UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Caio Mário Henriques Silva da Rocha Mesquita

Scenario generation for financial data: a machine learning dynamic copula approach based on realized volatility and correlation

Belo Horizonte
2024

Caio Mário Henriques Silva da Rocha Mesquita

Scenario generation for financial data: a machine learning dynamic copula approach based on realized volatility and correlation

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Adriano César Machado Pereira
Co-Advisor: Cristiano Arbex Valle

Belo Horizonte
2024

Mesquita, Caio Mário Henriques Silva da Rocha

M582s Scenario generation for financial data: [recurso eletrônico] a machine learning dynamic copula approach based on realized volatility and correlation / Caio Mário Henriques Silva da Rocha Mesquita – 2024.
1 recurso online (110 f. il, color.) : pdf.

Orientador: Adriano César Machado Pereira
Coorientador: Cristiano Arbex Valle
Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f.101-110

1. Computação – Teses. 2. Aprendizado do computador – Teses. 3.– Gestão de riscos – Teses. 4. Mercado financeiro Portifólios - Teses. I. Pereira, Adriano César Machado. II. Valle, Cristiano Arbex. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. IV.Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Scenario generation for financial data: a machine learning dynamic copula approach based on realized volatility and correlation

CAIO MÁRIO HENRIQUES SILVA DA ROCHA MESQUITA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO CÉSAR MACHADO PEREIRA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. CRISTIANO ARBEX VALLE - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. PAULO ANDRÉ LIMA DE CASTRO
Divisão de Ciência da Computação - ITA

Heitor Soares Ramos Filho:78751241404

Assinado de forma digital por Heitor Soares Ramos Filho:78751241404
Dados: 2024.05.21 10:26:18 -03'00'

PROF. HEITOR SOARES RAMOS FILHO
Departamento de Ciência da Computação - UFMG

GLPappa

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

EWanner

PROFA ELIZABETH FIALHO WANNER
Departamento de Computação - CEFETMG

Belo Horizonte, 22 de fevereiro de 2024.

Dedico este trabalho à minha mãe, ao meu pai e ao meu irmão, que sempre me apoiaram incondicionalmente a alcançar todos os meus sonhos.

Acknowledgments

Agradeço a Deus e minha família por todo apoio, incentivo e suporte ao longo de todos esses anos. A Stephany por todo carinho e companheirismo.

Um agradecimento especial aos meus orientadores Adriano César e Cristiano Arbex por toda a caminhada e aprendizado ao longo do trabalho.

Agradeço também ao Programa de Pós Graduação em Ciência da Computação da Universidade Federal de Minas Gerais pela oportunidade. Aos professores do programa e amigos que contribuíram muito a minha formação. A CAPES pelo apoio financeiro ao longo da pós-graduação.

Enfim, a todos que contribuíram de alguma forma para que eu chegasse até aqui.

Muito obrigado!

“*A persistência é o caminho do êxito.*”
(Charles Chaplin)

Resumo

A otimização de portfólio é uma questão fundamental em finanças quantitativas, e as técnicas de geração de cenários desempenham um papel vital na simulação do comportamento futuro de ativos para uso em estratégias de alocação. Na literatura, diversas abordagens existem para gerar cenários, que variam de observações históricas a modelos que preveem a volatilidade dos ativos. Nesta tese, propomos uma metodologia inovadora para gerar cenários discretos um dia à frente, os quais são então utilizados como entrada para alocação de portfólio. Nossa abordagem emprega algoritmos supervisionados de aprendizado de máquina como modelos de previsão para estimar a variância realizada e a correlação intradiária de Kendall dos ativos. Com base nessas previsões, aplicamos uma abordagem de cópula com distribuições de valores extremos para simular a distribuição de probabilidade multivariada dos ativos. Nossos experimentos computacionais indicam que nossa abordagem pode proporcionar previsões de volatilidade e correlação mais precisas, bem como portfólios com melhor relação risco-recompensa em comparação com baselines tradicionais da literatura.

Palavras-chave: aprendizado de máquina; funções cópula; volatilidade realizada; otimização de portfólios.

Abstract

Portfolio optimization is a fundamental issue in quantitative finance, and scenario generation techniques play a vital role in simulating the future behavior of assets for use in allocation strategies. In the literature, various approaches exist for generating scenarios, ranging from historical observations to models predicting asset volatility. In this dissertation, we propose a novel methodology for generating discrete scenarios one day ahead, which are then used as input for portfolio allocation. Our approach employs machine learning supervised algorithms as forecasting models to predict the realized variance and intraday Kendall correlation of assets. Using these predictions, we apply a copula approach with extreme value distributions to simulate the multivariate probability distribution of the assets. Our computational experiments indicate that our approach may yield more accurate volatility and correlation forecasts, as well as better risk-reward portfolios compared to traditional literature baselines.

Keywords: machine learning; copula functions; realized volatility; portfolio optimization.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Portfolio selection is a core problem in quantitative finance. It is a well-established research area that provides a theoretical and practical foundation to investors seeking to make informed investment decisions. The area has its roots in the seminal work of Markowitz [1952], which proposed the Modern Portfolio Theory (MPT). MPT states that an investor is risk-averse and has two conflicting goals when selecting a portfolio: maximizing expected return and minimizing risk. Markowitz introduced the mean-variance framework in which the risk measure of choice is the portfolio variance. With its subsequent developments, MPT has set the stage for theory and practice in finance for the past decades.

MPT, however, relies on assumptions that contradict well-known stylized facts observed in decades of studies of financial time series [Cont, 2001]. For instance, MPT assumes that asset returns follow a normal distribution. However, as observed extensively, such distributions are not necessarily symmetric and generally have heavier tails than expected in a normal distribution. Moreover, variance is not necessarily the most appropriate measure of risk as it penalizes “good volatility” (extreme positive returns). Finally, optimizing variance requires solving a quadratic programming program, which can lead to computational difficulties when, for example, exogenous constraints are included to simulate real-world trading conditions. These drawbacks have been well known for decades, and even Markowitz alternatively suggested using a downside risk measure, semi-variance [Markowitz, 1959].

Artzner et al. [1999] defined a set of desirable mathematical properties that risk measures should ideally satisfy. A risk measure that satisfies these properties is referred to as coherent. Both variance and Value-at-Risk (VaR) [Guldimann, 2000] violate some of these properties and are thus incoherent. A well-established coherent risk measure is the Conditional Value-at-Risk (CVaR), which has, since its inception, gained popularity as a preferential risk measure during the Basel III Convention [Chang et al., 2019].

Rockafellar and Uryasev [2000] proposed a linear programming model that selects the portfolio with minimum CVaR. Compared to the traditional mean-variance framework, which requires a vector of mean returns and a covariance matrix as input data, CVaR optimization relies on scenarios representing discrete multivariate distributions.

Historical data can represent scenarios. While historical data may implicitly keep properties such as heavy tails in marginal distributions and the overall correlation among assets, they do not necessarily reflect future behavior as the structure of markets is inherently dynamic.

1.1 Motivation

Generating scenarios that accurately reflect the future behavior of assets is a research area. It involves finding discrete multivariate distributions of asset returns that preserve both marginal moments and the dependence structure among them. As the literature shows, assets are generally not independent, and the correlations between them can also change over time.(Christodoulakis and Satchell [2002], Engle [2002], Tse and Tsui [2002]).

Several papers in literature have proposed scenario generation methods based on different techniques [Kaut and Wallace, 2003, Guastaroba et al., 2009]. A family of scenario generation techniques combine Generalized Auto-regressive Conditional Heteroskedasticity (GARCH) models [Bollerslev, 1986] with copula functions [Bai and Sun, 2007, Messaoud and Aloui, 2015, Chakkalakal et al., 2018]. GARCH models provide univariate predictions, and copulas are used to model asset dependence, allowing them to combine independent models.

In this context, it is possible to extend the use of copula functions along with different prediction models. More recently, in the literature, there is evidence of more efficient predictions of volatility using high-frequency data compared with daily historical data (Fleming et al. [2003], Caldeira et al. [2017]), and machine learning algorithms have not been much explored.

The motivation of the dissertation is to combine these techniques to generate future accurate scenarios. Generating reliable scenarios can benefit investors and portfolio managers who optimally allocate the weights according to some optimization criterion.

1.2 Research problem

The problem of the dissertation is to generate the future multivariate probability distribution of N assets in one day ahead horizon. This probability distribution is also known as a scenario matrix, where columns represent assets and rows represent different simultaneous simulations of the asset returns for an instant t can represent this distribution.

1.3 Objective

This dissertation proposes a new methodology capable of generating this future multivariate probability distribution of a set of assets based on machine learning regression models as forecasting predictors of future behavior (volatility and correlation). This new methodology uses machine learning algorithms and intraday measurements to extend the GARCH-Copula framework in the literature. We expected the proposed approach to generate more accurate scenarios than traditional baselines.

1.4 Research Hypothesis and Questions

The fundamental hypothesis of the dissertation is that, through machine learning algorithms, it is possible to achieve more accurate predictions compared to other classical models in the literature. Consequently, it is possible to obtain more precise future scenarios by employing superior predictions. Below there is a list of research questions related to the development of the dissertation:

1. Is the volatility of assets (represented by realized variance) dynamic over time?

2. Is the correlation structure between assets dynamic, and can we use intraday Kendall correlation to adjust copula functions?

3. Is it possible to increase the performance forecast of volatility and correlation of stocks using machine learning algorithms compared with traditional econometric forecasting models?

4. Can this methodology generate more accurate scenarios than other traditional literature techniques and achieve lower risk metrics using optimization portfolio models?

1.5 Main contribution

The main contribution of the dissertation is the development of a new methodology that combines realized volatility and correlation, machine learning regression algorithms, and a copula function to generate future scenarios to allocate the portfolio weights based on an optimization model. In the literature, several studies use some of these different techniques separately, but none use all of these cited techniques combined. This approach's potential benefit is supporting investment decisions with better portfolio allocation to minimize the risk.

1.6 Chapter organization

In Chapter 2, we review the literature on scenario generation models and intraday volatility forecasting. Next, in Chapter 3, we present the theoretical foundations, and in 4 we describe our methodology. In Chapter 5, we present our computational experiments, and in Chapter 6 we discuss the conclusion of the dissertation and future experiments.

Chapter 2

Literature review

In this chapter, we bring relevant papers in the context of the proposed methodology. In the first section, we focus on some of the more traditional econometric volatility modeling techniques. Further, we exploit the machine learning algorithms used to model volatility in the literature. Next, we mention the common techniques related to the problem of portfolio scenario generation, with an emphasis on copula GARCH approaches that are the core of our proposed methodology. In the final section, we compared our proposed methodology with the cited papers in the literature discussing similarities and main innovations.

2.1 Modeling and forecasting volatility

Over history, several models have been proposed to estimate and forecast future volatility using different approaches. Engle [1982] proposed the ARCH process assuming a normal distribution for the returns with constant mean and time-varying conditional variance. Bollerslev [1986] extended the idea with Generalized ARCH models (GARCH) by adding a lagged variance term in the conditional equation. This model became popular since it has few parameters, can generally explain the major stylized facts of returns [Cont, 2001], and has been empirically shown to produce forecasts with good accuracy [Taylor, 2007]. Several extensions of the traditional GARCH models have been proposed in the literature, hoping to replicate different properties observed in the asset return series [Nelson, 1991, Baillie and Mikkelsen, 1996].

The use and estimation of realized volatility emerged with the availability of high-frequency data when researchers began using the intraday sum of squared return as a proxy of assets. Andersen and Bollerslev [1998] showed that realized volatility was a better estimator of the true volatility of the assets and Andersen et al. [2001] demonstrated some of the statistical properties of the measure. The log of realized volatility was shown to have a distribution similar to normal and analysis of the auto-correlation function suggested a long memory process. For this reason, one of the first approaches to model realized volatility was using the autoregressive fractionally integrated moving average (ARFIMA) models. Later, Corsi [2009] proposed a heterogeneous autoregressive model (HAR), which also captured long-term memory while responding to short-term shocks. This model became popular because of its simplicity and empirically good performance. Some papers demonstrated superior accuracy in forecasting volatility using the realized volatility models against daily models [Pong et al., 2004, Izzeldin et al., 2019].

One important extension of the HAR model was proposed by Bollerslev et al. [2016] adding the realized quarticity measure to the model (HARQ), which increased the accuracy of forecasts. The economic value in forecasting volatility has been studied by Fleming et al. [1999], where the authors developed a methodology to evaluate the performance of a dynamic portfolio using forecasts of the covariance matrix and compared it with a static efficient portfolio. The authors showed that volatility timing strategies outperformed the static portfolio. The same methodology was used by Fleming et al. [2003] where the authors evaluated the performance of switching from daily to intraday returns to estimate the covariance matrix. The gains using realized volatility were substantial. Caldeira et al. [2017] compared the performance of covariance matrices forecasts using high frequency and low frequency in the Brazilian market. They used a multivariate GARCH framework and showed that realized covariance estimators performed significantly better than standard estimators.

Models that forecast covariance matrices must ensure that the estimates are positive definite. Different papers use different types of transformations to guarantee this property. Bauer and Vorkink [2011] developed a latent factor model using a matrix logarithmic transformation. The use of the transformation is also beneficial since it does not impose parameter restrictions for the predictions. Chiriac and Voev [2011] used a Cholesky matrix decomposition to forecast the Cholesky series of covariance matrices of a portfolio. The positivity of forecasts is ensured by squaring the reverse matrix transformation. They used a vector extension of the HAR and ARFIMA models. T. et al. [2018] also used the Cholesky decomposition and other transformations to extend the HARQ model for the multivariate case.

2.2 Forecasting volatility using machine learning algorithms

Some limitations of econometrics models are related to linear structure, parameter restrictions, and distribution assumptions. The fields of artificial intelligence and machine learning are getting attention with the development of nonlinear models, especially the use of neural networks that can be applied in this context by circumventing these disadvantages. There are a large number of papers that use hybrid neural networks along with GARCH models to improve the forecasts of volatility [Monfared and Enke, 2014, Kristjanpoller and Minutolo, 2018, 2014, Roh, 2007, Donaldson and Kamstra, 1997]. More recently, recurrent neural networks have achieved positive results forecasts in this context [Kim and Won, 2018, Liu, 2019].

In addition to the large number of studies using neural networks with daily volatility models over the past decades, there has been a recent surge in studies investigating the same concept in the context of realized volatility realized volatility, marking it as a current and evolving topic. Arnerić et al. [2018] compared the performance of HAR models with feed-forward neural networks. Björnsjö [2020] used different types of deep learning models that were implemented and compared with extensions of HAR models, while Vortelinos [2017] compared the performance of HAR against Principal Components Combining, neural networks, and a GARCH model. Although in-sample some neural network models performed well in all of these studies, the forecast performance in the out-of-sample was not superior to HAR models.

The benefit of combining neural networks with realized volatility can be seen in Maciel et al. [2017]. The authors proposed a hybrid neural fuzzy network with jump methodology to forecast the realized volatility of S&P 500, NASDAQ, FTSE, DAX, IBEX and Ibovespa indexes. The results showed that the nonlinear models outperformed the traditional linear regression approach of the HAR modeling. Also, Bucci [2020b] demonstrated the performance of a feed-forward and recurrent neural network with ARFIMAX and ARFIMA models for the logarithm of the Standard & Poor's (S&P) index. Using two different types of loss functions, the recurrent neural networks presented significantly better results.

In our literature review, we found few studies that used neural networks to forecast realized covariance matrices. Bucci [2020a] investigates the use of different types of neural networks to forecast the Cholesky factors of the realized covariance matrix, comparing the results with two vector auto-regressive models and the DCC GARCH model. The results showed statistically better performance using the neural networks against the baselines. Mesquita et al. [2020] used a multilayer perceptron neural network along with the HAR model to predict one day ahead of the covariance matrices. The predictions were employed in the optimization model of minimizing investment portfolio variance, and the results demonstrated superiority over the baselines. This study showcases the application and benefits of utilizing these enhanced predictions compared to other classical models in the literature.

Çepni et al. [2022] employed machine learning techniques to predict the realized variance of oil price returns. They showed that using aggregate economic-policy uncertainty as a predictor it was possible to improve the accuracy of forecasts of the realized variance/volatility of oil-price returns at intermediate and long forecast horizons. Christensen et al. [2022] compared several machine learning models with the HAR models and have shown that machine learning algorithms outperformed the baseline models. The gains were more significant at longer horizons, and the authors suggest that the high persistence in machine learning models facilitates capturing the long-term memory of realized variance. Zhang et al. [2023] also utilized various machine learning models and suggest that neural network models outperform linear regressions and tree-based models in terms of performance. This superiority is attributed to their capability to uncover and model complex latent interactions among variables. An interesting observation made in the study is that the results remain robust when they applied to new stocks not included in the training set, thereby offering novel empirical evidence supporting a universal volatility mechanism across stocks.

Gunnarsson et al. [2024] conduct a systematic literature review in the context of employing machine learning models to forecast realized volatility. They affirm that machine learning models, particularly memory-based neural networks (LSTM and GRM), stand among the top predictors, often being comparatively superior or equivalent to traditional baselines. This is largely attributed to the algorithms ability to capture nonlinear relationships, utilize diverse feature variables, and handle large volumes of data. However, a drawback is that these algorithms are black-box in nature, making it challenging to interpret the predictions. Souto and Moradi [2024] showed that neural basis expansion analysis with exogenous variables (NBEATSx) consistently showed statistically more accurate and robust forecasts than the other considered models (LSTM, HAR and GARCH models).

2.3 Scenario generation techniques

The drawbacks of the mean-variance framework previously discussed, have given rise to single-period portfolio selection models based on risk measures that can be optimized, in the case of discrete random variables, with linear programming (LP) models. The discrete random variables are asset returns defined by their executions under different scenarios. The first LP model for portfolio selection was proposed by Yitzhaki [1982], which used the Gini's mean (absolute) difference as a risk measure. However such models gained traction when Konno and Yamazaki [1991] proposed an LP model that optimized the mean-absolute deviation. Several other models, such as CVaR optimization [Rockafellar and Uryasev, 2000] and improvements regarding exogenous constraints have been since proposed. For more details, we refer the reader to Mansini et al. [2014].

Regardless of the risk measure chosen, the out-of-sample performance of a portfolio chosen via optimization depends strongly on its input data. Hence, another line of research that has become popular is scenario generation - how to best approximate the true multivariate distribution of asset returns with a discrete representation. The most common approach is using historical data as scenarios, which as discussed earlier may have some advantages but does not necessarily reflect the actual future distribution. There is also the problem of limited available data (only historical realizations), which has been previously dealt with the use of bootstrapping. There are however several alternative scenario generation methods, such as based on Monte Carlo simulation and moment-matching - for a review and comparative evaluation we refer the reader to Kaut and Wallace [2003], Guastaroba et al. [2009].

Monte Carlo-based methods assume that the underlying multivariate distribution is known. The most common approaches are to sample from a multivariate Normal distribution or, to account for heavy tails, a multivariate t-Student distribution with various degrees of freedom (depending on the asset). Other methods allow more general marginal distributions [Cario and Nelson, 1997, Lurie and Goldberg, 1998].

Moment-matching methods are used when the marginal distributions are not known, but its moments have been estimated [Vale and Maurelli, 1983, Smith, 1993, Kouwenberg, 2001]. Date et al. [2008] proposed a method for matching moments from partially specified distributions. This method was extended by Ponomareva et al. [2015] to account for asymmetric marginals. Høyland and Wallace [2001] proposed a nonlinear optimization model for generating discrete scenarios that minimize the square of the difference between a set of targets and the actual statistical properties of the scenarios. In subsequent work, Høyland et al. [2003] proposed a heuristic that generates discrete scenarios given target values for the first four marginal moments and the correlation matrix.

Alternatively, some methods attempt to predict the underlying multivariate distribution, usually via time-series forecasting models, then sample from the estimated distribution. Messina and Toscani [2008] proposed a scenario generation approach based on Hidden Markov Models that is capable of dynamically switching between multiple states, assuming that the underlying distribution is not necessarily described by a single model (as seen in bull and bear markets, for instance). In the univariate setting, the most common approaches are based on GARCH models [Bollerslev, 1986], which assume heteroskedasticity of returns time series. In these cases, researchers often assume that the data follows a normal or t-student distribution, and the dependence among assets is assumed to be linear and symmetric, modeled through a correlation matrix [Bauwens et al., 2006]. As is discussed in Embrechts et al. [2002] the use of linear correlation to model the dependence structure shows many disadvantages and limitations.

2.3.1 Volatility Copula models

Copulas functions Sklar [1959] have gained popularity in risk management as the dependence structure of the market can be isolated from the univariate marginal distributions [Joe, 1997, Nelsen, 2006]. In the literature, many papers use GARCH models with copula functions to model the volatility of assets and the non-linear and asymmetric dependence between them [Hsu et al., 2008, Huang et al., 2009, Liu and Luger, 2009, Koliai, 2016, Karmakar, 2017].

In Wang et al. [2010], Deng et al. [2011], Sahamkhadam et al. [2018] the authors combined extreme value theory, univariate GARCH models, and Copulas for modeling assets multivariate distributions. The GARCH model was applied in historical returns using the combination of two distributions to model the residuals: the Generalized Pareto Distribution for the upper and lower tail, and a Gaussin kernell for the middle part. They generated future distributions for each asset using GARCH models and modelled the correlation structure with copulas calibrated with historical data. They generated scenarios with Monte Carlo simulation for estimating risk. The use of extreme value theory can generate more flexible marginal distributions.

Fengler and Okhrin [2016] discuss that recently in the literature many studies have been exploring the theme of dynamic copula models [Dias et al., 2004, Patton, 2004, 2006, Chen and Fan, 2006, Jondeau and Rockinger, 2006, Enzo Giacomini and Spokoiny, 2009, Jin, 2009, Hafner and Manner, 2012, Creal et al., 2013, Härdle et al., 2013], where the parameters of the copula function vary over time and, in many cases, can be modeled by a time series process. Almeida and Czado [2012] propose a stochastic copula autoregressive model using inverse Fisher transformation of Kendall's tau for bivariate copulas. So and Yeung [2014] propose a vine-copula GARCH with dynamic conditional dependence. The authors developed a methodology to model the dependence between assets using any dependence measure and they performed experiments using linear correlation, rank correlation, and Kendall correlation. The use of Kendall correlation is an interesting approach because there is a direct relationship with a bivariate copula function [Alexander, 2008].

The vast majority of studies employ daily data when combining volatility forecasting models with copula functions. More recent papers extend these concepts by utilizing intraday data to model daily volatility and the correlation structure. De Lira Salvatierra and Patton [2015] propose a new class of dynamic copula models for daily asset returns that exploit information from high frequency data. The authors augmenting the generalized autoregressive score model (GAS) with realized correlation which improved the in-sample fit and out-of-sample density forecast. Fengler and Okhrin [2016] propose to predict the intraday covariance matrix based on HAR models. They estimate the copula parameters by covariance moment condition provided by Hoeffding's lemma. This approach allow the time dependecy structure to be time-varyng day by day. Both papers also apply their methodology in portfolio choice problem showing gains comparing with other classical approaches.

When analyzing the works cited in the literature, it is possible to perceive that the copula approach is quite interesting and flexible, allowing for:

1. Modeling non-linear dependences separately from the marginal distributions [Joe, 1997, Nelsen, 2006]

2. Utilization of different volatility forecasting models: GARCH Wang et al. [2010], Stochastic Volatility Hafner and Manner [2012], GAS [Creal et al., 2013], HAR [Fengler and Okhrin, 2016],

3. Deployment of models that use data from different frequencies: daily [Wang et al., 2010, Sahamkhadam et al., 2018, Almeida and Czado, 2012], intraday [De Lira Salvatierra and Patton, 2015, Fengler and Okhrin, 2016]

4. Implementation of dynamic copulas [Patton, 2006, Dias et al., 2004]

5. Adoption of various marginal distributions [Sahamkhadam et al., 2018]

6. Capable to generate the multivariate distribution and consequently use different optimization portfolio models [Sahamkhadam et al., 2018, Goel et al., 2019]

For these reasons the methodology we propose follows the approach of employing volatility forecasting models in conjunction with a dynamic copula function to generate a future multivariate distribution. Optimization models are then developed based on the generated distribution.

2.4 Comparative analysis of literature and our proposed approach

We propose the integration of four key elements for the new methodology: the utilization of machine learning algorithms as forecasting models, the incorporation of intraday data for predicting volatility and correlation, the application of a copula function to model the correlation structure, and the adoption of extreme value theory (EVT) for marginal distributions. In this manner, this represents the first methodology that combines all these elements [Mesquita et al., 2023]. The advantage is associated with each of the individual elements:

- The use of intraday data has greater predictive power compared to daily data.
- The utilization of machine learning exhibits greater predictive power in comparison to classic forecasting models.
- The application of a methodology for predicting intraday Kendall correlation ensures a dynamic correlation structure.
- The incorporation of marginal distributions through EVT distributions enhances flexibility.

Chapter 3

Theoretical foundation

In this chapter, we introduce the essential techniques employed in our methodology. The initial sections encompass the exposition of the GARCH model, Extreme Value Theory, and realized variance. These three concepts collectively serve as the foundation for modeling the univariate distributions of assets. Subsequently, we introduce copula functions to model the correlation structure between assets. Furthermore, we define the STARR ratio portfolio optimization model for effective asset allocation. Finally, we present the principles of supervised machine learning regression, which will be used in forecasting volatility and correlation.

3.1 GARCH model

The GARCH(1,1) model [Bollerslev, 1986] with normal distributions states the distribution of return for period t, conditional on all previous returns is

$$r_t \mid r_{t-}, r_{t-2}, \ldots \sim N(\mu, h_t) \tag{3.1}$$

with residuals

$$e_t = r_t - \mu \tag{3.2}$$

and the standardized residuals

$$z_t = \frac{e_t}{\sqrt{h_t}}. \tag{3.3}$$

The formal definition of the model GARCH(1,1) with conditional normal distributions is

$$r_t = \mu + h_t^{\frac{1}{2}} z_t, \tag{3.4}$$

$$z_t \sim i.i.d.\ N(0,1), \tag{3.5}$$

$$h_t = \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1} \tag{3.6}$$

with variance h_t, parameters $\alpha, \beta, \omega, \mu$, and constraints $\beta \geq 0$, $\alpha \geq 0$ $\omega \geq 0$ required to ensure positive variance. The variance h_t is a function of the previous squared residual and previous variance.

As discussed in [Taylor, 2007], empirical evidence has contradicted the assumption that returns have conditional normal distributions. The distribution of estimated standardized residuals obtained from observed returns and parameters has excess kurtosis and the assumption $z_t \sim N(0,1)$ is unideal for the satisfactory of return process. Some extensions of the model use other distributions (standardized t-distribution, generalized error distribution [Nelson, 1991]), which can be denoted by $D(0,1)$ and $z_t \sim i.i.d.\ D(0,1)$.

3.2 Extreme Value Theory - EVT

As is discussed in Longin [2017], Extreme Value Theory (EVT) serves as the foundation for investigating the asymptotic distribution of extreme or rare events, which are considerably larger in magnitude compared to the majority of observations. Grounded in a solid theoretical framework, EVT provides the basis for constructing parametric or semiparametric statistical models designed to handle rare events. It is particularly well-suited for modeling and quantifying events that occur with a very low probability. EVT has demonstrated its effectiveness as a robust and valuable tool for describing unusual scenarios that could have a significant impact across various application domains, especially in situations where understanding the behavior of the tail of the distribution is crucial.

