UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

RAFAEL RICCO

REALIZED SEMICOVARIANCES: EMPIRICAL APPLICATIONS TO VOLATILITY FORECASTING AND PORTFOLIO OPTIMIZATION

Porto Alegre
2021

RAFAEL RICCO

REALIZED SEMICOVARIANCES: EMPIRICAL APPLICATIONS TO VOLATILITY FORECASTING AND PORTFOLIO OPTIMIZATION

Dissertação submetida ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título de Mestre em Economia, área de concentração: Economia Aplicada.

Orientador: Prof. Dr. Flávio A. Ziegelmann

Porto Alegre
2021

CIP - Catalogação na Publicação

Ricco, Rafael
Realized Semicovariances: Empirical Applications to Volatility Forecasting and Portfolio Optimization / Rafael Ricco. -- 2021.
31 f.
Orientador: Flávio Ziegelmann.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Programa de Pós-Graduação em Economia, Porto Alegre, BR-RS, 2021.

1. Previsão de Volatilidade realizada. I. Ziegelmann, Flávio, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

RAFAEL RICCO

REALIZED SEMICOVARIANCES: EMPIRICAL APPLICATIONS TO VOLATILITY FORECASTING AND PORTFOLIO OPTIMIZATION

Dissertação submetida ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título de Mestre em Economia, área de concentração: Economia Aplicada.

BANCA EXAMINADORA:

Prof. Dr. Flávio Augusto Ziegelmann - Orientador
PPGE/UFRGS

Prof. Dr. Osvaldo Cândido Filho
UCB

Prof. Dr. Márcio Laurini
USP-RP

Prof. Dr. Tiago Filomena
PPGA/UFRGS

RESUMO

Este é um estudo empírico dividido em duas partes com o objetivo de explicar e prever a volatilidade realizada da carteira e realizar a otimização da carteira no mercado financeiro brasileiro usando semicovariâncias que usam dados de alta frequência. Na primeira parte do artigo, pretendemos prever a volatilidade realizada de uma carteira igualmente ponderada formada pelos retornos dos ativos da Bovespa brasileira, enquanto na segunda parte do artigo buscamos e encontramos uma carteira ótima. Em ambas as partes, utilizamos dados de alta frequência de dez ativos em diferentes segmentos e entre os mais negociados no mercado financeiro da Bovespa de julho de 2018 a janeiro de 2021. Os resultados mostram que os semicovariâncias ajudam a explicar melhor a variância da carteira de ações realizada e sob diferentes regimes, a relação entre eles pode mudar. Além disso, mostramos que usando todos semicovariâncias realizadas dentro de um modelo HAR podem levar a um "overfitting"e sob períodos de rebalanceamento de maior frequência, elas trazem melhorias no desempenho do portfólio ótimo de variância mínima.

Palavras-chave: Data de alta frequência. Previsão de volatilidade. Semivariâncias realizadas. Otimização de portfólio. Markov switching. LASSO. Performance econômica.

ABSTRACT

This is an empirical study split in two parts aimed to explain and forecast realized portfolio volatility and perform portfolio optimization in the Brazilian financial market using realized semicovariances that use high-frequency data. In the first part of the paper, we aim to forecast the realized volatility of an equally weighted portfolio formed by Brazilian Bovespa asset returns, whereas in the second part of the paper we search and find an optimum portfolio. In both parts, we use high-frequency data of ten assets from different segments and among the most negotiated in Bovespa financial market from July 2018 to January 2021. The results show that the realized semicovariances help to explain better the realized stock portfolio variance and under different regimes the relation among them can change. Also, we show that using all realized semicovariances within a HAR Model can lead to "overfitting"and under higher frequency rebalancing periods, they bring improvements on the minimum variance portfolio performance.

Keywords: High-frequency data. Volatility forecasting. Realized semicovariances. Portfolio optimization. Markov switching. LASSO. Economic performance.

SUMÁRIO

1 INTRODUÇÃO

O artigo incluído nesta dissertação consiste em um estudo empírico das medidas de semicovariância realizadas descritas no artigo do Bollerslev *et al.* (2020). A volatilidade é uma medida do grau de variação dos retornos de um ativo financeiro. Modelar e prever a volatilidade dos ativos financeiros é de grande interesse para muitos profissionais de finanças devido ao seu uso na gestão de risco, alocação de ativos, precificação de opções, entre outros. Muitos modelos têm sido usados para prever a variância dos retornos dos ativos, como GARCH em Bollerslev (1986) e suas derivações, que prevê a variância condicional dos retornos.

O advento dos dados de alta frequência permitiu a introdução de estimadores empíricos da variação quadrática para medir a variação ex-post dos preços dos ativos, ver Andersen e Bollerslev (1998), Andersen *et al.* (2001) e Barndorff-Nielsen e Shephard (2002). Hansen e Lunde (2010) lista seis maneiras pelas quais os dados de alta frequência melhoraram a previsão de volatilidade: i) os dados de alta frequência melhoram nossa compreensão das propriedades dinâmicas da volatilidade, que é a chave para previsão; ii) as medidas realizadas são preditores valiosos da volatilidade futura em modelos de forma reduzida; iii) as medidas realizadas permitiram o desenvolvimento de novos modelos de volatilidade que fornecem previsões mais precisas; iv) os dados de alta frequência melhoraram a avaliação das previsões de volatilidade de maneiras importantes; v) as medidas realizadas podem facilitar e melhorar a estimativa de modelos complexos de volatilidade, como modelos de volatilidade em tempo contínuo e vi) os dados de alta frequência melhoraram nossa compreensão das forças motrizes da volatilidade e sua importância relativa. Por exemplo, os dados de alta frequência permitiram uma análise detalhada dos anúncios de notícias e seus efeitos nos mercados financeiros.

Nesse contexto, Hansen *et al.* (2012) apresenta o GARCH realizado, um modelo GARCH que incorpora medidas realizadas como covariáveis. Patton e Sheppard (2015), através de um arcabouço empírico, mostra que a volatilidade futura está mais fortemente relacionada para a volatilidade dos retornos negativos anteriores do que para os retornos positivos e que o impacto de um salto de preço na volatilidade depende do sinal do salto, com saltos negativos (positivos) levando a uma maior (menor) volatilidade futura.

De acordo com Bollerslev *et al.* (2020), no contexto multivariado, um recente crescimento a literatura tem defendido vigorosamente o uso de dados intradiários de alta frequência para estimar de forma mais confiável matrizes de covariância de retorno de baixa frequência como em Andersen *et al.* (2003), Barndorff-Nielsen e Shephard (2004) e Barndorff-Nielsen *et al.* (2011).

Partindo do pressuposto de que os investidores se preocupam mais com a perda do que com o ganho, Barndorff-Nielsen *et al.* (2010) introduz a semivariância realizada, uma medida de risco que leva em conta o sinal de retorno e motiva Bollerslev *et al.* (2020) a propor uma decomposição da covariância realizada matriz em três componentes da matriz de semicovariância realizada ditada pelos sinais do retornos de alta frequência subjacentes. Segundo os autores, as matrizes de semicovariância realizadas podem ser vistas como uma extensão multivariada de alta frequência das semivariâncias originalmente propostas em Harry M Markowitz (1959), Mao

(1970), Hogan e Warren (1972) e Fishburn (1977).

Usando dados de alta frequência para uma grande seção transversal de ações dos EUA, Bollerslev *et al.* (2020) descobre que os modelos que incorporam as medidas de semicovariância realizada têm um desempenho de previsão superior a modelos que empregam as medidas de semivariância realizada ou apenas a matriz de covariância realizada.

No artigo que compõe o Capítulo 2 propomos um estudo empírico em duas partes aplicando as medidas de semicovariâncias realizadas descritas em Bollerslev *et al.* (2020)[1]. Na primeira parte do artigo, pretendemos prever a volatilidade realizada de uma carteira igualmente ponderada formada por ativos da Bolsa brasileira, enquanto na segunda parte do artigo, procuramos e encontramos um portfólio ideal usando as medidas realizadas descritas acima.

[1] o leitor interessado em se aprofundar nos conceitos teóricos pode consultar o Apêndice A onde extraímos os principais resultados obtidos em Bollerslev *et al.* (2020)

2 REALIZED SEMICOVARIANCES: EMPIRICAL APPLICATIONS TO VOLATILITY FORECASTING AND PORTFOLIO OPTIMIZATION

Rafael Ricco[1] and Flavio A. Ziegelmann [2]

September, 2021

Abstract. We propose a two-fold empirical study applying the concept of realized semicovariances as introduced by Bollerslev *et al.* (2020): in the first part of the paper we aim to forecast the realized volatility of an equally weighted portfolio formed by Brazilian Bovespa asset returns, whereas in the second part of the paper we search and find an optimum portfolio. In both parts we use high frequency data of ten assets from different segments and among the most negotiated in Bovespa financial market from July 2018 to January 2021. In addition, we investigate whether a Markov Switching strategy fits well to our modeling approach considering that our observed data starts some time before the Covid-19 pandemic and spans well into the pandemic period. The results suggest that the realized semicovariances help to explain better the realized stock portfolio variance and under different regimes the relation among them can change. Also, we show that using all realized semicovariances within a HAR Model can lead to "overfitting" and under higher frequency rebalancing periods, they bring improvements on the minimum variance portfolio performance.

