Universidade de São Paulo
Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto
Departamento de Economia
Programa de Pós-graduação em Economia
Área: Economia Aplicada

Volatilidade realizada multivariada: uma análise via aprendizado de máquina para dados do mercado brasileiro

Leonardo Ieracitano Vieira

Orientador: Prof. Dr. Márcio Poletti Laurini

Ribeirão Preto

2021

Prof. Dr. Vahan Agopyan

Reitor da Universidade de São Paulo

Prof. Dr. André Lucirton Costa

Diretor da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto

Prof. Dr. Sérgio Kannebley Júnior

Chefe do Departamento de Economia

Prof. Dr. Luciano Nakabashi

Coordenador do Programa de Pós-Graduação em Economia Aplicada

Leonardo Ieracitano Vieira

Volatilidade realizada multivariada: uma análise via aprendizado de máquina para dados do mercado brasileiro

Dissertação apresentada ao Programa de Pós-Graduação em Economia – Área de Concentração: Economia Aplicada, da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Universidade de São Paulo – USP

Versão corrigida. A original encontra-se disponível na FEA-RP/USP.

Orientador: Prof. Dr. Márcio Poletti Laurini

Ribeirão Preto

2021

Ficha Catalográfica

Leonardo Ieracitano Vieira
Volatilidade realizada multivariada: uma análise via aprendizado de máquina para dados do mercado brasileiro – Ribeirão Preto, 2021
52 p. : il.; 30 cm.
Orientador: Prof. Dr. Márcio Poletti Laurini

Dissertação de Mestrado – Universidade de São Paulo – Faculdade de Economia, Admnistração e Contabilidade – FEA/USP – Campus Ribeirão Preto; Departamento de Economia
Programa de Pós-Graduação em Economia; Área de Concentração: Economia Aplicada, 2021.

I. Orientador: Prof. Dr. Márcio Poletti Laurini. II. Universidade de São Paulo – USP – Campus Ribeirão Preto. III. Faculdade de Economia, Administração e Contabilidade. IV. Volatilidade realizada multivariada: uma análise via Machine Learning para dados do mercado brasileiro

1. Volatilidade Realizada. 2. Aprendizado de máquina. 3. Alta dimensão.

Agradecimentos

Ao meu orientador, Márcio Poletti Laurini, pelas disciplinas oferecidas no mestrado, pela excelente orientação e por todo o conhecimento transferido durante o curso.

Aos professores da FEA-RP que contribuíram na minha formação.

Aos meus professores da graduação na UFABC: Guilherme de Oliviera Lima Cagliari Marques, Bruno de Paula Rocha e Ricardo Buscariolli Pereira.

Ao Programa de Aperfeiçoamento de Ensino (PAE), pela oportunidade de exercer as monitorias de Econometria I e Econometria III.

À fundação CAPES e ao departamento de economia da FEA-RP, pelo apoio financeiro.

À minha família, namorada e aos amigos do mestrado.

Resumo

VIEIRA, L. I. **Volatilidade realizada multivariada: uma análise via aprendizado de máquina para dados do mercado brasileiro.**. Dissertação (Mestrado) - Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2021.

O presente trabalho é um exercício preditivo para a dinâmica da matriz de covariância incondicional de ativos do mercado brasileiro. Levamos em consideração métodos distintos para o cálculo da matriz, esquivando-se da matriz de covariância amostral e avaliamos o impacto preditivo no uso de regressões com encolhimento na estimação da matriz de covariância explicada pelo seu passado - um formato autorregressivo, portanto. Diferentemente do mundo univariado, o estudo de matriz de covariância tornou-se custoso devido à maldição da dimensionalidade. Tradicionalmente, via VAR, o exercício proposto traria problemas de especificação e também de dimensão, devido ao grande número de covariadas. Os resultados encontrados mostram que não necessariamente temos pior desempenho preditivo ao reduzir o número de séries, mas com a metodologia do MCS não rejeitamos a hipótese de mesma habilidade preditiva entre modelos que selecionam variáveis e que não o fazem. Diante do exercício proposto investigamos quais dificuldades e padrões estão inseridos nos dados no contexto do mercado brasileiro: trata-se de um mercado pouco líquido e que mesmo em ativos mais negociados temos problemas de dados faltantes e de concentração setorial nos ativos mais negociados. Do ponto de vista econômico encontramos resultados em linha com a literatura de referência, mostrando maior dinâmica intra setorial para processos de variância e do ponto de vista preditivo não encontramos um padrão claro para os processos de covariância.

Palavras-chave: Volatilidade realizada. Aprendizado de máquina. Alta dimensão.

Abstract

VIEIRA, L. I. **Multivariate realized volatility: an machine learning analysis for Brazilian market data.**. Dissertation (Master Degree) - School of Economics, Business and Accounting at Ribeirão Preto, University of São Paulo, Ribeirão Preto, 2021.

The present work is a predictive exercise for the unconditional covariance matrix dynamics of Brazilian market assets. We consider different methods for calculating the matrix, dodging the sample covariance matrix, and evaluate the predictive impact of using shrinkage regressions in estimating the covariance matrix explained by its past - an autoregressive format, therefore. Unlike the univariate world, the covariance matrix study has become costly due to the curse of dimensionality. Traditionally, via VAR, the proposed exercise would bring specification and dimension problems, due to the large number of covariates. The results found show that we do not necessarily have a worse predictive performance when reducing the number of series, but with the MCS methodology we do not reject the hypothesis of the same predictive ability between models that select variables and that do not. In view of the proposed exercise, we investigate which difficulties and patterns are inserted in the data in the context of the Brazilian market: it is a little liquid market and that even in the most traded assets we have problems with missing data and sector concentration in the most traded assets. From an economic point of view, we found results in line with the reference literature, showing greater intra-sector dynamics for processes of variance and, from a predictive point of view, we did not find a clear pattern for the processes of covariance.

Keywords: Realized volatility. Machine learning. High dimension.

Lista de Figuras

Lista de Tabelas

Conteúdo

1 Introdução

O presente trabalho é um exercício preditivo para a dinâmica da matriz de covariância incondicional de ativos do mercado brasileiro. A partir de dados intradiários de retorno, fizemos a construção de matrizes de covariância incondicional em uma frequência diária e por uma metodologia autorregressiva realizamos um exercício de previsão fora da amostra. Adicionalmente, temos em nosso exercício a incorporação de regressões com penalidades para contornar o problema de dimensão e a avaliação do impacto preditivo de modelos que selecionam variáveis. Buscamos investigar quais dificuldades e padrões são possíveis de detectar em um exercício multivariado para dados do mercado brasileiro.

Destacamos que apesar dos desafios que englobam o tema e nossa estrutura de dados, o objetivo do trabalho consiste na estimação e previsão da matriz de covariância realizada incondicional para um abundante volume de ativos, reservando discussões como ruídos de microestrutura de mercado. Com o exercício proposto, poderemos investigar em ativos da bolsa brasileira como são governadas as variâncias e covariâncias dos ativos em um mercado de capitais com idiossincrasias desafiadoras como a baixa liquidez e alta correlação.

Os dados em alta frequência permitem a construção da Matriz de Covariância Realizada, computando uma *proxy* preferível para a volatilidade verdadeira. Não há, até onde sabemos, registros na literatura que tenham lidado com este problema para dados do mercado brasileiro. Adicionalmente, com o uso das penalizações podemos lidar com o problema dimensional, que por muito limitou a literatura de volatilidade multivariada devido à complexidade computacional.

O entendimento da volatilidade se tornou fundamental para o apreçamento de ativos, seleção de portfólio e para o gerenciamento de risco. Historicamente, as literaturas de Finanças e Econometria sempre concentraram esforços e desenvolveram as mais diversas metodologias para a construção de melhores medidas para tal. No entanto, dois problemas são intrínsecos à discussão: o fato de que a volatilidade verdadeira não pode ser observada e o problema de dimensionalidade dos modelos multivariado.

O trabalho é dividido em 5 seções incluindo a presente introdução. Fazemos a revisão de literatura na seção 2. A metodologia é apresentada na terceira seção. Na quarta temos a análise empírica e concluímos na quinta.

2 Literatura

A revisão de literatura proposta busca conectar os trabalhos dos seguintes grandes tópicos, intrínsecos à nossa pergunta de pesquisa: resgatamos da literatura de finanças o trabalho seminal de Markowitz (1952), que introduziu a necessidade de estimações de matrizes de variância e covariância entre ativos; apresentamos os modelos de volatilidade trazidos à literatura por Bollerslev (1986) e o amadurecimento deste campo, tal como a plausível performance dos modelos univariados e a respectiva dificuldade prática dos modelos multivariados. Adicionalmente apresentamos uma vasta literatura de estimação de matrizes de covariância de alta dimensão e, finalmente, apresentamos o trabalho de Fleming et al. (2003) que justifica o uso de dados em alta frequência do ponto de vista econômico.

O trabalho de Markowitz (1952) diz respeito à seleção de portfólio. Trata-se de um estudo em que analiticamente é mostrado que o risco de uma carteira é dado não somente pela média dos riscos individuais, pois precisamos considerar a correlação existente entre os ativos. É dentro deste panorama que sua fundamentação sistematizou o comportamento de diversificação entre os agentes econômicos. Na época em que o trabalho fora divulgado, o risco de um ativo era medido pelo desvio-padrão de seu retorno. Em sua teoria, os retornos dos ativos são entendidos como variáveis aleatórias. O objetivo do investidor é a escolha dos pesos, isto é, o quanto percentualmente será alocado em cada ativo, de forma ótima. Neste contexto teórico a escolha ótima se dá pelo vetor de pesos em que se minimiza a volatilidade, medida pela variância da carteira.

Historicamente, temos em Engle (1982) o trabalho seminal para a estimativa de variância por meio da classe de modelos *Autoregressive Conditional Heteroscedasticity* de ordem q - ARCH(q). Trata-se do primeiro modelo para estimação de volatilidade, cuja motivação foi a observação empírica de propriedades de séries financeiras, tal como a maior importância da modelagem da estrutura de autocorrelação da variância, em relaçâo à estrutura de dependência da média. A especificação do modelo é dada por:

$$
\begin{aligned}
y_t &= u_t \\
u_t &\sim \mathcal{N}(0; \sigma_t^2) \\
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_j u_{t-j}^2
\end{aligned}
$$

Temos que u_t é um termo de perturbação e σ_t^2 é a variância do retorno condicionada na infor-

mação passada. À época, os modelos tradicionais estavam construídos sob hipótese de variância constante. O modelo ARCH surge, portanto, para que se pudesse generalizar uma nova classe de processos estocásticos de média zero, serialmente não autocorrelacionados e com variâncias condicionais não constantes, quando condicionadas no passado, porém constantes ao considerarmos o formato não condicional. A estimação sugerida é pelo método da máxima verossimilhança.

Apesar do desempenho e parcimônia do ARCH(q), percebe-se que este configura uma variância com memória que se interrompe no *lag* q. Isto sugere que um processo que exibe memória longa na variância demandaria a estimação um modelo com maior dimensão. Uma solução natural para este problema foi a especificação do modelo com defasagens da própria variância:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i u_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

Esta foi a solução proposta por Bollerslev (1986), cuja especificação se trata de um ARCH(p) generalizado, isto é, o *Generalized Autoregressive Conditional Heteroskedasticity* - GARCH(p, q). A continuação da literatura concentrou esforços para adaptações deste modelo, de tal forma que o que foi estabelecido fora uma família de modelos GARCH.

Temos em Bollerslev et al. (1988) uma mudança de paradigma com a reestruturação do *Capital Asset Pricing Model* (CAPM). De acordo com o modelo original, o preço de ativos está associado com a incerteza do retorno. O prêmio para induzir o investidor é proporcional ao risco não diversificável, medido pela covariância do retorno do ativo com o retorno do portfólio de mercado. Em um modelo usando dados de título público dos EUA e ações, encontra-se que a covariância condicional é ligeiramente variante no tempo e é determinante significativo do prêmio de risco. Adicionalmente, os *betas* implícitos são também variantes no tempo.

Hansen and Lunde (2005) posteriormente registram a plausibilidade do GAR-CH(1, 1), ao estimarem 330 modelos e evidenciam para dados de câmbio americano e do retorno de ações da *International Business Machines Corporation* (IBM) a melhor capacidade preditiva do modelo e sua estrutura parciomoniosa, computacionalmente eficiente. Entretanto, a extensão natural destes modelos, a versão multivariada, não computou a mesma reputação devido ao revés prático: modelos de volatilidade multivariados possuem dificuldade computacional de implementação devido ao problema dimensional, que mais tarde fora formalizado pela literatura como *curse of dimensionality*. A apresentação dos modelos sugeridos pela literatura é apresentada em Martin et al. (2012).

Dois são os problemas fundamentais da extensão multivariada dos modelos de volatidade: *(i)* a matriz de covariância entre os ativos deve ser definida positiva e *(ii)* o número de parâmetros

desconhecidos que governa as variâncias e covariâncias cresce exponencialmente conforme a dimensão do modelo (número de séries) cresce. Quatro são os modelos tradicionais:

- O modelo VECH, que se trata de uma generalização do GARCH(p, q) para o universo multivariado.
- O modelo BEKK (Engle and Kroner (1995)), que reduz a dimensão computada no VECH e possui a restrição matemática de que a matriz de covariância seja definida positiva.
- O modelo DCC (Engle (2002)), que reduz a dimensão dos modelos BEKK, tornando mais factível o uso de dimensões maiores até então.
- O modelo DECO (Engle and Kelly (2011)), que simplifica a especificação do DCC restringindo como numericamente idênticas as correlações contemporâneas.