Following the paper of Scarrott and MacDonald [2012], Balkema and de Haan [1974], Pickands [1975] demonstrated that if there is a non-degenerate limiting distribution for appropriately linearly rescaled excesses of a sequence of independent and identically distributed observations $X1, ..., Xn$ above a threshold u, then the limiting distribution will be a Generalized Pareto Distribution (GPD). In practical applications, the GPD is employed as a tail approximation to the population distribution, from which a sample of excesses $x - u$ above some sufficiently high threshold u is observed. The GPD is parameterized by scale and shape parameters $\sigma_u > 0$ and ξ, and can be alternatively specified in terms of threshold excesses $x - u$ or, as presented here, exceedances $x > u$

$$G(x|u, \sigma_u, \xi) = Pr(X < x|X > u) = \begin{cases} 1 - [1 + \xi(\frac{x-u}{\sigma_u})]_+^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - \exp[-(\frac{x-u}{\sigma_u})]_+, & \xi = 0 \end{cases} \tag{3.7}$$

where $y_+ = \max(y, 0)$. Implicitly underlying the GPD is a third parameter required for estimation of quantities like return levels, the proportion of threshold excesses $\phi_u =$

$Pr(x > u)$ used to calculate the unconditional survival probability

$$Pr(X > x) = \phi_u[1 - Pr(X < x|X > u)]. \tag{3.8}$$

3.3 Realized variance

Following Bollerslev et al. [2016], consider that the price series P_{it} of asset $i = 1, \ldots, N$ at time t follows a stochastic differential equation:

$$d\log(P_{it}) = \mu_{it}\,dt + \sigma_{it}\,dW_t, \tag{3.9}$$

where t represents time, μ represents the drift, σ represent the asset volatility and W_t is a Brownian motion. We would like to predict σ_{it+1}. The one-day integrated variance (IV) of asset i is defined as,

$$\mathrm{IV}_{it} = \int_{t-1}^{t} \sigma_{is}^2\,ds\,. \tag{3.10}$$

IV_{it} cannot be predicted as it is not directly observable. It can however be approximated by using high-frequency returns to calculate the realized variance (RV), which is defined as

$$\mathrm{RV}_{it} = \sum_{j=1}^{M} r_{itj}^2 \tag{3.11}$$

where $M = 1/\Delta$ and the Δ-period intraday return is defined as

$$r_{it,j} = \log(P_{i,t-1+j\Delta}) - \log(P_{i,t-1+(j-1)\Delta}). \tag{3.12}$$

3.4 Copula functions

A copula is a multivariate cumulative distribution function with each marginal following a uniform distribution. In finance, they are commonly used to model the dependence between random variables (assets), with applications in portfolio optimization models.

Following Alexander [2008], consider two random variables X_1 and X_2 with continuous marginal distributions $F_1(x_1)$ and $F_2(x_2)$, and $u_i = F_i(x_i)$, for $i = 1, 2$. The copula function must ensure four properties [Alexander, 2008]:

1. $C : [0,1] \times [0,1] \rightarrow [0,1]$;
2. $C(u_1, 0) = C(0, u_2) = 0$;
3. $C(u_1, 1) = u_1$ and $C(1, u_2) = u_2$;
4. $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$, for every $u_1, u_2, v_1, v_2 \in [0,1]$ with $u_1 \leq u_2$ and $v_1 \leq v_2$.

The Sklar Theorem [Sklar, 1959] shows that for any joint distribution function $F(x_1, x_2)$ there is a unique copula function $C : [0,1] \times [0,1] \rightarrow [0,1]$ that

$$F(x_1, x_2) = C(F(x_1), F(x_2)), \tag{3.13}$$

and distinct copulas define distinct joint densities.

It is possible to express any multivariate joint distribution using univariate marginal distribution functions and an associated copula function that delineates the dependence structure among the variables.

The copula conditional distribution of X_1 given X_2 is defined as

$$C_{1|2}(u_1 \mid u_2) = P(U_1 < \mu_1 \mid U_2 = \mu_2) = \frac{\partial C(\mu_1, \mu_2)}{\partial \mu_2}. \tag{3.14}$$

All the concepts above can be generalized to n dimensions.

For n random variables, the Gaussian copula is defined as

$$C(u_1, ..., u_n; \Sigma) = \boldsymbol{\Phi}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)), \tag{3.15}$$

where $\boldsymbol{\Phi}$ and Φ are the multivariate and univariate standard normal distribution functions respectively, and Σ is the correlation matrix between the random variables.

Another elliptical copula is the Student t can be represented as

$$C(u_1, ..., u_n; \Sigma) = \mathbf{t_v}(t_v^{-1}(u_1), ..., t_v^{-1}(u_n)), \tag{3.16}$$

where $\mathbf{t_v}$ and t_v are multivariate and univariate Student t distribution functions with degrees v of freedom, respectively.

The Clayton copula is assimetric copula function defined as

$$C(u_1, ..., u_n; \theta) = ((u_1)^{-\theta}, ..., (u_n)^{-\theta} - n + 1)^{-\frac{1}{\theta}}, \tag{3.17}$$

with $\theta \neq 0$ representing the copula parameter.

3.4.1 Calibrating the copulas

Following Demarta and McNeil [2005], there is a simple way of calibrating the correlation matrix of the elliptical copulas using Kendall's tau empirical estimates for each bivariate margin of the copula.

Rank correlations are non parametric dependence measures based on ranked data [Alexander, 2008]. If the data is composed of continuous variables, they are converted to the ranked form, retaining only the ranks of the observations. At time t, Consider the joint ranked intraday returns of two assets i and k as $(r_{it1}, r_{kt1}), (r_{it2}, r_{kt2}), ..., (r_{itM}, r_{ktM})$. Two pairs of observations l and m are said to be concordant if $(r_{itl} - r_{itm}) \times (r_{ktl} - r_{ktm}) > 0$. Accordingly the pairs are discordant if $(r_{itl} - r_{itm}) \times (r_{ktl} - r_{ktm}) < 0$. The Kendall correlation is calculated by comparing all possible pairs of observations $\{(r_{itl}, r_{ktl}), (r_{itm}, r_{ktm})\}$ for $l \neq m$, and is defined as:

$$\tau_{ikt} = 2 * \frac{N_C - N_D}{M(M-1)} \tag{3.18}$$

where N_C is the number of concordant pairs, N_D is the number of discordant pairs and M is the number of observations.

It can be shown that a bivariate copula $C(u_1, u_2)$ and the Kendall correlation are related by

$$\tau = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1 \tag{3.19}$$

and any elliptical copula has a correlation parameter equal to

$$\rho = sin(\frac{\pi}{2}\tau). \tag{3.20}$$

In the case of the Clayton copula function, the Kendall correlation is related to the parameter of the function as follows:

$$\theta = 2\tau(1-\tau)^{-1}. \tag{3.21}$$

3.4.2 Tail dependency

Tail dependence examines the concordance in the extreme values of the joint distribution. Karmakar [2017] discuss that the tail dependency measures the probability that two variables are in the lower or upper joint tails. In this context, the tail dependence

Table 3.1: Tail dependence coefficient of t-student copula function

v/p	**-0.5**	**0**	**0.5**	**0.9**	**1**
2	0.06	0.18	0.39	0.72	1
4	0.01	0.08	0.25	0.63	1
10	0.0	0.01	0.08	0.46	1
∞	0	0	0	0	1

coefficient serves as an indicator of the inclination of markets to experience simultaneous crashes or booms.

As discussed in Demarta and McNeil [2005], the Gaussian copula has zero tail dependence (for $\rho < 1$) while the Student t copula has a positive value. The tail dependence coefficient is given by:

$$\lambda = 2t_{v+1}(-\sqrt{v+1}\sqrt{1-\rho}/\sqrt{1+\rho}) \tag{3.22}$$

where ρ is the off-diagonal element of correlation matrix and v is the number of degrees of freedom. In Table 3.1 we show the tail dependence coefficient for different values of v and ρ. We observe that higher values of correlation and low degrees of freedom increase the coefficient of tail dependence in the t-Student copula function.

The Clayton copula is known for its ability to capture strong dependence in the lower tails, making it suitable for modeling instances where extreme events tend to occur together. It has zero tail dependence in the upper tail and a positive lower tail dependence coefficient, when $\theta > 0$, with

$$\lambda^l = \begin{cases} 2^{-\frac{1}{\theta}}, & \theta > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.23}$$

3.5 STARR ratio optimization model

Let N be the number of assets in which we can invest and let $R \in \mathbb{R}^N$ be a multivariate random variable representing asset returns. Suppose that R is discrete and composed of S equiprobable scenarios; each scenario s being represented by a vector $\mathbf{r_s} = r_{1s}, ..., r_{Ns}$. Then given portfolio weights $\mathbf{w} \in \mathbb{R}^N$, we have that portfolio returns are given by

$$r_s^p = \mathbf{w}^T\mathbf{r_s} = \sum_{i=1}^{N} w_i r_{is}, \forall s = 1, ..., S \tag{3.24}$$

Let $r_{(s)}, s = 1, .., S$ denote the sorted portfolio returns in each scenario in increasing order. Then for a given α (interpreted as confidence level $100(1-\alpha)$, the Conditional-Value-at-Risk (CVaR) [Rockafellar and Uryasev, 2000] for discrete distributions with S scenarios is given by

$$CVaR_\alpha(\mathbf{w}^T\mathbf{R}) = \frac{1}{\lfloor S_\alpha \rfloor}\sum_{s=1}^{\lfloor S_\alpha \rfloor}(-1)r_s^p \tag{3.25}$$

where $\lfloor \alpha \rfloor$ is the largest integer smaller or equal to a 1.

The STARR ratio [Martin et al., 2003] is the reward-risk measure associated with CVaR, and is defined as

$$STARR(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{u} - r_f}{CVaR_\alpha(\mathbf{w}^T\mathbf{R})} \tag{3.26}$$

The STARR ratio can be optimized as

$$\begin{aligned} \text{maximise} \quad & \sum_{i=1}^{N} u_i \hat{w}_i \\ \text{subject to} \quad & \theta + \frac{1}{S_\alpha}\sum_{s=1}^{S} d_s \leq 1 \\ & d_s \geq -\sum_{i=1}^{N} r_{is}\hat{w}_i - \theta, \quad \forall s = 1\ldots, S \\ & d_s \geq 0, \quad \forall s = 1\ldots, S \\ & \sum_{i=1}^{N} w_i = t \\ & t \geq 0 \end{aligned} \tag{3.27}$$

where decision variable θ is the negative value of the Value-at-Risk (VaR), and variables d_s represent either the positive difference between a return and the VaR in the case that return is worse than the VaR. If the return is equal or better than VaR d_s is equal to zero.

3.6 Supervised machine learning regression

It is a machine learning task that tries to learn a continuous input-output mapping from a limited number of examples. These inputs are a set of $(X, y)^m$ pairs where $X \in R^d$ and y is a continuous output target. This set is called training and each dimension d is a different feature. The goal is to learn a function $f : X \to y$ to predict correctly on new inputs of X (generalization).

Different machine learning algorithms can be used to perform the regression task, and each one has particular approaches with different hyperparameters that need to be chosen. Commonly in the training stage of these algorithms, the hyperparameters are optimized by minimizing a specific loss function.

3.6.1 Neural Networks

The idea behind artificial neural networks emerged in Rosenblatt [1958] when researchers at the time were trying to simulate the functioning of brain cells computationally. A neural network is a mathematical model with the ability to learn and generalize by the use of previous training examples.

Neural networks are composed of parallel nodes (neurons) responsible for calculating a certain mathematical function related to the learning of the network. These neurons are organized in layers that generally interconnect in a unidirectional way through different connections. Most of the time, each neuron is associated with a weight that stores the model learning and is responsible for weighing the input received to each neuron in the network.

To make the predictions it is necessary to train the network with a series of data so that learning occurs and the network can then infer the next values. Artificial neural networks can approximate functions, are robust and fault-tolerant. Because of these characteristics, this kind of model become good candidates for forecasting nonlinear systems and non-stationary time series, as is the case with some financial time series.

3.6.2 Random Forest

Random Forest was proposed by Breiman [2001], and it is a type of ensemble machine learning algorithm that combines multiple trees and calculates the average prediction in the regression task. It uses the bootstrap aggregation (bagging) technique which selects random samples with the replacement of the training and fits the trees with these samples. The advantage of this approach is to reduce the variance of the model since each tree received different training examples.

Another aspect of the algorithm is that each tree received a random subset of the features to perform the training. In many cases, for a total of N features, $\sqrt{N}$ features are used in each split.

3.6.3 eXtreme Gradient Boosting (XGBoost)

The XGBoost is end-to-end tree boosting system proposed by Chen and Guestrin [2016]. The boosting technique consists of building a strong predictor by iterative combining several weak predictors.

In the case of XGBoost, the weak predictors are regression trees using a gradient descent algorithm to minimize the loss function. Each tree is added in an iterative way using a subset of the training data, which helps to reduce model variance. The optimization function of the algorithm uses regularization techniques (adding penalty as model complexity increases) which also helps to prevent overfitting.

Chapter 4

Methodology

This chapter describes our approach to modeling the asset return process to simulate the one-day-ahead scenario matrices. We use the simulations as input to the STARR Ratio optimization model. We divided the methodology into the in-sample set and the out-of-sample set. To fit the models, we use the in-sample and out-of-sample set is responsible for forecasting and simulating the scenario matrices.

Figure 4.1 illustrates the methodology. Step 1 consists of collecting the intraday historical data of the assets. Step 2 consists of preparing the data (clean, transform, feature engineering) and splitting in-sample (training and validation sets) and out-of-sample (test set). In step 3, we fit the distributions of standardized returns using extreme value theory. Step 4 uses machine learning regression algorithms to predict the time-varying volatility and correlation of assets. In step 5, we combine the marginal distributions with the machine learning predictions and a copula function to generate discrete scenarios. This future multivariate probability distribution simulation is represented as step 6 of the Figure. The last step is to apply a portfolio optimization model in the simulation and find the optimum weights. More specifically, for each different day of the out-of-sample set, steps 4–7 comprise the following procedure:

1. For each asset, predict their realized variance one day ahead,
2. For each pair of assets, predict their respective intraday Kendall correlation one day ahead,
3. Adjust the correlation matrix parameter of the copula function using the Kendall correlation estimates,
4. Generate discrete scenarios based on the marginal distributions from step 3, with variances equal to the realized variance predictions, and the estimated copula,
5. Apply the STARR Ratio optimization model to calculate portfolio weights to be employed the next day.

We expect that more accurate forecasts, if achievable, will better approximate the actual optimal weights of the portfolio. In the following subsections, we present each step in more detail.

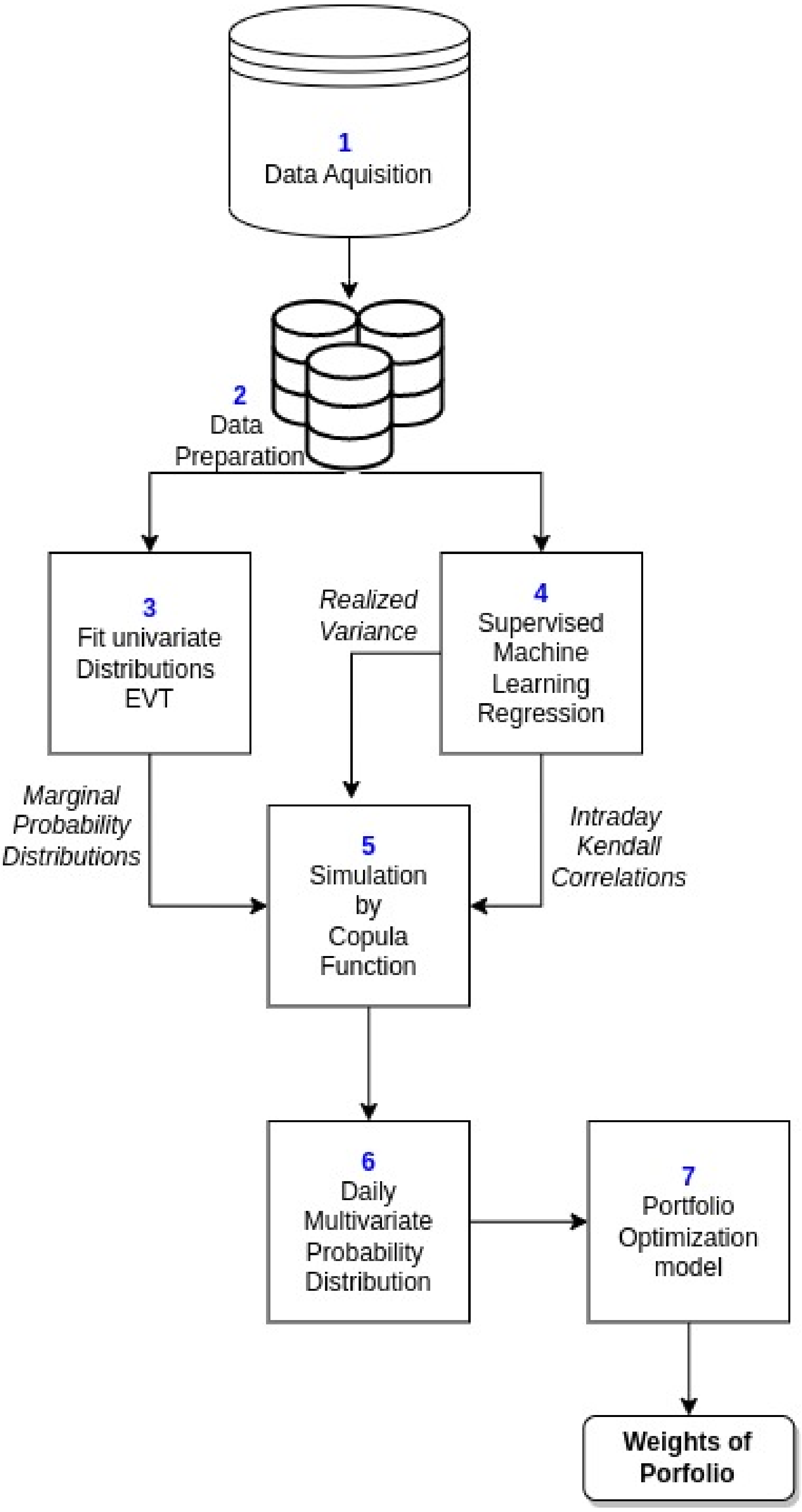


Figure 4.1: Steps of methodology

4.1 Data Preparation

We collect historical data on 29 B3 assets between 2008 and 2022, a total of approximately 14 years. We chose these assets because they represent different sectors of the Brazilian economy and exhibit higher liquidity in the market. All assets have periodicity of 5 minutes. We use the MetaTrader platform to collect the data.

The list of assets is: Ambev (ABEV3), Alpargatas (ALPA4), Banco do Brasil (BBAS3), Banco Bradesco (BBDC4), Braskem (BRKM5), Grazziotin (CGRA4), Companhia Energética de Minas Gerais (CMIG4), CPFL Energia (CPFE3), Cia Paranaense De Energia Copel (CPLE6), Atacadao (CRFB3), Companhia Siderurgica Nacional (CSNA3), Eletrobras (ELET3), Embraer (EMBR3), Energisa (ENGI4, ENGI11), Gerdau (GGBR4), Itau (ITSA4), Itau Unibanco (ITUB4), Light (LIGT3), Lojas Renner (LREN3), Petrobras (PETR4), Raia Drogasil (RADL3), Companhia de Saneamento Bsc DEDSP (SBSP3), Suzano (SUZB3), TIM Brasil Serviços e Participações (TIMS3), Unipar Carbocloro (UNIP6), Usinas Siderurgicas de Minas Gerais (USIM5), Vale (VALE3), Telefonica Brasil (VIVT3), and the Brazilian ETF index BOVA11.

The collected data corresponds to the opening, closing, high, low prices, volume, and tick-volume of each 5-minute candle $(\mathrm{PO}_{t,j}, \mathrm{PC}_{t,j}, \mathrm{PH}_{t,j}, \mathrm{PL}_{t,j}, \mathrm{v}_{t,j}, \mathrm{tv}_{t,j})$. The t index represents the day, while the j index represents an instant of 5 minutes during the day. The period from 2008 to 2020 was used as an in-sample to adjust the models (which we discuss in more detail in the following sections). Moreover, the out-of-sample period corresponds to 2021 to 2022.

We construct features that the machine learning algorithms use to predict the intraday moments and Kendall correlations. We define:

$$\mathrm{oc}_{t,j} = \log(\mathrm{PO}_{t,j}) - \log(\mathrm{PC}_{t,j}) \tag{4.1}$$

$$\mathrm{hc}_{t,j} = \log(\mathrm{PH}_{t,j}) - \log(\mathrm{PC}_{t,j}) \tag{4.2}$$

$$\mathrm{lc}_{t,j} = \log(\mathrm{PL}_{t,j}) - \log(\mathrm{PC}_{t,j}) \tag{4.3}$$

which represents the intraday logarithmic difference between open, high, and low prices with the asset's closing price. In Table 4.1, we defined some features based on the moments of the intraday probability distributions of these measures. Next, we define the intraday log return using the closing price of the assets.:

$$r_{t,j} = \log(\mathrm{PC}_{t,j}) - \log(\mathrm{PC}_{t,j-1}) \tag{4.4}$$

Table 4.1: Features using candle measurements: volume, tickvol, open, high and low prices

index	Feature description	Feature Definition
1	volume intraday mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{v}_{t,j}$
2	volume intraday variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{v}_{t,j} - \bar{\mathrm{v}}_t)^2$
3	volume intraday skewness	$\frac{\sum_{j=1}^{M}(\mathrm{v}_{t,j}-\bar{\mathrm{v}}_t)^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{v}_{t,j}-\bar{\mathrm{v}}_t)^2]^{3/2}}$
4	volume intraday kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{v}_{t,j}-\bar{\mathrm{v}}_t)^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{v}_{t,j}-\bar{\mathrm{v}}_t)^2]^2} - 3$
5	volume intraday variance with zero mean	$\sum_{j=1}^{M} \mathrm{v}_{t,j}^2$
6	tickvol intraday mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{tv}_{t,j}$
7	tickvol intraday variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{tv}_{t,j} - \bar{\mathrm{tv}}_t)^2$
8	tickvol intraday skewness	$\frac{\sum_{j=1}^{M}(\mathrm{tv}_{t,j}-\bar{\mathrm{tv}}_t)^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{tv}_{t,j}-\bar{\mathrm{tv}}_t)^2]^{3/2}}$
9	tickvol intraday kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{tv}_{t,j}-\bar{\mathrm{tv}}_t)^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}-\bar{\mathrm{tv}}_t)^2]^2} - 3$
10	tickvol intraday variance with zero mean	$\sum_{j=1}^{M} \mathrm{tv}_{t,j}^2$
11	intraday open close difference mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{oc}_{t,j}$
12	intraday open close difference variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{oc}_{t,j} - \bar{\mathrm{oc}}_t)^2$
13	intraday open close difference skewness	$\frac{\sum_{j=1}^{M}(\mathrm{oc}_{t,j}-\mathrm{r},\bar{\mathrm{oc}}_t)^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{oc}_{t,j}-\bar{\mathrm{oc}}_t)^2]^{3/2}}$
14	intraday open close difference kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{oc}_{t,j}-\mathrm{r},\bar{\mathrm{oc}}_t)^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{oc}_{t,j}-\bar{\mathrm{oc}}_t)^2]^2} - 3$
15	intraday open close difference variance with zero mean	$\sum_{j=1}^{M} \mathrm{oc}_{t,j}^2$
16	intraday high close difference mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{hc}_{t,j}$
17	intraday high close difference variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{hc}_{t,j} - \bar{\mathrm{hc}}_t)^2$
18	intraday high close difference skewness	$\frac{\sum_{j=1}^{M}(\mathrm{hc}_{t,j}-\bar{\mathrm{hc}}_t)^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}-\bar{\mathrm{hc}}_t)^2]^{3/2}}$
19	intraday high close difference kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{hc}_{t,j}-\bar{\mathrm{hc}}_t)^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{hc}_{t,j}-\bar{\mathrm{hc}}_t)^2]^2} - 3$
20	intraday high close difference variance with zero mean	$\sum_{j=1}^{M} \mathrm{hc}_{t,j}^2$
21	intraday low close difference mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{lc}_{t,j}$
22	intraday low close difference variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{lc}_{t,j} - \bar{\mathrm{lc}}_t)^2$
23	intraday low close difference skewness	$\frac{\sum_{j=1}^{M}(\mathrm{lc}_{t,j}-\bar{\mathrm{lc}}_t)^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{lc}_{t,j}-\bar{\mathrm{lc}}_t)^2]^{3/2}}$
24	intraday low close difference kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{lc}_{t,j}-\bar{\mathrm{lc}}_t)^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{lc}_{t,j}-\bar{\mathrm{lc}}_t)^2]^2} - 3$
25	intraday low close difference variance with zero mean	$\sum_{j=1}^{M} \mathrm{lc}_{t,j}^2$

Table 4.2: Moments features based on intraday log return

index	Feature description	Feature Definition
26	intraday log return mean (μ_t)	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{r}_{t,j}$
27	intraday log return variance (σ_t)	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j} - \bar{\mathrm{r}}_t)^2$
28	intraday log return skewness ($B_{1,t}$)	$\frac{\sum_{j=1}^{M}(\mathrm{r}_{t,j}-\bar{\mathrm{r}_t})^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}-\bar{\mathrm{r}_t})^2]^{3/2}}$
29	intraday log return kurtosis ($B_{2,t}$)	$\frac{\sum_{j=1}^{M}(\mathrm{r}_{t,j}-\bar{\mathrm{r}_t})^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}-\bar{\mathrm{r}_t})^2]^2} - 3$
30	intraday positive log return mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{r}_{t,j}^{+}$
31	intraday positive log return variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{+} - \bar{\mathrm{r}}_t^{+})^2$
32	intraday positive log return skewness	$\frac{\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{+}-\bar{\mathrm{r}}_t^{+})^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{+}-\bar{\mathrm{r}}_t^{+})^2]^{3/2}}$
33	intraday positive log return kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{+}-\bar{\mathrm{r}}_t^{+})^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{+}-\bar{\mathrm{r}}_t^{+})^2]^2} - 3$
34	intraday negative log return mean	$\frac{1}{M}\sum_{j=1}^{M} \mathrm{r}_{t,j}^{-}$
35	intraday negative log return variance	$\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{-} - \bar{\mathrm{r}}_t^{-})^2$
36	intraday negative log return skewness	$\frac{\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{-}-\bar{\mathrm{r}}_t^{-})^3}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{-}-\bar{\mathrm{r}}_t^{-})^2]^{3/2}}$
37	intraday negative log return kurtosis	$\frac{\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{-}-\bar{\mathrm{r}}_t^{-})^4}{[\frac{1}{M}\sum_{j=1}^{M}(\mathrm{r}_{t,j}^{-}-\bar{\mathrm{r}}_t^{-})^2]^2} - 3$

where $PC_{t,j}$ is the closing price of the asset on day t at some instant j and $PC_{t,j-1}$ is the closing price five minutes before. The daily returns r_t are the sum of M intraday returns

$$r_t = \sum_{j=1}^{M} r_{t,j} \tag{4.5}$$

With the intraday log return, we construct several features for the machine learning algorithms. In Table 4.2, we define the intraday moments, while in Table 4.3, we define intraday dispersion metrics.

Regarding the features involving pairs of assets, we collect the data at the same instant of time. As discussed in Massimo Guidolin [2018], we use synchronous time steps to avoid covariance bias. In Table 4.4, we define the intraday association features between pairs of assets.

4.2 Data modeling

In this section, we address the proposed modeling for the multivariate return process of the assets. The goal is to construct a model to sample from a distribution with the same statistical properties as the observed returns.