Keywords. High-frequency data. Volatility forecasting. Realized semicovariances. Portfolio optimization. Markov switching. LASSO. Economic performance.

JEL Classifications. C53, E37

[1] Graduate Program in Economics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil, e-mail: rafaricco@gmail.com

[2] Department of Statistics and Graduate Program in Economics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil, e-mail: flavioaz@mat.ufrgs.br

2.1 INTRODUCTION

Volatility is a measure of the degree of variation of the returns of a financial asset. Modeling and forecasting the volatility of financial assets is of great interest to many practitioners in finance due to their usage in risk management, asset allocation, option pricing, among others. Many models have been used to predict the variance of the assets' returns such as GARCH in Bollerslev (1986) and its derivations, that predicts the conditional variance of the returns.

The advent of high-frequency data has permitted the introduction of empirical estimators of the quadratic variation to measure the ex-post variation of asset prices, see Andersen and Bollerslev (1998), Andersen *et al.* (2001) and Barndorff-Nielsen and Shephard (2002). Hansen and Lunde (2010) list six ways that high-frequency data has improved volatility forecasting: i) high-frequency data improve our understanding of the dynamic properties of volatility which is key for forecasting; ii) realized measures are valuable predictors of future volatility in reduced-form models; iii) realized measures have enabled the development of new volatility models that provide more accurate forecasts; iv) high-frequency data has improved the evaluation of volatility forecasts in important ways; v) realized measures can facilitate and improve the estimation of complex volatility models, such as continuous-time volatility models and vi) high-frequency data has improved our understanding of the driving forces of volatility and their relative importance. For instance, high-frequency data have enabled a detailed analysis of news announcements and their effect on the financial markets.

The realized variance is the most commonly used realized measure constructed by adding up squared intraday returns. There is an extensive statistical theory for this subject that is derived in papers by Barndorff-Nielsen and Shephard (2002), Meddahi (2002), Andersen *et al.* (2003) and Mykland and Zhang (2009), among others.

In this context, Hansen *et al.* (2012) introduce the Realized GARCH, a GARCH model that incorporates realized measures as covariates. Patton and Sheppard (2015), through an empirical framework, show that future volatility is more strongly related to the volatility of past negative returns than to that of positive returns and that the impact of a price jump on volatility depends on the sign of the jump, with negative (positive) jumps leading to higher (lower) future volatility.

According to Bollerslev *et al.* (2020), in the multivariate context, a rapidly-growing recent literature has forcefully advocated the use of high-frequency intraday data to more reliably estimate lower-frequency return covariance matrices as in Andersen *et al.* (2003), Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen *et al.* (2011).

Noureldin *et al.* (2012) propose the HEAVY models, a new class of multivariate volatility models that utilizes high-frequency data. Their empirical results suggest that the HEAVY model outperforms the multivariate GARCH model in an out-of-sample analysis, with the gains being particularly significant at shorter forecast horizons. Borges *et al.* (2015) show that covariance matrices based on higher frequency data lead to better performance indicators for a Brazilian stock portfolio. Aıt-Sahalia and Xiu (2016) analyze that the crisis period of 2007-2010 did indeed result in an increase in quadratic variation in all the assets considered by them, however, it did not cause a significant change in the breakdown between their respective Brownian and jump contributions, with both moving consistently with one another. Bollerslev *et al.* (2018) depict through an empirical framework the relationship between volume, volatility and public news announcements.

Under the assumption that investors care more about loss than gain, Barndorff-Nielsen *et al.* (2010) introduce the realized semivariance, a measure of risk that takes the return sign into account, and motivate Bollerslev *et al.* (2020) to propose a decomposition of the realized covariance matrix into three realized semicovariance matrix components dictated by the signs of the underlying high-frequency returns. According to the authors, the realized semicovariance matrices may be seen as a high-frequency multivariate extension of the semivariances originally proposed in Harry M Markowitz (1959), Mao (1970), Hogan and Warren (1972) and Fishburn (1977).

Using high-frequency data for a large cross-section of U.S. equities, Bollerslev *et al.* (2020) find that models that incorporate the realized semicovariance measures have a superior forecast performance than models that employ the realized semivariance measures or just the realized covariance matrix.

Set against this background and given the importance of the covariance matrix of asset returns for portfolio management, this paper is aimed to use the realized semicovariance measures described in Bollerslev *et al.* (2020). In section 2.2 we describe the variance components of asset returns in the univariate and multivariate contexts. In section 2.3, we describe the statistical models used in the present work in order to analyze the relationship between the portfolio realized variance and its components and in section 2.4 we apply these models to evaluate how much the realized semicovariances help to explain and predict an equally weighted stock portfolio variance in the Brazilian financial market. In section 2.5, we present the portfolio optimization theoretical framework and in section 2.6 we apply the realized measures presented in Bollerslev *et al.* (2020) to evaluate if their use brings improvements in the economic performance for stock portfolios. Finally, we make conclusion remarks in section 2.7.

2.2 REALIZED VOLATILITY MEASURES

The importance of measuring correlations between assets of a portfolio goes back to the early 1950s with H. Markowitz (1952). Since then, many works have studied the estimation and prediction of asset returns covariance matrices as in Kendall and Hill (1953), Elton and Gruber (1973) and Bauwens *et al.* (2006) due to their importance to risk and pricing evaluation.

As in Barndorff-Nielsen *et al.* (2010), a number of researchers have been interested in measuring downside risk, the risk of observing returns in the left tail of their probability density function, using specific information based on negative returns. This has been made by quantities such as value at risk, expected shortfall, and semivariance, which were typically estimated using daily returns.

Within the high-frequency data context and following Barndorff-Nielsen *et al.* (2010), the Realized Variance (RV), defined as

$$RV = \sum_{j=1}^{n} (Y_{t_j} - Y_{t_{j-1}})^2 , \tag{2.1}$$

where $0 = t_0 < t_1 < \cdots < t_n = 1$ are the times (ticks) at which prices are available, estimates consistently the quadratic variation of log asset prices Y_t over a fixed time period, suppose $[0,1]$, that is,

$$\text{p-lim}_{n\to\infty} \sum_{j=1}^{n} (Y_{t_j} - Y_{t_{j-1}})^2 = [Y]_1 \text{ , with } [Y]_t = \int_0^t \sigma_s^2 ds ,$$

when Y_t is a Brownian semimartingale, for instance. Note that the signs of the returns are irrelevant in the limit.

Furthermore, if there are jumps in the underlying generating process, the second term that shows up in the quadratic variation below, representing the source of quadratic variation due to those jumps, takes no information from the signs of returns:

$$[Y]_t = \int_0^t \sigma_s^2 + \sum_{s\leq t} (\Delta Y_s)^2 . \tag{2.2}$$

Motivated by the previous reasoning, Barndorff-Nielsen *et al.* (2010) introduce the downside realized semivariance (RS^-), given by

$$RS^- = \sum_{j=1}^{t_j \leq 1} (Y_{t_j} - Y_{t_{j-1}})^2 \mathbb{I}_{Y_{t_j} - Y_{t_{j-1}} \leq 0} , \tag{2.3}$$

and

$$RS^+ = \sum_{j=1}^{t_j \leq 1} (Y_{t_j} - Y_{t_{j-1}})^2 \mathbb{I}_{Y_{t_j} - Y_{t_{j-1}} \geq 0} , \tag{2.4}$$

where $RV = RS^- + RS^+$.

They prove that under in-fill asymptotics

$$RS^- \xrightarrow{p} \frac{1}{2} \int_0^1 \sigma_s^2 ds + \sum_{s\leq 1} (\Delta Y_s)^2 \mathbb{I}_{\Delta Y_s \leq 0} \tag{2.5}$$

and

$$RS^+ \xrightarrow{p} \frac{1}{2} \int_0^1 \sigma_s^2 ds + \sum_{s\leq 1} (\Delta Y_s)^2 \mathbb{I}_{\Delta Y_s \geq 0} . \tag{2.6}$$

Inspired by Barndorff-Nielsen *et al.* (2010), Bollerslev *et al.* (2020) extend their work to the multivariate context. Let $\boldsymbol{X}_t = (X_{1,t}, \ldots, X_{d,t})^\top$ denote a d-dimensional log-price process, sampled on a regular time grid $0 = t_0 < t_1 < \cdots < t_n = T$ over some fixed time span $T > 0$ and let the ith return be denoted by $\boldsymbol{\Delta}_i \boldsymbol{X} = \boldsymbol{X}_{t_i} - \boldsymbol{X}_{t_{i-1}}$. The realized covariance matrix is defined as

$$\widehat{RC} = \sum_{i=1}^{n} (\boldsymbol{\Delta}_i \boldsymbol{X})(\boldsymbol{\Delta}_i \boldsymbol{X})^\top . \tag{2.7}$$