É na literatura empírica em que se encontram os maiores esforços para a solução deste problema dimensional, de tal forma que uma série de trabalhos é dedicada à estimação de matrizes de covariância de alta dimensão. Os trabalhos de maior respaldo estão em Fan et al. (2008), Fan et al. (2011a), Fan et al. (2011b), Fan et al. (2012) e Fan et al. (2012). Apresentaremos a seguir alguns dos trabalhos mais notórios, introduzindo com os dois trabalhos que serão de maior referência: Medeiros et al. (2016) e Medeiros et al. (2018).

Medeiros et al. (2016) utlizando dados de 30 ativos da *Dow Jones* abordaram a modelagem e previsão de matrizes de covariância realizadas de alta dimensão a partir da estimação de um VAR penalizado. Os autores consideraram estimadores do tipo *Least Absolute Shrinkage and Selection Operator* (LASSO) para a redução do problema dimensional e para que se forneça forte garantia teórica do mecanismo para a previsão. É demonstrado que se pode prever a matriz de covariância de forma quase tão precisa como no cenário em que se conheça a dinâmica verdadeira que governa a série. Os dados são diários agregados a cada 5 minutos e vão de 2006 até 2012.

Em Medeiros et al. (2018) o problema dimensional foi melhor enfrentado, pois os autores realizam o exercício para todos os 500 ativos do S&P 500 em uma base diária. A ideia é a estimação de uma matriz de covariância esparsa a partir de uma decomposição de fatores por nível de empresa (tamanho, valor de mercado, geração de lucro) e através de restrições setoriais na matriz de covariância residual. O modelo restrito é estimado por *Vector Heterogeneous Autoregressive* (VHAR), penalizando a seleção de variáveis via LASSO. O exercício gera uma melhoria nas estimativas de portfólios de mínima variância.

Seguindo o modelo de Fatores (Chamberlain and Rothschild (1983)), o excesso de retorno de um ativo i, $r_{i,t}$, satisfaz:

$$r_{i,t} = \beta_{i1,t} f_{1,t} + \cdots + \beta_{iK,t} f_{K,t} + \varepsilon_{i,t} \tag{2.1}$$

Em que $f_{1,t}, \ldots, f_{K,t}$ são os excessos de retorno dos K fatores, $\beta_{ik,t}, k = 1, \ldots, K$ são os efeitos marginais e $\varepsilon_{i,t}$ é o termo de erro. Para N ativos, o conjunto de equações pode ser escrito no formato matricial $\mathbf{r}_t = \mathbf{B}_t' \mathbf{f}_t + \boldsymbol{\varepsilon}_t$. É assumido que $E(\boldsymbol{\varepsilon}_t | \mathbf{f}_t) = \mathbf{0}$. Os fatores usados são combinações lineares dos ativos, isto é, formam-se carteiras de ações de curto e de longo prazo onde as empresas são classificadas de acordo com características idiossincráticas e setoriais.

Os autores partem de uma decomposição da matriz de covariância realizada em dois componentes: uma matriz de fatores e uma segunda matriz residual. Seja $\boldsymbol{\Sigma}_t$ a matriz de covariância realizada de retornos no instante t, isto é, $\boldsymbol{\Sigma}_t = \text{cov}(\mathbf{r}_t)$. Baseados nas considerações feitas anteriormente, podemos escrever:

$$\begin{aligned} \boldsymbol{\Sigma}_t &= \text{cov}(\mathbf{B}_t' \mathbf{f}_t) + \text{cov}(\boldsymbol{\varepsilon}_t) \\ &= \mathbf{B}_t' \Sigma_{\mathbf{f},t} \mathbf{B}_t + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t} \end{aligned} \tag{2.2}$$

Ledoit and Wolf (2004a) alertaram, de maneira teórica e prática, sobre os riscos para fins de otimização de portfólio via matriz de covariância amostral. A mesma possui erros de estimativa mais prováveis a perturbarem o otimizador média-variância. Sugerem que se substituia a matriz de covariância amostral através de uma transformação denotada por *shrinkage*. Esta abordagem computa estimativas que antes seriam valores extremos para valores centrais, reduzindo de maneira sistemática os erros de estimação onde mais importa. Estatisticamente isto é obtido através do desafio de se conhecer a intensidade para uso do *shrinkage*, cujo racional é trazido no *paper*.

Por fim, encontramos em Fleming et al. (2003) a justificativa para uso de dados em alta frequência. Baseados na literatura, recente à época, investigam empiricamente se há ganhos de precisão nas estimações de volatilidade diária a partir de dados *intraday*. Os autores analisam o valor econômico da chamada volatilidade realizada em um contexto de tomada de decisão de investidores. Os resultados apontam que ao substituir estimativas via dados diários por estimativas via dados *intraday* há melhoras substanciais: um investidor avesso ao risco estaria disposto a pagar de 50 a 200 pontos-base por ano para capturar os ganhos observados no desempenho do portfólio. Adicionalmente, estes ganhos são robustos para custos de transação.

3 Metodologia

Assim como em Medeiros et al. (2016) e Medeiros et al. (2018), denotamos como $\mathbf{\Sigma}_t$ a matriz de covariância no instante de tempo t. Cada entrada da matriz é potencialmente uma função das entradas passadas, isto é, escrevemos $\mathbf{\Sigma}_t$ como função de $\mathbf{\Sigma}_{t-1}, \ldots, \mathbf{\Sigma}_{t-p}$. Formalmente:

$$\mathbf{y}_t = \boldsymbol{\omega} + \sum_{i=1}^{p} \boldsymbol{\beta}_i \mathbf{y}_{t-i} + \boldsymbol{\epsilon}_t \tag{3.1}$$

$\mathbf{y}_t = \text{vech}(\mathbf{\Sigma}_t)$, em que $\text{vech}(\cdot)$ é a operação de vetorização, transformando $\mathbf{\Sigma}_t$ em um vetor coluna de entradas únicas. $\boldsymbol{\beta}_i$, $i = 1, \ldots, p$, é a matriz que capta a dinâmica entre $\mathbf{\Sigma}_t$ e seu passado, $\boldsymbol{\epsilon}_t$ é um termo de erro e $\boldsymbol{\omega}$ é um vetor de constantes. Trata-se, portanto, de uma estrutura Vetorial Autoregressiva, de ordem p - VAR(p).

Um importante fator a ser considerado é que levaremos em consideração na estimação a transformação logarítmica de $\mathbf{\Sigma}_t$. Isto é importante, pois para as séries de nossa previsão, a transformação exponencial irá nos garantir que o que é previsto é uma estrutura semi definida positiva.

Para uma matriz de covariância de n ativos teremos $n(n+1)/2$ entradas distintas. Um processo do tipo VAR(p) neste caso implicaria em um total de $n(n+1)(p+1)/2$ parâmetros. Ou seja, tanto no cálculo da matriz como na especificação do VAR recorreríamos a problemas de dimensionalidade, pois em ambos exercícios o número de parâmetros tem crescimento exponencial. Adicionalmente esta configuração computa um número maior de potenciais preditores do que observações - alta dimensão, portanto. Tibshirani et al. (2015) mostram que neste cenário a estimação tradicional via OLS implica *overfitting* e há erro de especificação na medida em que a solução não será única.

Em Laurini and Ohashi (2015) temos uma discussão quanto às limitações do uso da matriz de covariância amostral para processos dependentes. De fato, no *intraday* as séries de retorno são mais dependentes se compararmos com agregações temporais maiores. Adicionalmente, devido à ruídos de microestrutura de mercado o preço observado não é o verdadeiro, há erro de medida. Assuma que P_t é o preço *intraday* de um ativo genérico:

$$P_t = P_t^V + \eta_t \tag{3.2}$$

P_t^V é o valor do preço verdadeiro e η_t é o erro de medida. Perceba que:

$$P_t - P_{t-1} = P_t^V + \eta_t - P_{t-1} \tag{3.3}$$

$$\Delta P_t = P_t^V + \eta_t - P_{t-1}^V - \eta_{t-1} \tag{3.4}$$

$$= \Delta P_t^V + \eta_t - \eta_{t-1} \tag{3.5}$$

A primeira diferença da série gera contaminação do tipo MA(1).

Nosso trabalho busca incorporar estratégias para que se enfrentem os problemas em ambas as etapas da análise: anterior à estimação precisamos do cálculo de $\mathbf{\Sigma}_t$ e conforme apresentamos, prosseguir pela covariância amostral não seria a melhor forma. Apresentamos uma estrutura flexível que permite na manipulação dos dados a especificação de diversas alternativas à covariância amostral. Posteriormente formalizamos uma proposta de estimação equação por equação para (3.1) mais adequada para alta dimensão, com modelos tradicionais da literatura de *Machine Learning* (Tibshirani et al. (2001)).

Conforme discutido em Ledoit and Wolf (2004a) o uso da covariância amostral pode ser prejudicial para fins de otimização de carteira. A documentação deste fato estilizado é feita em Jobson and Korkie (1980). A abordagem tradicional de reunir dados históricos de retorno e gerar a matriz de covariância amostral é sensível ao número de ativos, devido a amplos erros de estimação. Isto implica que coeficientes mais extremos na matriz assumem valores maiores devido ao maior erro de previsão embutido. Tal fato em problemas de otimização condiciona pesos maiores nestes coeficientes. Este fenômeno ficou registrado como *error-maximization*, vide Michaud (1989).

Encontramos em Fan et al. (2016) uma revisão de métodos para estimação de matrizes de covariância e de precisão em alta dimensão. Assim como o problema numérico comentado anteriormente, temos adicionalmente que em alta dimensão a matriz se torna singular, não inversível portanto, e a agregação dos erros de estimação conduz a impactos na acurácia. Motivada a discussão, os autores segregam as estratégias de estimação. A prática moderna consiste em estimações consistentes baseadas em regularização. Um pressuposto básico é de que a matriz de interesse é esparsa.

Para realizar esta estimação é feita uma penalização ℓ_1 na função de máxima verossimilhança. Neste tipo de abordagem para que se reduza o viés é possível impor penalização não convexa. Por outro lado, existe uma literatura complementar que se baseia em abordagens baseadas no posto da matriz. São métodos que flexibilizam a distribuição dos dados para um cenário que não seja Gaussiano e com caudas pesadas, o que é recorrente em séries financeiras. Entretanto, a esparsidade nem sempre é empiricamente razoável, sobretudo em dados econô-

micos. Em um contexto de agregados econômicos, por exemplo, é irreal admitir ausência de correlação. Uma extensão natural para contornar este problema é a classe de modelos baseada em esparsidade condicional, em que se utilizam fatores comuns e se admite que a matriz de covariância dos componentes restantes é esparsa.

Para permitirmos flexibilização, iremos nos basear em Ardia et al. (2017). Com esta biblioteca podemos selecionar métodos variados para calcular $\mathbf{\Sigma}_t$. Além de garantir o não uso da covariância amostral, acreditamos que mais de um método pode ser uma forma de garantir robustez às estimações. Abaixo uma descrição das estruturas que utilizaremos:

- Matriz do tipo `ewma`: aqui computamos a matriz de covariância baseada no *Exponential Weighting Moving Average* (EWMA), documentada no tradicional *RiskMetrics*, Morgan et al. (1996). Formalmente é denotar a matriz e uma constante $0 < \lambda < 1$, tal que:

$$\mathbf{\Sigma}_{t+1} = \lambda \mathbf{\Sigma}_t + (1 - \lambda)\mathbf{r}_t \mathbf{r'}_t$$

  Considerando uma amostra de N ativos, $\mathbf{r}_t$ é um vetor $N \times 1$ de retornos no instante t. Por convenção adotamos $\lambda = 0.94$. No entanto, temos em Medeiros et al. (2016) que $\lambda = 0.96$.

- Matriz do tipo `cor`: trata-se de uma média ponderada da matriz de covariância amostral e de um *shrinkage target*, que nesta especificação se trata de uma estrutura de correlação constante entre os pares. Esta é a matriz definida em Ledoit and Wolf (2004a).