Table 4.3: Dispersion features based on intraday log return

index	Feature description	Feature Definition
38	Realized variance (RV$_t$)	$\sum_{j=1}^{M} r_{t,j}^2$
39	Logarithmic of realized variance	$\log(\text{RV}_t)$
40	5 days moving average	$\frac{1}{5}\sum_{lag=0}^{5}\log(\text{RV}_{t-lag})$
41	20 days moving average	$\frac{1}{20}\sum_{lag=0}^{20}\log(\text{RV}_{t-lag})$
42	Bipower variation (BPV$_t$)	$(\sqrt{2/\pi})^{-2}\sum_{j=1}^{M-1} \mid r_{t,j} \mid\mid r_{t,j+1} \mid$
43	Logarithmic of bipower variation	$\log(\text{BPV}_t)$
44	5 days moving average ($\log(\text{RV}_{t-5\mid t})$)	$\frac{1}{5}\sum_{lag=0}^{5}\log(\text{BPV}_{t-lag})$
45	20 days moving average ($\log(\text{RV}_{t-20\mid t})$)	$\frac{1}{20}\sum_{lag=0}^{20}\log(\text{BPV}_{t-lag})$
46	Realized variance of positive returns (RV_t^+)	$\sum_{j=1}^{M} r_{t,j}^2\Pi_{\{r_{t,j}>0\}}$
47	Logarithmic of RV_t^+	$\log(\text{RV}_t^+)$
48	5 days moving average ($\log(\text{RV}^+_{t-5\mid t})$)	$\frac{1}{5}\sum_{lag=0}^{5}\log(\text{RV}^+_{t-lag})$
49	20 days moving average (($\log(\text{RV}^+_{t-20\mid t})$))	$\frac{1}{20}\sum_{lag=0}^{20}\log(\text{RV}^+_{t-lag})$
50	Realized variance of negative returns (RV_t^-)	$\sum_{j=1}^{M} r_{t,j}^2\Pi_{\{r_{t,j}<0\}}$
51	Logarithmic of RV_t^-	$\log(\text{RV}_t^-)$
52	5 days moving average $\log(\text{RV}^-_{t-5\mid t})$)	$\frac{1}{5}\sum_{lag=0}^{5}\log(\text{RV}^-_{t-lag})$
53	20 days moving average $\log(\text{RV}^-_{t-20\mid t})$)	$\frac{1}{20}\sum_{lag=0}^{20}\log(\text{RV}^-_{t-lag})$
54	Jump variation (J$_t$)	$\max[\text{RV}_t - \text{BPV}_t, 0]$
55	Logarithmic of J$_t$	$\log(\text{J}_t)$
56	5 days moving average ($\log(\text{J}_{t-5\mid t})$)	$\frac{1}{5}\sum_{lag=0}^{5}\log(\text{J}_{t-lag})$
57	20 days moving average ($\log(\text{J}_{t-20\mid t})$)	$\frac{1}{20}\sum_{lag=0}^{20}\log(\text{J}_{t-lag})$
58	Realized quarticity (RQ$_t$)	$\frac{M}{3}\sum_{j=1}^{M} r_{t,j}^4$
59	Logarithmic of RQ$_t$	$\log(\text{RQ}_t)$
60	5 days moving average ($\log(\text{RQ}_{t-5\mid t})$)	$\frac{1}{5}\sum_{lag=0}^{5}\log(\text{RQ}_{t-lag})$
61	20 days moving average ($\log(\text{RQ}_{t-20\mid t})$)	$\frac{1}{20}\sum_{lag=0}^{20}\log(\text{RQ}_{t-lag})$
62	Moving average indicator rule	$\begin{cases} 1, & \log(\text{RV}_{t-5\mid t}) > \log(\text{RV}_{t-20\mid t}), \\ 0, & \log(\text{RV}_{t-5\mid t}) <= \log(\text{RV}_{t-20\mid t}) \end{cases}$
63	Moving average indicator rule	$\begin{cases} 1, & \log(\text{BPV}_{t-5\mid t}) > \log(\text{BPV}_{t-20\mid t}), \\ 0, & \log(\text{BPV}_{t-5\mid t}) <= \log(\text{BPV}_{t-20\mid t}) \end{cases}$
64	Moving average indicator rule	$\begin{cases} 1, & \log(\text{J}_{t-5\mid t}) > \log(\text{J}_{t-20\mid t}), \\ 0, & \log(\text{J}_{t-5\mid t}) <= \log(\text{J}_{t-20\mid t}) \end{cases}$
65	Moving average indicator rule	$\begin{cases} 1, & \log(\text{RV}^+_{t-5\mid t}) > \log(\text{RV}^+_{t-20\mid t}), \\ 0, & \log(\text{RV}^+_{t-5\mid t}) <= \log(\text{RV}^+_{t-20\mid t}) \end{cases}$
66	Moving average indicator rule	$\begin{cases} 1, & \log(\text{RV}^-_{t-5\mid t}) > \log(\text{RV}^-_{t-20\mid t}), \\ 0, & \log(\text{RV}^-_{t-5\mid t}) <= \log(\text{RV}^-_{t-20\mid t}) \end{cases}$
67	Moving average indicator rule	$\begin{cases} 1, & \log(\text{RQ}^-_{t-5\mid t}) > \log(\text{RQ}^-_{t-20\mid t}), \\ 0, & \log(\text{RQ}^-_{t-5\mid t}) <= \log(\text{RQ}^-_{t-20\mid t}) \end{cases}$

Table 4.4: Intraday association features of pairs of assets i, k

index	Feature description	Feature Definition
68	Realized covariance ($RCov_{t,ik}$)	$\sum_{j=1}^{M} r_{t,ij} \times r_{t,kj}$
69	Realized correlation ($RCor_{t,ik}$)	$\frac{RCov_{t,ik}}{\sqrt{RV_{t,i} \times RV_{t,k}}}$
70	Intraday Kendall correlation ($\tau_{t,ik}$)	$\frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$
71	Intraday Spearman correlation	$1 - \frac{6 \sum_{j=1}^{M} (R(r_{t,ij}) - R(r_{t,kj}))^2}{M(M^2 - 1)}$
72	Intraday distance correlation	$\frac{dCov(X,Y)}{\sqrt{dVar(X) \times dVar(Y)}}$
73	Intraday mutual information	$\sum_{y \in Y} \sum_{x \in X} P_{(X,Y)}(x,y) log(\frac{P_{(X,Y)}(x,y)}{P_X(x) \times P_{Y(y)}})$

We base the univariate return process of each asset on the GARCH model. The variance is a function of observed returns, and for each day, the asset return follows a probability distribution represented by a parametric function with estimated mean and variance. In this study, the difference is that we propose to model the variance using supervised machine learning regression algorithms instead of the econometric approach. We use past observed features to input the algorithms to obtain the one-day-ahead variance.

Further, we propose to model the multivariate assets' return process, preserving the existing dependence between assets. To achieve this goal, we use a copula function that can separately model the individual's marginal distributions from the joint distribution. In this step, we also use machine learning regression algorithms to predict the intraday Kendall correlation between each pair of assets. This approach is essential to capture the dependencies from assets and adjust the copula function parameters. Once again, past observed features are used as input to the algorithms to obtain one day ahead of the correlation estimations.

With this data modeling, we simulate the one-day ahead scenario matrix. The simulated scenario matrix preserves the statistical proprieties from the multivariate assets. It is an ideal dynamic approach since variance and correlation are not constant over time. In the following subsections, we describe each step of the proposed modeling in detail.

4.2.1 The proposed modeling

The proposed modeling extends the GARCH-EVT-Copula framework of the literature, using realized variance, intraday Kendall correlation, and machine learning algorithms. The realized variance is assumed to be the variance of return on day t. We use machine learning algorithms to learn a function f_1 of realized variance. Following the GARCH model definition in 3.1 we define:

$$r_t = \sqrt{RV} z_t, \tag{4.6}$$

$$z_t \sim i.i.d.\ G(z), \tag{4.7}$$

$$RV_t = f_1(v_{1,t-1}) \tag{4.8}$$

where $v_{1,t-1}$ is a vector of past features defined in Tables 4.1, 4.2 and 4.3, z_t is the standardized returns ($\frac{r_t}{\sqrt{RV_t}}$), and $G(z)$ is the probability distribution of standardized returns. To model this distribution, we follow the same approach of Wang et al. [2010], Sahamkhadam et al. [2018] using the extreme value theory (EVT). We model the distribution's tails by a generalized Pareto using the method peak over the threshold. To model the middle of the distribution, we use a Gaussian kernel

$$G(z) = \begin{cases} \frac{N_{u^L}}{N}\{1 + \xi^L \frac{u^L - z}{\beta^L}\}^{-\frac{1}{\xi^L}}, & z < -u^L \\ \phi(z), & u^L < z < u^R \\ 1 - \frac{N_{u^R}}{N}\{1 + \xi^L \frac{u^R - z}{\beta^R}\}^{-\frac{1}{\xi^R}}, & z > -u^R \end{cases} \tag{4.9}$$

where ξ, β, u^L, u^R denote: shape, scale, upper and lower thresholds respectively.

This approach models each asset independently. To model the dependency between assets, we use a time-varying copula approach where for each day t we adjust the copula function parameters by the estimates of intraday Kendall correlation τ_t given by a second learning function $\tau_t = f_2(v_{2,t-1})$ where $v_{2,t-1}$ is a second vector of past features defined in Table 4.4 used to train the algorithms to predict one day ahead the intraday Kendall correlation.

Embrechts et al. [2002] discuss that Pearson's correlation is a linear measure of association, and it is inflexible to capture non-linear dependencies. The author identified several problems associated with this metric; for example, feasible values for correlation depend on the marginal distribution of the random variables. Further, the paper also discusses that many management systems use correlation to model dependencies where distributions are not Gaussian, so Pearson's correlation is misleading. Only when the distributions of returns are multivariate normal, or t-student Pearson's correlation can be justified as a measure of dependence.

For these reasons, we obtain the parameters of the copula functions by estimating the Kendall correlation of intraday returns using Equation 3.20.

4.2.2 Supervised machine learning regression

Machine learning algorithms are the core step of the methodology, responsible for producing the accurate predictions of realized variance and intraday Kendall correlation of the assets in the out-of-sample set. We use these predictions to construct the scenario matrix and find the optimum weights of the portfolio.

We use the in-sample set in the training process to learn two different functions f_1, and f_2 that map the features into the target variables:

$$\begin{aligned}
&RV_t = f_1(v_{1,t-1}) \\
&v_{1,t-1} = (\text{feature}_{1,t-1}, \text{feature}_{2,t-1}, ..., \text{feature}_{67,t-1}), \\
&\tau_t = f_2(v_{2,t-1}) \\
&v_{2,t-1} = (\text{feature}_{26,t-1}, ..., \text{feature}_{29,t-1}, \text{feature}_{68,t-1}, ..., \text{feature}_{73,t-1})
\end{aligned} \tag{4.10}$$

where all the features are lagged for at least one day from the target variable, and correspond to the Tables 4.1, 4.2, 4.3 and 4.4.

We use features related to the statistical properties of the intraday probability distributions and lagged values of these variables to predict the realized variance (features 1 to 67 of the Tables 4.1, 4.2, 4.3). We aggregated each asset's training data, forming a large, unified training dataset. We employed this approach to leverage substantial data for training machine learning algorithms. In the case of predicting intraday Kendall correlation, we use the same approach of stacking training data for each pair of assets. The difference in the training process in this case pertains to the input features, involving variables related to asset association measures and moments of intraday distribution (features 26 to 29 of the Table 4.2, and 68 to 73 of Table 4.4).

We use the grid search technique to optimize and tune the hyperparameters of the algorithms. This approach involves an exhaustive search across the hyperparameter space, selecting the combination that results in the most minor error related to cross-validation.

After the training process, we have the adjusted machine learning algorithms to make out-of-sample set predictions.

4.2.3 Generating the scenario matrix

With the estimates of realized variance, the residual distribution of the assets and the intraday Kendall correlations of the assets, it is possible to generate the scenario matrix following the algorithm suggested by Alexander [2008] :

1. Generate simulations $\{u_1, u_2, ..., u_N\}$ from independent uniform random variables.
2. Fix $u_1^* = u_1$ and then apply the inverse conditional copula $u_2^* = C_{2|1}^{-1}(u_2 \mid u_1^*)$
3. Fix $u_3^* = C_{3|1,2}^{-1}(u_3 \mid u_1^*, u_{_2}^*)$
4. Fix $u_N^* = C_{N|1,2,3,...,N-1}^{-1}(u_N \mid u_1^*, u_{_2}^*, u_{_3}^*, ..., u_{_{N-1}}^*)$
5. Feed the simulations $\{u_{_1}^*, u_{_2}^*, u_{_3}^*, ..., u_{_N}^*\}$ into the marginals to obtain a corresponding simulation $\{D_1^{-1}(u_1^*), D_2^{-1}(u_2^*), D_3^{-1}(u_3^*), ..., D_N^{-1}(u_{_N}^*)\}$, where D_i corresponds to the residual distributions of each asset with zero mean and variance equal the estimated realized variance $y_{t,i}$.

We repeat these steps 1000 times, generating a scenario matrix with dimensions $1000 \times N$. Each row corresponds to a realization of the multivariate distribution predicted for the one-day-ahead asset returns.

The entire covariance matrix of the elliptical copulas (with more than two dimensions) can be obtained by calculating the empirical Kendall's τ matrix and then constructing the estimator. However, Demarta and McNeil [2005] also mentions that this procedure is not guaranteed to produce a positive definite correlation matrix. In this case, the eigen method of Rousseeuw and Molenberghs [1993] can be applied. In the case of the t-student copula function, we still must, however, arbitrarily choose a value of degrees of freedom v. Since the best choice for v is unclear, and may vary over time, ideally, it should also be predicted. We, however, opted to defer this to future work and, instead, use a fixed value of $v = 2$ which gives the higher values of tail dependence for this function.

In the case of the Clayton copula function, we utilize the approach discussed in Kojadinovic and Yan [2010], where we average the forecasts of Kendall correlations and then employ equation 3.21 to calculate the corresponding parameter of the copula function.

4.2.4 Portfolio Optimization

After generating the scenario matrix, we use the transformation of e^x to each log return element of the matrix. This step is necessary since the optimization models use simple returns instead of log returns. With the transformed matrix, we use it as input in a portfolio selection model looking for a vector of weights $(\hat{w}_{1t}, \ldots, \hat{w}_{Nt})$ that maximizes the STARR Ratio model.

We chose the STARR Ratio allocation model because it is a model that takes into consideration the trade-off between return and risk. Additionally, it employs Conditional

Value at Risk as a risk measure, which is a coherent measure aligned with the ideal properties for a risk function. If we used the CVaR minimization model, it would be necessary to incorporate a portfolio return constraint, as return holds significance within this context. Depending on the chosen threshold, the model might fail to converge to a solution. Thus, by utilizing the STARR Ratio, we circumvent the issue of return selection, as the model maximizes the portfolio return divided by the CVaR.

Chapter 5

Experiments and Results

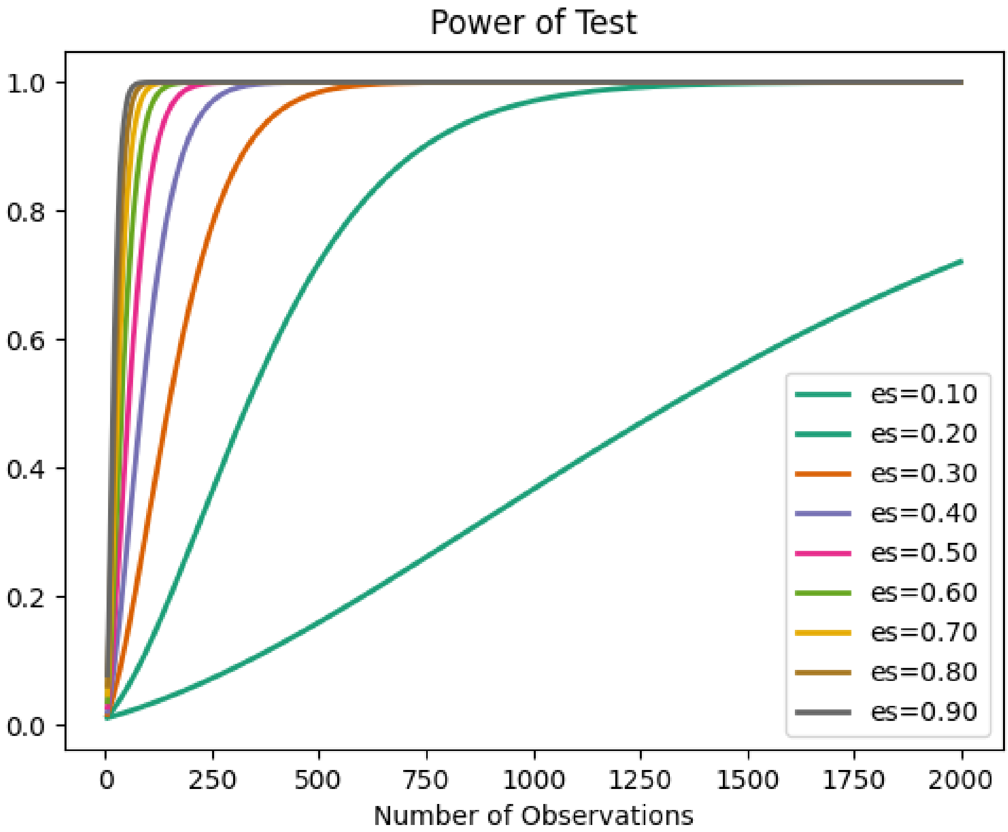


Figure 5.1: Power of the statistic test

In this section, we present our computational results by applying every step of the methodology described in the previous section. We implemented the code in R and Python and used the R packages `rugarch`, `rmgarch`, `Copula`, `multDM`, `MCS`, PerformanceAnalytics and the Python package `H2O Automl`.

Throughout the chapter, we use hypothesis tests to assess the statistical significance of the various results. The p-value obtained in the statistical test represents the probability of obtaining the result, assuming the null hypothesis to be true. A significance level α is then defined. We reject a null hypothesis if the p-value exceeds α. To avoid type 2 errors in the test (failing to reject the null hypothesis even when there is a significant effect), we can calculate the test's power, representing the probability that the test correctly rejects the null hypothesis, 1− type 2 error.

To calculate the test's power, it is necessary to specify the sample size, the significance level α, and the effect size. Figure 5.1 illustrates the test power values for a $\alpha = 1\%$ as the sample size increases for different effect size values. We will use hypothesis tests with samples of 2000 and 353 in our experiments. Assuming an effect size greater than or equal to 0.4, the probability of encountering a type 2 error is approximately zero.

5.1 In-sample analysis

In this section, we used the in-sample (training and validation sets) to analyze the assets' statistical properties and to train and calibrate the machine-learning algorithms.

5.1.1 Realized variance and intraday Kendall correlation time series analysis

Figure 5.2 shows the time series of realized variance of PETR4, VALE3, ITSA4, and ABEV3 over nearly ten years and the autocorrelation functions of these assets for the same period. It is possible to observe some stylized facts already mentioned in the literature. We notice the volatility clustering effect, where high-volatility events tend to cluster in time. There is also substantial positive autocorrelations of the first 30 lags with slow decay, indicating a long memory effect.

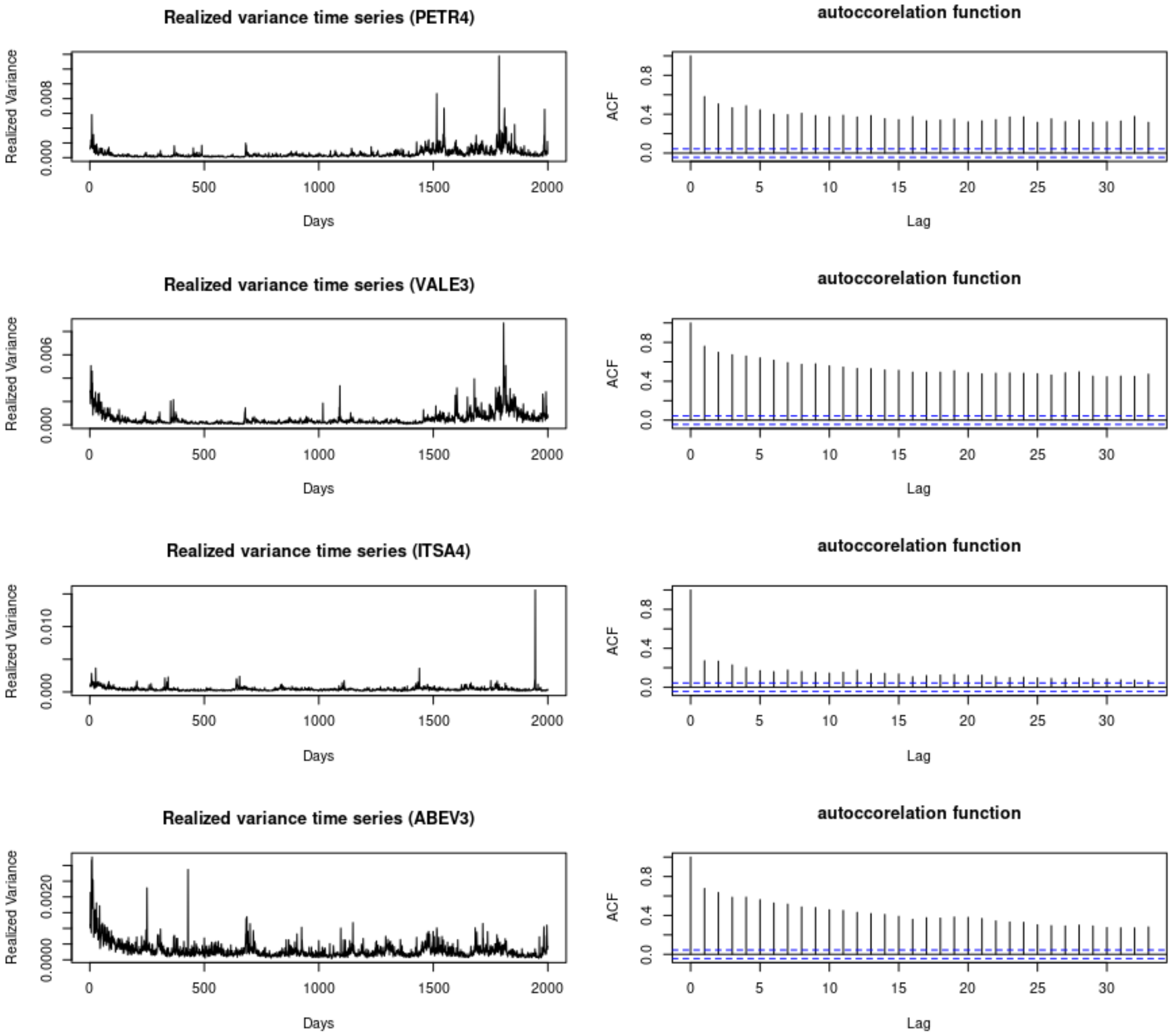


Figure 5.2: Time series of realized variance and the corresponding autocorrelation functions.

Figures 5.3 and 5.4 show the time series of intraday Kendall correlation between these pairs of assets and the autocorrelation function for the same period. We observe that this correlation is dynamic. Also, we notice it is not an i.i.d. presenting serial correlation and the same characteristic of the long memory effect with a significant correlation with the first 30 lags.

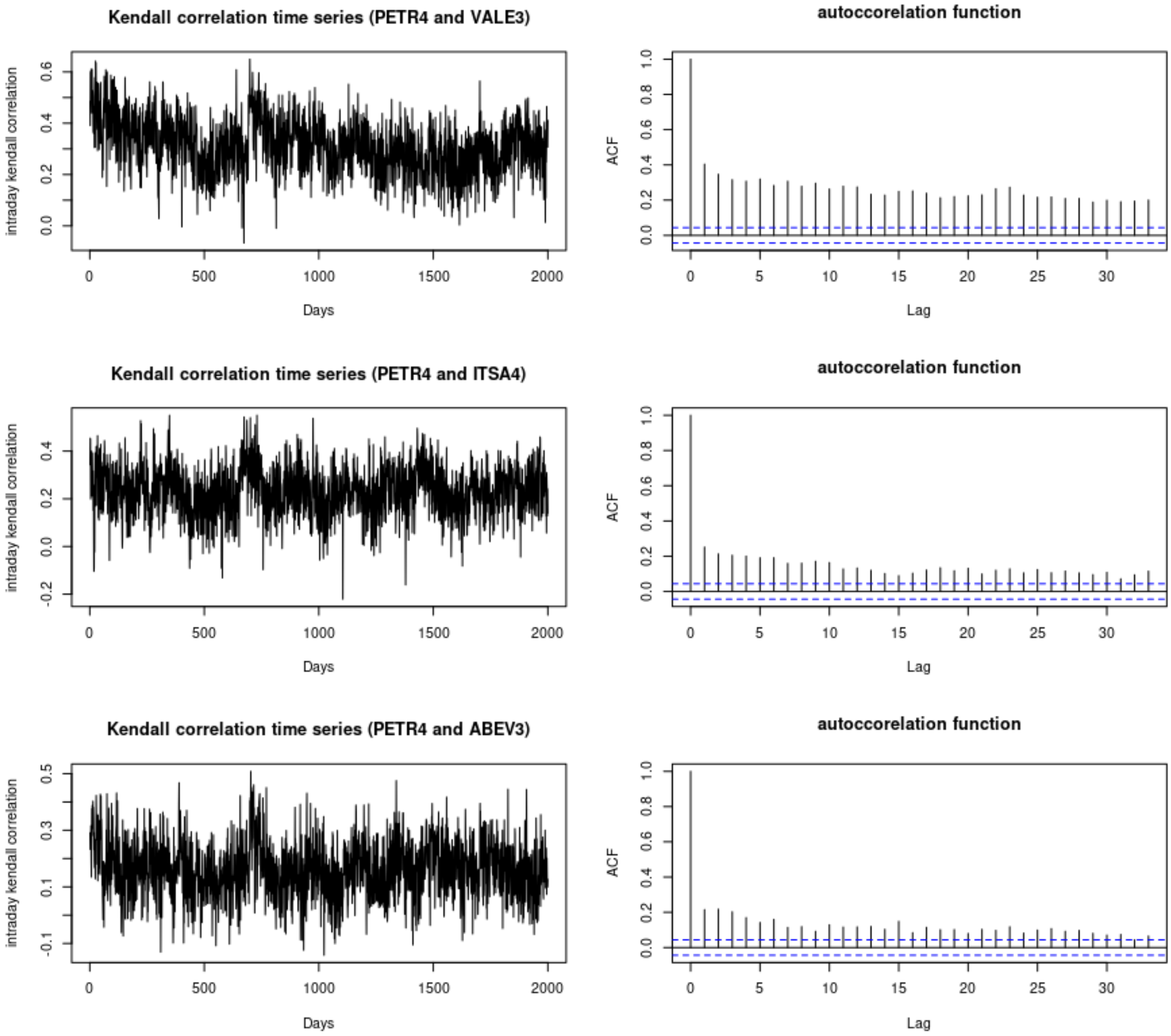


Figure 5.3: Time series of intraday Kendall correlation and the autocorrelation function

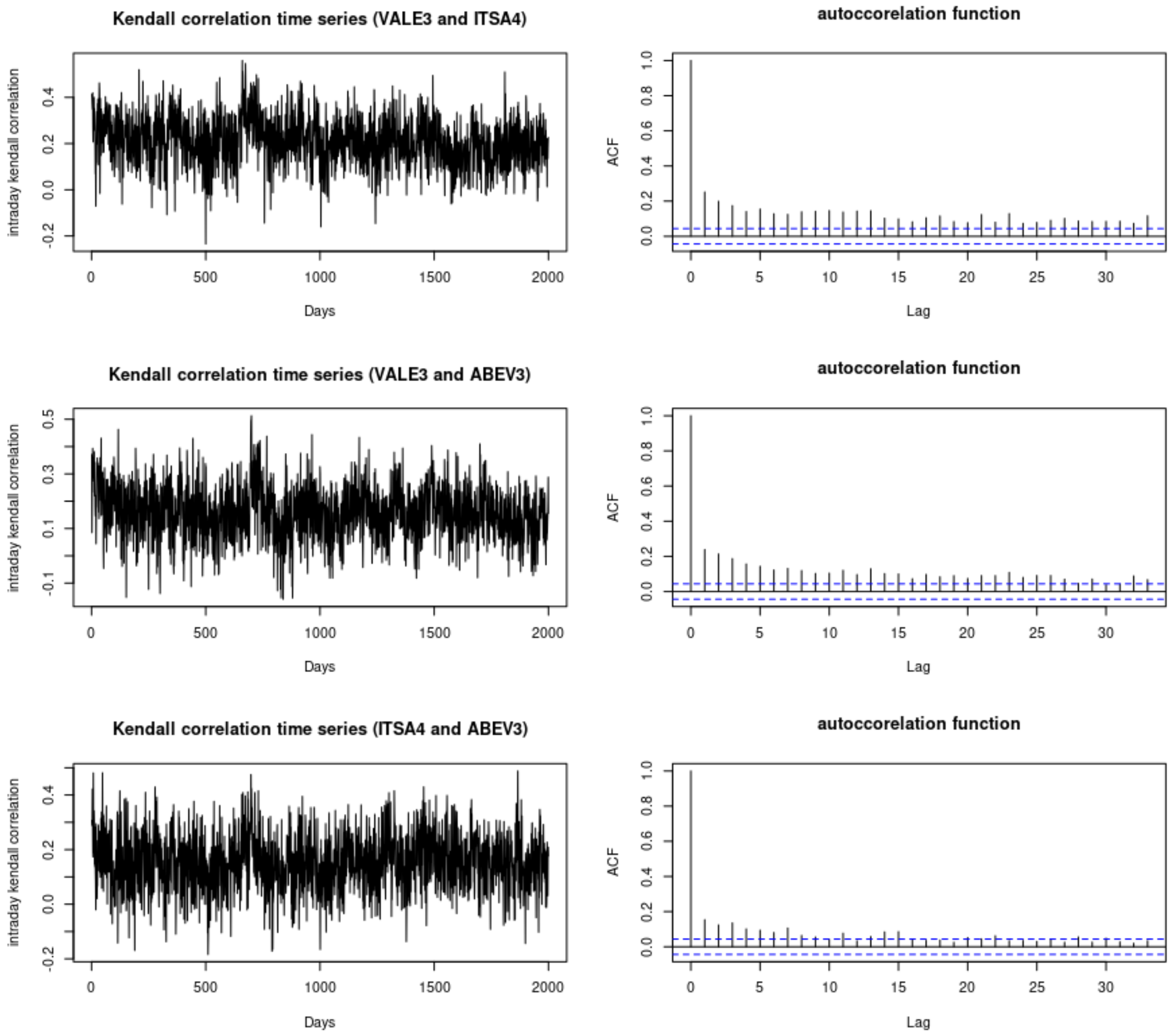


Figure 5.4: Time series of intraday Kendall correlation and the autocorrelation function

These results suggest that it might be possible to develop a model to predict future observations of these measurements based on past values since there is a significant correlation. Also, the behavior of the series is dynamic. Therefore, we suggest that accurate forecasts can approximate the actual future behavior of asset distributions.