Letting $p(x) = \max\{x, 0\}$ and $n(x) = \min\{x, 0\}$ denote the component-wise positive and negative elements of the real vector x, the corresponding “positive", “negative" and “mixed" realized semicovariance matrices are then simply defined as

$$\begin{aligned}
\widehat{RSC}_p &= \sum_{i=1}^{n} p(\boldsymbol{\Delta}_i \boldsymbol{X}) p(\boldsymbol{\Delta}_i \boldsymbol{X})^\top, \\
\widehat{RSC}_n &= \sum_{i=1}^{n} n(\boldsymbol{\Delta}_i \boldsymbol{X}) n(\boldsymbol{\Delta}_i \boldsymbol{X})^\top, \\
\widehat{RSC}_m &= \sum_{i=1}^{n} [p(\boldsymbol{\Delta}_i \boldsymbol{X}) n(\boldsymbol{\Delta}_i \boldsymbol{X})^\top + n(\boldsymbol{\Delta}_i \boldsymbol{X}) p(\boldsymbol{\Delta}_i \boldsymbol{X})^\top] .
\end{aligned} \tag{2.8}$$

Note that $\widehat{RC} = \widehat{RSC}_p + \widehat{RSC}_n + \widehat{RSC}_m$ and $\widehat{RSC}_p, \widehat{RSC}_n$ are defined as sums of vector outer-products and thus are positive semidefinite, whereas $\widehat{RSC}_m$ is indefinite.

Motivated by empirical observations from each of the 30 Dow Jones Industrial Average (DJIA) stocks returns on two different days, Bollerslev *et al.* (2020) note that estimates of $\widehat{RSC}_p$ and $\widehat{RSC}_n$ can diverge in response to the content of the news/event. The reader can find the theoretical framework that illustrate the distinction between the information carried on by the positive semicovariance and the negative semicovariance matrices in Bollerslev *et al.* (2020).

2.3 VOLATILITY FORECASTING

Many models have been developed to forecast the realized variance of a portfolio and we describe some of them below. The realized variance of a portfolio with portfolio weights $\boldsymbol{w}$ may be expressed as

$$\widehat{RV}^{port} = \boldsymbol{w}^\top \widehat{RC} \boldsymbol{w} , \tag{2.9}$$

where $\widehat{RC}$ is defined in (2.7). Since $\widehat{RC} = \widehat{RSC}_p + \widehat{RSC}_n + \widehat{RSC}_m$, we have

$$\begin{aligned}
\widehat{RV}^{port} &= \boldsymbol{w}^\top \widehat{RC} \boldsymbol{w} \\
&= \boldsymbol{w}^\top \widehat{RSC}_p \boldsymbol{w} + \boldsymbol{w}^\top \widehat{RSC}_n \boldsymbol{w} + \boldsymbol{w}^\top \widehat{RSC}_m \boldsymbol{w} \\
&= \widehat{P}^{port} + \widehat{N}^{port} + \widehat{M}^{port} .
\end{aligned} \tag{2.10}$$

A widely used model to forecast the realized variance is the HAR model of Corsi (2009), defined by

$$\widehat{RV}^{port}_{t+1|t} = \alpha_0 + \alpha_d \widehat{RV}^{port}_t + \alpha_w \widehat{RV}^{port}_{t-1:t-4} + \alpha_m \widehat{RV}^{port}_{t-5:t-21} , \tag{2.11}$$

where $\widehat{RV}^{port}_{t-l:t-k} = \frac{1}{k-l+1} \sum_{s=l}^{k} \widehat{RV}^{port}_{t-s}$. As in Bollerslev *et al.* (2020), we consider it as our benchmark model to evaluate forecast performance.

In addition to HAR, borrowing the idea from Patton and Sheppard (2015), we consider a HAR extension, the Semivariance HAR (SHAR), which includes the semivariances to the explanatory variables set as below:

$$\widehat{PSV} = \sum_{i=1}^{n} [p(\boldsymbol{w}^\top \boldsymbol{\Delta}_i \boldsymbol{X})]^2$$
$$\widehat{NSV} = \sum_{i=1}^{n} [n(\boldsymbol{w}^\top \boldsymbol{\Delta}_i \boldsymbol{X})]^2 ,$$

where $\boldsymbol{w}$ represents the portfolio weights, $\boldsymbol{\Delta}_i \boldsymbol{X} = \boldsymbol{X}_{t_i} - \boldsymbol{X}_{t_{i-1}}$ is the i-th intraday return of a d-dimensional log-price process $\boldsymbol{X}_t$ and $p(x) = \max\{x, 0\}$ and $n(x) = \min\{x, 0\}$ denote the component-wise positive and negative elements of the real vector x. The forecasting scheme results in

$$\widehat{RV}^{port}_{t+1|t} = \alpha_0 + \alpha_{d,p}\widehat{PSV}_t + \alpha_{d,n}\widehat{NSV}_t + \alpha_w \widehat{RV}^{port}_{t-1:t-4} + \alpha_m \widehat{RV}^{port}_{t-5:t-21} \ . \tag{2.12}$$

The model above allows us to verify whether the semivariances bring additional information to forecasting the portfolio realized variance.

Within the context of HAR extensions, following Bollerslev *et al.* (2020), we consider another one, the SemiCovariance HAR (SCHAR) that includes the semicovariance components depicted in (2.10). The one-step ahead forecast for the portfolio realized variance results in

$$\begin{aligned} \widehat{RV}^{port}_{t+1|t} = \alpha_0 &+ \alpha_{d,p}\widehat{P}^{port}_t + \alpha_{w,p}\widehat{P}^{port}_{t-1:t-4} + \alpha_{m,p}\widehat{P}^{port}_{t-5:t-21} \\ &+ \alpha_{d,n}\widehat{N}^{port}_t + \alpha_{w,n}\widehat{N}^{port}_{t-1:t-4} + \alpha_{m,n}\widehat{N}^{port}_{t-5:t-21} \\ &+ \alpha_{d,m}\widehat{M}^{port}_t + \alpha_{w,m}\widehat{M}^{port}_{t-1:t-4} + \alpha_{m,m}\widehat{M}^{port}_{t-5:t-21} . \end{aligned} \tag{2.13}$$

Additionally, in order to select the best predictors while avoiding over-fitting, we apply the Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996) in (2.13), obtaining what we name as SCHAR-lasso-in. Our LASSO estimator, considering α_0 as pre-estimated and $\widehat{RV}^{port}_t$ corrected by its mean, is given by the solution of

$$\widehat{\boldsymbol{\alpha}} = \arg\min \left[RSS(\alpha_1, \ldots, \alpha_9) + \lambda \sum_{j=1}^{9} |\alpha_j| \right] , \tag{2.14}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_9) = (\alpha_{d,p}, \ldots, \alpha_{m.m})$, $RSS(\alpha_1, ..., \alpha_9) = \sum_{t=22}^{T} (\widehat{RV}^{port}_t - \alpha_1 \widehat{P}^{port}_{t-1} - \cdots - \alpha_9 \widehat{M}^{port}_{t-6:t-22})^2$. The tuning parameter λ is usually chosen by data-driven techniques such as cross-validation[3] or information criteria.

As a final model specification, we introduce a two-regime Markov-Switching (MS) SCHAR model. The state variable S_t is an unobservable Markov chain and the estimation process is based on Kim's filter as described in Kim and Nelson (1999). Given the probability of the realized volatility is far away from the normal distribution the coefficients are estimated by the quasi-maximum likelihood estimate introduced in Lindsay (1988). Mathematically,

$$\ell = \sum_{t=1}^{T} \log \left(\sum_{S_t=0}^{1} f(\widehat{RV}^{port}_t | S_t, \psi_{t-1}) Pr[S_t | \psi_{t-1}] \right) . \tag{2.15}$$

[3] In the present work, we use a 10-fold cross-validation.

where $f(\widehat{RV}_t^{port}|S_t,\psi_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_{S_t}^2}} exp(-\frac{(\widehat{RV}_t^{port} - \alpha_{0,S_t} - \alpha_{d,p,S_t}\widehat{P}_{t-1}^{port} - \dots - \alpha_{m,m,S_t}\widehat{M}_{t-6:t-22}^{port})^2}{2\sigma_{S_t}^2})$. As the regimes S_t are unobservable, to evaluate the log-likelihood in (2.15) we need to calculate the weights $Pr[S_t|\psi_{t-1}]$ for $S_t = 0$ and $S_t = 1$ (two states). We do so here via Kim's filter (see Kim and Nelson (1999)).

2.4 EMPIRICAL EXERCISE 1

In this section we forecast the realized volatility of an equally weighted stock portfolio using the models described in Section 2.3. Our portfolio is composed of 10 BOVESPA index stocks depicted in Table 1, chosen to represent a high level of negotiation volume as well as to ensure heterogeneity among the different sectors of the economy.