  Seja S a matriz de covariância amostral e F um estimador altamente estruturado. A ideia é propor uma combinação convexa $\delta F + (1 - \delta)S$, tal que $0 < \delta < 1$. É uma técnica de *shrinkage* na medida em que "encolhemos" S na direção de F. δ é conhecido como constante de *shrinkage*. Para introduzirmos F, teremos a seguinte notação, assim como no trabalho original: seja y_{it}, $1 \leq i \leq N, 1 \leq t \leq T$. A análise irá assumir que os retornos são independentes e identicamente distribuídos ao longo do tempo e com quartos momentos finitos. Aqui, $\bar{y}_i = T^{-1} \sum_{i=1}^{T} y_{it}$. Defina que Σ é a verdadeira matriz de covariância e S é a matriz de covariância amostral. Teremos que σ_{ij} representa as entradas de Σ e que s_{ij} representa as entradas de S. As correlações populacional e amostral são, respectivamente:

$$\rho_{i,j\,\Sigma} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \tag{3.6}$$

$$\rho_{i,j\,S} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \tag{3.7}$$

Adicionalmente, podemos tomar a média de tais medidas:

$$\bar{\rho}_{\Sigma} = \frac{2}{(N-1)N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho_{i,j\,\Sigma} \tag{3.8}$$

$$\bar{\rho}_{S} = \frac{2}{(N-1)N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho_{i,j\,S} \tag{3.9}$$

A matriz F, definida como *shrinkage target*, terá como entrada na diagonal e fora da diagonal, respectivamente:

$$f_{ii} = s_{ii} \tag{3.10}$$

$$f_{ij} = \bar{\rho}_S \sqrt{s_{ii} s_{jj}} \tag{3.11}$$

Finalmente, encontramos δ a partir de um problema de otimização. Aqui, a função perda a ser otimizada é intuitiva e não demanda a inversa de S: trata-se da distância quadrática entre a verdadeira matriz de covariância e a estimada, baseada na norma de Frobenius. A norma de Frobenius de uma matriz simética $N \times N$ de entradas z_{ij} é definida por $||Z||^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} z_{ij}^2$. O objetivo é encontrar a constante de *shrinkage* que minimiza o valor esperado abaixo:

$$\hat{\delta}^* = \arg\min\{E(||\delta F + (1-\delta)S - \Sigma||)^2\} \tag{3.12}$$

- Matriz do tipo `large`: este é o estimador proposto em Ledoit and Wolf (2004b). Aqui, temos que o *shrinkage target* é dado por um modelo de um fator. O fator é igual à média transversal de todas as variáveis. O peso, também chamado de *shrinkage intensity* é escolhido a partir da minimização da perda quadrática, medida pela norma de Frobenius. Trata-se de um método mais adequado para alta dimensão e que, adicionalmente, é bem condicionado no sentido que a inversão é garantida (matriz não singular) e não somos induzidos a erros de estimação. Trata-se de uma regularização nos autovalores da matriz de tal forma que os autovalores sofrem um direcionamento forçado para valores mais centrais.

  Aqui, o objetivo é encontrar

  $$\Sigma^* = \rho_1 \mathbf{I} + \rho_2 S \tag{3.13}$$

  que minimiza $E\Big(||\Sigma^* - \Sigma||^2\Big)$. $\mathbf{I}$ é a matriz identidade, $S = XX'/n$ é a matriz de covariância amostral, em que X é uma matriz $p \times n$ de n observações independentes e identicamente distribuídas com média zero e variância Σ. No trabalho original, sob amostras finitas, os autores formulam o seguinte problema:

  $$\min_{\rho_1, \rho_2} E\Big(||\Sigma^* - \Sigma||^2\Big) \tag{3.14}$$
  $$s.t. : \Sigma^* = \rho_1 \mathbf{I} + \rho_2 S$$

A solução é dada por:

$$\Sigma^* = \frac{\beta^2}{\delta^2}\mu\mathbf{I} + \frac{\alpha^2}{\delta^2}S \tag{3.15}$$

$$E\Big(||\Sigma^* - \Sigma||^2\Big) = \frac{\alpha^2\beta^2}{\delta^2} \tag{3.16}$$

- $\mu = \langle \Sigma, \mathbf{I} \rangle$.
- $\alpha^2 = ||\Sigma - \mu\mathbf{I}||^2$.
- $\beta^2 = E\Big(||S - \Sigma||^2\Big)$.
- $\delta^2 = E\Big(||S - \mu\mathbf{I}||^2\Big)$.

Nosso objetivo é colocarmos as diversas maneiras de se calcular a matriz para avaliar se há concorrência.

Nossa proposta é a estimação equação por equação de (3.1), a partir da biblioteca desenvolvida por Friedman et al. (2010). Diferentemente do proposto em nossos trabalhos de referência, Medeiros et al. (2016) e Medeiros et al. (2018), vamos estabelecer uma estrutura flexível que a partir do desempenho preditivo de uma classe de modelos irá designar se a estimação deve ser realizada via *Ridge Regression* (Hoerl and Kennard (2000)), LASSO (Tibshirani (1996)) ou *Elastic Net* (Zou and Hastie (2005)).

Seja y_i um elemento de $\mathbf{\Sigma}_t$ e x_i o conjunto que contém $\mathbf{\Sigma}_t$ defasada. A estratégia de estimação consiste em resolver:

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta' x_i) + \lambda\Big[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\Big]\right\} \tag{3.17}$$

Trata-se de um Modelo Linear Geral com uma estrutura de máxima verossimilhança penalizada. Em (3.17), N é o número de observações temporais, λ é um parâmetro de ajuste e $l(\cdot)$ é a contribuição da observação i para a função de log-verossimilhança. Aqui, $0 \leq \alpha \leq 1$, tal que teremos uma especificação do tipo LASSO se $\alpha = 1$, *Ridge* para $\alpha = 0$ e *Elastic Net* para $0 < \alpha < 1$.

O mecanismo *Ridge* é o único método que não seleciona variáveis: um conjunto de covariadas correlacionadas terá coeficientes numericamente próximos. Trata-se de um estimador com fórmula fechada, pois se trata de um problema de programação quadrática. No LASSO existe a possibilidade de seleção de variáveis, reduzindo a dimensão do problema. Por fim, *Elastic Net* é uma estrutura híbrida, pois leva em conta ambas as regularizações, ℓ_1 do LASSO e ℓ_2 da *Ridge Regression*. Os algoritmos da biblioteca `glmnet` usam a descida cíclica de coordenadas, que otimiza sucessivamente a função objetivo sobre cada parâmetro com outros fixos e alterna

repetidamente até a convergência. O pacote também faz uso das regras rígidas para restrição eficiente do conjunto ativo.

Iremos partir de uma especificação que admite erros Gaussianos. Isto faz com que nosso problema seja escrito por:

$$\min_{\beta_0,\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i')^2+\lambda\Big[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1\Big]\right\} \tag{3.18}$$

$\lambda \geq 0$, tradicionalmente, é encontrado por validação cruzada. No entanto, não optamos resolver o problema desta forma. O uso de validação cruzada não é recomendável em problemas que envolvam série de tempo, devido à estrutura de dependência. Por isso partiremos de uma adaptação da biblioteca e encontraremos λ a partir de um critério de informação. Três são critérios utilizados na literatura:

$$\text{AIC} = 2k - 2\ln\hat{L} \tag{3.19}$$

$$\text{BIC} = k\ln n - 2\ln\hat{L} \tag{3.20}$$

$$\text{HQ} = -2\hat{L} + 2k\ln(\ln n) \tag{3.21}$$

Em que n é o número de observações, k é o número de parâmetros e $\hat{L}$ é o valor máximo da função de verossimilhança. Em Hamilton and Press (1994) encontramos a seguinte relação, para $n \geq 16$:

$$\text{BIC} \leq \text{HQ} \leq \text{AIC} \tag{3.22}$$

Ou seja, HQ é um critério de informação intermediário, enquanto que o AIC é o mais flexível e BIC é o mais rigoroso, isto é, penaliza mais severamente a inclusão de variáveis no modelo. Em nosso exercício optamos pelo critério de Hannan-Quinn (HQ). Esta adaptação está implantada na função `ic.glmnet` [1]

Conforme dito anteriormente:

- Se $\alpha = 0$, resolvemos o seguinte problema:

$$\min_{\beta_0,\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i')^2+\lambda||\beta||_2^2/2\right\} \tag{3.23}$$

  Esta é a estrutura de uma *Ridge Regression*. Perceba que se $\lambda = 0$, estaríamos em um problema tradicional de Mínimos Quadrados. Adicionalmente, é possível demonstrar que se $\hat{\boldsymbol{\beta}}$ resolve o problema acima, então $\lim_{\lambda\to\infty}\hat{\boldsymbol{\beta}} = \mathbf{0}$.

[1] Disponível em https://github.com/gabrielrvsc/HDeconometrics.

- Se $\alpha = 1$, resolvemos o seguinte problema:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i')^2 + \lambda ||\beta||_1 \right\} \tag{3.24}$$

  Esta é a estrutura do método LASSO, que se trata de um método de *shirinkage*. Quando estamos diante de um cenário de valores elevados, podemos via LASSO selecionar variáveis e, adicionalmente, produzir soluções esparsas. Ou seja, é possível realizarmos uma redução de dimensão no problema, encontrando uma matriz de coeficientes que utiliza menos *features* (preditores) do que uma solução tradicional de Mínimos Quadrados ou via *Ridge*. Finalmente, como podemos ver em Tibshirani et al. (2015), em amostras finitas temos boa performance e contornamos o problema clássico do *trade-off* de viés e variância.

- Se $0 < \alpha < 1$, resolvemos o seguinte problema:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i')^2 + \lambda \Big[(1-\alpha)||\beta||_2^2/2 + \alpha ||\beta||_1 \Big] \right\} \tag{3.25}$$

  Esta é a estrutura do *Elastic Net*, conforme Zou and Hastie (2005). Trata-se de uma formulação híbrida, em que se computa tanto penalização ℓ_1 como ℓ_2. Tal formato é benéfico e surge na literatura como resposta para algumas limitações teóricas que são consequência do LASSO: *(i)* o LASSO em um cenário de alta dimensão, isto é, mais preditores (k) do que observações (n), tem a capacidade de selecionar no máximo n variáveis, *(ii)* para variáveis que, aos pares, são altamente correlacionadas, o LASSO somente irá selecionar uma delas e *(iii)* em alta dimensão, para *features* altamente correlacionadas, o desempenho preditivo na *Ridge Regression* é superior ao LASSO.

Uma vez que nossa estrutura é flexível, demandaremos um método de escolha para o modelo a partir de algum critério. Neste trabalho vamos propor o algoritmo *Model Confidence Set* (MCS), proposto em Hansen et al. (2011). A ideia é partirmos de um *grid* discreto, tal que $\alpha_{\text{grid}} = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$. Isto é, iremos por equação, estimar (3.18), tomar um $\alpha \in \alpha_{\text{grid}}$ e avaliar o modelo "vencedor" a partir de um desempenho preditivo. Separamos 20% da amostra para estabelecer um *training set* e *test set*. Isto é, 20% dos dados foram excluídos da amostra para computarmos o desempenho preditivo.

O MCS se trata da construção de um conjunto de modelos tal que o melhor modelo, do ponto de vista preditivo, seja elemento deste conjunto dado um nível de confiança. É um algoritmo que testa de forma sequencial a hipótese nula de que os modelos possuem acurácia idêntica. Baseado em um critério de eliminação, seleciona o melhor modelo ou conjunto de modelos. É, portanto, uma maneira inferencial de seleção de modelos, pois se baseia em métodos globais, diferentemente de avaliarmos medidas pontuais.

Para implementarmos o MCS iremos nos basear em Bernardi and Catania (2014). Seja Y_t nossa série temporal no instante de tempo t e $\hat{Y}_{i,t}$ o *fit* do modelo i, em t. O primeiro passo é definir uma função perda $\ell_{i,t}$ que está associada ao i-ésimo modelo, tal que:

$$\ell_{i,t} = \ell(Y_t, \hat{Y}_{i,t}) \tag{3.26}$$

O procedimento é iniciado a partir de um conjunto $M = \hat{M}^0$ de modelos de dimensão m. Para um dado nível de confiança teremos o retorno de um conjunto menor, $\hat{M}^*$, que é o *Superior Set of Models* (SSM), de dimensão $m^* \leq m$. Podemos encontrar em SSM um conjunto de modelos equivalentes, tal que o cenário ideal é de $m^* = 1$. Vamos definir como $d_{ij,t}$ a diferença entre $\ell(\cdot)$ avaliada nos modelos i e j:

$$d_{ij,t} = \ell_{i,t} - \ell_{j,t} \tag{3.27}$$

$$i, j = 1, \ldots, m \tag{3.28}$$

$$t = 1, \ldots, n \tag{3.29}$$

Assuma que

$$d_{i,t} = \frac{1}{(m-1)} \sum_{j \in M} d_{ij,t}$$

é a perda associada ao modelo i relativamente a qualquer outro modelo j no instante t. A hipótese de igualdade de acurácia pode ser formulada por:

$$\mathrm{H}_0 : E(d_i) = 0, \forall i, i = 1, \ldots, m \tag{3.30}$$

$$\mathrm{H}_A : E(d_i) \neq 0, \exists i, i = 1, \ldots, m \tag{3.31}$$

Aqui, $E(d_i)$ é assumida como finita e não dependente no tempo. Para prosseguir o teste, duas estatísticas são construídas:

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\hat{\mathrm{var}}(\bar{d}_{ij})}} \tag{3.32}$$

$$t_i = \frac{\bar{d}_i}{\sqrt{\hat{\mathrm{var}}(\bar{d}_i)}} \tag{3.33}$$

Em que $\bar{d}_i = (m-1)^{-1} \sum_{j \in M} \bar{d}_{ij}$ é a perda do i-ésimo modelo comparada com a perda média entre os modelos de M; $\bar{d}_{ij} = m^{-1} \sum_{t=1}^{m} d_{ij,t}$ mede a perda amostral entre o i-ésimo e j-ésimo modelo, $\hat{\mathrm{var}}(\bar{d}_i)$ e $\hat{\mathrm{var}}(\bar{d}_{ij})$ são estimativas por *bootstrap* das variâncias de $\bar{d}_i$ e $\bar{d}_{ij}$, respectivamente.