5.1.2 Analysis of the probability distribution of the standardized returns

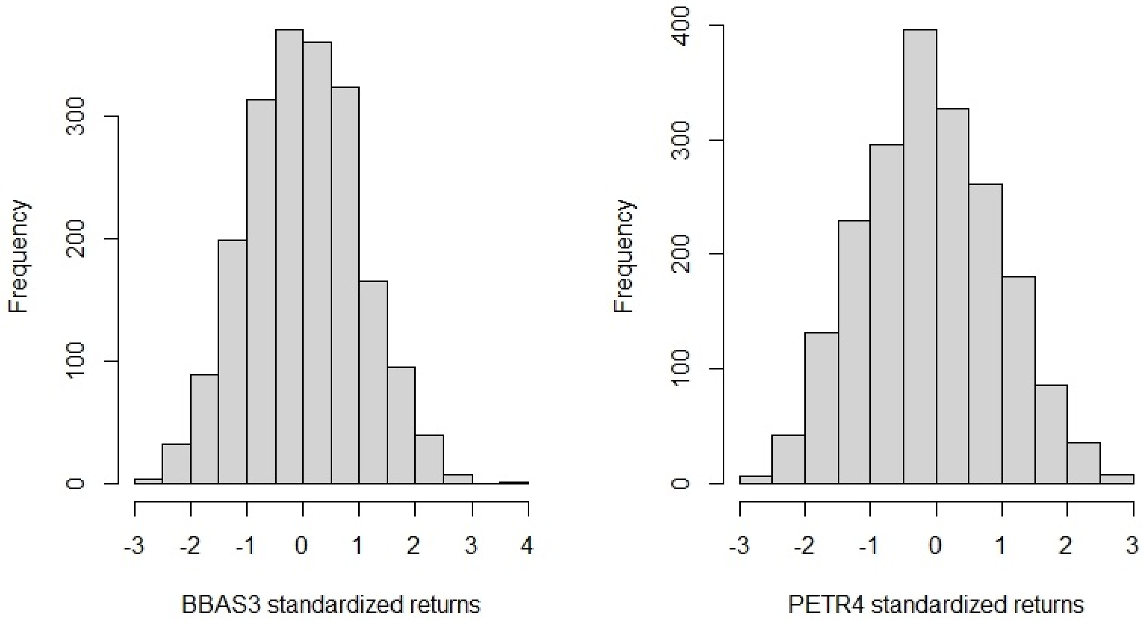


Figure 5.5: Histograms of the standardized returns.

Next, we analyze the distribution of standardized returns ($\frac{r_t}{\sqrt{RV_t}}$). Table 5.1 shows the first four moments of these distributions. In all assets, we observe the mean close to zero and the standard deviation close to 1. Also, the skewness is close to zero, and kurtosis is near 3. Figure 5.5 shows the histograms of BBAS3 and PETR4. By analyzing the moments and histograms, we observe another stylized fact already studied in the literature: the distribution of standardized returns is almost normal [Taylor, 2007].

In Table 5.2, we present the values of the statistic and p-value from the Kolmogorov-Smirnov test. We compare the standardized return distributions with standardized normal and a fitted EVT distribution using maximum likelihood estimation to calibrate the parameters. This test quantifies a distance between the empirical distribution functions of two samples with the null hypothesis that the samples are drawn from the same distribution [Massey Jr, 1951]. Considering $F(x)$ the empirical distribution function and $G(x)$ another empirical distribution function

Table 5.1: First four moments

Asset	Mean	Std	Skewness	Kurtosis
ABEV3	5.20E-02	8.47E-01	-1.45E-03	3.02E+00
ALPA4	4.92E-03	7.86E-01	-8.66E-02	3.00E+00
BBAS3	5.95E-03	9.76E-01	6.84E-02	2.72E+00
BBDC4	1.64E-02	9.10E-01	9.60E-02	3.04E+00
BRKM5	-2.39E-02	8.94E-01	2.12E-02	3.07E+00
CGRA4	1.57E-01	8.10E-01	-2.73E-01	2.01E+00
CMIG4	-1.11E-02	9.07E-01	-1.02E-02	3.09E+00
CPFE3	-8.31E-03	8.34E-01	-1.81E-03	2.88E+00
CPLE6	2.03E-02	8.78E-01	1.35E-01	3.14E+00
CRFB3	1.11E-01	8.35E-01	-4.69E-02	2.82E+00
CSNA3	-8.97E-02	9.32E-01	1.58E-01	3.02E+00
ELET3	-6.12E-02	9.28E-01	1.17E-01	3.07E+00
EMBR3	-1.72E-02	8.88E-01	3.49E-02	3.02E+00
ENGI4	3.01E-02	8.03E-01	-6.33E-01	2.18E+00
ENGI11	1.96E-02	8.19E-01	-3.61E-01	2.26E+00
GGBR4	-9.60E-02	9.58E-01	1.46E-01	2.69E+00
ITSA4	1.50E-02	7.77E-01	3.88E-02	2.84E+00
ITUB4	2.16E-02	9.23E-01	9.29E-02	2.87E+00
LIGT3	-3.28E-02	8.22E-01	1.11E-02	3.21E+00
LREN3	2.51E-02	8.95E-01	1.02E-01	2.89E+00
PETR4	-8.31E-02	1.02E+00	1.01E-01	2.60E+00
RADL3	5.58E-03	8.23E-01	3.59E-02	2.82E+00
SBSP3	4.54E-02	8.32E-01	-1.36E-02	3.19E+00
SUZB3	-6.00E-02	8.80E-01	1.02E-01	3.01E+00
TIMS3	3.83E-02	8.64E-01	6.78E-03	2.99E+00
UNIP6	-1.34E-02	8.26E-01	1.07E-01	5.12E+00
USIM5	-1.16E-01	9.80E-01	1.40E-01	2.92E+00
VALE3	-4.12E-02	9.93E-01	1.02E-01	2.84E+00
VIVT3	7.90E-02	8.30E-01	-1.19E-01	2.69E+00

$$D_{n,m} = sup|F(x) - G(x)| \tag{5.1}$$

where *sup* is the supremum function. $D_{n,m}$ correspond the statistic used to test the null hypothesis $F(x) = G(x)$. For large samples, the null hypothesis is rejected at level α if

$$D_{n,m} > \sqrt{-ln(\frac{\alpha}{2}) \times \frac{1+\frac{m}{n}}{2m}} \tag{5.2}$$

where n and m are the sizes of the samples.

The results indicate that several assets reject the null hypothesis of normality. In all cases, the EVT distribution better fits the distribution of the standardized returns.

Figure 5.6 shows the cumulative distributions of PETR4 and the cumulative distributions from simulated normal and EVT distributions. Analyzing the Figure, it can

Table 5.2: Kolmogorov-Smirnov statistic and p-value

Asset	**KS EVT**		**KS normal**	
	D	**p-value**	**D**	**p-value**
ABEV3	2.05E-02	7.95E-01	5.60E-02	3.78E-03
ALPA4	2.85E-02	3.91E-01	8.20E-02	2.89E-06
BBAS3	2.10E-02	7.70E-01	2.95E-02	3.49E-01
BBDC4	2.10E-02	7.70E-01	4.00E-02	8.15E-02
BRKM5	2.30E-02	6.65E-01	4.55E-02	3.18E-02
CGRA4	2.40E-02	6.12E-01	1.33E-01	1.11E-15
CMIG4	1.55E-02	9.70E-01	4.25E-02	5.40E-02
CPFE3	1.30E-02	9.96E-01	5.50E-02	4.72E-03
CPLE6	2.05E-02	7.95E-01	6.20E-02	9.17E-04
CRFB3	1.45E-02	9.99E-01	8.35E-02	1.84E-04
CSNA3	1.65E-02	9.48E-01	5.65E-02	3.38E-03
ELET3	1.15E-02	9.99E-01	5.90E-02	1.89E-03
EMBR3	1.85E-02	8.84E-01	4.50E-02	3.48E-02
ENGI4	2.20E-02	7.18E-01	3.09E-01	2.20E-16
ENGI11	2.00E-02	8.19E-01	1.47E-01	2.20E-16
GGBR4	2.15E-02	7.44E-01	5.80E-02	2.39E-03
ITSA4	2.65E-02	4.84E-01	8.45E-02	1.26E-06
ITUB4	1.85E-02	8.84E-01	3.30E-02	2.26E-01
LIGT3	2.15E-02	7.44E-01	7.50E-02	2.60E-05
LREN3	1.60E-02	9.60E-01	3.65E-02	1.39E-01
PETR4	1.65E-02	9.48E-01	6.20E-02	9.17E-04
RADL3	2.45E-02	5.86E-01	6.40E-02	5.54E-04
SBSP3	2.65E-02	4.84E-01	7.60E-02	1.92E-05
SUZB3	2.00E-02	8.19E-01	7.15E-02	7.25E-05
TIMS3	1.75E-02	9.87E-01	3.20E-02	5.02E-01
UNIP6	1.85E-02	8.84E-01	2.32E-01	2.20E-16
USIM5	8.50E-03	1.00E+00	6.35E-02	6.29E-04
VALE3	1.70E-02	9.35E-01	3.45E-02	1.85E-01
VIVT3	1.60E-02	9.60E-01	7.75E-02	1.21E-05

be noticed that the cumulative distribution of PETR4 returns is quite close to the cumulative distribution of EVT, whereas there is a greater distance for the cumulative normal distribution.

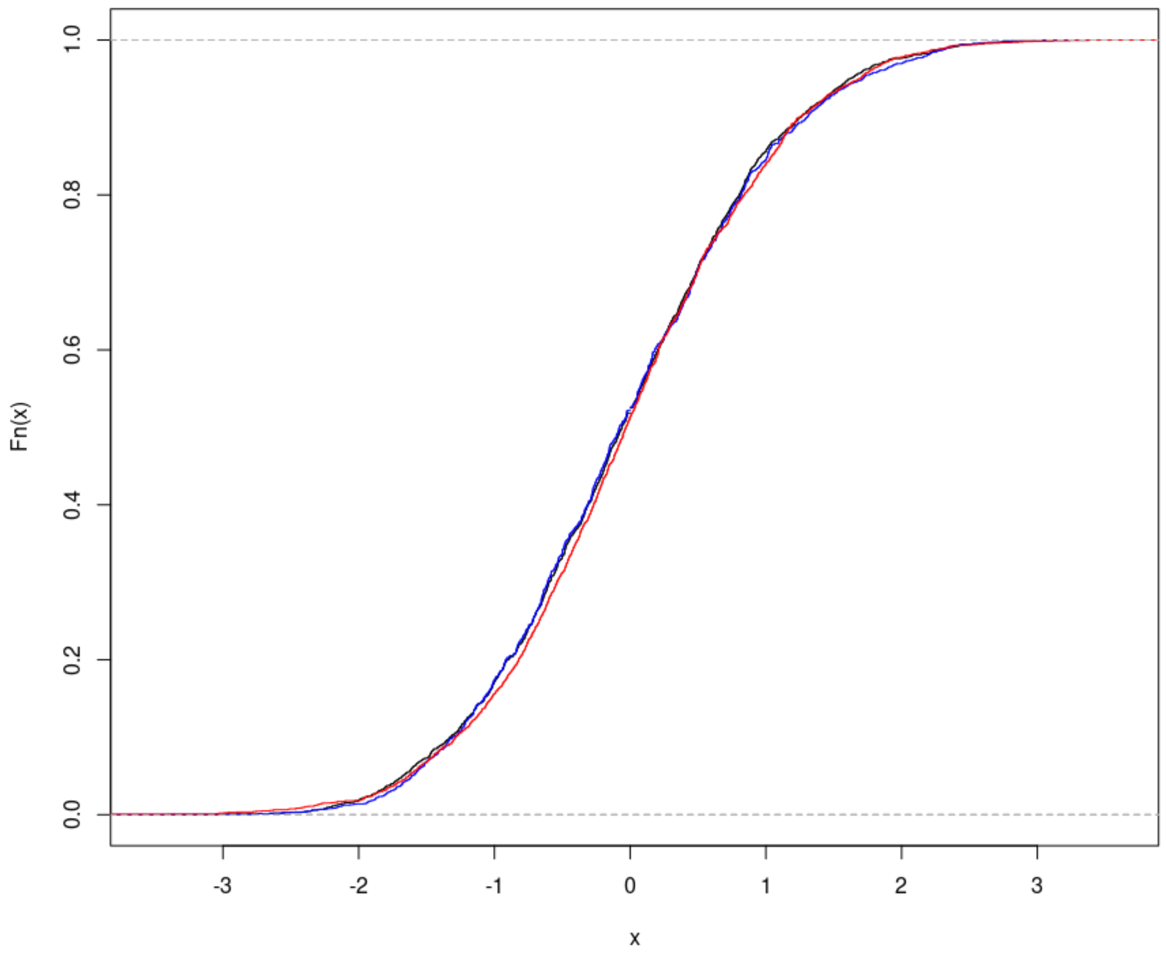


Figure 5.6: Comparison of the cumulative distribution functions. The black curve shows the standardized PETR4 returns, the red curve shows the standard normal, and blue curve shows the EVT.

5.1.3 Simulated returns analysis

In this section, we used the observed realized variance and the fitted EVT probability distributions of the standardized returns to simulate the returns of each asset. For 2000 days, we extracted random samples of the EVT distributions and multiplied them by the values of realized variance.

We used the nonparametric Kolmogorov-Smirnov test of goodness fit to evaluate each asset simulation. Table 5.3 shows the p-values and statistics of the test. It is possible to observe that only three of the simulated assets reject the null hypothesis of

Table 5.3: Kolmogorov-Smirnov test between each simulated asset and corresponding asset

Asset	**D statistic**	***p*-value**
ABEV3	2.35E-02	6.39E-01
ALPA4	3.45E-02	1.85E-01
BBAS3	1.90E-02	8.63E-01
BBDC4	3.30E-02	2.26E-01
BRKM5	2.40E-02	6.12E-01
CGRA4	8.10E-02	4.00E-06
CMIG4	1.45E-02	9.85E-01
CPFE3	1.90E-02	8.63E-01
CPLE6	1.55E-02	9.70E-01
CRFB3	2.10E-02	9.80E-01
CSNA3	2.55E-02	5.34E-01
ELET3	1.90E-02	8.63E-01
EMBR3	1.70E-02	9.35E-01
ENGI4	1.32E-01	1.89E-15
ENGI11	1.04E-01	9.92E-10
GGBR4	1.10E-02	1.00E+00
ITSA4	2.25E-02	6.92E-01
ITUB4	2.10E-02	7.70E-01
LIGT3	2.15E-02	7.44E-01
LREN3	2.55E-02	5.34E-01
PETR4	2.15E-02	7.44E-01
RADL3	2.40E-02	6.12E-01
SBSP3	2.10E-02	7.70E-01
SUZB3	1.70E-02	9.35E-01
TIMS3	4.50E-02	2.63E-01
UNIP6	3.25E-02	2.41E-01
USIM5	1.65E-02	9.48E-01
VALE3	2.45E-02	5.86E-01
VIVT3	2.25E-02	6.92E-01

equal distributions.

Figure 5.7 shows the histograms, and Figure 5.8 shows the cumulative distribution functions from PETR4 and the simulated corresponding asset. As we notice in the Kolmogorov-Smirnov test, the simulation distributions are close to the actual asset. In Figure 5.9, we show the returns time series, the autocorrelation function of the returns, and the autocorrelation function of the squared returns from PETR4 and compare them with the simulated asset. We can observe that the simulated time series has similar statistical properties to the autocorrelation functions of the assets.

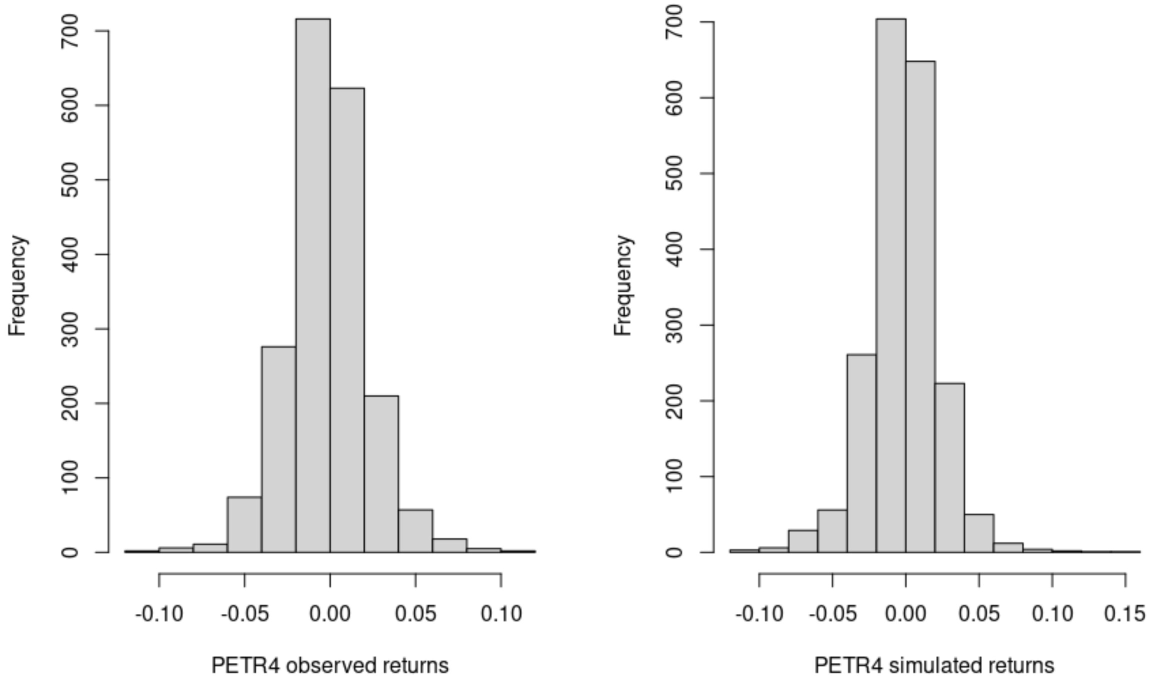


Figure 5.7: Histograms of the PETR4 asset and the simulated corresponded distribution

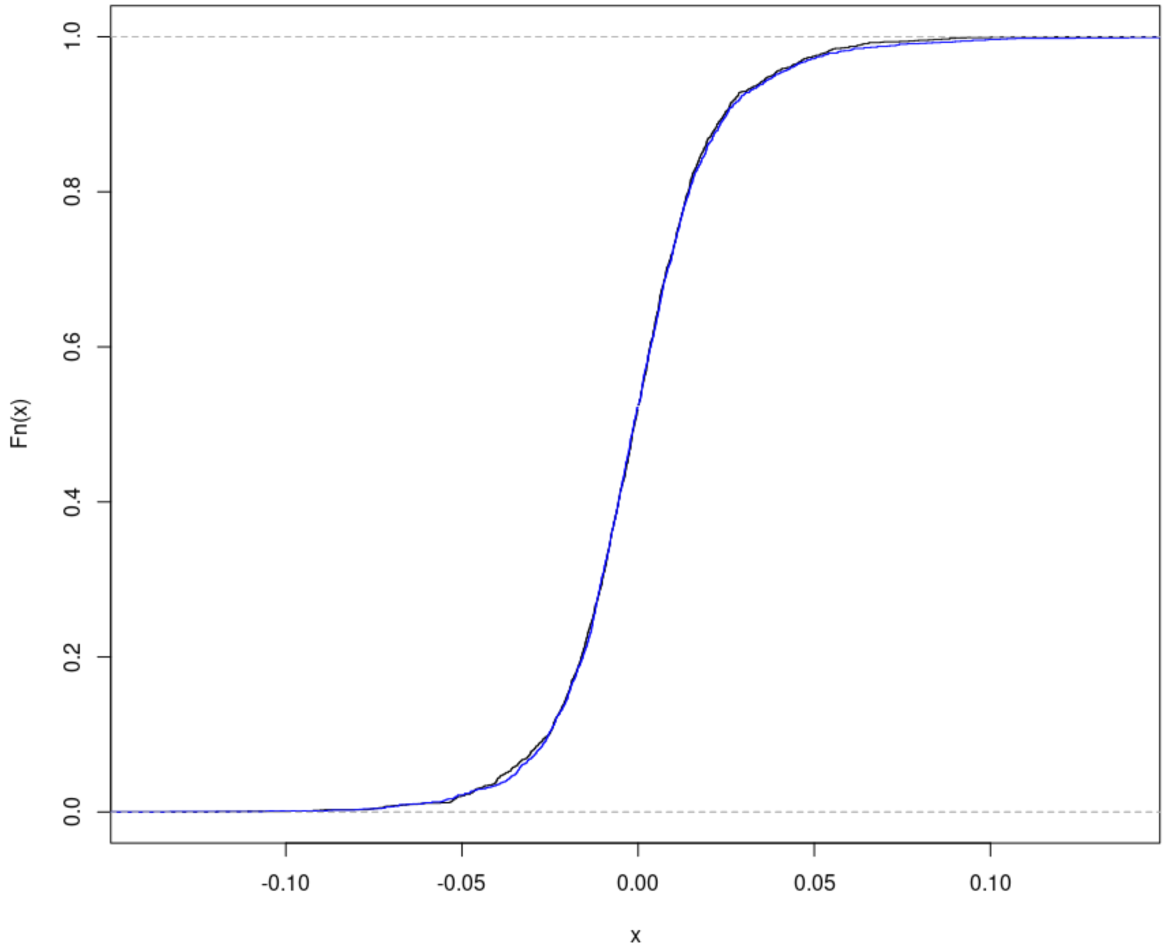


Figure 5.8: Comparison of the cumulative distribution functions. Black curve shows PETR4 returns and the blue curve the corresponding simulated returns

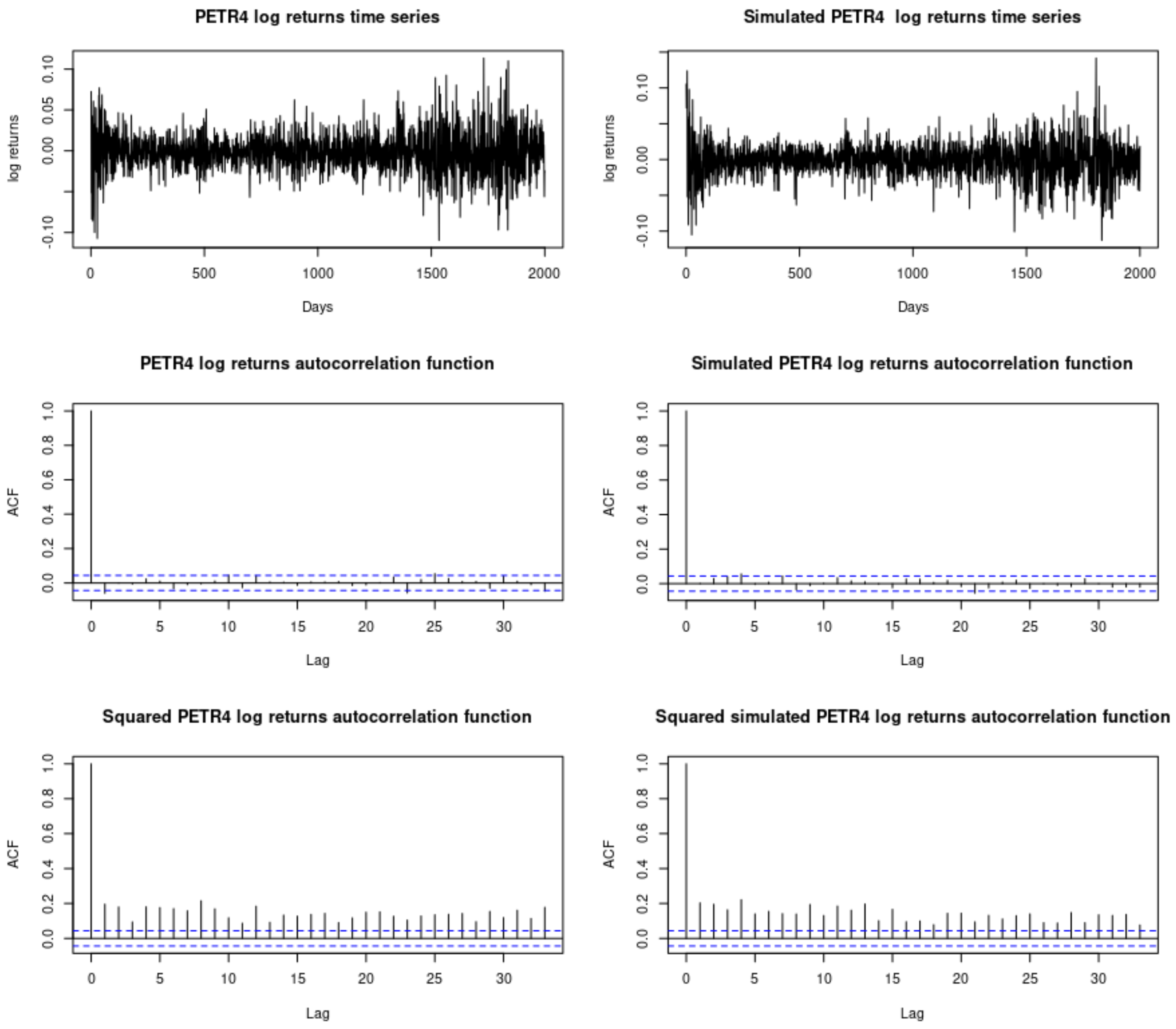


Figure 5.9: Comparison of the returns time series, and autocorrelation functions of PETR4 and the corresponded simulation

5.1.4 Training of machine learning algorithms

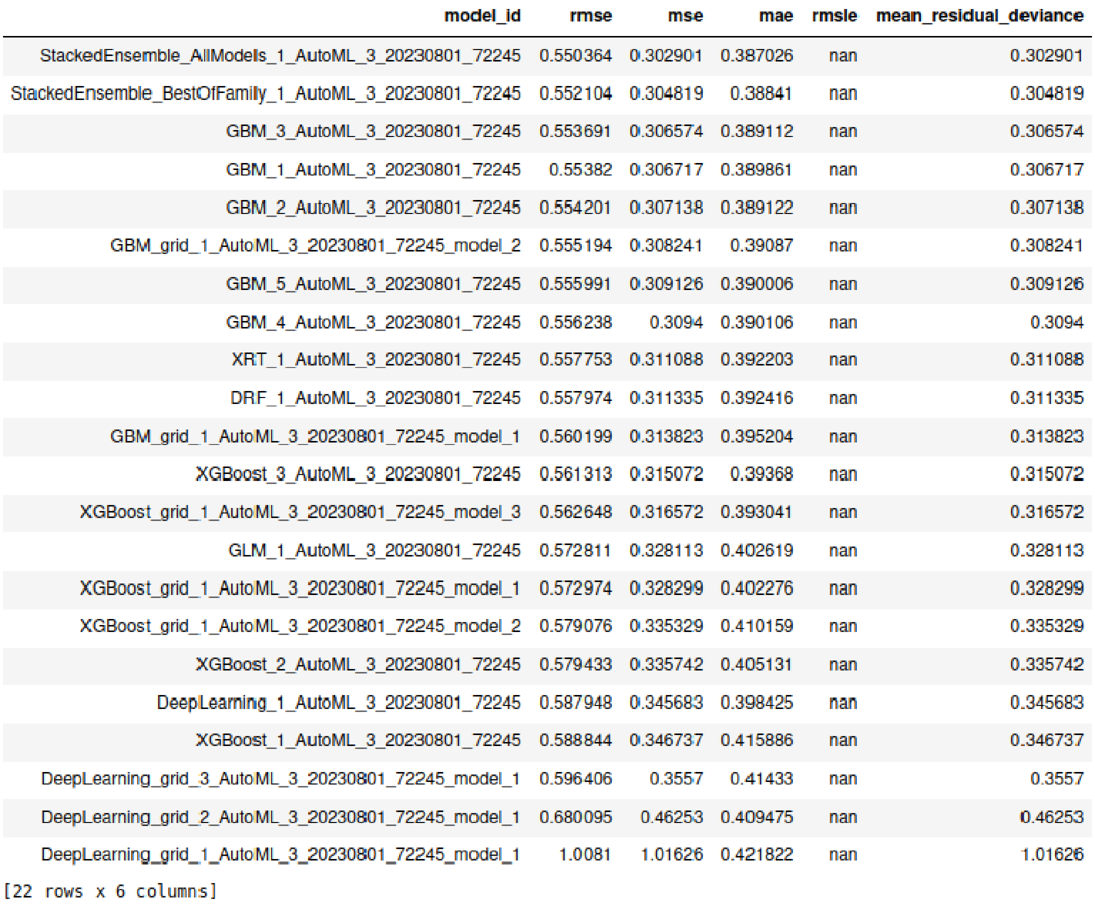
To train the machine learning models we use the h20 package automl `https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html` [LeDell and Poirier, 2020]. With this tool, we were able to automatically train the following machine learning algorithms:

- Random Forest,

- Extremely Randomized Forest,
- Gradient Boosting Machine (GBM),
- eXtreme Gradient Boosting (XGBoost),
- Deep Neural Net.