Company's Name	Ticker Symbol	Sector
Ambev S.A.	ABEV3	Consumer Staples
B3	B3SA3	Financials
Bradesco S.A.	BBDC4	Financials
Intermedica S.A.	GNDI3	Health Care
Itaú S.A.	ITUB4	Financials
JBS S.A.	JBSS3	Consumer Staples
Magazine Luiza S.A.	MGLU3	Consumer Discretionary/Information Technology
Petrobras S.A.	PETR4	Oil & Gas
Suzano S.A.	SUZB3	Industrials
Vale S.A.	VALE3	Industrials Materials

Table 1 – Stocks included in the portfolio.

We consider a sample period from July 2018 to January 2021, comprising a total of 624 trading days. Besides, we use 5-minute intra-day returns to construct the realized measures, excluding the overnight returns[4].

2.4.1 In-sample Analysis

To generate the in-sample analysis, we first estimate the following three variations of HAR Models seen previously: HAR itself, SHAR and SCHAR.

Table 2 shows us the estimation results for each model aforementioned. We can observe that all coefficients are statistically significant at the 5% level in the first model. In the second model we can see that the daily positive semivariance ($\widehat{PSV}_{t-1}$) and both weekly and monthly realized covariances ($\widehat{RV}_{t-2:t-5}^{port}$, $\widehat{RV}_{t-6:t-22}^{port}$) are the main drivers of the realized portfolio variance ($\widehat{RV}_t$). Lastly, in the third model , we can notice that all coefficients, but monthly negative semicovariance and monthly positive semicovariance ($\widehat{N}_{t-6:t-22}^{port}$ and $\widehat{P}_{t-6:t-22}^{port}$), are statistically significant at 5%, which contrasts with the analysis in Bollerslev *et al.* (2020), where all negative

[4] The dataset can be fetched on https://github.com/rricco/realvol/tree/master/data

semicovariances are significant at 5% level. Looking at the Adjusted R^2 measures for each model (0.703 for the first model against 0.761 and 0.807 for the second and third model respectively), our findings imply that both realized semivariance and realized semicovariance components bring improvements for the explanation of the realized portfolio variance.

	Dependent variable:		
	$\widehat{RV}_t^{port}$		
	HAR	SHAR	SCHAR
$\widehat{RV}_{t-1}^{port}$	0.612^{***} (0.040)		
$\widehat{PSV}_{t-1}$		6.639^{***} (0.509)	
$\widehat{NSV}_{t-1}$		0.601^{***} (0.455)	
$\widehat{RV}_{t-2:t-5}^{port}$	0.306^{***} (0.047)	0.210^{***} (0.041)	
$\widehat{RV}_{t-6:t-22}^{port}$	-0.069^{**} (0.034)	-0.096^{***} (0.030)	
$\widehat{P}_{t-1}^{port}$			1.232^{***} (0.093)
$\widehat{P}_{t-2:t-5}^{port}$			1.726^{***} (0.252)
$\widehat{P}_{t-6:t-22}^{port}$			-0.561 (0.860)
$\widehat{N}_{t-1}^{port}$			-0.509^{***} (0.074)
$\widehat{N}_{t-2:t-5}^{port}$			-1.408^{***} (0.212)
$\widehat{N}_{t-6:t-22}^{port}$			1.228 (0.772)
$\widehat{M}_{t-1}^{port}$			-0.937^{***} (0.222)
$\widehat{M}_{t-2:t-5}^{port}$			-1.249^{**} (0.573)
$\widehat{M}_{t-6:t-22}^{port}$			2.655^{***} (0.796)
Observations	602	602	602
R^2	0.704	0.763	0.809
Adjusted R^2	0.703	0.761	0.807
Residual Std. Error	0.0003 (df = 598)	0.0002 (df = 597)	0.0002 (df = 592)
F Statistic	474.398^{***} (df = 3; 598)	479.963^{***} (df = 4; 597)	279.366^{***} (df = 9; 592)

Table 2 – Parameter estimates and their respective standard errors in brackets for each model in (2.11), (2.12) and (2.13), respectively. The first column shows us the covariates used across models. $***$, $**$ and $*$ represent whether a coefficient is significant at 1%, 5% or 10% levels, respectively.

In addition, Table 3 shows us the coefficients selected by the Least Absolute Shrinkage and Selection Operator (LASSO) from the SCHAR Model. As we can see, the shrinkage method chooses four covariates ($\widehat{P}_{t-1}^{port}$, $\widehat{P}_{t-2:t-5}^{port}$, $\widehat{M}_{t-1}^{port}$, $\widehat{M}_{t-6:t-22}^{port}$) as best predictors of the portfolio realized variance.

2.4.2 In-sample Analysis under different regimes

Our sample spans from 2 July 2018 to 8 January 2021, that is, it enters well the COVID-19 pandemic, which sharply increases volatility in the financial markets. To overcome this situation, we firstly split our dataset into a pre-pandemic period (from 2 July 2018 to 28 February 2020) with 388 observations and a post-pandemic period (from 2 March 2020 to 8 January 2021) with

	Dependent variable:
	$\widehat{RV}_t^{port}$
$\widehat{P}_{t-1}^{port}$	0.9933
$\widehat{P}_{t-2:t-5}^{port}$	0.370
$\widehat{P}_{t-6:t-22}^{port}$	0
$\widehat{N}_{t-1}^{port}$	0
$\widehat{N}_{t-2:t-5}^{port}$	0
$\widehat{N}_{t-6:t-22}^{port}$	0
$\widehat{M}_{t-1}^{port}$	-0.5523
$\widehat{M}_{t-2:t-5}^{port}$	0
$\widehat{M}_{t-6:t-22}^{port}$	0.1529

Table 3 – Lasso parameter estimates for model (2.13). The covariates chosen by lasso are $\widehat{P}_{t-1}^{port}$, $\widehat{P}_{t-2:t-5}^{port}$, $\widehat{M}_{t-1}^{port}$, $\widehat{M}_{t-6:t-22}^{port}$.

214 observations. Then we apply the Least Absolute Shrinkage and Selection Operator (LASSO) in the SCHAR Model for each sub-period. Table 4 shows us the results for each period.

	Dependent variable: $\widehat{RV}_t^{port}$	
	SCHAR-lasso-in-pre	SCHAR-lasso-in-post
$\widehat{P}_{t-1}^{port}$	0.3258	0.9780
$\widehat{P}_{t-2:t-5}^{port}$	0	0.3763
$\widehat{P}_{t-6:t-22}^{port}$	0	0
$\widehat{N}_{t-1}^{port}$	0.4056	0
$\widehat{N}_{t-2:t-5}^{port}$	0.1538	0
$\widehat{N}_{t-6:t-22}^{port}$	0	0
$\widehat{M}_{t-1}^{port}$	0	-0.6121
$\widehat{M}_{t-2:t-5}^{port}$	0	0
$\widehat{M}_{t-6:t-22}^{port}$	-0.3422	0.2565

Table 4 – Lasso parameter estimates for model (2.13) for each sub-period. The covariates chosen by lasso for the pre-pandemic period are $\widehat{P}_{t-1}^{port}$, $\widehat{N}_{t-1}^{port}$, $\widehat{N}_{t-2:t-5}^{port}$, $\widehat{M}_{t-6:t-22}^{port}$. On the other hand, the covariates chosen by lasso for the post-pandemic period are $\widehat{P}_{t-1}^{port}$, $\widehat{P}_{t-2:t-5}^{port}$, $\widehat{M}_{t-1}^{port}$, $\widehat{M}_{t-6:t-22}^{port}$

Selecting those variables with non-zero coefficients in Table 4 for either of the two sub-periods ($\widehat{P}_{t-1}^{port}$,$\widehat{P}_{t-2:t-5}^{port}$, $\widehat{N}_{t-1}^{port}$, $\widehat{N}_{t-2:t-5}^{port}$, $\widehat{M}_{t-1}^{port}$ and $\widehat{M}_{t-6:t-22}^{port}$), we run a Markov Switching model for the full sample. Based on the results of this analysis, in a next step we set to vary only those coefficients that were statistically significant for both regimes ($\widehat{P}_{t-1}^{port}$ and $\widehat{P}_{t-2:t-5}^{port}$). Table 5 shows us the estimates MS-SCHAR Model. It is worth noticing that only the first two coefficients ($\widehat{P}^{port}_{t-1}$,$\widehat{P}^{port}_{t-2:t-5}$) vary according to the state variable (S_t) and that all covariates coefficients are statistically significant.

Looking at Figures 1 and 2 we can see a prominent regime change from February 2020 to July 2020 which coincides with the COVID-19 pandemic outbreak over the world, suggesting that in stressed scenarios the relation between the portfolio realized variance and its daily positive realized semicovariance component ($\widehat{P}^{port}_{t-1}$) gets stronger as depicted in Table 5 where the coefficient of the daily positive semicovariance under regime 1 ($\widehat{P}^{port}_{t-1,1}$) is greater than the coefficient of the daily positive semicovariance under regime 0 ($\widehat{P}^{port}_{t-1,0}$). Finally, the reader can see the transition probabilities from the Markov Switching Model in Table 6. Under a first-order Markov chain with two possible states (regimes), the smoothed probabilities indicate a state ($S_t = 0$) that is characterized by a low portfolio's volatility and another state ($S_t = 1$) presenting a high portfolio's volatility (stressed scenario) caused by the COVID-19 pandemic.