Duas estatísticas são computadas para testar a hipótese nula de igual capacidade preditiva: $T_{R,M}$ e $T_{\max,M}$, onde:

$$T_{R,M} = \max\{|t_{ij}|\} \tag{3.34}$$

$$T_{\max,M} = \max\{t_i\} \tag{3.35}$$

O algoritmo se baseia na seguinte regra de eliminação:

$$e_{R,M} = \arg\max_{i} \left\{ \sup \frac{\bar{d}_{ij}}{\sqrt{\hat{\text{var}}(\bar{d}_{ij})}} \right\} \tag{3.36}$$

$$e_{\max,M} = \arg\max_{i \in M} \left\{ \frac{\bar{d}_{i}}{\sqrt{\hat{\text{var}}(\bar{d}_{i})}} \right\} \tag{3.37}$$

Se não conseguimos rejeitar H_0, o algoritmo é finalizado e concluímos que todos modelos pertencem ao MCS. Na direção contrária temos a eliminação do modelo com pior performance e o algoritmo reinicia a execução com $M - 1$ modelos.

Para a execução do algoritmo em nosso exercício, estamos considerando uma matriz de perda modular, isto é: seja $\hat{\sigma}_{t+1}$ a nossa previsão um passo à frente do desvio padrão de um ativo genérico e seja σ_{t+1} o desvio padrão um passo à frente observado. A matriz de perda irá calcular, por observação, $|\hat{\sigma}_{t+1} - \sigma_{t+1}|$. Funções deste tipo são menos sensíveis a *outliers*. Adicionalmente, temos um nível de significância de 20%, a estatística TR e 2000 replicações por *bootstrap*.

4 Aplicação Empírica

4.1 Dados

Utilizamos dados intradiários de negociação dos ativos da B3 (Brasil, Bolsa e Balcão), a bolsa de valores oficial do Brasil. Para a coleta, nos baseamos na biblioteca desenvolvida por Perlin and Ramos (2016). Dados de negociação são mais recorrentes na literatura, uma vez que para o exercício proposto não há a necessidade da construção de livros de ordem. O horizonte temporal de nossa pesquisa é 02/07/2018 até 05/02/2020 (393 dias) e os retornos estão agregados de 5 em 5 minutos.

As transações selecionadas vão desde às 10:15 até às 16:40. Esta janela de horário foi escolhida pois nos baseamos no horário de funcionamento da B3 das 10:00 às 16:55, que é o horário limite para negociação. Das 16:55 às 17:00 ocorre o período de *closing call*. Neste período não necessariamente todos os papeis da B3 estão à disposição para o leilão de fechamento. Nestes horários está permitida a compra e venda de ações, contratos futuros e derivativos.

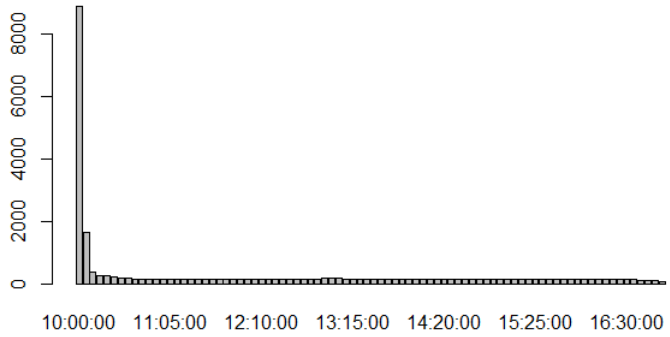


Figura 1 – *O gráfico agrupa a amostra por horário de transação e realiza a contagem de dados faltantes. Aqui a amostra é completa, desde às 10:00 até às 16:55. Os horários estão no eixo horizontal e a quantidade de dados faltantes no eixo vertical.*

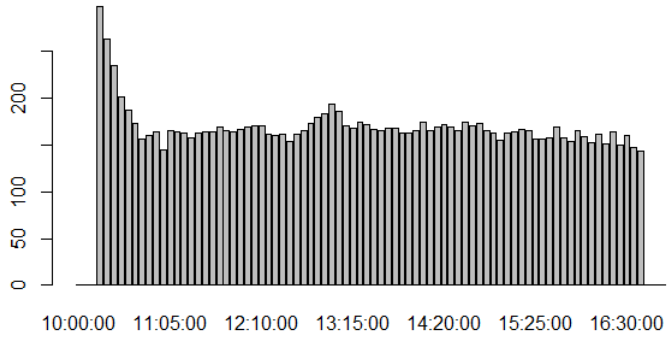


Figura 2 – *O gráfico agrupa a amostra por horário de transação e realiza a contagem de dados faltantes. Aqui temos a amostra filtrada, desde às 10:15 até às 16:40. Os horários estão no eixo horizontal e a quantidade de dados faltantes no eixo vertical.*

Pesquisamos todo o horário de funcionamento da B3 e isto irá auxiliar na leitura dos dois últimos gráficos. No período das 09:30 às 09:45 ocorre o período de cancelamento: trata-se do período onde a bolsa deixa habilitado o cancelamento de ofertas residuais de sessões anteriores, por exemplo. Das 09:45 às 10:00 ocorre a chamada pré-abertura: apenas são registradas as ofertas de compra e de venda, porém somente no período de negociação é feita a transação. Trata-se de um período que fornece base informacional para o preço dos ativos. Das 10:00 às 16:55 temos as negociações de fato, a *call* de fechamento das 16:55 até 17:00 e, por fim, o período de *after market* das 17:30 às 18:00.

Este recorte na amostra, 15 minutos após o período inicial e 15 minutos antes do período final tem o benefício de conter uma quantidade de dados faltantes reduzida. De fato, períodos de abertura e fechamento de mercado são mais problemáticos, conforme a figura (1) evidencia. Trata-se de uma contagem simples, onde agrupamos a base por horário de transação e somamos a quantidade de valores ausentes.

Inicialmente somamos a quantidade de negociações por ativo em todo o período. Em seguida ordenamos nossa base pela quantidade de negociação e selecionamos os 50 ativos mais negociados. Nas discussões a seguir buscamos explorar mais a amostra, mostrando que tipos de ativos sobraram em nosso recorte, o quão distante está o primeiro do último ativo e outras informações adicionais.

Código do Ativo	**Quantidade**	%
3	36	72
4	9	18
5	2	4
6	1	2
11	2	4

Tabela 1 – Amostra classificada de acordo com o código do ativo.

Os ativos na B3 são classificados dentro de um certo padrão de código. O número ao final do ativo irá trazer as informações necessárias. Em nossa amostra temos que 72% dos ativos são de final 3 (VALE3, por exemplo), isto é, ações ordinárias que dão ao acionista direito a voto em caso de assembleias; 18% dos ativos são de final 4 (PETR4, por exemplo), que são ações preferenciais cujo impacto é a garantia da preferência na distribuição de dividendos e na distribuição de juros sobre capital próprio; 4% da amostra é de final 5 (USIM5, por exemplo), que também são ações preferenciais, porém de classe A; 2% da amostra é de final 6 (ELET6, por exemplo), que são ações preferenciais de classe B e, finalmente, 4% da amostra é de final 11 (BOVA11, por exemplo), indicando que o ativo é um *Brazilian Deposit Receipts* (BDR) - certificados de depósitos de ações de companhias no exterior - ou uma *unit*, que são ativos compostos por mais de uma ação, como fundos de índices. Em nossos dados somente 3 pares de ativos refletem a mesma empresa: BBDC3 e BBDC4, ELET3 e ELET6, PETR3 e PETR4.

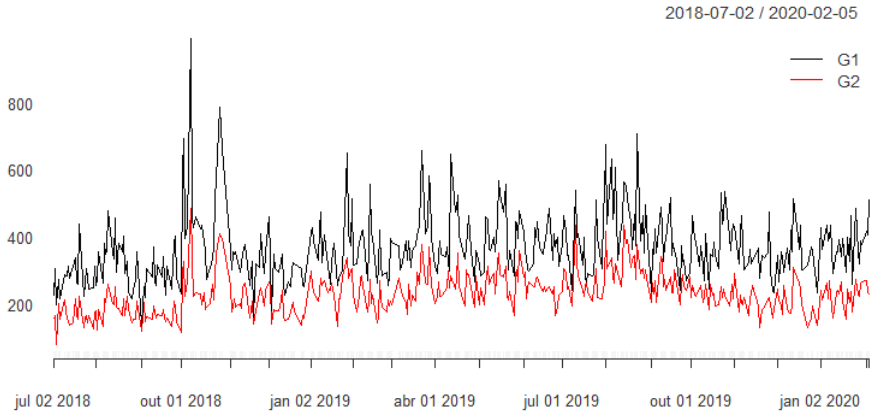


Figura 3 – *Comparação da quantidade média diária de negociações entre ativos. G1 denota o grupo que corresponde aos 10 primeiros ativos de nossa lista. G2 abrange desde o 11º ativo até o 20º. A linha preta mostra a quantidade média de negociações diárias para os ativos pertencentes ao G1. A linha vermelha mostra a quantidade média de negociações diárias para os ativos pertencentes ao G2.*

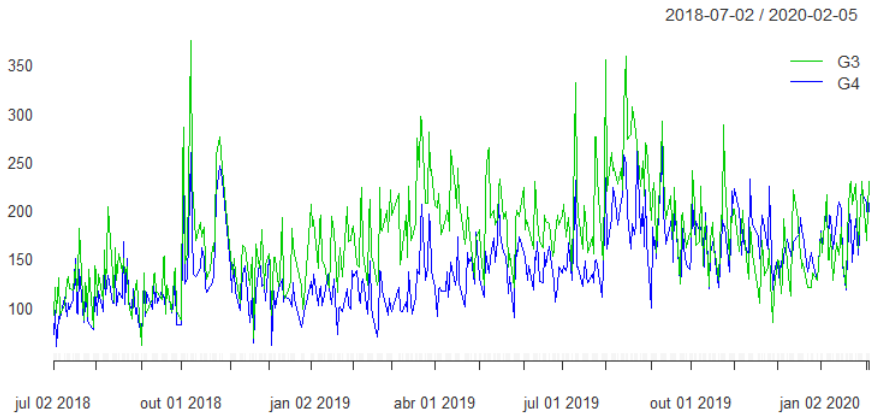


Figura 4 – *Comparação da quantidade média diária de negociações entre ativos. G3 abrange desde o 21º ativo até o 30º. G4 abrange desde o 31º ativo até o 40º. A linha verde mostra a quantidade média de negociações diárias para os ativos pertencentes ao G3. A linha azul mostra a quantidade média de negociações diárias para os ativos pertencentes ao G4.*

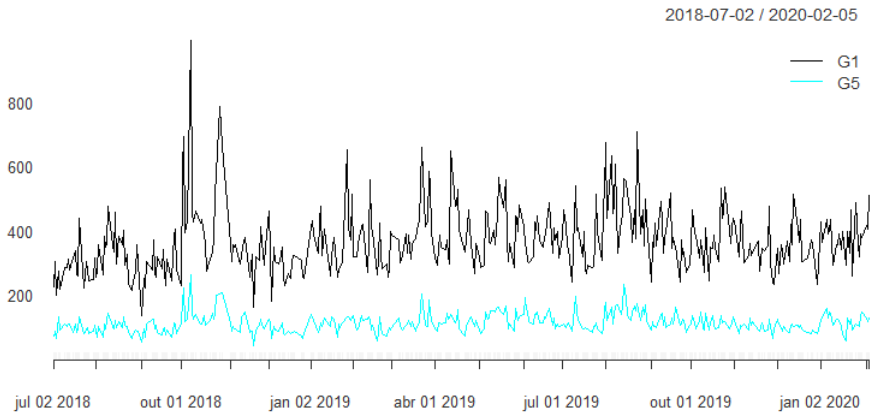


Figura 5 – *Comparação da quantidade média diária de negociações entre ativos. G1 denota o grupo que corresponde aos 10 primeiros ativos de nossa lista. G5 abrange desde o 41º ativo até o 50º. A linha preta mostra a quantidade média de negociações diárias para os ativos pertencentes ao G1. A linha verde-água mostra a quantidade média de negociações diárias para os ativos pertencentes ao G5.*

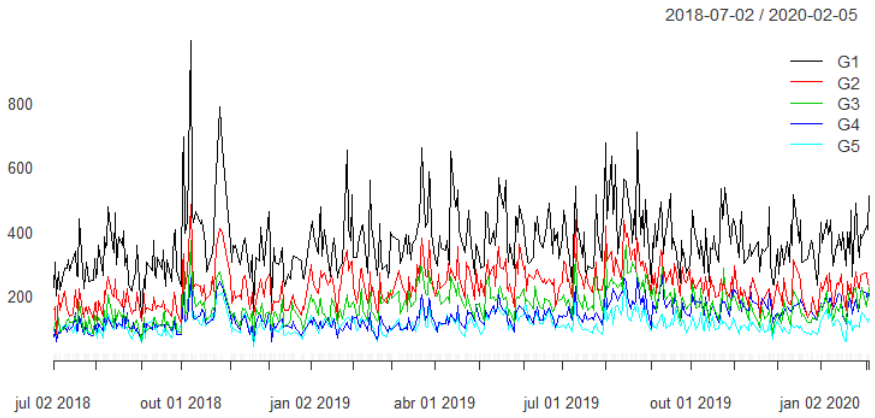


Figura 6 – *Comparação da quantidade média diária de negociações entre ativos. Todos os grupos.*

Em mais um exercício para investigar nossos dados, questionamos o quão distante estão os ativos que lideram a lista em relação aos ativos de posição mais inferior. Nos gráficos (3), (4), (5) e (6) temos uma possível leitura para esta pergunta. Apesar de não formal, percebe-se comportamento similar na trajetória das curvas entre grupos não tão distantes entre si, como por exemplo G1 e G2. Ao compararmos os grupos G1 e G5, notamos uma distância maior, porém não aberrante. Isto simultaneamente traz evidência que se trata de uma amostra heterogênea, mas que adicionalmente nos permite trabalhar com a quantidade de ativos desejada, uma vez que são ativos presentes em toda a amostra e em uma quantidade relevante. Nas próximas discussões iremos apresentar a estratégia para lidar com dados faltantes, no entanto a visualização destas figuras nos assegura, por exemplo, que não estamos lidando com grupos onde há dados faltantes em um período muito longo.