We use the grid search technique To optimize the hyperparameters of the algorithms. With this approach, we test different combinations of hyperparameters and select the one that results in the lowest error. Subsequently, we combine the trained models to obtain stacked models, where the final prediction is determined by aggregating the individual predictions.

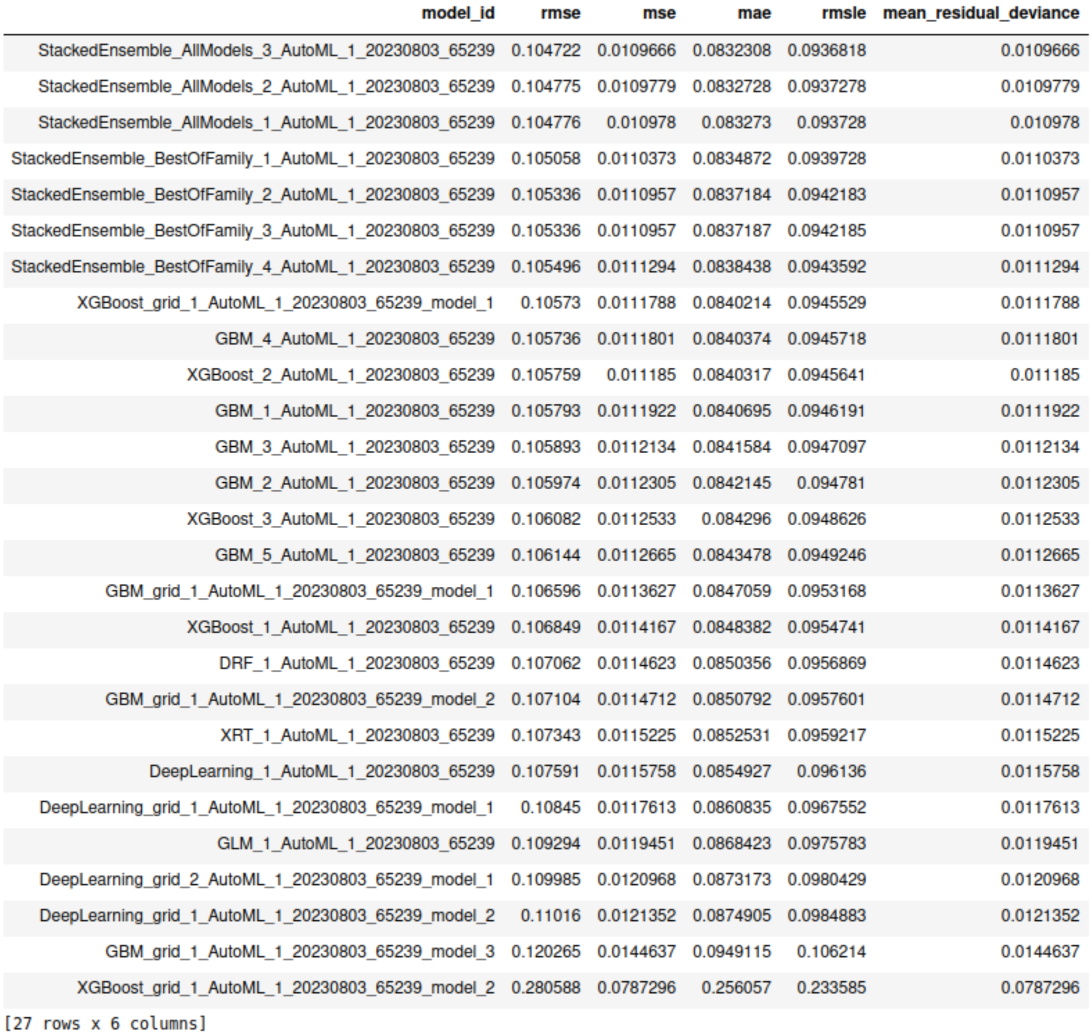
Figures 5.10 and 5.11 illustrate the ranking of trained algorithms based on the obtained error. In both cases, we obtain the best predictions from combining individual algorithms.



model_id	rmse	mse	mae	rmsle	mean_residual_deviance
StackedEnsemble_AllModels_1_AutoML_3_20230801_72245	0.550364	0.302901	0.387026	nan	0.302901
StackedEnsemble_BestOfFamily_1_AutoML_3_20230801_72245	0.552104	0.304819	0.38841	nan	0.304819
GBM_3_AutoML_3_20230801_72245	0.553691	0.306574	0.389112	nan	0.306574
GBM_1_AutoML_3_20230801_72245	0.55382	0.306717	0.389861	nan	0.306717
GBM_2_AutoML_3_20230801_72245	0.554201	0.307138	0.389122	nan	0.307138
GBM_grid_1_AutoML_3_20230801_72245_model_2	0.555194	0.308241	0.39087	nan	0.308241
GBM_5_AutoML_3_20230801_72245	0.555991	0.309126	0.390006	nan	0.309126
GBM_4_AutoML_3_20230801_72245	0.556238	0.3094	0.390106	nan	0.3094
XRT_1_AutoML_3_20230801_72245	0.557753	0.311088	0.392203	nan	0.311088
DRF_1_AutoML_3_20230801_72245	0.557974	0.311335	0.392416	nan	0.311335
GBM_grid_1_AutoML_3_20230801_72245_model_1	0.560199	0.313823	0.395204	nan	0.313823
XGBoost_3_AutoML_3_20230801_72245	0.561313	0.315072	0.39368	nan	0.315072
XGBoost_grid_1_AutoML_3_20230801_72245_model_3	0.562648	0.316572	0.393041	nan	0.316572
GLM_1_AutoML_3_20230801_72245	0.572811	0.328113	0.402619	nan	0.328113
XGBoost_grid_1_AutoML_3_20230801_72245_model_1	0.572974	0.328299	0.402276	nan	0.328299
XGBoost_grid_1_AutoML_3_20230801_72245_model_2	0.579076	0.335329	0.410159	nan	0.335329
XGBoost_2_AutoML_3_20230801_72245	0.579433	0.335742	0.405131	nan	0.335742
DeepLearning_1_AutoML_3_20230801_72245	0.587948	0.345683	0.398425	nan	0.345683
XGBoost_1_AutoML_3_20230801_72245	0.588844	0.346737	0.415886	nan	0.346737
DeepLearning_grid_3_AutoML_3_20230801_72245_model_1	0.596406	0.3557	0.41433	nan	0.3557
DeepLearning_grid_2_AutoML_3_20230801_72245_model_1	0.680095	0.46253	0.409475	nan	0.46253
DeepLearning_grid_1_AutoML_3_20230801_72245_model_1	1.0081	1.01626	0.421822	nan	1.01626

```
[22 rows x 6 columns]
```

Figure 5.10: Training performance predicting one day ahead realized variance. Comparison of validation error for different machine learning algorithms



model_id	rmse	mse	mae	rmsle	mean_residual_deviance
StackedEnsemble_AllModels_3_AutoML_1_20230803_65239	0.104722	0.0109666	0.0832308	0.0936818	0.0109666
StackedEnsemble_AllModels_2_AutoML_1_20230803_65239	0.104775	0.0109779	0.0832728	0.0937278	0.0109779
StackedEnsemble_AllModels_1_AutoML_1_20230803_65239	0.104776	0.010978	0.083273	0.093728	0.010978
StackedEnsemble_BestOfFamily_1_AutoML_1_20230803_65239	0.105058	0.0110373	0.0834872	0.0939728	0.0110373
StackedEnsemble_BestOfFamily_2_AutoML_1_20230803_65239	0.105336	0.0110957	0.0837184	0.0942183	0.0110957
StackedEnsemble_BestOfFamily_3_AutoML_1_20230803_65239	0.105336	0.0110957	0.0837187	0.0942185	0.0110957
StackedEnsemble_BestOfFamily_4_AutoML_1_20230803_65239	0.105496	0.0111294	0.0838438	0.0943592	0.0111294
XGBoost_grid_1_AutoML_1_20230803_65239_model_1	0.10573	0.0111788	0.0840214	0.0945529	0.0111788
GBM_4_AutoML_1_20230803_65239	0.105736	0.0111801	0.0840374	0.0945718	0.0111801
XGBoost_2_AutoML_1_20230803_65239	0.105759	0.011185	0.0840317	0.0945641	0.011185
GBM_1_AutoML_1_20230803_65239	0.105793	0.0111922	0.0840695	0.0946191	0.0111922
GBM_3_AutoML_1_20230803_65239	0.105893	0.0112134	0.0841584	0.0947097	0.0112134
GBM_2_AutoML_1_20230803_65239	0.105974	0.0112305	0.0842145	0.094781	0.0112305
XGBoost_3_AutoML_1_20230803_65239	0.106082	0.0112533	0.084296	0.0948626	0.0112533
GBM_5_AutoML_1_20230803_65239	0.106144	0.0112665	0.0843478	0.0949246	0.0112665
GBM_grid_1_AutoML_1_20230803_65239_model_1	0.106596	0.0113627	0.0847059	0.0953168	0.0113627
XGBoost_1_AutoML_1_20230803_65239	0.106849	0.0114167	0.0848382	0.0954741	0.0114167
DRF_1_AutoML_1_20230803_65239	0.107062	0.0114623	0.0850356	0.0956869	0.0114623
GBM_grid_1_AutoML_1_20230803_65239_model_2	0.107104	0.0114712	0.0850792	0.0957601	0.0114712
XRT_1_AutoML_1_20230803_65239	0.107343	0.0115225	0.0852531	0.0959217	0.0115225
DeepLearning_1_AutoML_1_20230803_65239	0.107591	0.0115758	0.0854927	0.096136	0.0115758
DeepLearning_grid_1_AutoML_1_20230803_65239_model_1	0.10845	0.0117613	0.0860835	0.0967552	0.0117613
GLM_1_AutoML_1_20230803_65239	0.109294	0.0119451	0.0868423	0.0975783	0.0119451
DeepLearning_grid_2_AutoML_1_20230803_65239_model_1	0.109985	0.0120968	0.0873173	0.0980429	0.0120968
DeepLearning_grid_1_AutoML_1_20230803_65239_model_2	0.11016	0.0121352	0.0874905	0.0984883	0.0121352
GBM_grid_1_AutoML_1_20230803_65239_model_3	0.120265	0.0144637	0.0949115	0.106214	0.0144637
XGBoost_grid_1_AutoML_1_20230803_65239_model_2	0.280588	0.0787296	0.256057	0.233585	0.0787296

```
[27 rows x 6 columns]
```

Figure 5.11: Training performance predicting one day ahead intraday Kendall correlation. Comparison of validation error for different machine learning algorithms

5.2 Out-of-sample analysis

In this section, we analyze the results for the out-of-sample set. First, we analyze the performance of the machine learning algorithms to predict one day ahead of the realized variance compared with econometric models. Next, we analyze the performance of predicting one day ahead of the intraday Kendall correlation. Further, we simulate the scenario matrix with the predictions and use the STARR Ratio optimization model to build the portfolios and analyze the financial returns obtained.

5.2.1 Forecasting one day ahead realized variance

In this section, we compare the predictions of fitted machine learning algorithms with the predictions from some models found in the literature. We chose six baselines to forecast the realized variance: the GARCH model with a normal distribution defined in Section 3.1 and five econometric models based on intraday data:

- Exponential moving average (EMA)
- Heterogeneous Autoregression (HAR)
- Heterogeneous Autoregression with Jumps (HAR-J)
- Semivariance-HAR (SHAR)
- Continuous Heterogeneous Autoregression (CHAR)
- Heterogeneous Autoregression Quarticity (HARQ)

We chose the GARCH model with daily data as it is the most widely used model in the literature in this context, as well as the HAR model and its extensions in the context of intraday models. We also employed the exponential moving average technique, given its widespread usage in industry.

The EMA can be defined as:

$$EMA_{t+1} = \alpha \times RV_t + (1 - \alpha) \times EMA_{t-1} \tag{5.3}$$

where $\alpha = 0.9$ is the weighting factor and $EMA_1 = RV_1$.

The HAR model, proposed by Corsi [2009], can be defined as:

$$\mathrm{RV}_{t+1} = c + \beta_1 \mathrm{RV}_t + \beta_2 \mathrm{RV}_{t-5|t} + \beta_3 \mathrm{RV}_{t-22|t} + \epsilon_t \tag{5.4}$$

where $\mathrm{RV}_{t-5|t}$ is the weekly average of the lagged 5 values of RV series, $\mathrm{RV}_{t-22|t}$ is the monthly average of the lagged 22 values of RV series, and $c, \beta_1, \beta_2, \beta_3$ are scalar parameters.

The HAR-J and CHAR models incorporate discontinuous (Jumps) and continuous (Bi-power variation) parts of the total variation of intraday returns [Andersen et al., 2007]. The HAR-J model is defined as

$$\mathrm{RV}_{t+1} = c + \beta_1 \mathrm{RV}t + \beta_2 \mathrm{RV}_{t-5|t} + \beta_3 \mathrm{RV}_{t-22|t} + \beta_4 \mathrm{J}_t + \epsilon_t \tag{5.5}$$

while the CHAR model is defined as

$$\mathrm{RV}_{t+1} = c + \beta_1 \mathrm{BPV}_t + \beta_2 \mathrm{BPV}_{t-5|t} + \beta_3 \mathrm{BPV}_{t-22|t} + \epsilon_t. \tag{5.6}$$

The SHAR model decomposes the total variance in the variation related to negative and positive intraday returns:

$$\mathrm{RV}_{t+1} = c + \beta_1^{+}\mathrm{RV}_t^{+} + \beta_1^{-}\mathrm{RV}_t^{-} + \beta_2\mathrm{RV}_{t-5|t} + \beta_3\mathrm{RV}_{t-22|t} + \epsilon_t \tag{5.7}$$

The HARQ model [Bollerslev et al., 2016] extends the HAR model by including the error measures related to the realized variance. The HARQ model is defined as

$$\begin{aligned} \mathrm{RV}_{t+1} &= c + \beta_{1,t}\mathrm{RV}_t + \beta_2\mathrm{RV}_{t-5|t} + \beta_3\mathrm{RV}_{t-22|t} + \epsilon_t \\ \beta_{1,t} &= (\beta_1 + \beta_{1Q}\mathrm{RQ}_t^{1/2}) \end{aligned} \tag{5.8}$$

where β_1, β_{1Q} are scalar parameters, RQ_t is the realized quarticity $RQ_t = \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4$. The β_{1t} parameter will vary with the estimated measurement error variance. In these models, the parameters can be obtained using ordinary least squares.

To compare the performance among the different models, we used the Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{353}\sum_{t=1}^{353}(y_t - \hat{y}_t)^2} \tag{5.9}$$

in which y_t represents the actual value, and $\hat{y}_t$ represents the predicted value.

Table 5.4 presents the comparison results of the root mean square error (RMSE) for various variance forecasting models over a 1-day ahead horizon. Overall, the results suggest that the ML model exhibited the smallest forecast errors, followed by the intraday models from the literature. Among the 29 examined assets, the ML model achieved lower RMSE compared with the HAR, CHAR, SHAR, HAR-J, HAR-Q, EMA, and GARCH models in 23 (79.31%), 23 (79.31%), 22 (75.86%), 26 (89.65%), 23 (79.31%), 23 (79.31%), and 29 (100%) assets, respectively.

To assess the statistical significance of the forecast differences, we employed the Diebold-Mariano test [Diebold and Mariano, 1995]. Considering $e_{1,t} = y_t - \hat{y}_{1,t}$ and $e_{2,t} = y_t - \hat{y}_{2,t}$ the residuals from two different models. And $d_t = e_{1,t}^2 - e_{2,t}^2$ the loss differential between the two forecasts. The test states the two forecasts have equal accuracy if and only if the loss differential has zero expectation for all t. The null hypothesis of the test is $H_0 : E(d_t) = 0,\ \forall t$.

$$\begin{aligned} \bar{d} &= \frac{1}{n}\sum_{1}^{T} d_t \\ \gamma_k &= \frac{1}{n}\sum_{t=k+1}^{n}(d_t - \bar{d})(d_{t-k} - \bar{d}) \\ DM &= \frac{d}{\sqrt{[\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k]\frac{1}{n}}},\ DM \sim N(0,1). \end{aligned}$$

Table 5.4: Comparison of Realized Variance RMSE from different models

	ML	**HAR**	**CHAR**	**SHAR**	**HAR-Q**	**HAR-J**	**EMA**	**GARCH**
ABEV3	**1.68E-04**	1.72E-04	1.87E-04	1.72E-04	1.77E-04	1.72E-04	1.79E-04	2.28E-04
ALPA4	5.01E-04	5.10E-04	**5.00E-04**	5.09E-04	5.20E-04	5.10E-04	5.05E-04	5.77E-04
BBAS3	**1.87E-04**	1.95E-04	2.02E-04	1.97E-04	2.03E-04	1.95E-04	1.96E-04	5.86E-04
BBDC4	**1.73E-04**	1.82E-04	1.82E-04	1.82E-04	1.90E-04	1.82E-04	1.80E-04	1.85E-04
BRKM5	**3.88E-04**	3.94E-04	3.89E-04	3.94E-04	4.03E-04	3.95E-04	3.97E-04	4.50E-04
CGRA4	**1.41E-03**	1.49E-03	1.55E-03	1.49E-03	1.49E-03	1.49E-03	1.37E-03	1.69E-03
CMIG4	**2.19E-04**	2.25E-04	2.73E-04	2.29E-04	2.32E-04	2.25E-04	2.30E-04	3.59E-04
CPFE3	**2.17E-04**	2.22E-04	2.32E-04	2.23E-04	2.24E-04	2.22E-04	2.26E-04	2.73E-04
CPLE6	**1.90E-04**	2.09E-04	1.96E-04	1.97E-04	2.01E-04	2.09E-04	1.96E-04	4.20E+00
CRFB3	3.06E-04	3.01E-04	**3.00E-04**	3.05E-04	3.08E-04	3.02E-04	3.09E-04	3.15E-04
CSNA3	3.71E-04	3.67E-04	**3.62E-04**	3.67E-04	3.74E-04	3.68E-04	3.75E-04	3.89E-04
ELET3	8.53E-04	8.39E-04	8.46E-04	8.44E-04	8.48E-04	8.39E-04	**8.16E-04**	2.52E-03
EMBR3	4.82E-04	4.97E-04	4.92E-04	4.97E-04	5.15E-04	4.97E-04	**4.76E-04**	5.33E-04
ENGI4	**2.36E-04**	2.39E-04	2.38E-04	2.38E-04	2.46E-04	2.39E-04	2.45E-04	2.81E-04
ENGI11	1.19E-03	1.17E-03	1.17E-03	**1.16E-03**	**1.16E-03**	1.17E-03	1.20E-03	1.29E-03
GGBR4	**2.28E-04**	2.33E-04	2.35E-04	2.33E-04	2.44E-04	2.33E-04	2.33E-04	2.78E-04
ITSA4	**1.37E-04**	1.44E-04	1.57E-04	1.44E-04	1.50E-04	1.44E-04	1.42E-04	1.83E-04
ITUB4	**1.89E-04**	2.00E-04	1.96E-04	1.99E-04	2.10E-04	2.00E-04	2.00E-04	2.10E-04
LIGT3	4.48E-04	4.48E-04	4.49E-04	4.49E-04	4.58E-04	4.49E-04	**4.46E-04**	5.41E-04
LREN3	**4.14E-04**	4.17E-04	**4.14E-04**	4.19E-04	4.29E-04	4.17E-04	4.21E-04	4.78E-04
PETR4	**2.99E-04**	3.05E-04	3.03E-04	3.01E-04	3.13E-04	3.05E-04	3.08E-04	3.33E-04
RADL3	**2.83E-04**	2.91E-04	**2.83E-04**	2.85E-04	2.92E-04	2.91E-04	2.93E-04	1.94E+00
SBSP3	**3.34E-04**	3.41E-04	3.40E-04	3.41E-04	3.47E-04	3.41E-04	3.52E-04	3.65E-04
SUZB3	**2.19E-04**	**2.19E-04**	2.22E-04	**2.19E-04**	2.26E-04	**2.19E-04**	2.21E-04	2.49E-04
TIMS3	**2.16E-04**	2.17E-04	2.17E-04	**2.16E-04**	2.34E-04	**2.16E-04**	2.18E-04	2.60E-04
UNIP6	4.87E-04	4.78E-04	4.79E-04	**4.76E-04**	4.89E-04	4.81E-04	4.83E-04	5.62E-04
USIM5	**3.37E-04**	3.40E-04	3.39E-04	3.39E-04	3.53E-04	3.41E-04	3.46E-04	3.68E-04
VALE3	**1.41E-04**	1.48E-04	1.81E-04	1.49E-04	1.60E-04	1.48E-04	1.51E-04	4.46E-03
VIVT3	1.37E-04	**1.28E-04**	1.52E-04	1.29E-04	1.32E-04	1.29E-04	1.29E-04	1.43E-04

Table 5.5 displays the p-values resulting from individual comparisons between the literature and ML models. In Table 5.6, we present the percentage of assets the machine learning algorithm had higher forecast accuracy according to the DM test for different α levels. In Table 5.7 we present the same results using the Bonferroni correction. These findings suggest that the ML model yields superior predictions to the other approaches examined for the data within this period.

5.2.2 Forecasting Kendall's tau correlation

Regarding intraday Kendall correlation forecasting, we computed the RMSE for each pair of stocks, resulting in a total of $\frac{29\times 28}{2} = 406$ distinct pairs. We compared the performance of the ML model against predictions from the following baselines:

Table 5.5: p-values of Diebold-Mariano test comparing the realized variance predictions of machine learning algorithm with some baselines

	HAR	**CHAR**	**SHAR**	**HAR-J**	**HAR-Q**	**EMA**	**GARCH**
ABEV3	4.69E-02	1.02E-10	6.11E-02	5.08E-02	1.31E-03	2.11E-03	3.97E-12
ALPA4	6.47E-02	1.95E-01	8.06E-02	6.77E-02	1.57E-03	1.18E-01	1.35E-05
BBAS3	2.02E-04	4.55E-10	7.92E-05	1.96E-04	1.11E-07	9.42E-03	4.92E-70
BBDC4	1.03E-03	5.83E-04	9.02E-04	1.02E-03	1.46E-06	8.27E-05	2.53E-02
BRKM5	1.60E-03	1.04E-02	1.96E-03	9.95E-04	1.81E-05	5.98E-03	6.76E-06
CGRA4	2.90E-04	5.84E-08	7.76E-04	2.74E-04	8.01E-04	1.54E-01	7.56E-11
CMIG4	2.65E-06	7.84E-14	1.05E-06	7.80E-06	3.33E-07	9.85E-05	5.65E-10
CPFE3	3.31E-03	6.43E-10	7.42E-04	3.22E-03	2.93E-04	1.69E-03	2.02E-03
CPLE6	5.08E-07	3.17E-02	3.71E-03	4.06E-07	7.60E-03	3.56E-02	0.00E+00
CRFB3	2.92E-01	6.42E-02	1.07E-01	2.28E-01	3.48E-03	3.07E-02	9.55E-02
CSNA3	3.80E-01	2.37E-01	3.73E-01	3.54E-01	1.20E-02	1.03E-01	5.30E-02
ELET3	5.06E-01	1.52E-05	9.12E-02	4.35E-01	5.93E-03	2.40E-01	4.09E-75
EMBR3	2.44E-03	5.05E-02	3.18E-03	3.25E-03	6.50E-06	6.63E-03	1.61E-04
ENGI4	7.06E-02	1.70E-03	9.06E-02	8.80E-02	8.88E-03	2.76E-04	1.32E-06
ENGI11	9.93E-01	6.17E-01	9.81E-01	9.77E-01	9.66E-01	1.84E-02	8.71E-08
GGBR4	3.02E-03	2.90E-04	3.70E-03	2.92E-03	2.42E-06	7.64E-03	3.62E-05
ITSA4	5.14E-02	1.32E-17	3.79E-02	2.34E-02	9.82E-04	3.11E-01	1.61E-11
ITUB4	1.40E-03	8.30E-06	8.33E-04	1.20E-03	5.18E-06	1.49E-04	1.28E-03
LIGT3	2.78E-01	2.18E-01	2.95E-01	2.78E-01	1.48E-02	9.09E-02	4.41E-03
LREN3	1.58E-01	8.32E-02	1.19E-01	1.57E-01	3.92E-03	7.28E-03	1.25E-04
PETR4	1.22E-01	8.60E-03	1.65E-01	1.25E-01	1.16E-03	1.69E-03	2.45E-02
RADL3	3.01E-02	1.87E-01	2.03E-01	3.01E-02	1.87E-03	1.38E-04	0.00E+00
SBSP3	3.64E-04	1.88E-05	2.48E-04	5.27E-04	1.64E-05	4.86E-06	2.64E-05
SUZB3	2.88E-02	6.39E-04	1.90E-02	4.49E-02	4.13E-04	8.70E-03	1.55E-05
TIMS3	2.27E-01	7.84E-02	3.40E-01	3.60E-01	1.32E-02	2.04E-01	3.79E-06
UNIP6	8.91E-02	1.16E-03	1.27E-01	2.04E-02	4.64E-04	7.44E-02	2.80E-02
USIM5	8.84E-04	2.60E-04	8.84E-04	6.52E-04	1.88E-05	1.51E-03	7.23E-04
VALE3	5.23E-03	6.53E-15	2.98E-04	4.69E-03	1.41E-06	2.43E-03	4.08E-171
VIVT3	1.00E+00	6.83E-15	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.79E-01

- EMA,
- zero correlation (zero),
- the last observed intraday correlation as the prediction (τ_{t-1}),
- average of historical intraday Kendall correlation samples (intraday mean),
- average of historical daily Kendall correlation samples (daily mean)

In Figure 5.12, we compare the RMSE of Machine Learning versus EWA for each pair of assets.

Out of the total 406 pairs of stocks, the ML model exhibited lower RMSE errors in 370 (91.13%), 378 (93.56%), 338 (83.25%), 319 (78.96%) and 245 (60.34%) cases when compared to the zero correlation, τ_{t-1}, daily mean, intraday mean, and EMA, respectively.

