIVMIVD-K	1,1590322	-	0,6139501	-
MIVMIVD-U	1,2889568	0,11	0,2333865	(0,62)

Tabela 3.4: Comparação Uhclust e K-means

	IS	%	TO_t	%
DJI				
MIV	1,5270665	-	0,0887114	-
MIVEW-U	1,1551623	(0,24)	0,0315683	(0,64)
MIVMIVD-U	1,1635825	(0.23)	0,0310255	(0,65)
COMPSTAT				
MIV	0,5115118	-	1,3327332	-
MIVEW-U	0,7986696	0,54	0,1807911	(0,86)
MIVMIVD-U	0,7958215	0,55	0,1634314	(0,87)
SP500				
MIV	0,4345132	-	0,8916669	-
MIVEW-U	1,2838724	1,95	0,2608947	(0,70)
MIVMIVD-U	1,2889568	1,96	0,2333865	(0,73)

Tabela 3.5: Comparação Uhclust e MIV

4 Conclusão

Para obter um portfólio mais robusto e preciso através do modelo de otimização de portfólios de Markowitz, foi aplicado dois métodos de agrupamento, além do pré-processamento dos dados, das medidas de distância/dissimilaridade e das diferentes formas de estimar a matriz de covariânciam, além da metodologia de modelos mistos. Para verificar o desempenho do Uhclust neste contexto e comparar com outros modelos, 3 conjuntos de dados foram selecionados (DJI, COMPSTAT e SP500) e balanceados. Os resultados verificam que, em comparação com o método de agrupamento K-means, o método de agrupamento Uhclust se mostrou superior na maioria dos casos, os modelos que melhor performaram foram MIVEW-U e MIVMIVD-U, o modelo EWMIV não se mostrou útil. Para os próximos passos da pesquisa será estudado um critério com mais embasamento para o parâmetro que define o numero de grupos do K-means.

Referências Bibliográficas

Amenc, N. e Le Sourd, V. (2003). *Portfolio theory and performance analysis*. John Wiley and Sons Ltd.

Best, M. J. e Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. *Review of Financial Studies*, 4(2):315–42.

Caiado, J., Crato, N., e Peña, D. (2009). Comparison of time series with unequal length in the frequency domain. MPRA Paper 15310, University Library of Munich, Germany.

Cybis, G., Valk, M., e Lopes, S. (2018). Clustering and classification problems in genetics through u -statistics. *Journal of Statistical Simulation and Computation*, 88:1882–1902.

DeMiguel, V., Garlappi, L., e Uppal, R. (2007). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5):1915–1953.

DeMiguel, V. e Nogales, F. J. (2009). Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577.

Hartigan, J. A. e Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Ledoit, O. e Wolf, M. (2003). Honey, I shrunk the sample covariance matrix. Economics Working Papers 691, Department of Economics and Business, Universitat Pompeu Fabra.

Markowitz, H. (1952). Portfolio selection*. *The Journal of Finance*, 7(1):77–91.

Martini, M. F. G. (2013). Renda fixa versus renda variável: uma análise descritiva entre as rentabilidades dos investimentos. *Revista On-Line IPOG*, 1(5):1.

Massahi, M., Mahootchi, M., e Khamseh, A. A. (2020). Development of an efficient cluster-based portfolio optimization model under realistic market conditions. *Empirical Economics*, 59(5):2423–2442.

Mendes, B. V. M. e Leal, R. P. C. (2005). Robust multivariate modeling in finance. *International Journal of Managerial Finance*, 1:95–106.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42.

Santos, A. A. P. e Tessari, C. (2012). Técnicas quantitativas de otimização de carteiras aplicadas ao mercado de ações brasileiro. *Revista Brasileira de Financas*, 10(3):369–393.

Soon, L.-K. e Lee, S. H. (2007). An empirical study of similarity search in stock data. In *Proceedings of the 2nd International Workshop on Integrating Artificial Intelligence and Data Mining - Volume 84*, AIDM '07, page 31–38, AUS. Australian Computer Society, Inc.

Tola, V., Lillo, F., Gallegati, M., e Mantegna, R. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258.

Valk, M. e Cybis, G. (2020). U-statistical inference for hierarchical clustering. *Forthcoming at Journal of Computational and Graphical Statistics*.