O mercado de capitais brasileiro apesar de continuamente se desenvolver, ainda possui baixa liquidez. O problema de liquidez se edivencia quando trabalhamos com a base de dados completa: detectamos anomalias em alguns ativos que possuem uma quantidade de negociação bastante reduzida. Mesmo com o uso de interpolação ou até mesmo técnicas mais avançadas para preenchimento deste conjunto de dados faltantes, lidar com ativos poucos líquidos é problemático. Abaixo um levantamento, por setor, dos ativos em nossa amostra:

Setor	**Ativos**	**(%)**
Financeiro	10	20%
Consumo Cíclico	9	18%
Materiais Básicos	8	16%
Bens Industriais	7	14%
Utilidade Pública	5	10%
Petróleo, Gás e Biocombustíveis	4	8%
Consumo não Cíclico	4	8%
Comunicações	2	4%
Saúde	1	2%

Tabela 2 – Amostra classificada por setor econômico. A classificação por setor é disponibilizada pela própria B3.

Para o preenchimento dos dados faltantes que restaram na amostra após o filtro optamos pelo uso de *splines* cúbicos. A interpolação por *splines* é uma abordagem pela qual o interpolante é um tipo particular de polinômio por partes, denominado *spline*. No lugar de ajustarmos um único polinômio de grau alto para todos os valores, ajustamos polinômios de baixo grau, no nosso caso grau 3, em pequenos subconjuntos dos valores. Trata-se de um método preferível à interpolação polinomial porque o erro de interpolação pode ser reduzido. Sejam $x_1 < x_2 < \cdots < x_n$ os pontos de interpolação. Um *spline* cúbico é uma função $s(x)$, definida no intervalo $[x_1, x_n]$ com as seguintes propriedades:

- $s(x), s'(x)$ e $s''(x)$ são funções contínuas no intervalo (x_1, x_n).
- Em cada subintervalo $[x_i, x_{i+1}]$, $s(x)$ é um polinômio cúbico tal que $s(x_i) = f_i = f(x_i)$ para $i = 1, \ldots, n$.

Em nosso exercício a interpolação é feita de forma univariada, em todas as séries de retorno. Via filtro de Kalman, séries de retorno são preenchidas com zeros e ao tentar interpolar as séries de preço encontramos problemas de convergência. Adicionalmente, em termos de performance computacional, destacamos a eficiência do preenchimento via *spline* cúbico.

Feita a seleção dos ativos, o passo seguinte é a construção da matriz de covariância diária. Conforme discutido na seção anterior, optamos por flexibilizar a escolha da matriz para que possamos incorporar métodos alternativos e evitar o uso da matriz de covariância amostral.

Como ilustração, tomemos o seguinte exemplo que demonstra a construção do *dataset* final. Para simplificação, vamos assumir um exemplo ingênuo com dois ativos e dois dias.

Dia	**Horário**	**Retorno Ativo 1**	**Retorno Ativo 2**
Dia 1	h_1	$r_{1_{1,1}}$	$r_{2_{1,1}}$
	h_2	$r_{1_{1,2}}$	$r_{2_{1,2}}$
	$\vdots$	$\vdots$	$\vdots$
	h_k	$r_{1_{1,k}}$	$r_{2_{1,k}}$
Dia 2	h_1	$r_{1_{2,1}}$	$r_{2_{2,1}}$
	h_2	$r_{1_{2,2}}$	$r_{2_{2,2}}$
	$\vdots$	$\vdots$	$\vdots$
	h_k	$r_{1_{2,k}}$	$r_{2_{2,k}}$

Tabela 3 – Tabulação da base de dados - etapa inicial.

Aqui, $h_1, \ldots, h_k$ são os horários de cada negociação. $r_{n_{t,k}}$ é o retorno do ativo n, no dia t e horário k. Esta é nossa amostra na etapa de coleta. Considerando dois ativos, uma matriz de covariância $\mathbf{\Sigma}_t$ terá o seguinte formato, no dia t:

$$\mathbf{\Sigma}_t = \begin{bmatrix} \sigma_{1,1_t} & \sigma_{1,2_t} \\ \sigma_{2,1_t} & \sigma_{2,2_t} \end{bmatrix}$$

σ_{m,n_t} é a covariância entre os ativos m e n, no dia t. O dado *intraday* é utilizado para a construção de uma *proxy* diária à volatilidade do ativo. Tomaremos vech($\mathbf{\Sigma}_t$) transposto para que possamos ter um vetor linha e adicionalmente eliminar entradas idênticas, pois se trata de uma matriz simétrica. O algoritmo trabalha com a exclusão do triângulo inferior da matriz. Os dados em sua versão final possuem a seguinte configuração:

	$\sigma_{1,1}$	$\sigma_{1,2}$	$\sigma_{2,2}$
Dia 1	$\sigma_{1,1_1}$	$\sigma_{1,2_1}$	$\sigma_{2,2_1}$
Dia 2	$\sigma_{1,1_2}$	$\sigma_{1,2_2}$	$\sigma_{2,2_2}$

Tabela 4 – *Tabulação da base de dados - etapa final.*

Generalizando para uma amostra de j dias e n ativos, nosso *dataset* terá j linhas e $n(n+1)/2$ colunas. Por fim, os dados estão todos padronizados.

4.2 Estimação

Neste trabalho, construímos a partir dos dados intradiários matrizes de covariância incondicional com frequência diária para dados do mercado brasileiro. A ideia é para cada componente da matriz propor uma estimação, equação por equação, tal que uma série de covariância ou variância será explicada pelo seu passado e o passado de todos os outros elementos da matriz. Avaliamos, em um exercício de previsão dentro da amostra, a diferença de resultado quando mudamos a forma do cálculo de matriz de covariância. Separamos 20% da amostra para a construção de um conjunto de teste e na base de treino estimamos um modelo com 1 defasagem, isto é, escrevemos $\mathbf{\Sigma}_t$ como função de $\mathbf{\Sigma}_{t-1}$.

Computamos os resultados para todos os formatos de cálculo de matriz de covariância que apresentamos, de tal forma que o exercício proposto consiste no seguinte algoritmo: para os dados de retornos, já preenchidos quando faltantes, calculamos a matriz de covariância incondicional diária, tal que o resultado final é uma tabela na qual as colunas são os elementos da matriz e as linhas são os dias de nossa amostra. A estimação é feita com a adaptação do `glmnet`, onde o critério de penalização é escolhido via critério de informação HQ. Para cada série estimamos 11 modelos, tal que cada modelo possui um valor de $\alpha \in \alpha_{\text{grid}}$. Estimados os modelos, construímos com dados fora da amostra a tabela com a função perda modular. Estes serão os dados para a execução do MCS. Após a execução do MCS, salvamos o modelo vencedor a partir do nível de significância estabelecido e elegemos o melhor modelo, ou conjunto de modelos.

Optamos na apresentação dos resultados em três formatos: *(i)* faremos, assim como em Medeiros et al. (2016), nossa principal referência, a seleção de variáveis por setor econômico. Este tipo de leitura permite identificarmos efeitos intra e entre setores. Adicionalmente, mostraremos *(ii)* quais os modelos elegíveis por setor econômico, a fim de identificar a existência de algum padrão. Faremos isto em duas óticas: na ótica de elementos dentro da diagonal principal da matriz de covariância, variâncias, e elementos de covariância, fora da diagonal principal. Uma importante observação a ser feita é a dificuldade de classificação de covariância entre ativos diferentes. Decidimos replicar a estratégia de nossa referência e pontuar que o processo de covariância entre um ativo a e b pertence, mutualmente, aos setores econômicos a e b, para setor de $a \neq$ setor de b. Por fim, iremos discutir os resultados olhando para as estimações no mundo completo com 50 ativos e em um formato compacto com apenas 10 ativos, os 10 ativos mais negociados no período no caso.

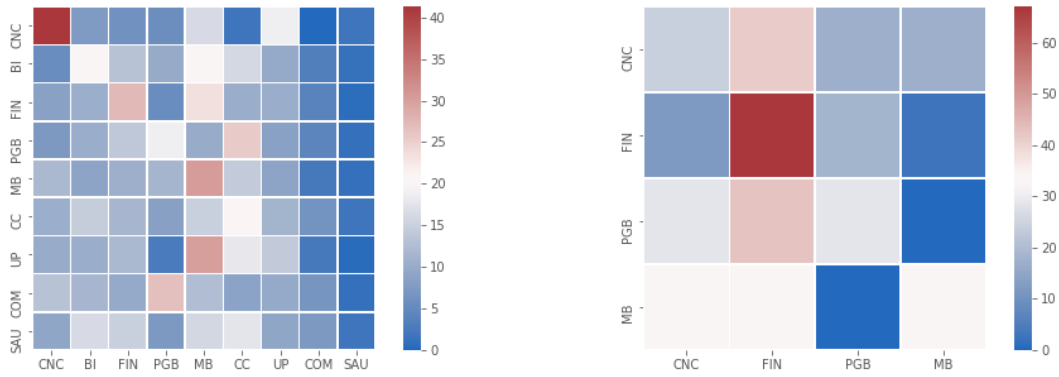
Usamos aqui como notação as iniciais dos setores para a construção dos gráficos: CNC denota consumo não cíclico, BI denota bens industriais, FIN setor financeiro, PGB petróleo,

gás e biocombustíveis, MB materiais básicos, CC consumo cíclico, UP utilidade pública, COM comunicações e, finalmente, SAU denota saúde.

Inicialmente apresentamos abaixo como se dá a seleção das variáveis para as equações de variâncias. Temos nas figuras (7), (8), (9), (13), (14) e (15) os resultados variando a forma que se calcula a matriz incondicional diária: especificação `lw`, `cor` e `ewma`, respectivamente. A ideia é investigar no exercício preditivo como setorialmente os modelos, dentro do MCS, realizam a seleção das variâncias e covariâncias defasadas. Nos gráficos à esquerda temos a amostra completa com 50 ativos e com 10 ativos na figura à direita.

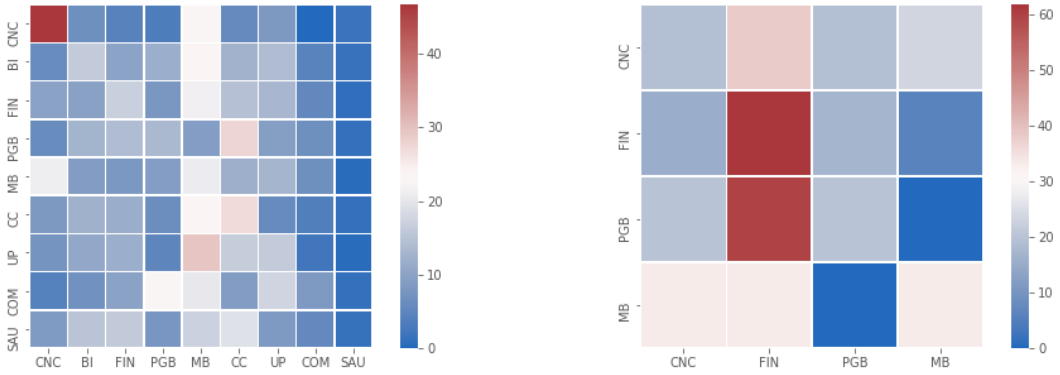
Posteriormente, para cada série de variância fazemos a classificação para os modelos selecionados no MCS. Isto nos permite investigar se há ou não ganho preditivo quando se altera a forma de calcular matriz de covariância e quando se muda o setor econômico em questão. Estes resultados estão compostos nas figuras (10), (11), (12), (16), (17) e (18). O padrão aqui se repete: temos à esquerda a amostra completa e parcial à direita, com 50 e 10 ativos, respectivamente.