Table 5.6: Percentage of assets that machine learning algorithm had rejected null hypothesis from DM-test against the baselines for different α levels

α	**HAR**	**CHAR**	**SHAR**	**HARQ**	**HARJ**	**EMA**	**GARCH**
0.01	44.83	62.07	44.83	44.83	79.31	79.31	58.62
0.05	55.17	68.97	51.72	58.62	79.31	79.31	68.97
0.10	68.97	79.31	65.52	68.97	79.31	79.31	75.86
0.15	72.41	79.31	75.86	72.41	79.31	79.31	82.76
0.20	75.86	79.31	75.86	75.86	79.31	79.31	86.21
0.25	79.31	79.31	75.86	79.31	79.31	79.31	93.10

Table 5.7: Percentage of assets that machine learning algorithm had rejected null hypothesis from DM-test against the baselines for different corrected α levels by Bonferroni

α	**corrected α**	**HAR**	**BPV**	**SHAR**	**HARJ**	**HARQ**	**EMA**	**GARCH**
0.01	**3.45E-04**	13.79	44.83	13.79	13.79	37.93	20.69	65.52
0.05	**1.72E-03**	31.03	58.62	31.03	31.03	62.07	31.03	72.41
0.10	**3.45E-03**	41.38	58.62	37.93	41.38	65.52	37.93	75.86
0.15	**5.17E-03**	41.38	58.62	44.83	44.83	72.41	37.93	79.31
0.20	**6.90E-03**	44.83	58.62	44.83	44.83	75.86	44.83	79.31
0.25	**8.62E-03**	44.83	62.07	44.83	44.83	79.31	51.72	79.31

In Table 5.8, we present the percentage of assets the machine learning algorithm had higher forecast accuracy according to the DM test for different α levels, and in Table 5.9 we present the same results using the Bonferroni correction. These findings also suggest that the ML model yields superior predictions for intraday Kendall correlation compared to the other approaches examined for the data within this period.

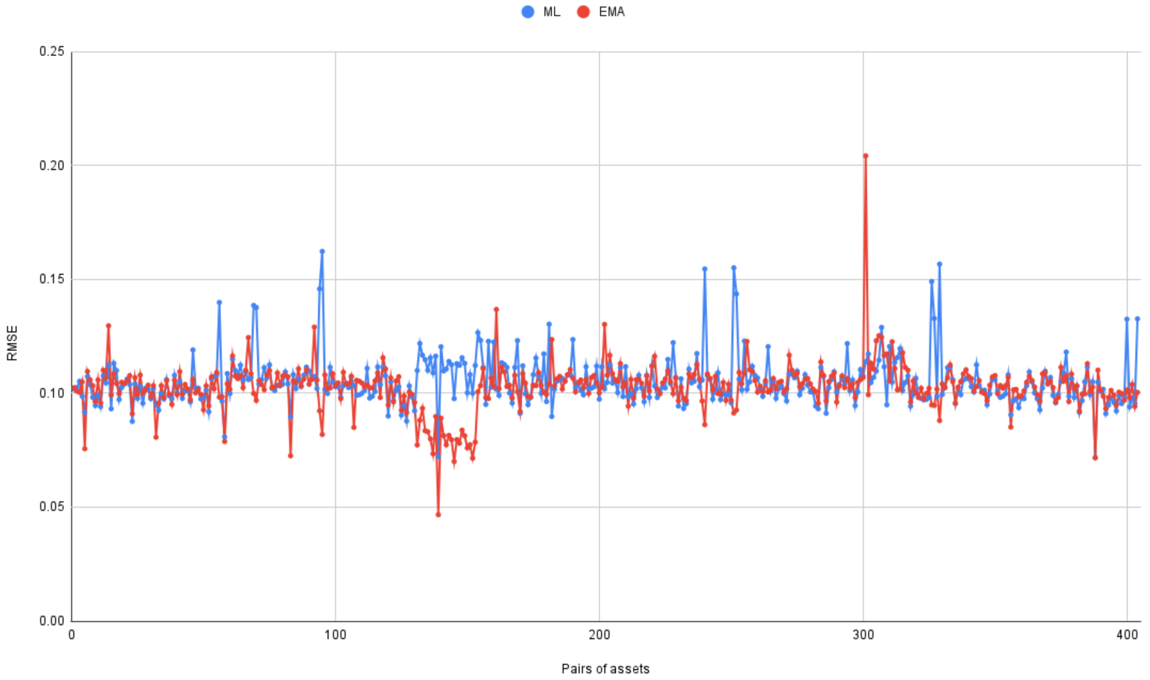


Figure 5.12: RMSE of Kendall correlations

Table 5.8: Percentage of pairs of assets that machine learning algorithm had rejected null hypothesis from DM-test against the baselines for different α levels

α	**Zero**	τ_{t-1}	**Daily Mean**	**Intraday Mean**	**EMA**
0.01	88.17	92.32	68.71	57.38	10.64
0.05	89.16	92.57	71.42	63.79	23.02
0.10	89.40	93.06	73.39	66.74	29.95
0.15	89.90	93.06	75.36	69.95	35.64
0.20	90.14	93.31	76.60	70.68	40.34
0.25	90.39	93.56	79.06	71.42	44.30

Table 5.9: Percentage of assets that machine learning algorithm had rejected null hypothesis from DM-test against the baselines for different corrected α levels by Bonferroni

Alpha	**Adjusted Alpha**	**Zero**	τ_{t-1}	**Daily Mean**	**Intraday Mean**	**EMA**
0.01	**2.46E-05**	85.96	87.62	58.86	44.08	2.97
0.05	**1.23E-04**	85.96	90.09	60.59	47.29	3.21
0.10	**2.46E-04**	86.69	91.08	63.05	48.76	3.46
0.15	**3.69E-04**	86.69	91.08	63.30	49.26	3.71
0.20	**4.92E-04**	86.94	91.08	64.03	49.50	3.71
0.25	**6.15E-04**	86.94	91.08	64.03	49.75	3.71

5.2.3 Scenario Generation and Portfolio Optimization

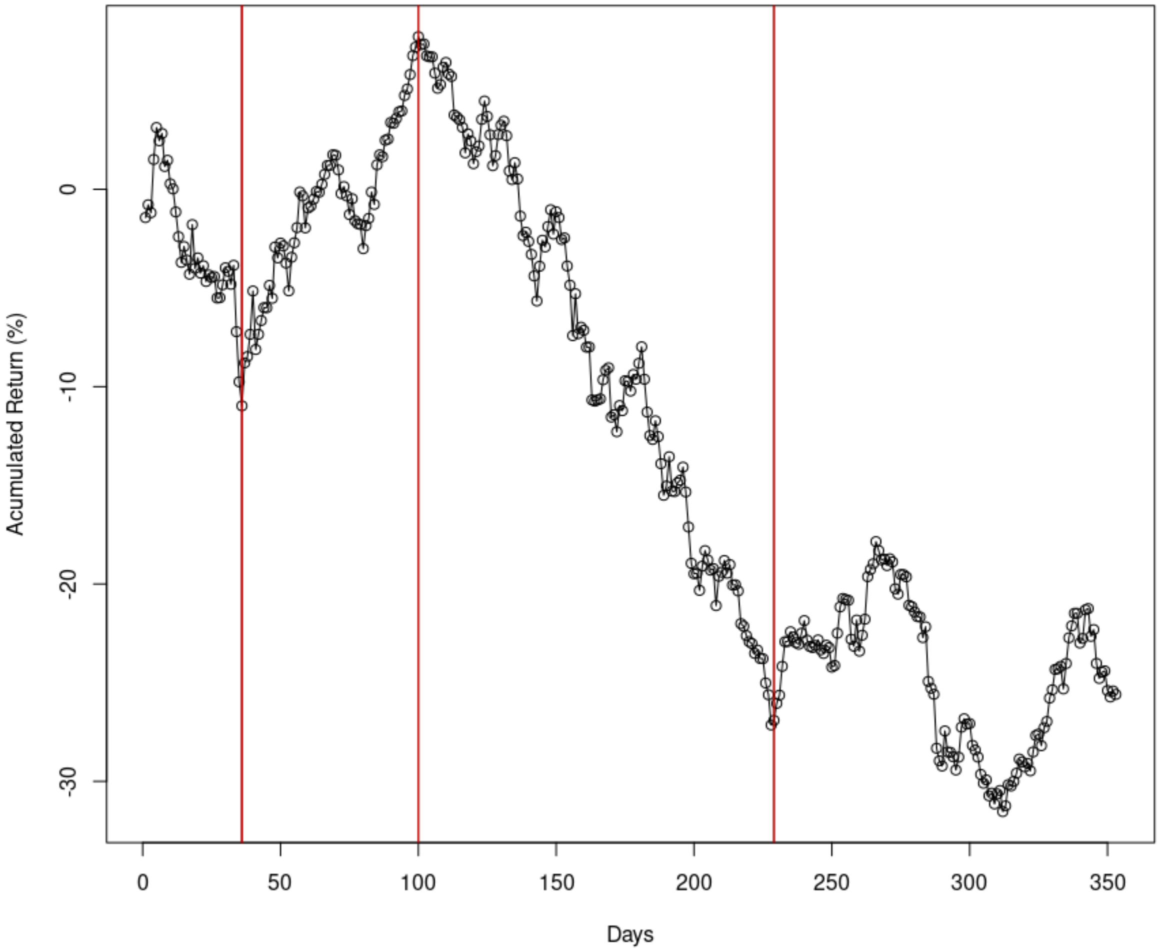


Figure 5.13: Index BOVA11 operation performance from 04/01/2021 to 13/09/2022

This section aims to apply all the steps of the proposed methodology to generate the future scenario matrix (1 day ahead) and optimize the portfolio weights through simulation. With the obtained weights, we simulate the operation in the market by buying the stocks at the beginning of the day (09:00 am) and selling them at the end of the day (6:00 pm). To compare the results obtained by the proposed methodology, we use other baselines from the literature to operate in the same manner. We compare the following strategies:

- Historical: This baseline uses the historical distribution of assets (using the last

1000 historical data observations) representing the future matrix. A sliding window is used for each new day to always consist of a matrix of scenarios for 1000 days Guastaroba et al. [2009].

- Bootstrap: This strategy uses historical observations representing the future matrix; however, each scenario (row of the matrix) corresponds to a random choice of a past day. We use bootstrap with repetition, meaning the same historical day can repeat. Each new day also adds this day to the pool of days to be drawn from historical observations using a sliding window. We use 1000 samples for the matrix Guastaroba et al. [2009].

- GARCH-EVT-Copula: This approach uses GARCH models adjusted from historical days to fit each asset. Each residual distribution follows a generalized Pareto distribution adjusted by the EVT technique. A copula function (adjusted with the historical data) models the correlation between assets. The GARCH model predicts the one-day-ahead volatility, adjusting each distribution to have the associated volatility Sahamkhadam et al. [2018], Wang et al. [2010].

- ML-EVT-Copula: The proposed methodology uses machine learning algorithms to make predictions of intraday volatility and correlation between assets (1 day ahead). We set the marginal distributions of each asset using the EVT technique and volatility predictions. We also set the copula parameters using the Kendall correlation predictions. This approach dynamically adapts the volatility and correlation of the multivariate distribution.

Each generated scenario matrix consists of 1000 scenarios (rows) and 29 columns (each asset) in all the different strategies. In the GARCH-EVT-Copula and ML-EVT-Copula strategies, we adjust the mean of each asset in the scenario matrix (for each different day) to be equal to the same mean obtained by the Bootstrap strategy. We apply this approach for two reasons. The first is that the historical and Bootstrap baselines will obtain mean values different from zero when selecting historical days. We adopt this criterion to prevent any bias related to the differences in the means from the results of a different strategy. The second reason is that since we use the STARR Ratio optimization model, the assets must have mean values different from zero.

In the GARCH-EVT-Copula and ML-EVT-Copula strategies, we use three different copula functions to evaluate the impacts on the financial returns: normal, t-student (with $v = 2$), and Clayton copula. Both normal and t-student copulas are symmetric however the normal copula has zero tail dependence while the t-student has positive values. On the other hand, the Clayton copula is asymmetric with tail dependence only in the negative tail.

Figure 5.13 shows the cumulative financial return of the BOVA11 index when using the strategy of buying (at 09:00) and selling (at 18:00) over the 353 days of the out-of-sample dataset (04/01/2021 - 13/09/2022). We observe market behaviors in different periods: growth, decline, and sideways movement. The proposed analysis involves evaluating the different strategies in the different individual periods and analyzing the 353 consecutive days.

5.2.3.1 First period: from 04/01/2021 to 01/03/2021 (decline movement)

Figures 5.14, 5.15, 5.16, 5.17, and Table 5.10, suggest that the ML-EVT-Clayton approach stood out, being the only one to show a positive mean (and consequently a positive cumulative return). Regarding downside risk measures and CVaR, it was the second-best strategy, trailing only behind ML-EVT-Tstudent. It also achieved the best values for the Sharpe ratio, Sortino Ratio, Starr Ratio, and Omega Ratio.

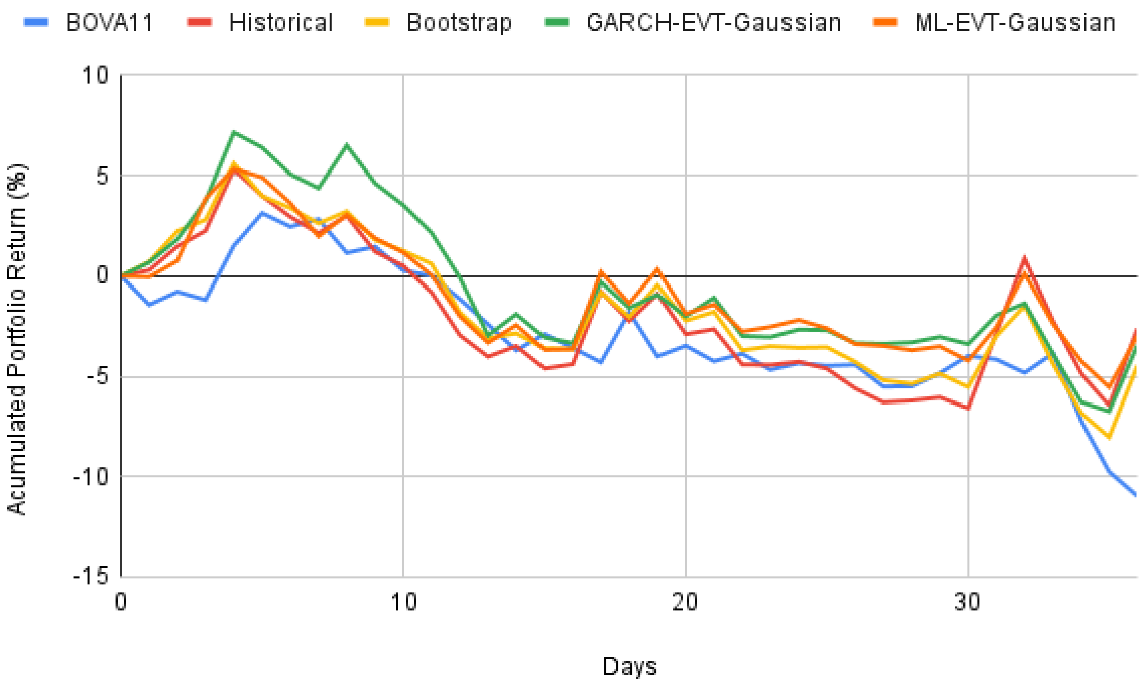


Figure 5.14: Comparison of accumulated portfolio return for different portfolio strategies. Period of decline movement of the market, days 1-36

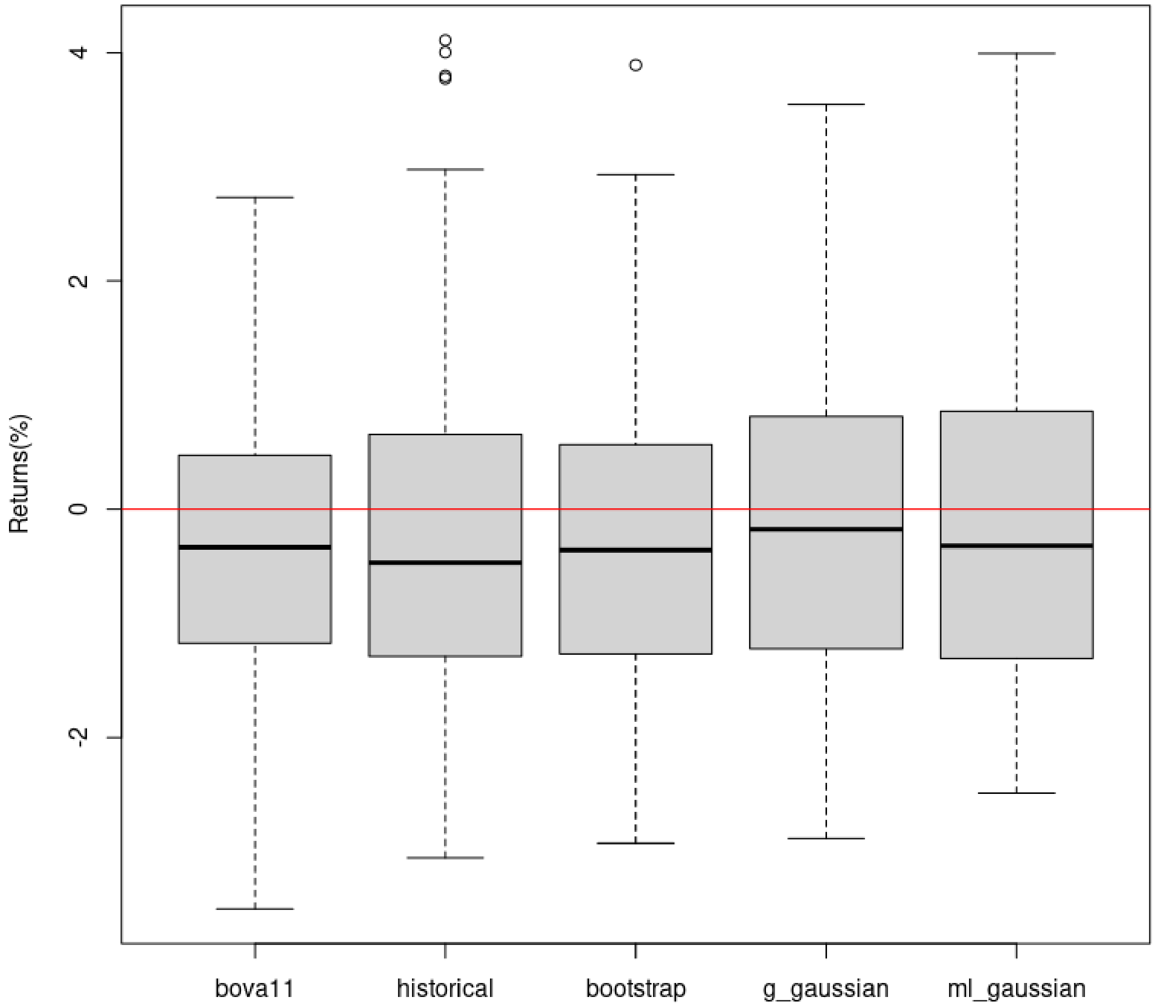


Figure 5.15: Comparison of return distributions for different portfolio strategies. Period of decline movement of the market, days 1-36

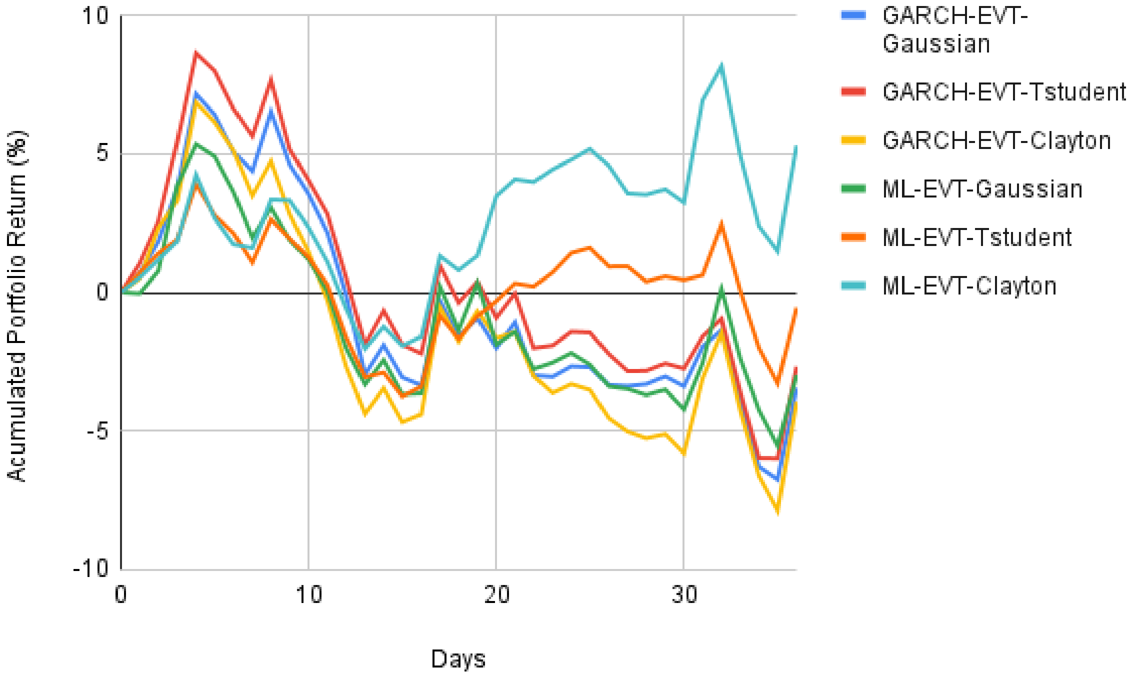


Figure 5.16: Comparison of accumulated portfolio return for different copula functions. Period of decline movement of the market, days 1-36

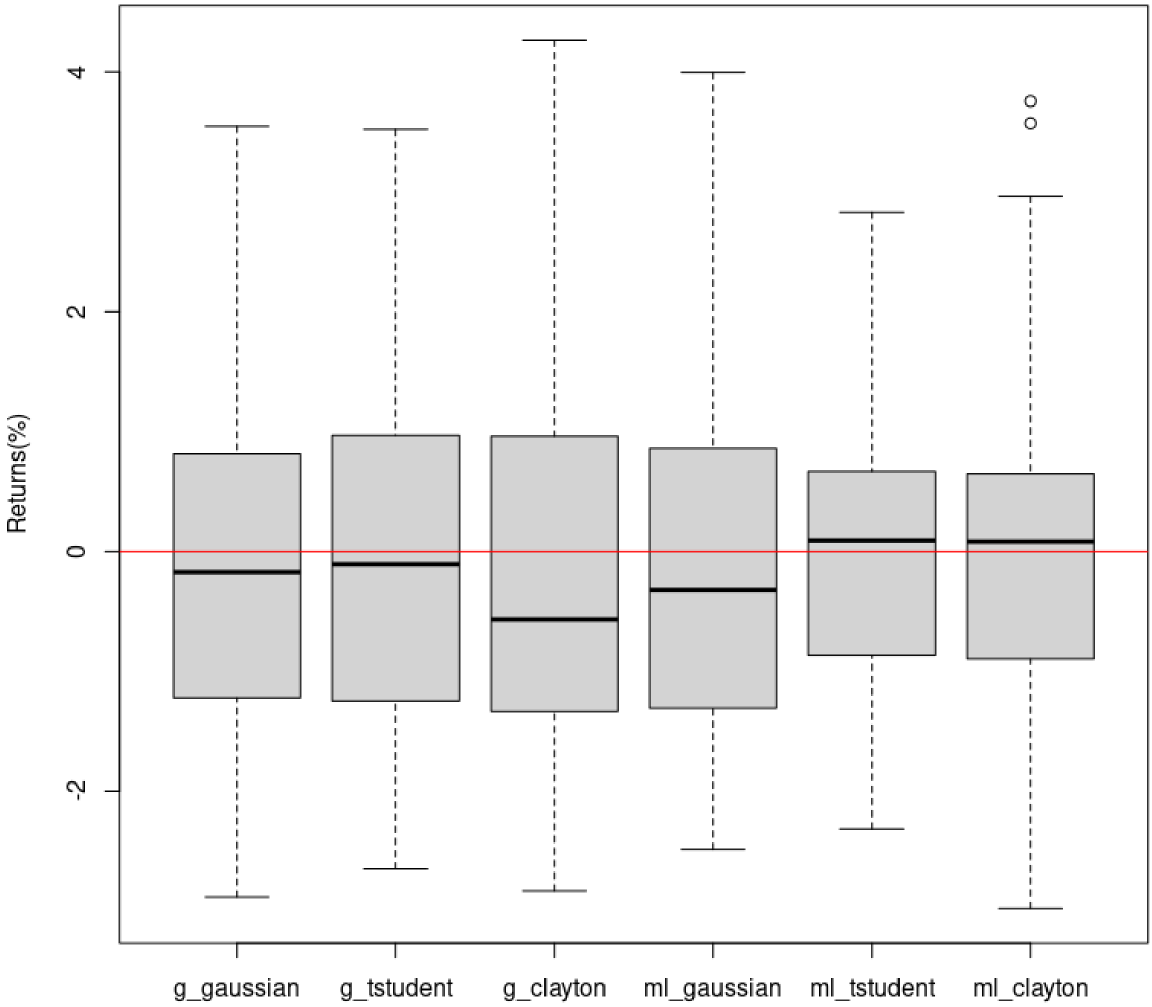


Figure 5.17: Comparison of return distributions for different copula functions. Period of decline movement of the market, days 1-36

Table 5.10: Comparison metrics of different portfolio strategies. Period of decline movement of the market, days 1-36

	Mean	**Std**	**Downside**	**CVaR**	**Sharpe**	**Sortino**	**Starr**	**Omega**
BOVA11	-0.31	1.31	1.08	-2.97	-0.24	-0.29	0.11	0.54
Historical	-0.06	1.87	1.14	-3.19	-0.03	-0.05	0.02	0.92
Bootstrap	-0.11	1.59	1.07	-2.83	-0.07	-0.11	0.04	0.83
G-Gaussian	-0.08	1.61	1.10	-2.92	-0.05	-0.08	0.03	0.88
G-Student	-0.06	1.66	1.11	-2.96	-0.04	-0.06	0.02	0.91
G-Clayton	-0.10	1.78	1.13	-3.08	-0.05	-0.09	0.03	0.87
ML-Gaussian	-0.07	1.59	1.01	-2.73	-0.05	-0.07	0.03	0.89
ML-Student	-0.01	1.22	0.81	-2.22	-0.01	-0.01	0.00	0.98
ML-Clayton	0.15	1.51	0.88	-2.50	0.10	0.18	-0.06	1.32

5.2.3.2 Second period: from 02/03/2021 to 07/06/2021(upswing movement)

Analyzing Figures 5.18, 5.19, 5.20, 5.21, and Table 5.11, it is noticeable that different strategies excel in various ways. In terms of mean and cumulative return, ML-EVT-Gaussian and GARCH-EVT-Clayton stood out. Concerning risk measures, the BOVA11 index had the lowest standard deviation, the most minor downside risk, and the CvaR. The best strategy for the Sharpe Ratio was BOVA11, while ML-EVT-GARCH excelled in the Sortino Ratio, Starr Ratio, and Omega Ratio.