	Dependent variable:
	$\widehat{RV}^{port}_t$
$\widehat{P}^{port}_{t-1,0}$	0.5848*** (0.0974)
$\widehat{P}^{port}_{t-1,1}$	1.2234*** (0.1730)
$\widehat{P}^{port}_{t-2:t-5,0}$	1.4123*** (0.1213)
$\widehat{P}^{port}_{t-2:t-5,1}$	2.0681*** (0.1719)
$\widehat{N}^{port}_{t-1}$	−0.5286*** (0.0595)
$\widehat{N}^{port}_{t-2:t-5}$	−1.3821*** (0.1755)
$\widehat{M}^{port}_{t-1}$	−0.9829*** (0.2029)
$\widehat{M}^{port}_{t-6:t-22}$	0.6257*** (0.1230)

Table 5 – Parameter estimates and their respective standard errors in brackets. $***$, $**$ and $*$ represent whether a coefficient is significant at t 1%, 5% or 10% levels, respectively.

	0	1
0	0.9908	0.0931
1	0.0092	0.9069

Table 6 – Estimated transition probabilities for $S_t = 0,1$.

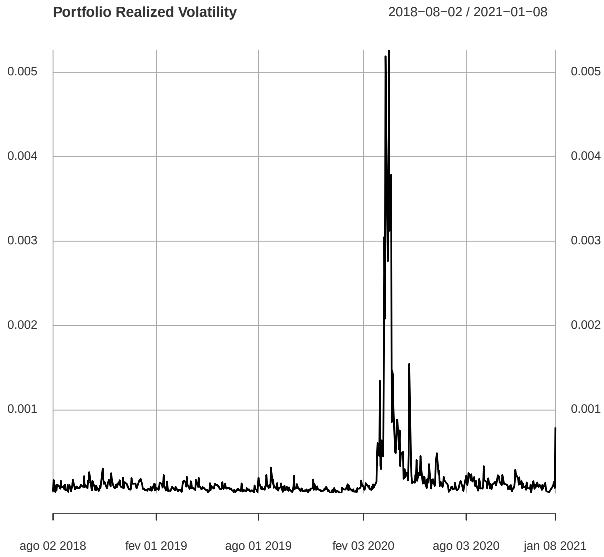

Figure 1 – Realized variance of the equally-weighted portfolio from April 2018 to January 2021

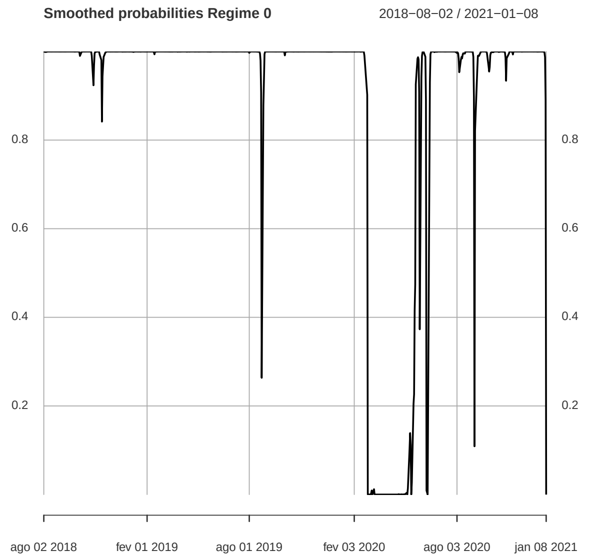

Figure 2 – Smoothed probabilities for regime 0 ($S_t = 0$) from the Markov Switching SCHAR Model.

2.4.3 Out-of-sample Analysis

Our out-of-sample analysis makes use of the four models described previously, namely, HAR, SHAR, SCHAR and SCHAR-lasso-in, using the same equally-weighted portfolio as used in the in-sample analysis. We construct rolling out-of-sample one-step ahead forecasts based on each of the different models, with model parameters re-estimated daily using the most recent 542 observations[5]. Given that in Bollerslev *et al.* (2020) the out-of-sample analysis suggests that the SCHAR model is “over-parameterized", we also include the SCHAR-lasso-out model in addition to the SCHAR-lasso-in model, that applies LASSO on each rolling window estimation.

5 The code that runs the rolling window analysis can be reached at https://github.com/rricco/realvol

In order to evaluate the forecast performance of the models, we rely on the model confidence set (MCS) procedure introduced by Hansen *et al.* (2011). The objective of the MCS is to determine the set of models containing the best models with a given probability based on a loss function such as MSE (mean squared error). The procedure employs a bootstrap implementation in order to compute the p-values for all models.

Table 7 shows us the forecast accuracy for each model measured by both MSE (mean-squared-error) and MAE (mean-absolute-error). The smallest MSE and MAE are indicated in bold. The cells with $*$ indicate the models that were chosen by the MCS procedure. As we can see, the simple HAR Model is the most accurate model, although in the previous section it was the one that least explained the dependent variable variance. It is worth noticing that SHAR, SCHAR-lasso-in and SCHAR-lasso-out were included in the MCS procedure. Moreover, considering that the sample size and the portfolio's dimension are not very large in the current work as in Bollerslev *et al.* (2020), our results are consistent with their work because:

- In Bollerslev *et al.* (2020), the difference among the models in the forecast accuracy is greater for large dimensional portfolios;
- The authors suggest that the SCHAR Model may be "over-parameterized" and as such it is likely to perform poorly in out-of-sample analysis. However, when we calibrate the "overfitting" from SCHAR Model through the LASSO method we reach out better results as depicted in Table 7.

Model	MSE	MAE
HAR	**1.00***	**1.00***
SHAR	1.02*	1.17*
SCHAR	1.38	1.59
SCHAR-lasso-in	1.01*	1.19*
SCHAR-lasso-out	1.02*	1.21*

Table 7 – Mean squared errors (MSE) and mean absolute errors (MAE) for the forecasts relative to the HAR Model. The values in bold indicate the method with lowest values of MSE and MAE. Cells with $*$ indicate that the method is included in the MCS constructed based on the T_{max} statistic using the squared/absolute errors with 5% of significance.

We also analyze which covariates are chosen at each rolling window estimation in SCHAR-lasso-out Model. Figure 3 shows us that the covariate with the highest frequency of choices by lasso is $\widehat{P}_t^{port}$ 60 times out of 60, followed by $\widehat{P}_{t-1:t-4}^{port}$ 27 times out of 60 and finally $\widehat{M}_t^{port}$ 3 times out of 60.

Figure 4 shows us the sample mean estimate for each covariate and its sample confidence interval using two sample standard deviations in the rolling window analysis. $\widehat{P}_t^{port}$ has a sample mean estimate of 0.8185, $\widehat{P}_{t-1:t-4}^{port}$ has a sample mean estimate of 0.0525 and $\widehat{M}_t^{port}$ has a sample mean estimate of -0.0595.

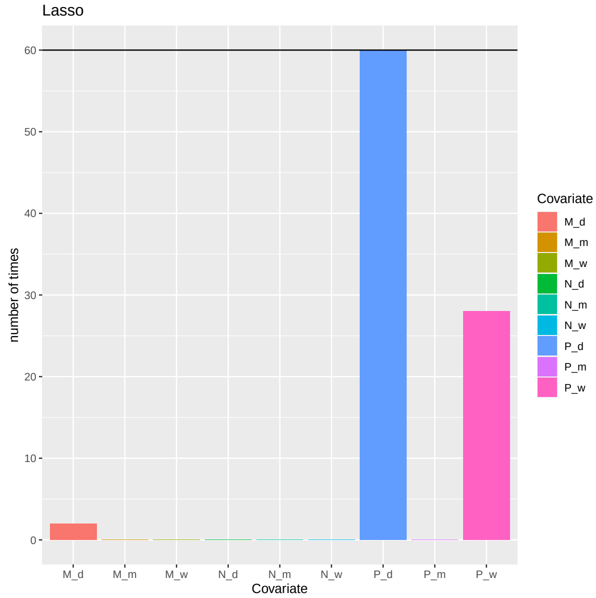

Figure 3 – Number of times each covariate is chosen by lasso from SCHAR-lasso-out Model. ($P_d = \widehat{P}_t^{port}$, $P_m = \widehat{P}_{t-5:t-21}^{port}$, $P_w = \widehat{P}_{t-1:t-4}^{port}$, $N_d = \widehat{N}_t^{port}$, $N_m = \widehat{N}_{t-5:t-21}^{port}$, $N_w = \widehat{N}_{t-1:t-4}^{port}$, $M_d = \widehat{M}_t^{port}$, $M_m = \widehat{M}_{t-5:t-21}^{port}$, $M_w = \widehat{M}_{t-1:t-4}^{port}$)