4.2.0.1 Elementos da diagonal principal



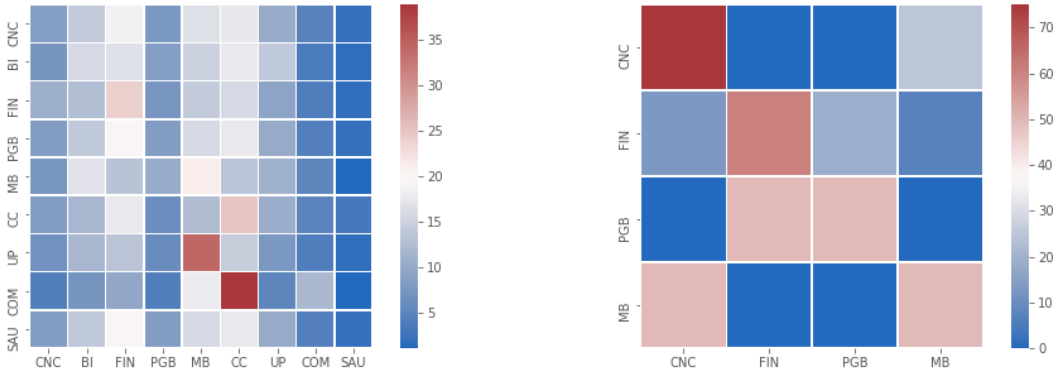
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 7 – *Resultado da estimação para elementos da diagonal. O heatmap indica o quanto percentualmente, em média, para variâncias do setor na coluna foram selecionadas variâncias e covariâncias do setor nas linhas. Matriz de covariância incondicional calculada no formato* `lw`.



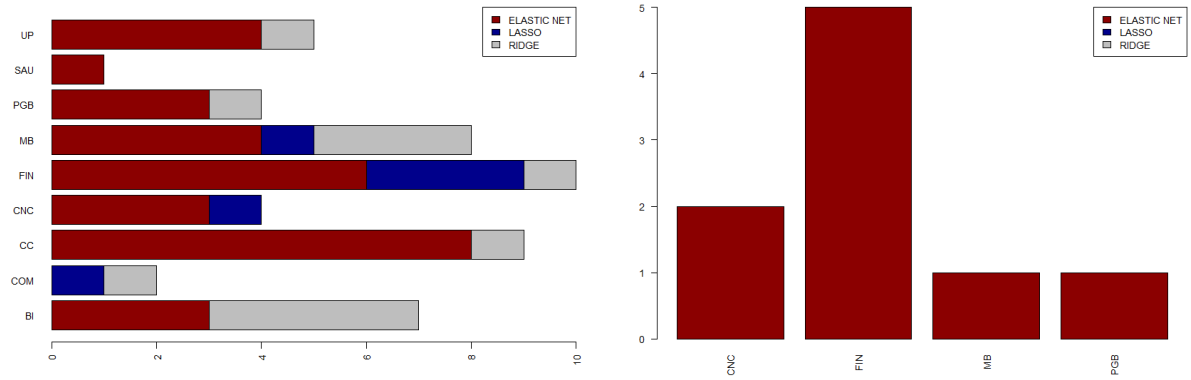
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 8 – *Resultado da estimação para elementos da diagonal. O heatmap indica o quanto percentualmente, em média, para variâncias do setor na coluna foram selecionadas variâncias e covariâncias do setor nas linhas. Matriz de covariância incondicional calculada no formato* `cor`.



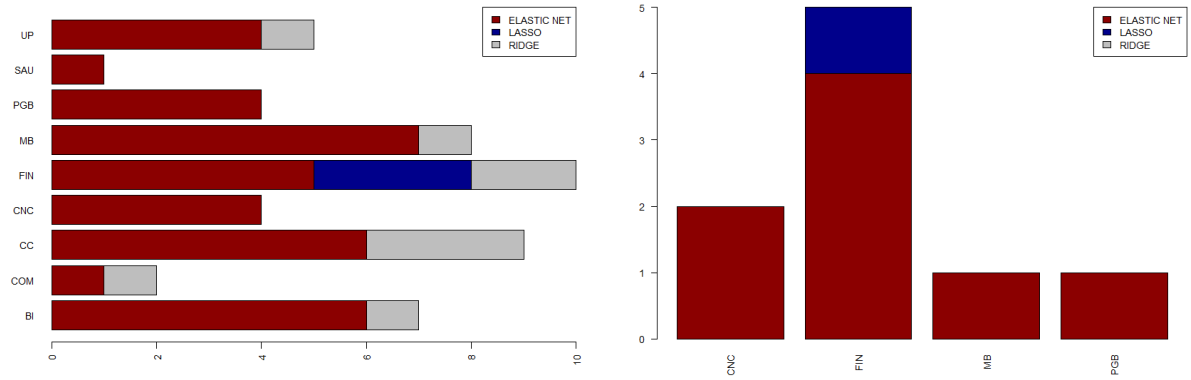
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 9 – *Resultado da estimação para elementos da diagonal. O heatmap indica o quanto percentualmente, em média, para variâncias do setor na coluna foram selecionadas variâncias e covariâncias do setor nas linhas. Matriz de covariância incondicional calculada no formato* `ewma`.



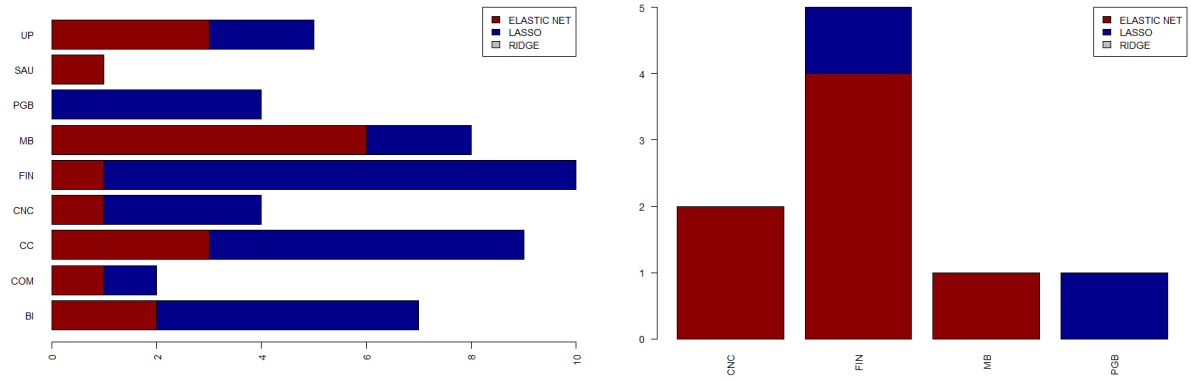
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 10 – *Tipos de estruturas selecionadas dentro do MCS. Resultado segregado por setor econômico. Matriz de covariância incondicional calculada no formato* `lw`.



(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 11 – *Tipos de estruturas selecionadas dentro do MCS. Resultado segregado por setor econômico. Matriz de covariância incondicional calculada no formato* `cor`.



(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

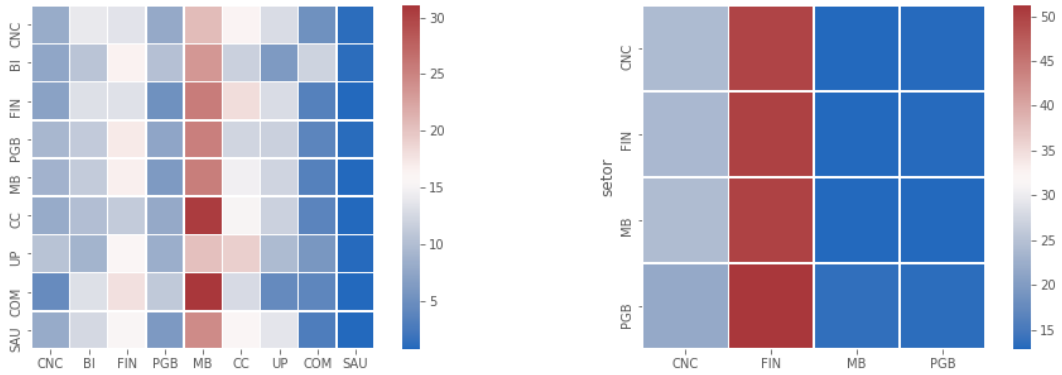
Figura 12 – *Tipos de estruturas selecionadas dentro do MCS. Resultado segregado por setor econômico. Matriz de covariância incondicional calculada no formato* `ewma`.

É possível identificar nos *heatmaps* que existe nas diagonais uma maior intensidade, revelando que para processos de variância existe uma dinâmica intra setorial. Nota-se tal fenômeno tanto na visão com 50 ativos como na versão do exercício com apenas 10 ativos. Por sinal, ambos os formatos revelam forte presença do setor financeiro. Neste ponto de vista não se encontram diferenças relevantes quando variamos a matriz nos formatos `lw` e `cor`. Porém é possível perceber no formato `ewma`, na aplicação com 10 ativos, uma maior presença de variáveis selecionadas dentro e fora do mesmo setor.

Nos formatos `lw` e `cor` notamos que migrando do cenário de 10 para 50 ativos se manifesta dentro do MCS a estrutura de regressão *Ridge*. Dentro do nosso contexto isto pode ser interpretado como uma equivalência preditiva entre as especificações de regressão. Apesar disto significar que não existe performance preditiva maior nos métodos com encolhimento, trata-se de uma evidência de que, por outro lado, não há perda por se escolher o formato mais parcimonioso: como estamos falando da dinâmica de uma matriz de covariância, mais ativos traria um crescimento exponencial no número de séries e parâmetros a serem estimados.

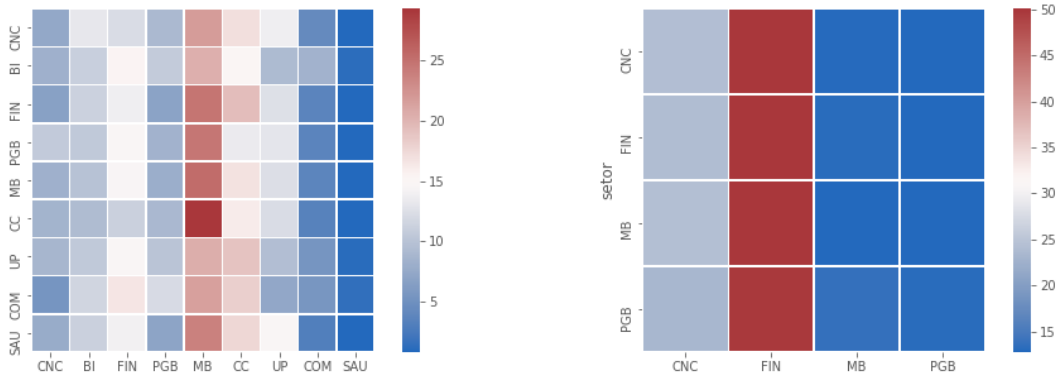
Em linhas gerais percebemos que o resultado é sensível ao formato em que se calcula a matriz de covariância é que com mais ativos torna-se maior a presença de modelos que não selecionam variáveis.

4.2.0.2 Elementos fora da diagonal principal



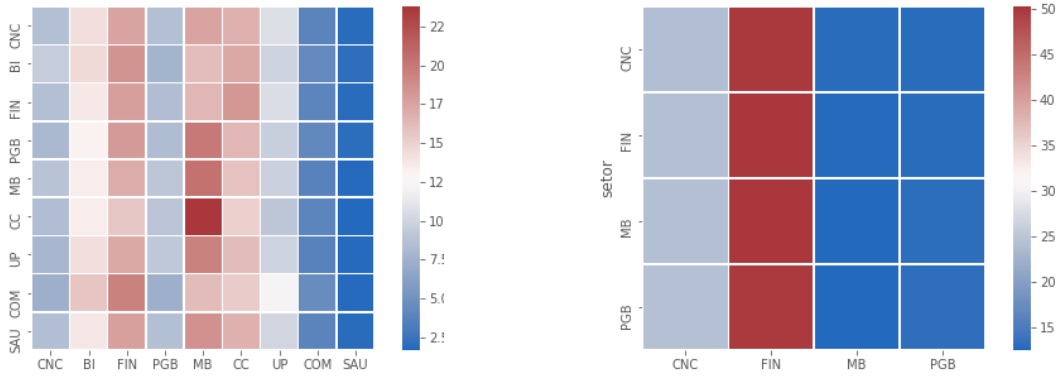
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 13 – *Resultado da estimação para elementos da diagonal. O heatmap indica o quanto percentualmente, em média, para covariâncias do setor na coluna foram selecionadas variâncias e covariâncias do setor nas linhas. Matriz de covariância incondicional calculada no formato* `lw`.



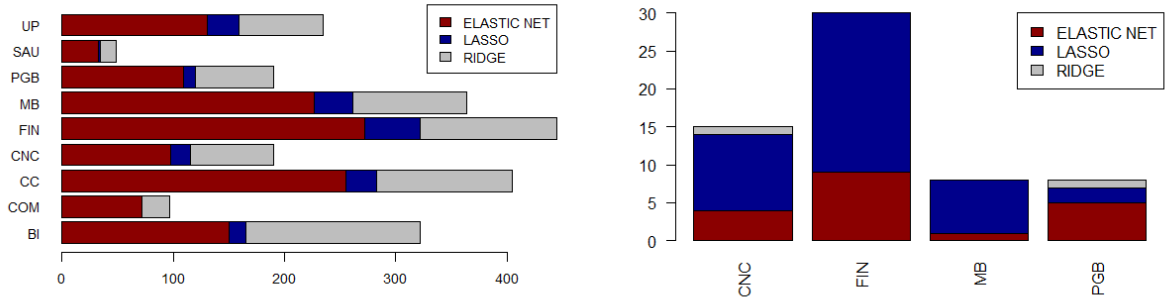
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 14 – *Resultado da estimação para elementos da diagonal. O heatmap indica o quanto percentualmente, em média, para covariâncias do setor na coluna foram selecionadas variâncias e covariâncias do setor nas linhas. Matriz de covariância incondicional calculada no formato* `cor`.