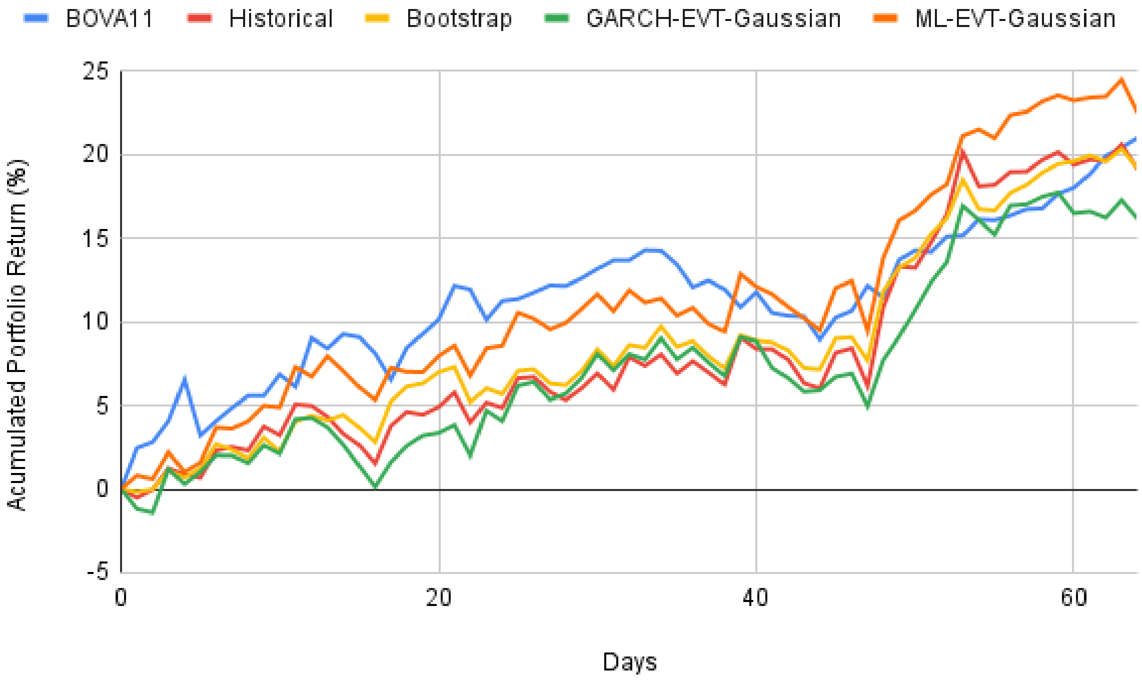


Figure 5.18: Comparison of accumulated portfolio return for different portfolio strategies. Period of decline upswing of the market, days 37-100

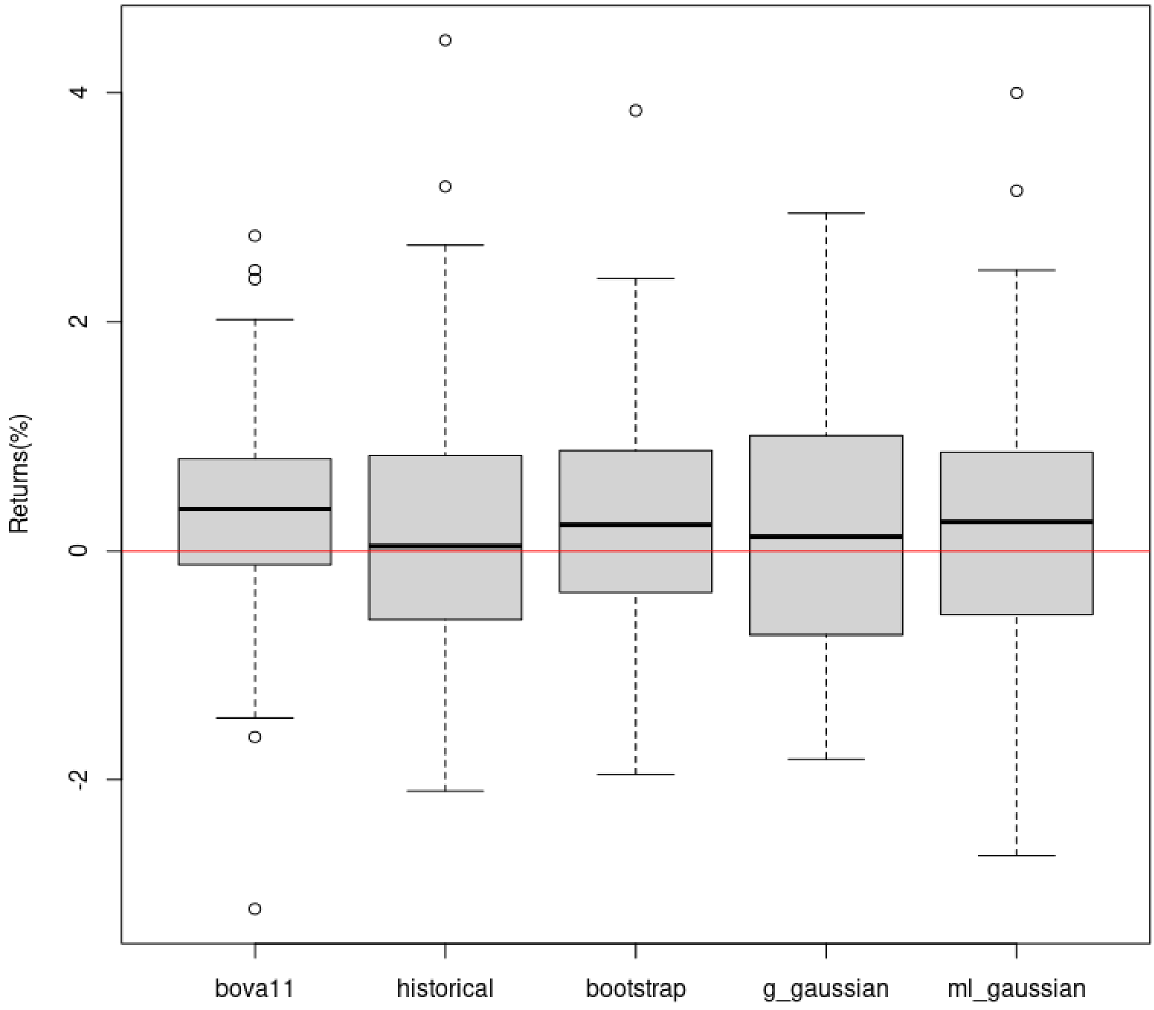


Figure 5.19: Comparison of return distributions for different portfolio strategies. Period of upswing movement of the market, days 37-100

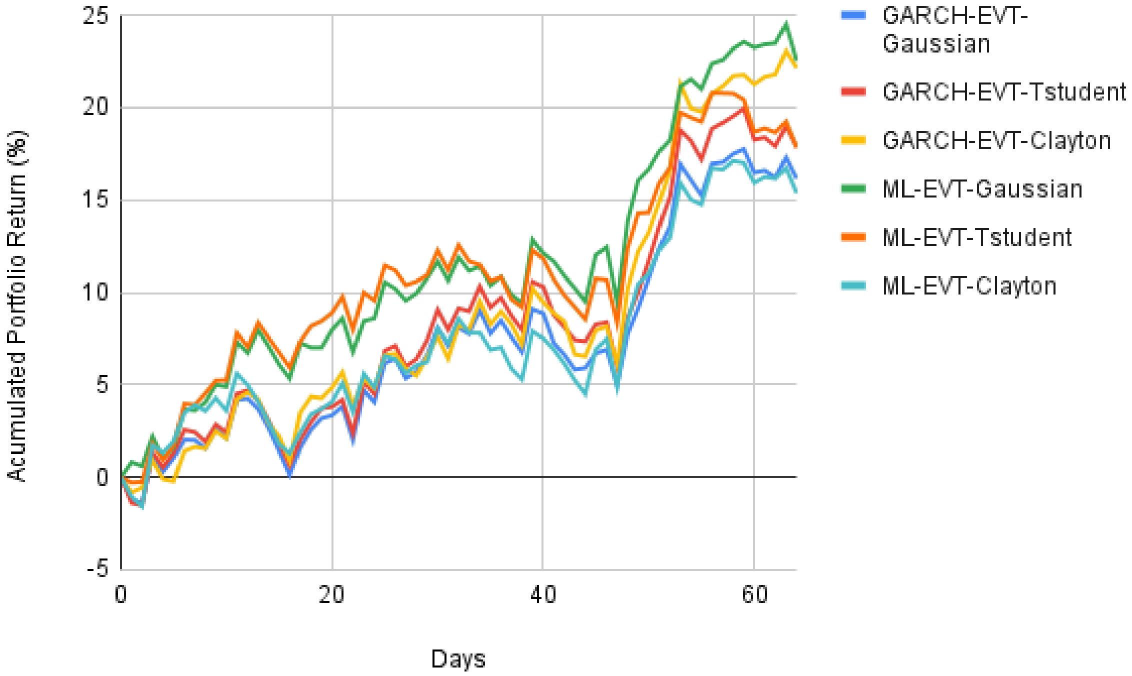


Figure 5.20: Comparison of acumulated portfolio return for different copula functions. Period of upswing movement of the market, days 37-100

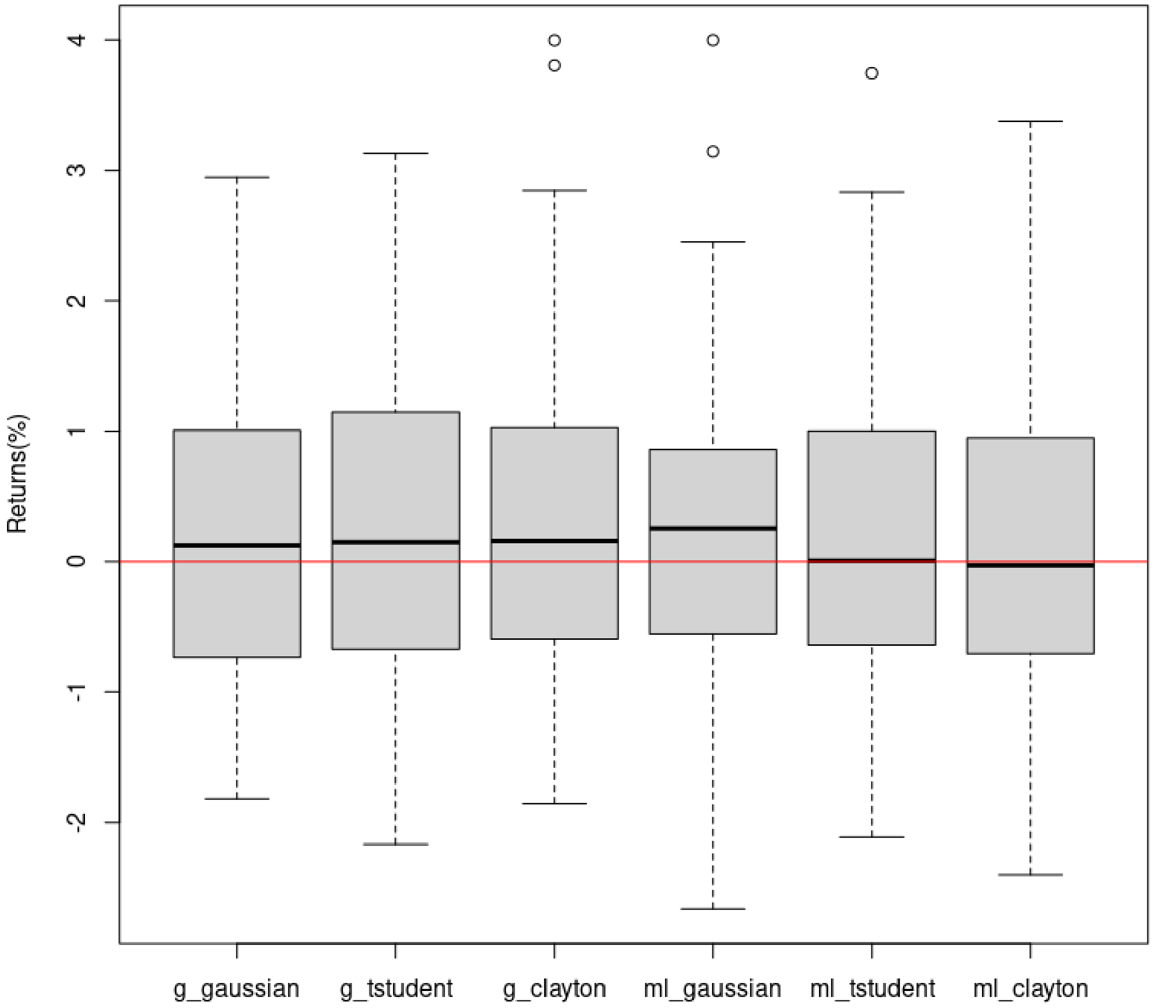


Figure 5.21: Comparison of return distributions for different copula functions. Period of decline upswing of the market, days 37-100

Table 5.11: Comparison metrics of different portfolio strategies. Period of upswing movement of the market, days 37-100

	Mean	**Std**	**Downside**	**CVaR**	**Sharpe**	**Sortino**	**Starr**	**Omega**
BOVA11	0.30	1.01	0.59	-2.18	0.30	0.51	-0.14	2.26
Historical	0.28	1.22	0.59	-1.74	0.23	0.48	-0.16	1.86
Bootstrap	0.28	1.02	0.50	-1.46	0.28	0.56	-0.19	2.07
G-Gaussian	0.24	1.17	0.62	-1.82	0.21	0.39	-0.13	1.67
G-Student	0.26	1.22	0.64	-1.91	0.22	0.41	-0.14	1.71
G-Clayton	0.32	1.25	0.59	-1.78	0.26	0.54	-0.18	1.97
ML-Gaussian	0.33	1.18	0.58	-1.75	0.28	0.56	-0.19	2.09
ML-Student	0.26	1.17	0.57	-1.72	0.22	0.46	-0.15	1.83
ML-Clayton	0.23	1.20	0.60	-1.81	0.19	0.39	-0.13	1.64

5.2.3.3 Third period: from 08/06/2021 to 06/01/2022 (decline movement)

In the third period (decline), all strategies showed a negative mean as per Figures 5.22, 5.23, 5.24, 5.25, and Table 5.12. The ML-EVT-Tstudent strategy achieves the best mean. Regarding risk measures, BOVA11 had the lowest standard deviation, while ML-EVT-Gaussian exhibited the best downside risk value, and GARCH-EVT-Gaussian had the best CvaR value. The GARCH-EVT-Clayton strategy achieved the best Omega Ratio.

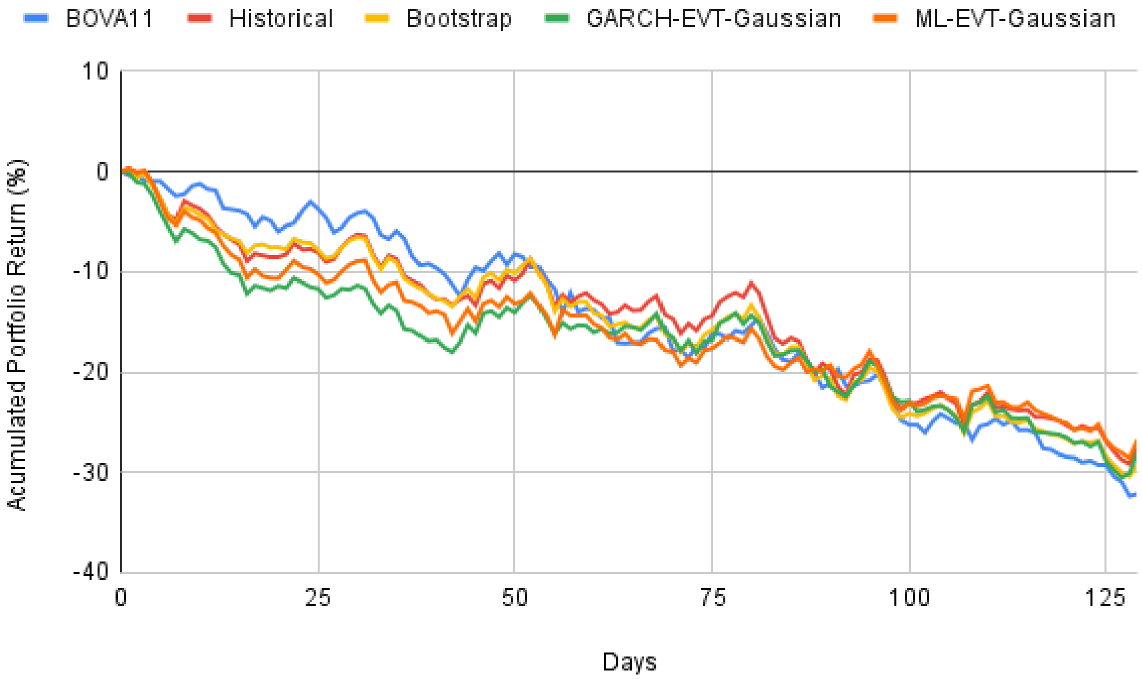


Figure 5.22: Comparison of accumulated portfolio return for different portfolio strategies. Period of decline movement of the market, days 101-229

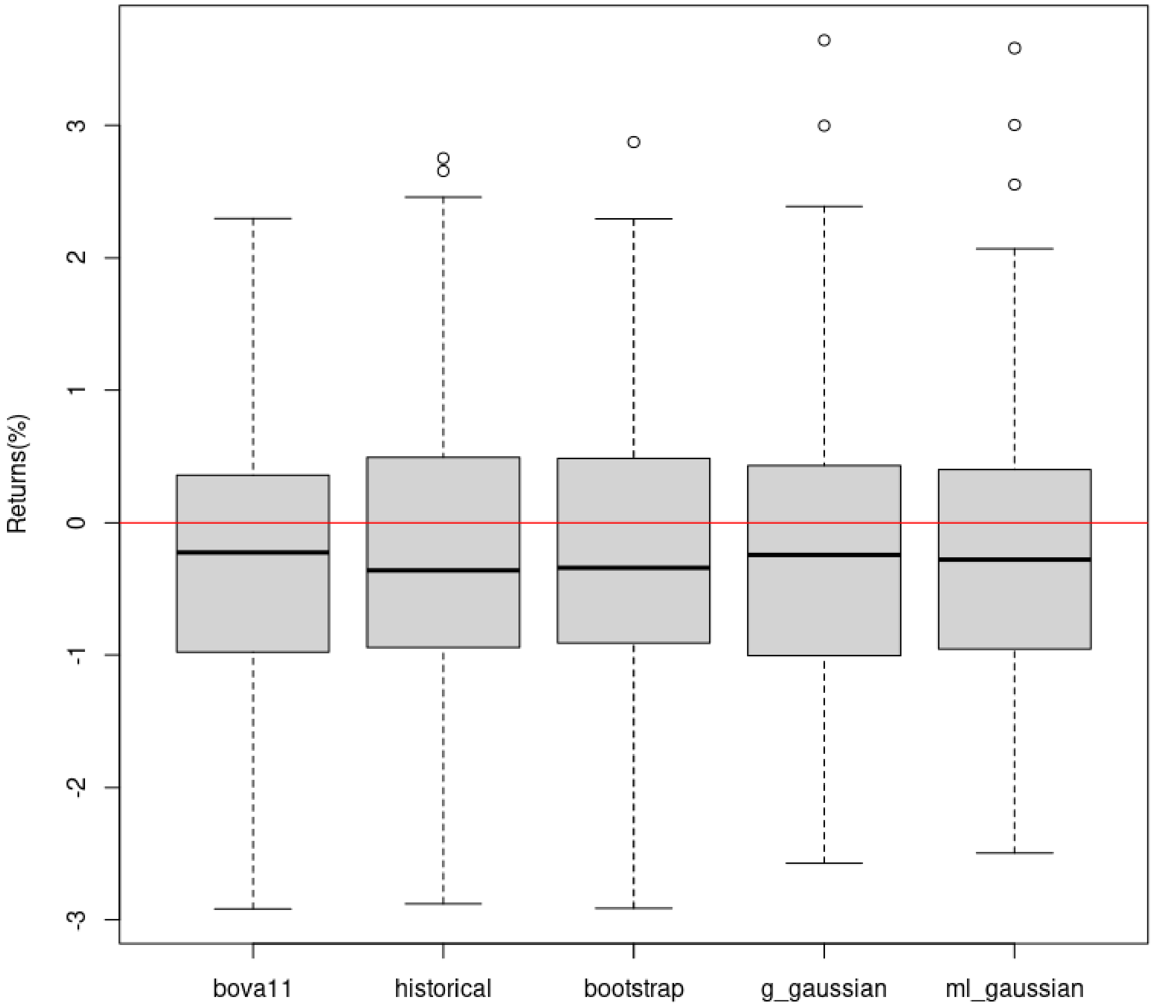


Figure 5.23: Comparison of return distributions for different portfolio strategies. Period of decline movement of the market, days 101-229

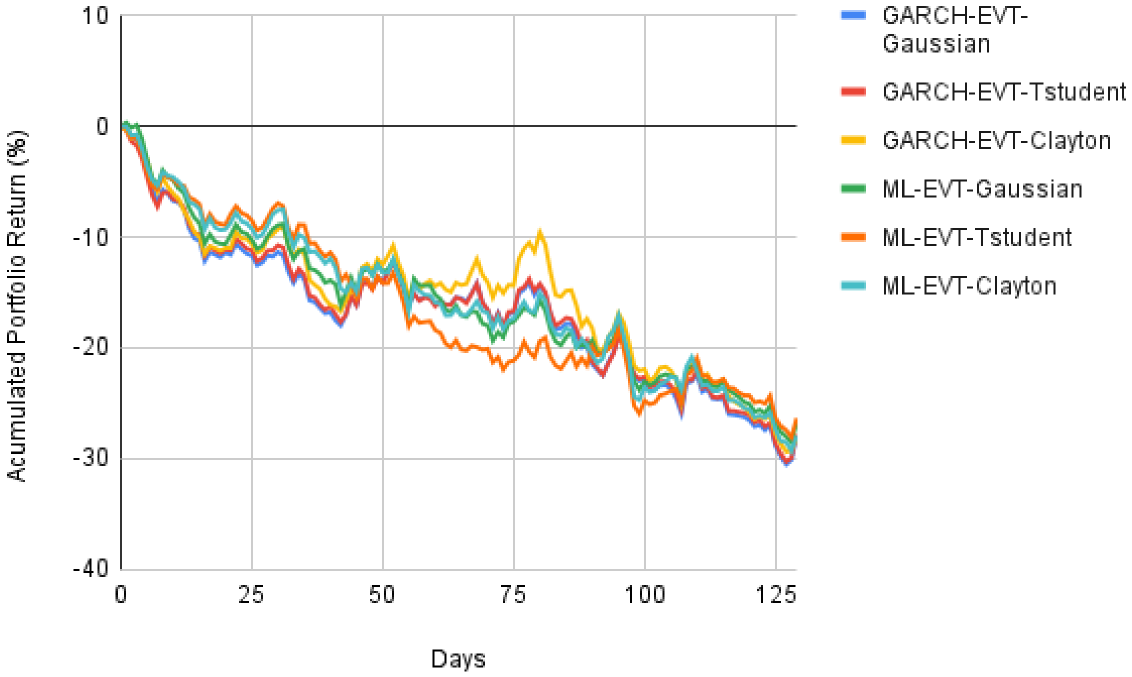


Figure 5.24: Comparison of acumulated portfolio return for different copula functions. Period of decline movement of the market, days 101-229

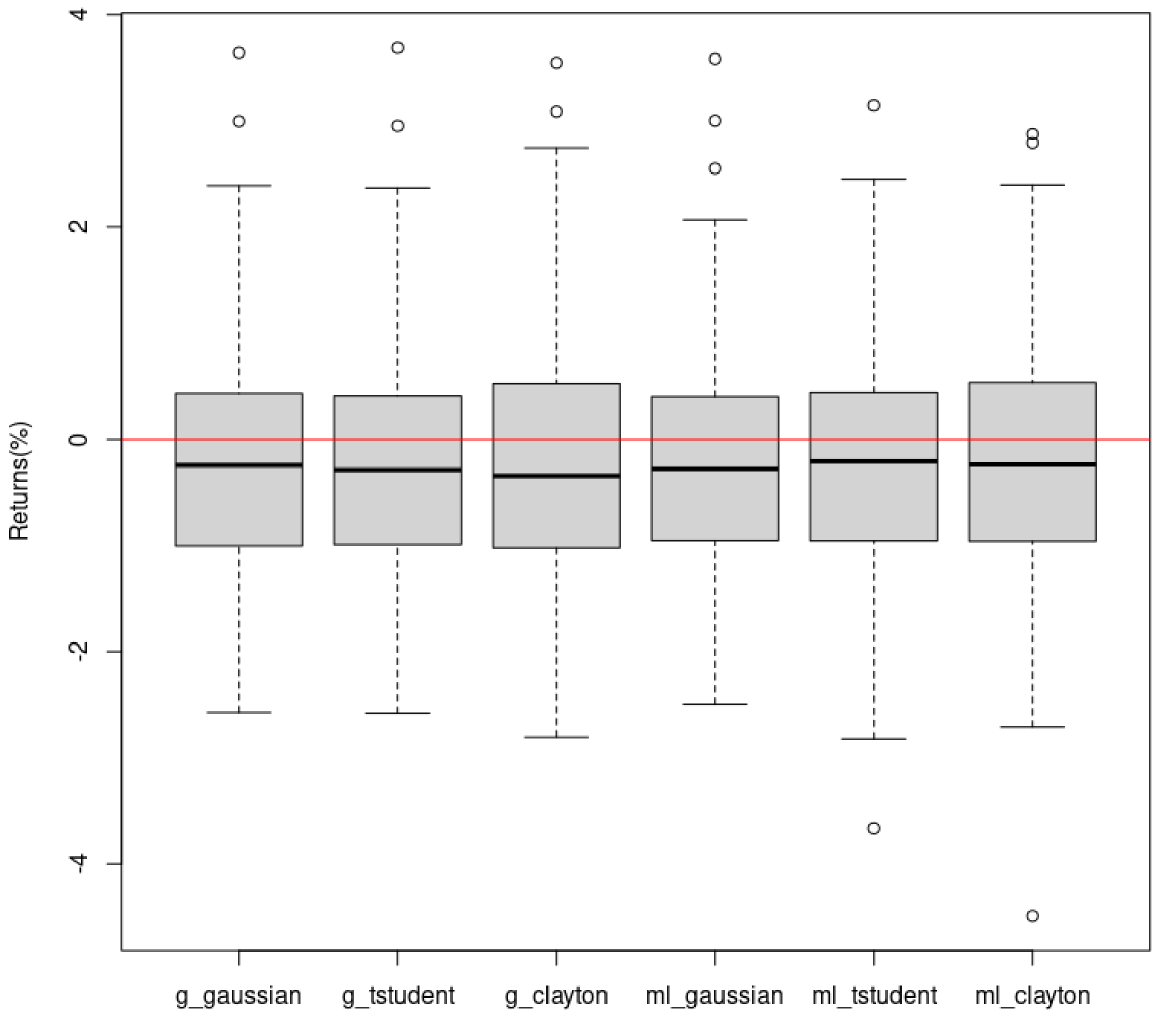


Figure 5.25: Comparison of return distributions for different copula functions. Period of decline upswing of the market, days 101-229

Table 5.12: Comparison metrics of different portfolio strategies. Period of decline movement of the market, days 101-229

	Mean	Std	Downside	CVaR	Sharpe	Sortino	Starr	Omega
BOVA11	-0.30	1.07	0.94	-2.51	-0.27	-0.31	0.12	0.49
Historical	-0.24	1.16	0.95	-2.53	-0.21	-0.26	0.10	0.59
Bootstrap	-0.26	1.10	0.92	-2.46	-0.24	-0.28	0.11	0.55
G-Gaussian	-0.25	1.13	0.90	-2.27	-0.22	-0.27	0.11	0.58
G-Student	-0.25	1.14	0.91	-2.28	-0.22	-0.27	0.11	0.58
G-Clayton	-0.25	1.21	0.97	-2.54	-0.21	-0.26	0.10	0.60
ML-Gaussian	-0.24	1.13	0.90	-2.33	-0.21	-0.26	0.10	0.59
ML-Student	-0.23	1.13	0.93	-2.64	-0.20	-0.25	0.09	0.59
ML-Clayton	-0.25	1.17	0.96	-2.82	-0.21	-0.26	0.09	0.58

5.2.3.4 Forth period: from 07/01/2022 to 13/09/2022 (sideway movement)

In the period of sideways movement, it is evident from Figures 5.26, 5.27, 5.28, 5.29, and Table 5.13 that the ML-EVT-Clayton strategy outperforms all others in terms of mean, downside risk, Sharpe Ratio, Sortino Ratio, and Omega Ratio. BOVA11 had the best standard deviation and CvaR, with ML-EVT-Tstudent (followed by ML-EVT-Clayton) achieving the second-best CvaR.