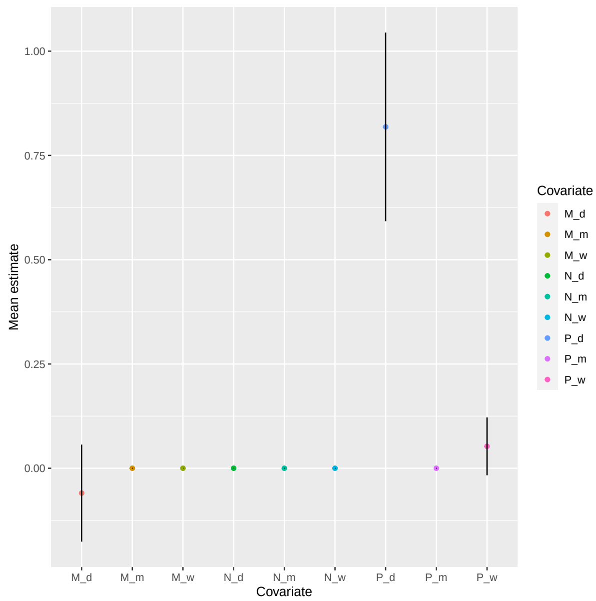

Figure 4 – Mean estimate for each covariate (dots on the plot) and its sample confidence interval using two sample standard deviations (vertical black line). ($P_d = \widehat{P}_t^{port}$, $P_m = \widehat{P}_{t-5:t-21}^{port}$, $P_w = \widehat{P}_{t-1:t-4}^{port}$, $N_d = \widehat{N}_t^{port}$, $N_m = \widehat{N}_{t-5:t-21}^{port}$, $N_w = \widehat{N}_{t-1:t-4}^{port}$, $M_d = \widehat{M}_t^{port}$, $M_m = \widehat{M}_{t-5:t-21}^{port}$, $M_w = \widehat{M}_{t-1:t-4}^{port}$)

2.5 PORTFOLIO OPTIMIZATION

The covariance matrix plays a key role in portfolio optimization theory. In the seminal work by H. Markowitz (1952), the problem becomes either minimizing the risk (variance) of the portfolio for a fixed mean return of it or maximizing the mean return for a fixed risk. Formally we can find the optimum portfolio as the solution of the following problem:

$$\min_{\boldsymbol{w}} \quad \boldsymbol{w}^{\mathsf{T}} C_{t|t-1} \boldsymbol{w} - \frac{1}{\lambda} \boldsymbol{w}^{\mathsf{T}} \mu_{t|t-1},$$

where $\boldsymbol{w}$ is the vector of portfolio weights, $C_{t|t-1}$ is the conditional covariance matrix, $\mu_{t|t-1}$ is the vector of the conditional mean and λ is the investor's risk aversion coefficient.

To circumvent issues related to estimation errors associated with estimating conditional return means, we focus on minimum variance portfolios as in Engle and Sheppard (2008) and Borges *et al.* (2015), among many others. Mathematically our problem becomes the following:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \boldsymbol{w}^{\mathsf{T}} C_{t|t-1} \boldsymbol{w} \\ & \text{subject to} \\ & \boldsymbol{w}^{\mathsf{T}} \mathbf{1} = 1 \ , \end{aligned} \tag{2.16}$$

where $\mathbf{1}$ is the vector of ones.

One of the pioneering works in empirical applications of portfolio allocation using intraday data is the paper by Fleming *et al.* (2003). The authors indicate that covariance matrices estimated by intradaily returns can generate gains in portfolio performance compared to those estimated by daily returns.

Another work in this topic is the paper by Liu (2009). His findings are that the gains generated by intradaily returns depend on the rebalancing frequency and the prediction horizon. According to the author, if an investor rebalances his portfolio monthly with at least the previous 12 months of data, daily and intradaily returns have similar performance results, however it is worth using intradaily data when the portfolio is rebalanced daily. Hautsch *et al.* (2013) show that large-scale portfolio covariance matrices based on high-frequency data generate a significantly lower portfolio volatility than methods employing daily returns.

In this context, Borges *et al.* (2015) compare different covariance matrix estimators based on intradaily or daily data for the Brazilian market. Their analysis suggests that conditional covariance estimates perform better than unconditional estimators and that the covariance matrix forecasts based on high-frequency data present lower portfolio volatility compared to using daily returns.

Given this scenario, here we evaluate the performance of a Brazilian stock portfolio comprised by 10 stocks from Ibovespa index, using the realized semicovariance measures of Bollerslev *et al.* (2020) and comparing them to using the standard realized covariance matrix.

Our portfolio analyzes are based on the realized covariance matrix, the positive semicovariance matrix and the negative semicovariance matrix measures defined in (2.7) and (2.8) respectively. We estimate the conditional realized measure (RM) for a single period as

$$\widehat{RM_{t|t-1}} = \frac{1}{\ell} \sum_{i=1}^{\ell} RM_{i,t-1} \ , \tag{2.17}$$

where ℓ is the length (in days) of the rebalancing period $t-1$ and $RM_{i,t-1}$ is the realized measure ($\widehat{RC}$, $\widehat{RSC}_p$ or $\widehat{RSC}_n$ as defined in section 2.2) for day i within the rebalancing period $t-1$.

The data used to run the analyzes is the same as that used in section 2.4. For each period we compute $\widehat{RM_{t|t-1}}$ in (2.16) to generate the minimum variance portfolio. In the sequence we evaluate the portfolio performance in a daily basis in terms of the Sharpe ratio (S_a), Sortino ratio (S_o) and turnover (T_o), for each realized measure and rebalancing period (daily, weekly or monthly). Assuming the risk free rate as zero, these statistics are calculated as follows:

$$S_a = \frac{\hat{\mu}}{\hat{\sigma}}, \qquad S_o = \frac{\hat{\mu}}{\hat{\sigma}_-}, \qquad T_o = \frac{1}{T}\sum_{t=1}^{T} |(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)|^\intercal \mathbf{1}$$

where T is the length of the out-of-sample period, N is the number of stocks, $\boldsymbol{w}_t$ is the $(1 \times N)$ vector of the portfolio weights at day t, $\boldsymbol{r}_t$ is the $(1 \times N)$ vector of the assets' returns at day t, $\mathbf{1}$ is the $(1 \times N)$ vector of ones and

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t^\intercal \boldsymbol{r}_t$$

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{w}_t^\intercal \boldsymbol{r}_t - \hat{\mu})^2$$

$$\hat{\sigma}_-^2 = \frac{1}{T}\sum_{t=1}^{T} [(\boldsymbol{w}_t^\intercal \boldsymbol{r}_t - \hat{\mu})\mathbb{I}_{\boldsymbol{w}_t^\intercal \boldsymbol{r}_t < \hat{\mu}}]^2 \quad \text{(Downside Semivariance)}$$

2.6 EMPIRICAL EXERCISE 2

In this section we compare the out-of-sample portfolio performance using different realized measures ($\widehat{RC}$, $\widehat{RSC}_p$ or $\widehat{RSC}_n$) for the covariance matrix, based on high-frequency data. The portfolios are rebalanced daily, weekly and monthly, and are analyzed according to their performance in terms of Sharpe ratio, Sortino ratio and turnover.

Tables 8 to 10 show the performances of the indicators mentioned above for daily, weekly and monthly rebalancing, respectively.

	Average return	Standard deviation	Sharpe ratio	Sortino ratio	Turnover
$\widehat{RSC}_n$	0.001	0.020	0.045*	0.060*	1.093
$\widehat{RSC}_p$	0.0005	0.020	0.024	0.033	1.109
$\widehat{RC}$	0.001	0.019	0.040	0.053	0.829*

Table 8 – Out-of-sample performance of the minimum variance portfolio using 10 assets traded at B3 stock exchange based on daily returns. The best results for Sharpe, Sortino and Turnover are with $*$.

The results in Table 8 indicate that, when the investor rebalances the portfolio daily, the $\widehat{RC}$ measure results in a portfolio with lower standard deviation. However, in terms of risk-adjusted returns, the $\widehat{RSC}_n$ measure presents the highest sharpe and sortino ratios.

In terms of transaction costs, the $\widehat{RC}$ measure presents the best performance (lowest turnover). Table 9 shows us the results when the investor rebalances the portfolio weekly. Again,

	Average return	Standard deviation	Sharpe ratio	Sortino ratio	Turnover
$\widehat{RSC}_n$	0.001	0.019	0.051	0.069	0.164
$\widehat{RSC}_p$	0.001	0.020	0.063	0.086*	0.166
$\widehat{RC}$	0.001	0.018	0.064*	0.085	0.112*

Table 9 – Out-of-sample performance of the minimum variance portfolio using 10 assets traded at B3 stock exchange based on daily returns. The best results for Sharpe, Sortino and Turnover are with $*$.