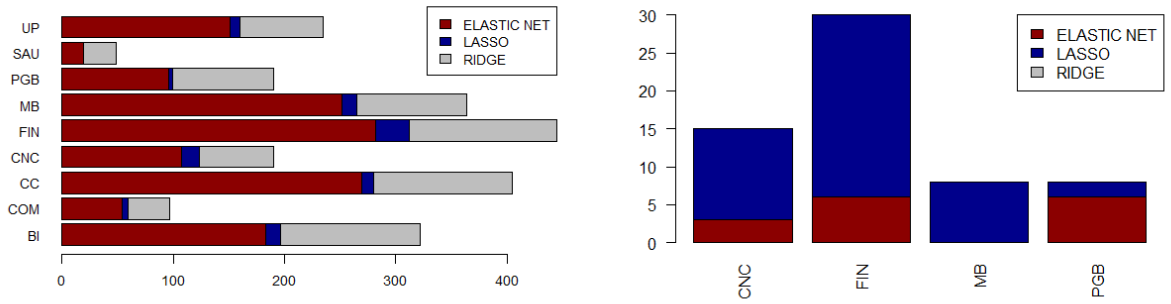
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 15 – *Resultado da estimação para elementos da diagonal. O heatmap indica o quanto percentualmente, em média, para covariâncias do setor na coluna foram selecionadas variâncias e covariâncias do setor nas linhas. Matriz de covariância incondicional calculada no formato* `ewma`.



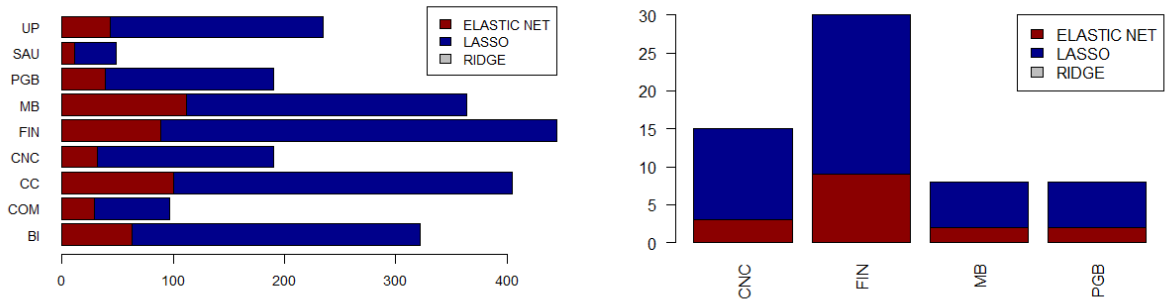
(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 16 – *Tipos de estruturas selecionadas dentro do MCS. Resultado segregado por setor econômico. Matriz de covariância incondicional calculada no formato* `lw`.



(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 17 – *Tipos de estruturas selecionadas dentro do MCS. Resultado segregado por setor econômico. Matriz de covariância incondicional calculada no formato* `cor`.



(a) *Resultado da estimação usando 50 ativos.* (b) *Resultado da estimação usando 10 ativos.*

Figura 18 – *Tipos de estruturas selecionadas dentro do MCS. Resultado segregado por setor econômico. Matriz de covariância incondicional calculada no formato* `ewma`.

Notamos dentro do universo com menos ativos que não há diferença de resultado quando levamos em conta as diferentes formas de se calcular matriz de covariância. Assim como relatamos para os processos de variância, notamos aqui novamente presença relevante do setor financeiro, porém o destaque para processos fora da diagonal está no setor de materiais básicos (MB). Este é um cenário que se opõem quando levamos em consideração as duas amostras: o setor financeiro perde relevância quando se aumenta o número de ativos e, no caso contrário, o setor de materiais básicos ganha relevância. Quanto aos tipos de estruturas dentro do MCS, notamos assim como no universo com 10 ativos para processos dentro da diagonal que não se altera o resultado quando mudamos o formato de cálculo de matriz.

Conclusão

Neste trabalho utilizamos dados intradiários do mercado brasileiro para o cálculo da matriz de covariância incondicional a partir de métodos distintos. Consideramos na amostra dados de negociação dos 50 ativos mais negociados no período e optamos por uma abordagem autorregressiva para captar como a matriz muda no tempo. Não encontramos até o momento um trabalho parecido para o contexto brasileiro do ponto de vista preditivo e de tamanho amostral - conseguimos realizar as estimações usando um abundante número de ativos. Enxergamos nas limitações a escolha ainda não formal para a defasagem do modelo e também a ainda difícil interpretabilidade econômica dos resultados encontrados. Adicionalmente destacamos que não houve tratamento de *outliers*. Propomos regressão com encolhimento para lidar com o desafio da dimensionalidade e avaliamos o impacto preditivo através do algoritmo do MCS em um algoritmo de estimação equação por equação. Na demonstração dos resultados investigamos por setor econômico qual o tipo de regressão se manifesta dentro do MCS e como setorialmente se dá a seleção das variáveis para os processos de variância e covariância. Notamos para processos de variância presença intra setorial e maior sensibilidade dos resultados frente ao tipo de matriz de covariância. Para processos de covariância não há sensibilidade grande quando se flexibiliza a forma de cálculo da matriz e temos maior presença do setor de materiais básicos. Destacamos como pesquisa futura maior rigor no tratamento dos dados faltantes e a aplicação do método proposto em exercícios de seleção de portfólio.

Apêndice

Devido à dificuldade em computar resultados no formato matricial, dedicamos ao apêndice um passo a passo do algoritmo para os 2 ativos mais negociados no período de nossa amostra: `PETR4` e `VALE3`. Lembrando que estamos olhando a contabilidade de *trades* para computar nossa ordenação. Adicionalmente, mostramos mais detalhadamente as séries, os resultados do MCS e das estimações fora da amostra. Nesta visão mais micro podemos explorar também a série de covariância, que possui bastante ruído na frequência diária. Vamos apresentar os resultados na mesma ordem, para os diferentes tipos de cálculo de matriz: `ewma`, `cor` e `large`. Importante ressaltar que aqui estamos diante do contexto de amostra completa, com 50 ativos.

Nas figuras (19), (20) e (21) temos para `PETR4` o resultado fora da amostra, para cada tipo de matriz de covariância. Abaixo destacamos para cada valor de α se o modelo foi selecionado no MCS na tabela (5). Nas figuras (22), (23) e (24) temos para `VALE3` o resultado fora da amostra, para cada tipo de matriz de covariância e, por fim, nas figuras (25), (26) e (27) temos o resultado de previsão para fora da amostra na série de covariância entre `PETR4` e `VALE3`. Pelas tabelas podemos visualizar, por tipo de matriz, como se deu a seleção de modelos dentro do MCS.

4.3 PETR4

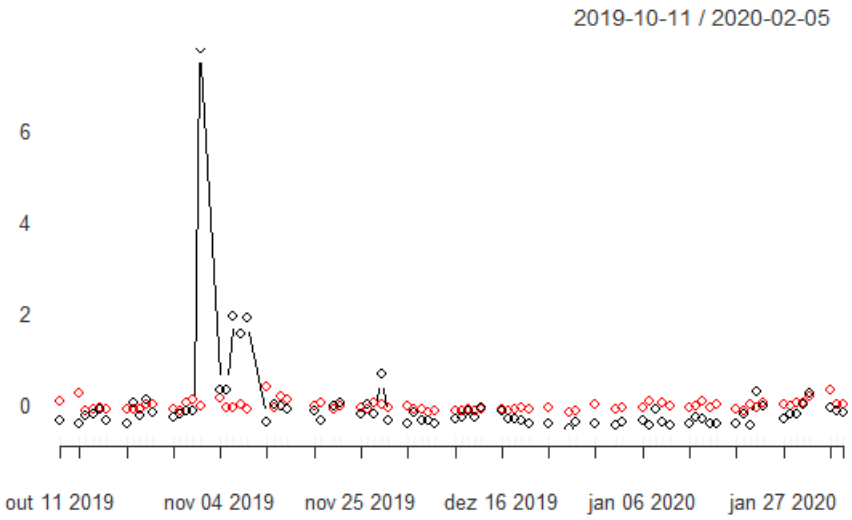


Figura 19 – *Resultado do algoritmo para PETR4. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `ewma`.

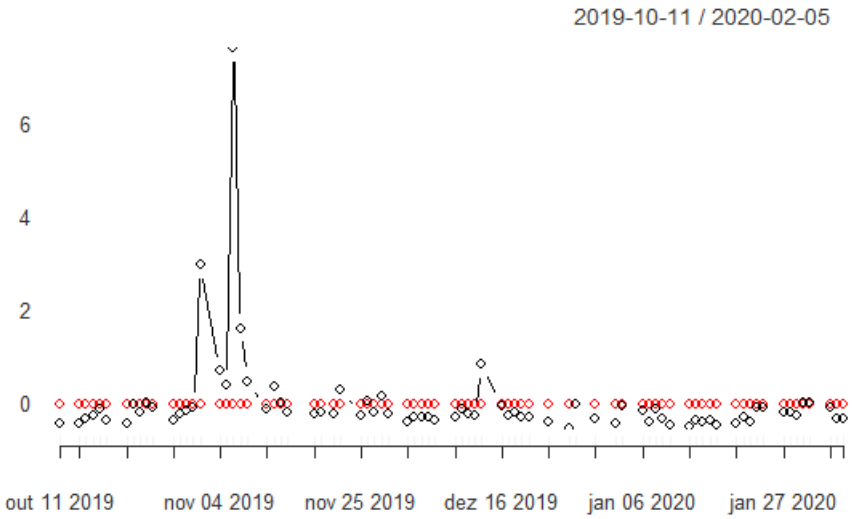


Figura 20 – *Resultado do algoritmo para PETR4. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `cor`.

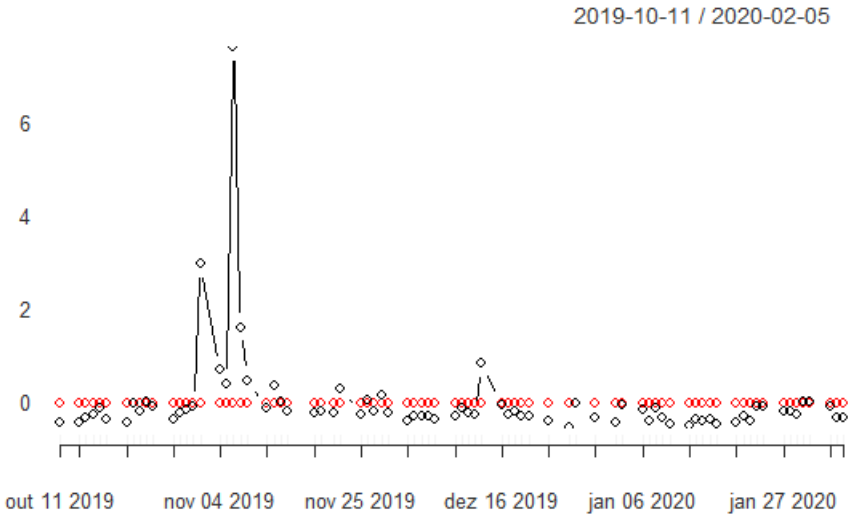


Figura 21 – *Resultado do algoritmo para PETR4. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `large`.

	`ewma`	`cor`	`large`
α =0	x	x	x
α =0.1	x	x	x
α =0.2	x	x	x
α =0.3		x	x
α =0.4		x	x
α =0.5		x	x
α =0.6			x
α =0.7			
α =0.8			
α =0.9			
α =1			

Tabela 5 – Tabela do MCS para `PETR4`. Quando marcada com x, temos que a especificação para o respectivo valor de α pertence ao MCS e, quando colorida em vermelho, indicamos a linha referente ao modelo vencedor.

4.4 VALE3

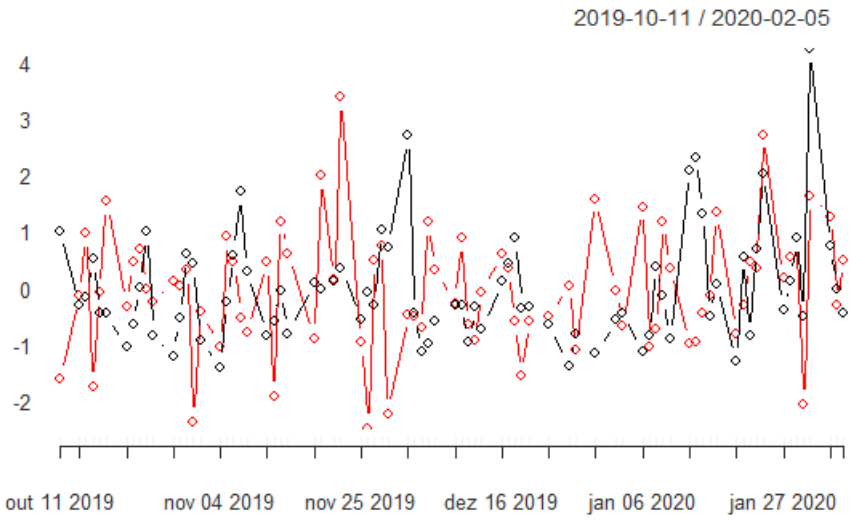


Figura 22 – *Resultado do algoritmo para VALE3. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `ewma`.