	Average return	Standard deviation	Sharpe ratio	Sortino ratio	Turnover
$\widehat{RSC}_n$	0.001	0.019	0.037	0.051	0.030
$\widehat{RSC}_p$	0.0005	0.020	0.024	0.032	0.029
$\widehat{RC}$	0.001	0.019	0.045*	0.060*	0.019*

Table 10 – Out-of-sample performance of the minimum variance portfolio using 10 assets traded at B3 stock exchange based on daily returns. The best results for Sharpe, Sortino and Turnover are with $*$.

the $\widehat{RC}$ measure presents the lowest standard deviation. In terms of risk-adjusted returns, the $\widehat{RC}$ measure presents the highest Sharpe ratio, whereas the $\widehat{RSC}_p$ has the highest Sortino ratio. In terms of transaction costs, the $\widehat{RC}$ measure, again, presents the lowest turnover.

Finally, Table 10 shows us the results when the investor rebalances the portfolio monthly. Alike the two previous rebalancing periods, the $\widehat{RC}$ measure presents the lowest standard deviation and in terms of risk-adjusted returns the $\widehat{RC}$ measure presents the highest Sharpe and Sortino ratios and the lowest turnover.

These results suggest that the realized components of the covariance matrix suit better with higher frequency rebalancing periods in terms of economic performance. Following the literature, the aforementioned results are consistent with Bollerslev *et al.* (2020) that find that realized semicovariance matrices ($\widehat{RSC}_p$, $\widehat{RSC}_n$) generally respond to new information faster than the realized covariance matrix ($\widehat{RC}$). Moreover, such feature helps us to justify the higher turnover presented by the semicovariance measures ($\widehat{RSC}_p$, $\widehat{RSC}_n$).

2.7 CONCLUSION

The scope of the present work is to propose the use of realized measures developed by Bollerslev *et al.* (2020) in volatility forecasting and portfolio optimization. We are particularly interested in addressing these issues for the Brazilian stock market.

In terms of volatility forecasting we draw the following lines. In our in-sample analysis, we come to the conclusion that the inclusion of semicovariance components in the model brings extra goodness of fit for the realized portfolio variance model. Moreover, we show that a Markov Switching Model makes sense for our comprised period of analysis, indicating that the beginning of Covid-19 pandemic is related to a higher volatility period and also that the relation between the realized portfolio variance and its semicovariance components can change under different regimes. Our out-of-sample analysis demonstrates that the SCHAR-lasso-in and SCHAR-lasso-out are part of the Model Confidence Set (MCS) whereas the SCHAR is not, which suggests that the SCHAR Model can suffer from “overfitting".

Finally, we see in the portfolio optimization analysis that under higher frequency rebalancing periods, minimum variance portfolios using the negative semicovariance matrices present better performances in terms of risk-adjusted returns compared to those that use the standard realized covariance matrices, which corroborates the analysis in Bollerslev *et al.* (2020) that suggests the semicovariance components respond faster to new information.

3 CONSIDERAÇÕES FINAIS

No artigo que integra esta dissertação propomos o uso de medidas realizadas descritas por Bollerslev *et al.* (2020) na previsão de volatilidade e otimização de portfólio voltados para o mercado de ações brasileiro.

Na análise “In-sample", nossos resultados sugerem que a inclusão de componentes de semicovariância traz uma qualidade extra de ajuste para o modelo explicativo da variância de portfólio realizado. Além disso, usando um modelo de “Markov switching" vemos que durante um período de alta volatilidade que caracterizou o início da pandemia de Covid-19 a relação entre a variância realizada do portfólio e seus componentes de semicovariância pode mudar sob diferentes regimes. Na análise “Out-of-sample" nossos resultados demonstram que os modelos SCHAR-lasso-in e o SCHAR-lasso-out entram do Model Confidence Set (MCS), enquanto o SCHAR não, o que sugere que o modelo SCHAR pode sofrer de “ overfitting ".

Finalmente, dentro do contexto de otimização de Portfólios, nossos resultados sugerem que em períodos de rebalanceamento de maior frequência, carteiras ótimas de variância mínima usando as matrizes de semicovariância negativas apresentam melhores desempenhos em termos de retornos ajustados ao risco em comparação com aquelas carteiras ótimas que usam as matrizes de covariância realizada padrão, o que corrobora a análise em Bollerslev *et al.* (2020) que sugere que os componentes de semicovariância respondem mais rápido a novas informações.

REFERÊNCIAS

AIT-SAHALIA, Y.; XIU, D. Increased correlation among asset classes: Are volatility or jumps to blame, or both? **Journal of Econometrics**, Elsevier, v. 194, n. 2, p. 205–219, 2016.

ANDERSEN, T. G.; BOLLERSLEV, T. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. **International Economic Review**, JSTOR, p. 885–905, 1998.

ANDERSEN, T. G.; BOLLERSLEV, T.; DIEBOLD, F. X.; LABYS, P. Modeling and forecasting realized volatility. **Econometrica**, Wiley Online Library, v. 71, n. 2, p. 579–625, 2003.

ANDERSEN, T. G.; BOLLERSLEV, T.; DIEBOLD, F. X.; LABYS, P. The distribution of realized exchange rate volatility. **Journal of the American Statistical Association**, Taylor & Francis, v. 96, n. 453, p. 42–55, 2001.

BARNDORFF-NIELSEN, O.; HANSEN, P. R.; LUNDE, A.; SHEPHARD, N. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. **Journal of Econometrics**, Elsevier, v. 162, n. 2, p. 149–169, 2011.

BARNDORFF-NIELSEN, O.; KINNEBROCK, S.; SHEPHARD, N. Measuring downside risk: realised semivariance. In: BOLLERSLEV, T.; RUSSEL, J.; WATSON, M. (Eds.). **Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle**. [S.l.]: Oxford University Press, 2010.

BARNDORFF-NIELSEN, O.; SHEPHARD, N. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. **Econometrica**, Wiley Online Library, v. 72, n. 3, p. 885–925, 2004.

BARNDORFF-NIELSEN, O.; SHEPHARD, N. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. **Journal of the Royal Statistical Society Series B**, Wiley Online Library, v. 64, n. 2, p. 253–280, 2002.

BAUWENS, L.; LAURENT, S.; ROMBOUTS, J. V. K. Multivariate GARCH models: a survey. **Journal of applied econometrics**, Wiley Online Library, v. 21, n. 1, p. 79–109, 2006.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. **Journal of Econometrics**, North-Holland, v. 31, n. 3, p. 307–327, 1986.

BOLLERSLEV, T.; LI, J.; PATTON, A. J.; QUAEDVLIEG, R. Realized semicovariances. **Econometrica**, Wiley Online Library, v. 88, n. 4, p. 1515–1551, 2020.

BOLLERSLEV, T.; LI, J.; XUE, Y. Volume, volatility, and public news announcements. **The Review of Economic Studies**, Oxford University Press, v. 85, n. 4, p. 2005–2041, 2018.

BORGES, B.; CALDEIRA, J. F.; ZIEGELMANN, F. A. Selection of minimum variance portfolio using intraday data: An empirical comparison among different realized measures for bm&fbovespa data. **Brazilian Review of Econometrics**, v. 35, n. 1, p. 23–46, 2015.

CORSI, F. A simple approximate long-memory model of realized volatility. **Journal of Financial Econometrics**, Oxford University Press, v. 7, n. 2, p. 174–196, 2009.

ELTON, E. J.; GRUBER, M. J. Estimating the dependence structure of share prices–implications for portfolio selection. **The Journal of Finance**, JSTOR, v. 28, n. 5, p. 1203–1232, 1973.

ENGLE, R.; SHEPPARD, K. **Evaluating the specification of covariance models for large portfolios**. [S.l.], 2008.

FISHBURN, P. C. Mean-risk analysis with risk associated with below-target returns. **The American Economic Review**, JSTOR, v. 67, n. 2, p. 116–126, 1977.

FLEMING, J.; KIRBY, C.; OSTDIEK, B. The economic value of volatility timing using "realized" volatility. **Journal of Financial Economics**, Elsevier, v. 67, n. 3, p. 473–509, 2003.

HANSEN, P. R.; HUANG, Z.; SHEK, H. H. Realized GARCH: a joint model for returns and realized measures of volatility. **Journal of Applied Econometrics**, Wiley Online Library, v. 27, n. 6, p. 877–906, 2012.

HANSEN, P. R.; LUNDE, A. Forecasting volatility using high frequency data. In: THE Oxford Handbook of Economic Forecasting. Oxford, England: Blackwell, 2010. P. 525–556.

HANSEN, P. R.; LUNDE, A.; NASON, J. M. The Model Confidence Set. **Econometrica**, v. 79, n. 2, p. 453–497, 2011.

HAUTSCH, N.; KYJ, L. M.; MALEC, P. **Do High-Frequency Data Improve High-Dimensional Portfolio Allocations?** [S.l.], 2013.

HOGAN, W. W.; WARREN, J. M. Computation of the efficient boundary in the ES portfolio selection model. **Journal of Financial and Quantitative Analysis**, JSTOR, p. 1881–1896, 1972.

KENDALL, M. G.; HILL, A. B. The analysis of economic time-series-part i: Prices. **Journal of the Royal Statistical Society Series A**, JSTOR, v. 116, n. 1, p. 11–34, 1953.

KIM, C. J.; NELSON, C. R. **State-space models with regime switching: classical and Gibbs-sampling approaches with applications**. [S.l.]: MIT Press, 1999.

LINDSAY, B. G. Composite likelihood methods. **Contemporary Mathematics**, v. 80, n. 1, p. 221–239, 1988.

LIU, Q. On portfolio optimization: How and when do we benefit from high-frequency data? **Journal of Applied Econometrics**, Wiley Online Library, v. 24, n. 4, p. 560–582, 2009.

MAO, J. C. T. Survey of capital budgeting: Theory and practice. **Journal of finance**, JSTOR, p. 349–360, 1970.

MARKOWITZ, H. Portfolio Selection, Journal of Finance. **Journal of Finance**, p. 77–91, 1952.

MARKOWITZ, Harry M. Portfolio Selection: Efficient Diversification of Investments. **Cowles Foundation Monograph**, v. 16, 1959.

MEDDAHI, N. A theoretical comparison between integrated and realized volatility. **Journal of Applied Econometrics**, Wiley Online Library, v. 17, n. 5, p. 479–508, 2002.

MYKLAND, P. A.; ZHANG, L. Inference for continuous semimartingales observed at high frequency. **Econometrica**, Wiley Online Library, v. 77, n. 5, p. 1403–1445, 2009.

NOURELDIN, D.; SHEPHARD, N.; SHEPPARD, K. Multivariate high-frequency-based volatility (HEAVY) models. **Journal of Applied Econometrics**, Wiley Online Library, v. 27, n. 6, p. 907–933, 2012.

PATTON, A. J.; SHEPPARD, K. Good volatility, bad volatility: Signed jumps and the persistence of volatility. **Review of Economics and Statistics**, MIT Press, v. 97, n. 3, p. 683–697, 2015.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

APPENDIX A – THEORETICAL FRAMEWORK

The material of this appendix is extracted from Bollerslev *et al.* (2020). Suppose that the log-price vector X_t is an Itô semimartingale as

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + J_t, \tag{A.1}$$

where b is the $\mathbb{R}^d$-valued drift process, W is a d-dimensional standard Brownian motion, σ is the $d \times d$ dimensional stochastic volatility matrix and J is a finitely active pure-jump process. We denote the spot covariance matrix of X by $c_t = \sigma_t \sigma_t^\top$ and further set

$$v_{j,t} = \sqrt{c_{jj,t}}, \quad \rho_{jk,t} = \frac{c_{jk,t}}{v_{j,t} v_{k,t}}. \tag{A.2}$$

where $v_{j,t}$ and $\rho_{jk,t}$ denote the spot volatility of asset j and the spot correlation coefficient between assets j and k, respectively. Moreover, Bollerslev *et al.* (2020) consider the following assumption,

Assumption 1. The process X is an Itô semimartingale defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ of the form (A.1) with $J_t = \int_0^t \int_{\mathbb{R}} \delta(s,u) \mu(ds,du)$, where the process b is locally bounded, the process σ is càdlàg and takes value in $\mathbb{R}^{d \times d}$, δ is a predictable function and μ is a Poisson random measure defined on $\mathbb{R}_+ \times \mathbb{R}$ with compensator $\nu(dt,du) = dt \bigotimes \lambda(du)$ for some finite measure λ on $\mathbb{R}$.

Consider the realized semicovariance estimators defined by equation (2.8) and let ΔX_s denote the price jump occurring at time s, if a jump occurred and set it to zero if no jump occurred at time s. Further define

$$\begin{aligned} RSC_p^J &= \sum_{s \leq T} p(\Delta X_s) p(\Delta X_s)^\top \\ RSC_n^J &= \sum_{s \leq T} n(\Delta X_s) n(\Delta X_s)^\top \\ RSC_m^J &= \sum_{s \leq T} [p(\Delta X_s) n(\Delta X_s)^\top + n(\Delta X_s) p(\Delta X_s)^\top] \end{aligned}$$

These measures characterize the discontinuous parts of the semicovariance measures. Bollerslev *et al.* (2020) demonstrate that

Theorem A.0.1. *Under Assumption 1,* $(\widehat{RSC}_p, \widehat{RSC}_n, \widehat{RSC}_m) \xrightarrow{\mathbb{P}} (RSC_p, RSC_n, RSC_m)$ *where* RSC_p, RSC_n *and* RSC_m *are* $d \times d$ *matrices with their (j,k) elements given by*

$$\begin{aligned} [RSC_p]_{jk} &= \int_0^T v_{j,s} v_{k,s} \psi(\rho_{jk,s}) ds + [RSC_p^J]_{jk}, \\ [RSC_n]_{jk} &= \int_0^T v_{j,s} v_{k,s} \psi(\rho_{jk,s}) ds + [RSC_n^J]_{jk}, \\ [RSC_m]_{jk} &= -2 \int_0^T v_{j,s} v_{k,s} \psi(-\rho_{jk,s}) ds + [RSC_m^J]_{jk}, \end{aligned}$$

and $\psi(.)$ *is defined as*

$$\psi(\rho) = (2\pi)^{-1} \left(\rho arccos(-\rho) + \sqrt{1 - \rho^2} \right), \tag{A.3}$$

Note that each semicovariance matrix contains both diffusive and jump covariation components. Moreover, the limiting variables RSC_p and RSC_n share exactly the same diffusive component, but their jump components differ,

$$\widehat{RSC}_p - \widehat{RSC}_n \xrightarrow{\mathbb{P}} RSC_p - RSC_n = RSC_p^J - RSC_n^J.$$

In this context, the first-order asymptotic behavior of the concordant semicovariance differential (CSD) is fully characterized by the "directional co-jumps". Then, when there are no jumps, one can not distinguish the information conveyed by $\widehat{RSC}_p$ and $\widehat{RSC}_n$. Hence, in order to reveal the differential information inherent in the realized measures, Bollerslev *et al.* (2020) derive a more refined second-order asymptotic analysis.

Following Bollerslev *et al.* (2020), we consider a bivariate setting to present the second-order asymptotic analysis for $[\widehat{RSC}_p]_{12} = [\widehat{RSC}_p]_{21}$ and $[\widehat{RSC}_n]_{12} = [\widehat{RSC}_n]_{21}$.

Consider that the stochastic volatility σ_t is also an Itô semimartingale as

$$\sigma_t = \sigma_0 + \int_0^t \tilde{b}_s ds + \int_0^t \tilde{\sigma}_s dW_s + \tilde{M}_t + \sum_{s \leq t} \Delta\sigma_s \mathbb{I}_{\{\|\Delta\sigma_s\| > \sigma^*\}}, \tag{A.4}$$

where $\tilde{b}$ is the drift, $\tilde{\sigma}$ is a $d \times d \times d$ tensor-valued process, $\tilde{M}$ is a local martingale that is orthogonal to the Brownian motion W. According to Bollerslev *et al.* (2020), the process $\tilde{\sigma}$ collects the loadings of the stochastic volatility matrix σ on the price Brownian shocks dW, and hence is considered as a "leverage effect" and $\tilde{M}$ collects "small" volatility jumps in the form of a purely discontinuous local martingale. Meanwhile, the term $\sum_{s\leq t} \Delta\sigma_s \mathbb{I}_{\{\|\Delta\sigma_s\| > \sigma^*\}}$ collects the "large" volatility jumps, which often occur in response to major news announcements as described in Bollerslev *et al.* (2018). Moreover, Bollerslev *et al.* (2020) consider the following assumption

Assumption 2. Assuming the Assumption 1 is satisfied, the process σ in (A.4) satisfies (1) σ_t is non-singular almost surely for all t; (2) $\tilde{b}$ is locally bounded; (3) $\tilde{\sigma}$ is a $d \times d \times d$ càdlàg process; (4) the process $\tilde{M}$ is a local martingale that is orthogonal to W with $\|\Delta\tilde{M}\| \leq \sigma^*$ for some constant $\sigma^* > 0$ and its predictable quadratic covariation process has the form $\langle \tilde{M}, \tilde{M} \rangle = \int_0^t \tilde{q}_s ds$ for some locally bounded process $\tilde{q}$; (5) the compensator of the pure-jump process $\sum_{s\leq t} \Delta\sigma_s \mathbb{I}_{\{\|\Delta\sigma_s\| > \sigma^*\}}$ has the form $\int_0^t q_s ds$ for some locally bounded process q.

Bollerslev *et al.* (2020) demonstrate that

Theorem A.0.2. *Under Assumption 2,*

$$\Delta_n^{-\frac{1}{2}} \begin{bmatrix} [\widehat{RSC}_p]_{12} \\ [\widehat{RSC}_n]_{12} \end{bmatrix} \xrightarrow{\mathcal{L}-s} \begin{bmatrix} B \\ -B \end{bmatrix} + \begin{bmatrix} L \\ -L \end{bmatrix} + \begin{bmatrix} \zeta \\ -\zeta \end{bmatrix} + \begin{bmatrix} \tilde{\zeta}_p \\ \tilde{\zeta}_n \end{bmatrix} + \begin{bmatrix} \tilde{\xi}_p \\ \tilde{\xi}_n \end{bmatrix}.$$