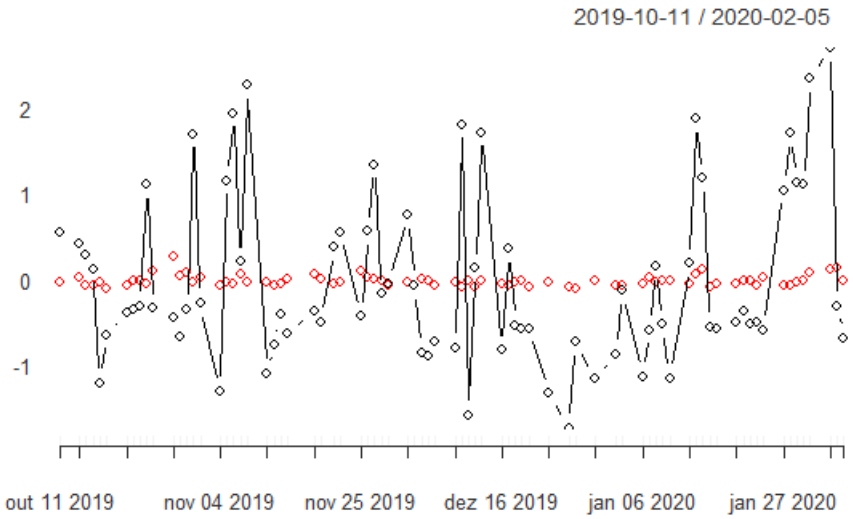


Figura 23 – *Resultado do algoritmo para VALE3. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `cor`.

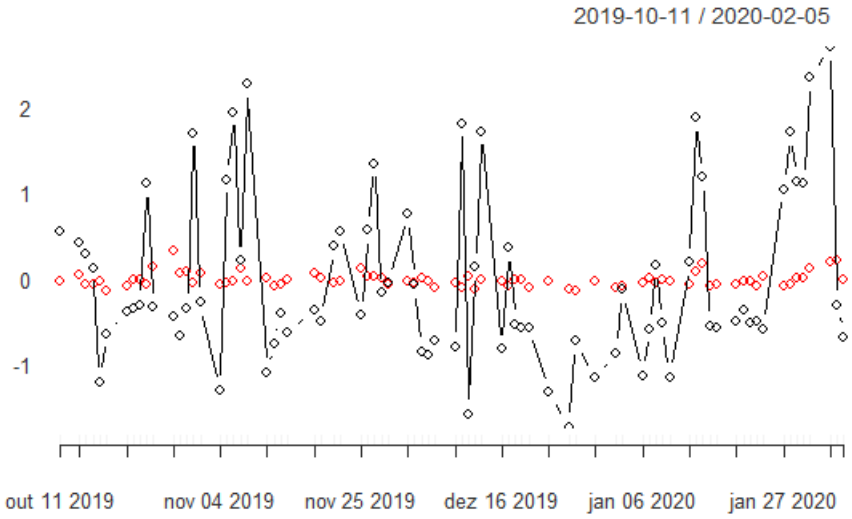


Figura 24 – *Resultado do algoritmo para VALE3. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `large`.

	ewma	**cor**	**large**
α =0			x
α =0.1		x	x
α =0.2		x	x
α =0.3	x	x	x
α =0.4		x	x
α =0.5		x	x
α =0.6		x	x
α =0.7		x	x
α =0.8			x
α =0.9			x
α =1			x

Tabela 6 – Tabela do MCS para `VALE3`. Quando marcada com x, temos que a especificação para o respectivo valor de α pertence ao MCS e, quando colorida em vermelho, indicamos a linha referente ao modelo vencedor.

4.5 Covariância entre PETR4 e VALE3

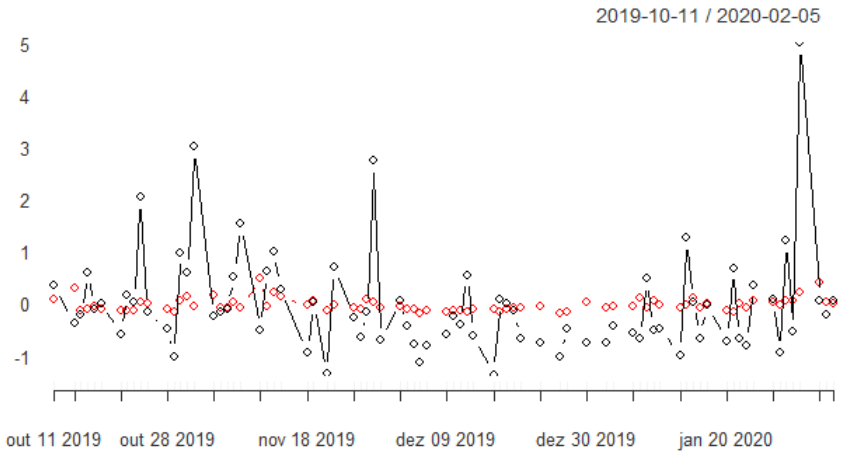


Figura 25 – *Resultado do algoritmo para a covariância entre os ativos. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `ewma`.

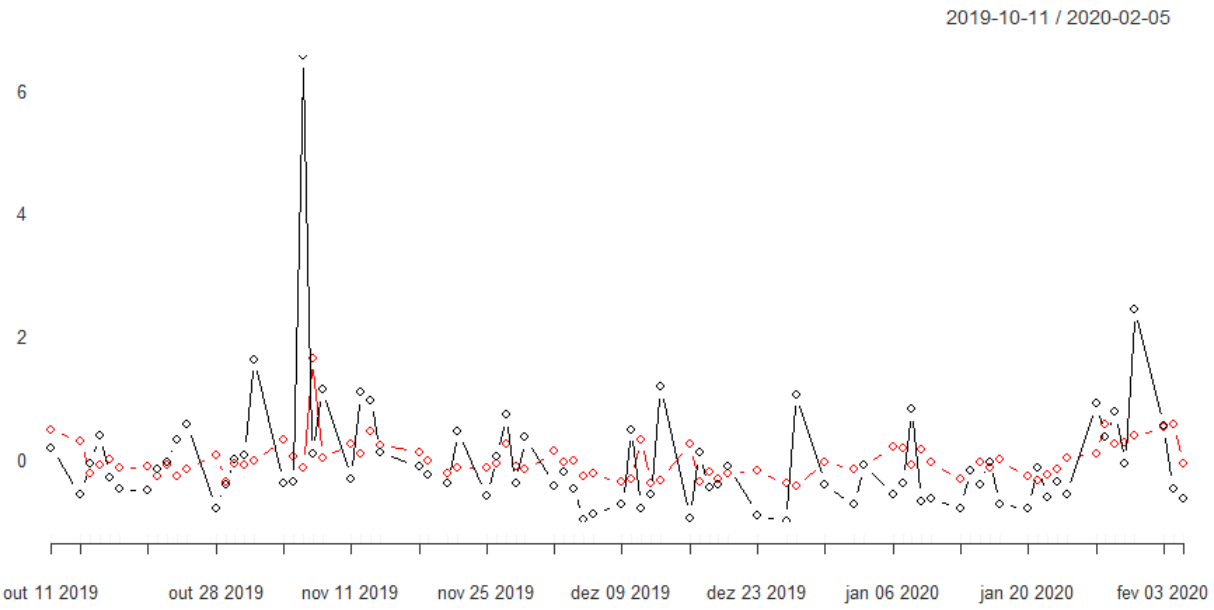


Figura 26 – *Resultado do algoritmo para a covariância entre os ativos. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `cor`.

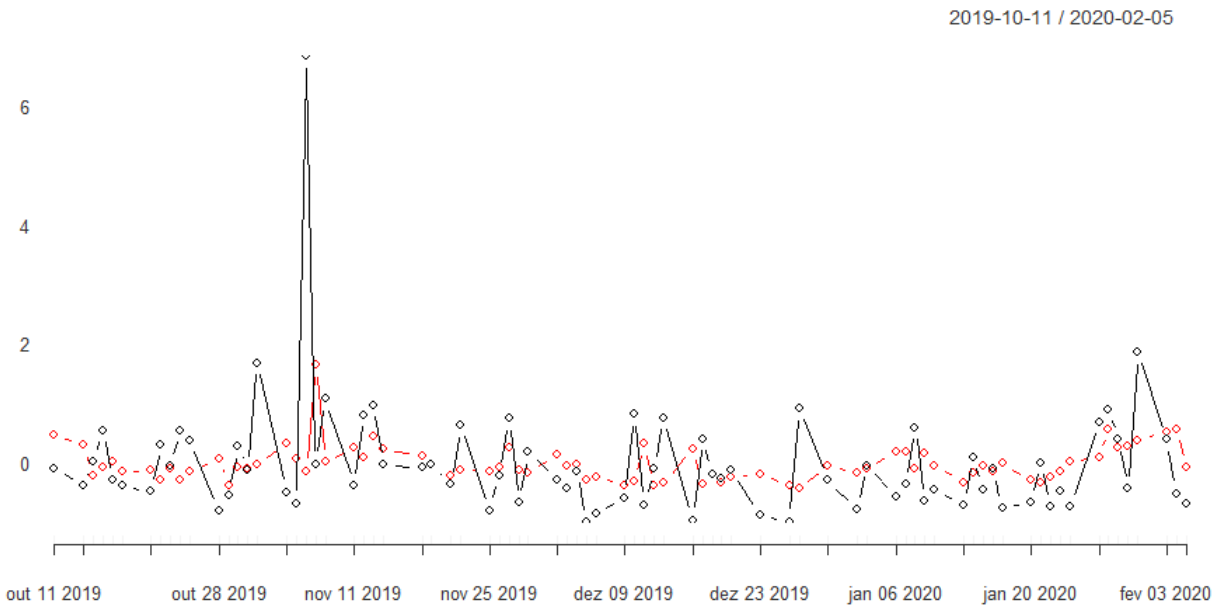


Figura 27 – *Resultado do algoritmo para a covariância entre os ativos. A linha em vermelho representa o previsto pelo modelo vencedor. A matriz de covariância é calculada via* `large`.

	ewma	**cor**	**large**
α =0	x	x	x
α =0.1	x		
α =0.2	x		
α =0.3	x		x
α =0.4			x
α =0.5		x	x
α =0.6		x	x
α =0.7		x	x
α =0.8		x	x
α =0.9			x
α =1			x

Tabela 7 – Tabela do MCS para a covariância entre as séries. Quando marcada com x, temos que a especificação para o respectivo valor de α pertence ao MCS e, quando colorida em vermelho, indicamos a linha referente ao modelo vencedor.

Bibliografia

Ardia, D., Boudt, K., and Gagnon-Fleury, J.-P. (2017). Riskportfolios: Computation of risk-based portfolios in r. *Journal of Open Source Software*, 2(10):171.

Bernardi, M. and Catania, L. (2014). The model confidence set package for r. *Innovation Finance Accounting eJournal*.

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

Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–131.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business Economic Statistics*, 20(3):339–350.

Engle, R. and Kelly, B. (2011). Dynamic equicorrelation. *Journal of Business Economic Statistics*, 30.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized arch. *Econometric Theory*, 11(1):122–150.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186 – 197. Econometric modelling in finance and risk management: An overview.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

Fan, J., Liao, Y., and Mincheva, M. (2011a). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Lv, J., and Qi, L. (2011b). Sparse high-dimensional models in economics. *Annual review of economics*, 3:291–317.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107:592–606.

Fleming, J., Kirby, C., and Ostdiek, B. (2003). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, 67(3):473 – 509.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22.

Hamilton, J. and Press, P. U. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, P. R. and Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a garch(1,1)? *Journal of Applied Econometrics*, 20(7):873–889.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86.

Jobson, J. D. and Korkie, B. (1980). Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554.

Laurini, M. P. and Ohashi, A. (2015). A noisy principal component analysis for forward rate curves. *Eur. J. Oper. Res.*, 246:140–153.

Ledoit, O. and Wolf, M. (2004a). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411.

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

Martin, V., Hurn, S., and Harris, D. (2012). *Econometric Modelling with Time Series: Specification, Estimation and Testing*. Themes in Modern Econometrics. Cambridge University Press.

Medeiros, M., Brito, D., and Ribeiro, R. (2018). Forecasting large realized covariance matrices: The benefits of factor models and shrinkage. *SSRN Electronic Journal*.

Medeiros, M., Callot, L., and Kock, A. (2016). Modeling and forecasting large realized covariance matrices and portfolio choice. *Journal of Applied Econometrics*, pages n/a–n/a.

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

Morgan, J., Limited, R., Ltd, R., and of New York, M. G. T. C. (1996). *RiskMetrics: Technical Document*. J. P. Morgan.

Perlin, M. and Ramos, H. (2016). *GetHFData: A R Package for Downloading and Aggregating High Frequency Trading Data from Bovespa*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Hastie, T., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Tibshirani, R., Hastie, T., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman Hall/CRC.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.