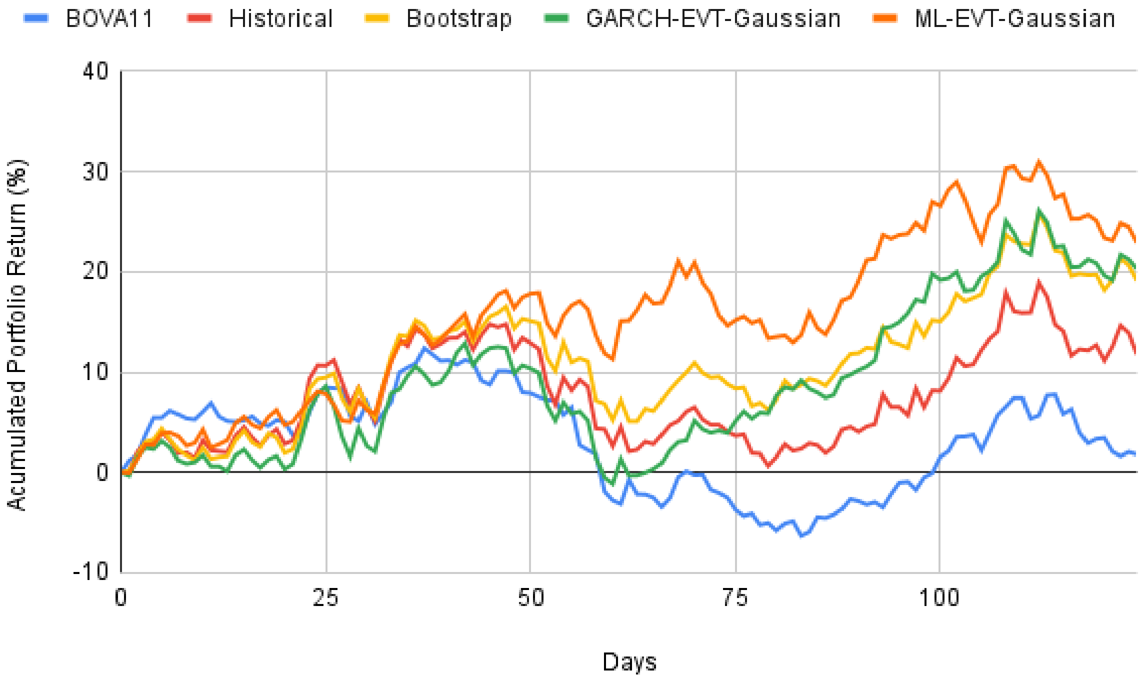


Figure 5.26: Comparison of accumulated portfolio return for different portfolio strategies. Period of sideway movement of the market, days 230-353

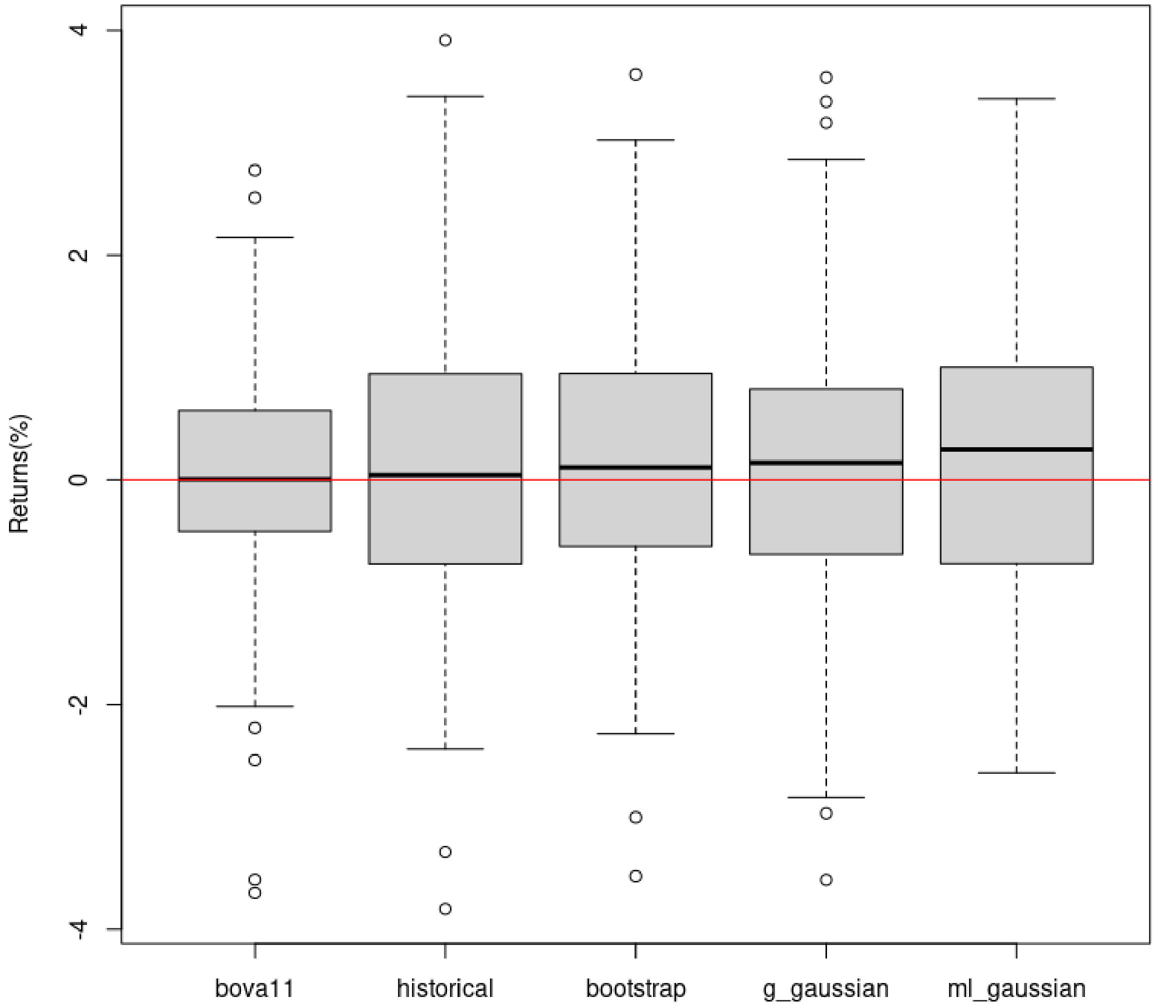


Figure 5.27: Comparison of return distributions for different portfolio strategies. Period of sideway movement of the market, days 230-353

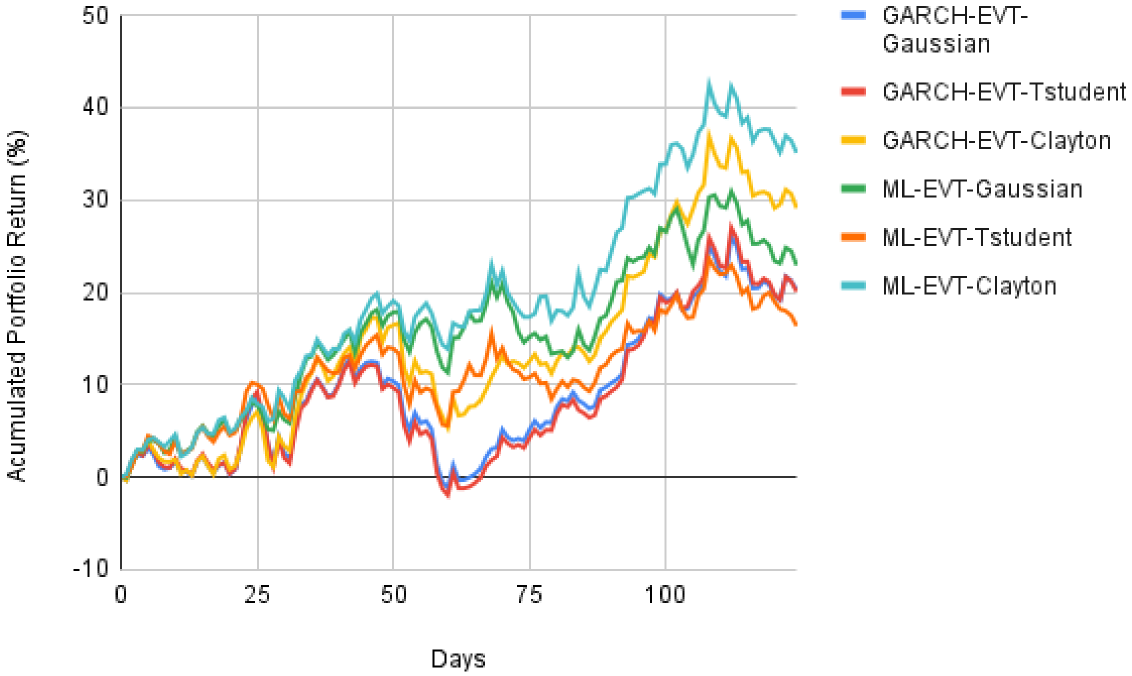


Figure 5.28: Comparison of accumulated portfolio return for different copula functions. Period of sideway movement of the market, days 230-353

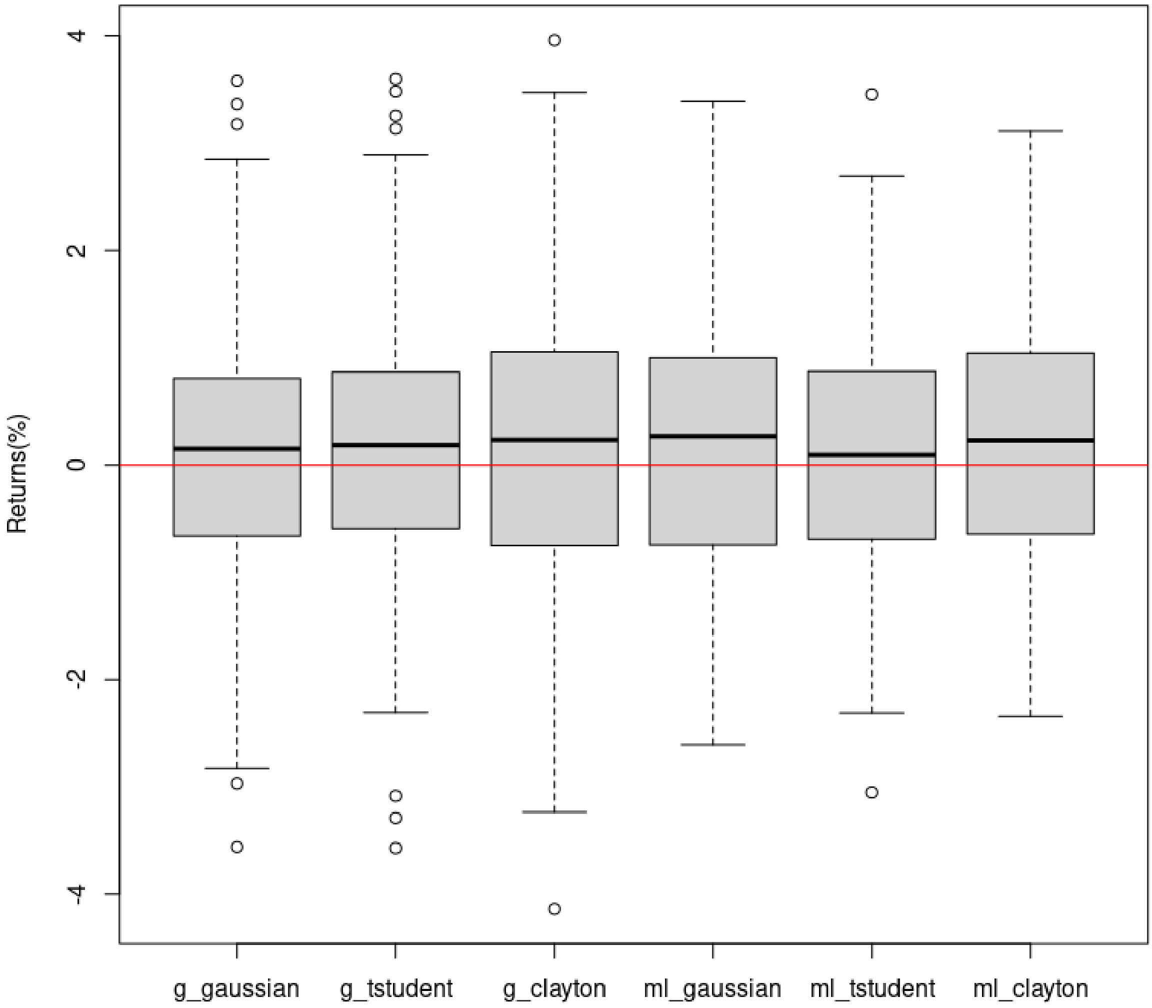


Figure 5.29: Comparison of return distributions for different copula functions. Period of sideway of the market, days 230-353

Table 5.13: Comparison metrics of different portfolio strategies. Period of sideway movement of the market days 230-353

	Mean	**Std**	**Downside**	**CVaR**	**Sharpe**	**Sortino**	**Starr**	**Omega**
BOVA11	0.02	1.12	0.81	-2.68	0.02	0.03	-0.01	1.05
Historical	0.10	1.37	0.90	-2.70	0.07	0.11	-0.04	1.20
Bootstrap	0.15	1.26	0.80	-2.45	0.12	0.19	-0.06	1.35
G-Gaussian	0.16	1.32	0.83	-2.52	0.12	0.19	-0.06	1.37
G-Student	0.16	1.37	0.87	-2.67	0.11	0.18	-0.06	1.35
G-Clayton	0.22	1.35	0.84	-2.73	0.16	0.26	-0.08	1.51
ML-Gaussian	0.17	1.21	0.75	-2.19	0.14	0.23	-0.08	1.42
ML-Student	0.13	1.18	0.73	-2.13	0.11	0.18	-0.06	1.32
ML-Clayton	0.25	1.24	0.72	-2.15	0.20	0.35	-0.12	1.67

5.2.3.5 Analysis of the entire period: from 04/01/2021 to 13/09/2022

The results above suggest that different strategies perform better for different comparison measures in different periods. To assess which strategy is more robust, ensuring performance across different market periods, we conducted the same analysis for the entire period. Figures 5.30, 5.31, 5.32, 5.33, and Table 5.14 demonstrate the results obtained.

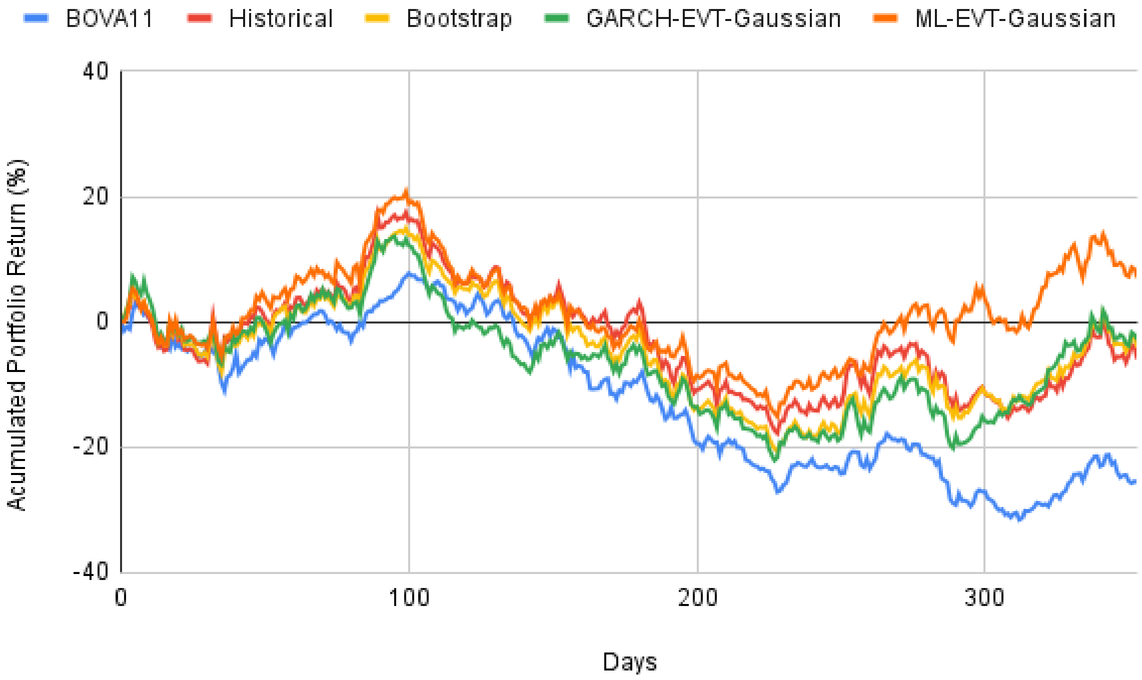


Figure 5.30: Comparison of accumulated portfolio return for different portfolio strategies for the entire period analyzed, days 1-353

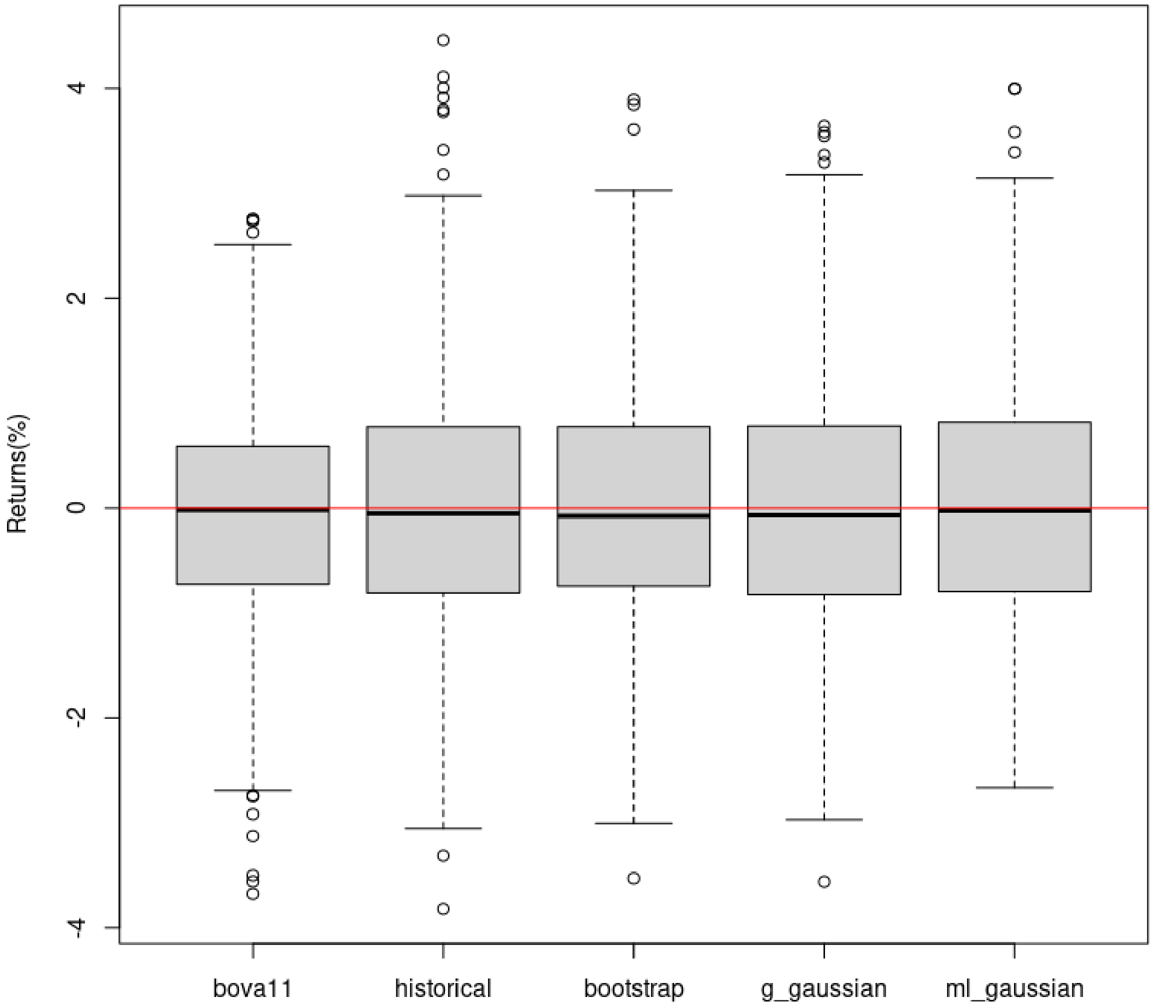


Figure 5.31: Comparison of return distributions for different portfolio strategies for the entire period analyzed, days 1-353

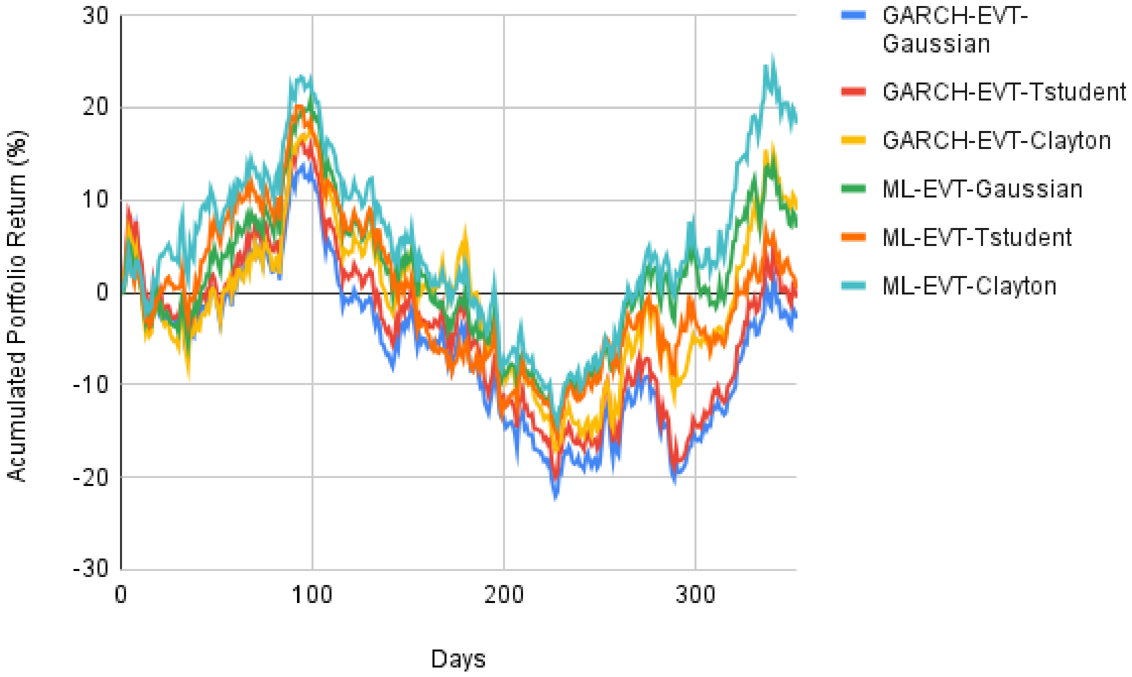


Figure 5.32: Comparison of accumulated portfolio return for different copula functions for the entire period analyzed, days 1-353

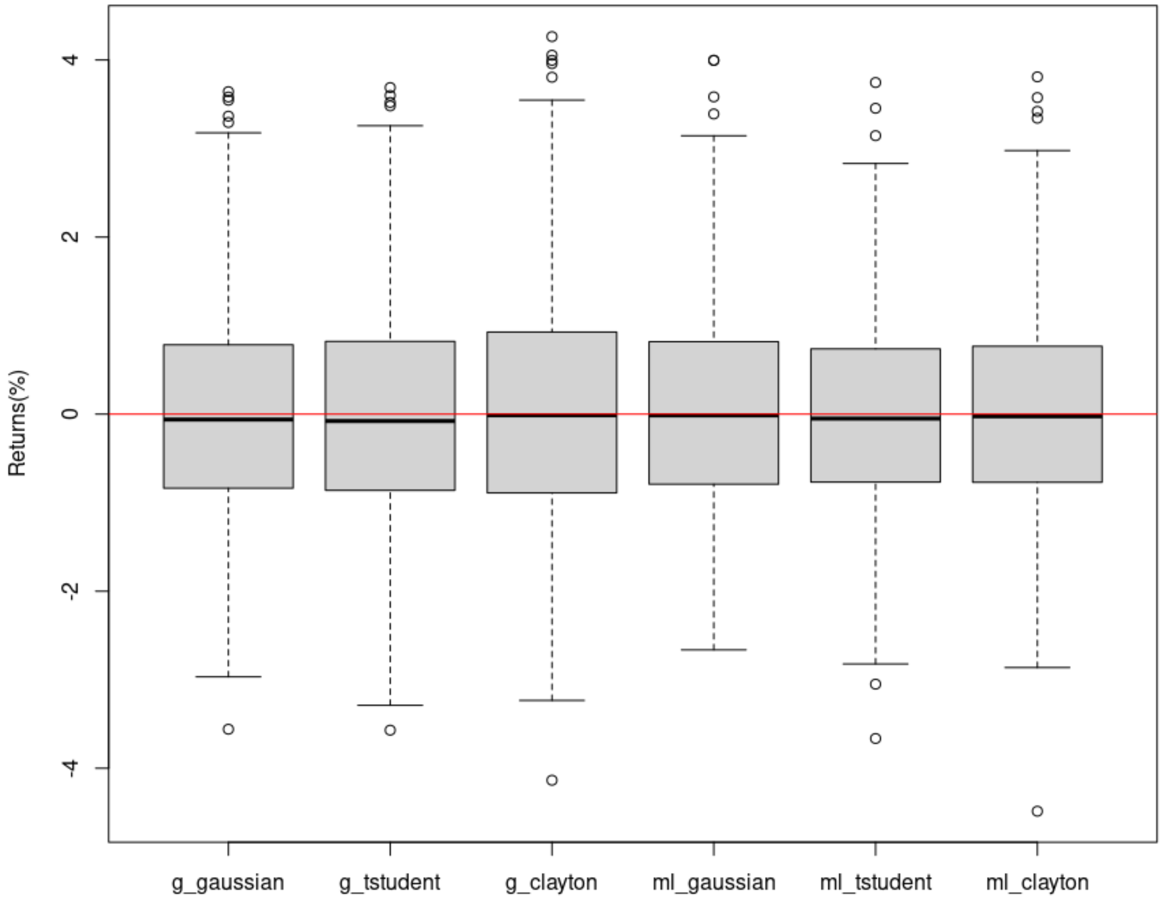


Figure 5.33: Comparison of return distributions for different copula functions for the entire period analyzed, days 1-353

Table 5.14: Comparison metrics of different portfolio strategies for the entire period analyzed, days 1-353

	Mean	**Std**	**Downside**	**CVaR**	**Sharpe**	**Sortino**	**Starr**	**Omega**
BOVA11	-0.08	1.12	0.86	2.61	-0.07	-0.09	-0.03	0.83
Historical	-0.01	1.34	0.90	2.48	-0.01	-0.01	-0.00	0.98
Bootstrap	-0.00	1.22	0.84	2.39	-0.00	-0.00	-0.00	0.99
G-Gaussian	-0.00	1.27	0.86	2.38	-0.00	-0.00	-0.00	1.00
G-Student	0.01	1.31	0.88	2.45	0.01	0.01	0.00	1.01
G-Clayton	0.03	1.35	0.89	2.50	0.02	0.04	0.01	1.06
ML-Gaussian	0.03	1.23	0.81	2.25	0.02	0.03	0.01	1.06
ML-Student	0.01	1.18	0.79	2.27	0.01	0.01	0.00	1.02
ML-Clayton	0.05	1.25	0.81	2.36	0.04	0.07	0.02	1.12

It is noticeable that, among all strategies, ML-EVT-Clayton stands out for the entire period, achieving the best values for Mean, Sharpe Ratio, Sortino Ratio, and Omega Ratio. Regarding risk measures, BOVA11 showed the lowest standard deviation, while ML-EVT-Tstudent and ML-EVT-Gaussian exhibited the lowest downside risk and CvaR values, respectively.

Table 5.15 displays the top 4 strategies for each comparison measure when considering the entire period. It is noteworthy that in all comparison measures, except for standard deviation, the three ML-EVT strategies are included in the top 4. This result highlights the financial advantage of using superior predictions compared to other strategies.

Table 5.15: Comparison of the top 4 strategies for each portfolio measure for the entire period (353 days). "ML-" stands for machine learning, while "G-" is the abbreviation for the GARCH model. The copula functions are abbreviated by the first letters G, T, C (Gaussian, t-student, Clayton).

Top	Mean	Std	Downside	CVaR	Sharpe	Sortino	Starr	Omega
1	ML-C	BOVA11	ML-T	ML-G	ML-C	ML-C	ML-C	ML-C
2	G-C	ML-T	ML-G	ML-T	G-C	G-C	G-C	G-C
3	ML-G	Bootstrap	ML-C	ML-C	ML-G	ML-G	ML-G	ML-G
4	ML-T	ML-G	Bootstrap	G-G	ML-T	ML-T	ML-T	ML-T

5.2.3.6 Operating with stop loss

In this section, we demonstrate how we can employ the proposed methodology in conjunction with the market operation stop-loss technique to enhance the financial outcomes achieved. The technique involves monitoring the returns obtained by the portfolio. If the cumulative negative return reaches a predefined threshold throughout the day, we sell the stocks at that moment without waiting until the end of the day.

Our experiments selected four threshold values for comparison with the same strategy without using stop loss: -4%, -3%, -2%, and -1%. We chose the ML-Clayton strategy to test the technique, as it yielded the best financial values in the previous section. In Figure 5.34, we present boxplot graphs of the strategy for different threshold values. Upon examining the distributions, it is evident that the negative extreme values decrease as we impose more stringent stop-loss limits.

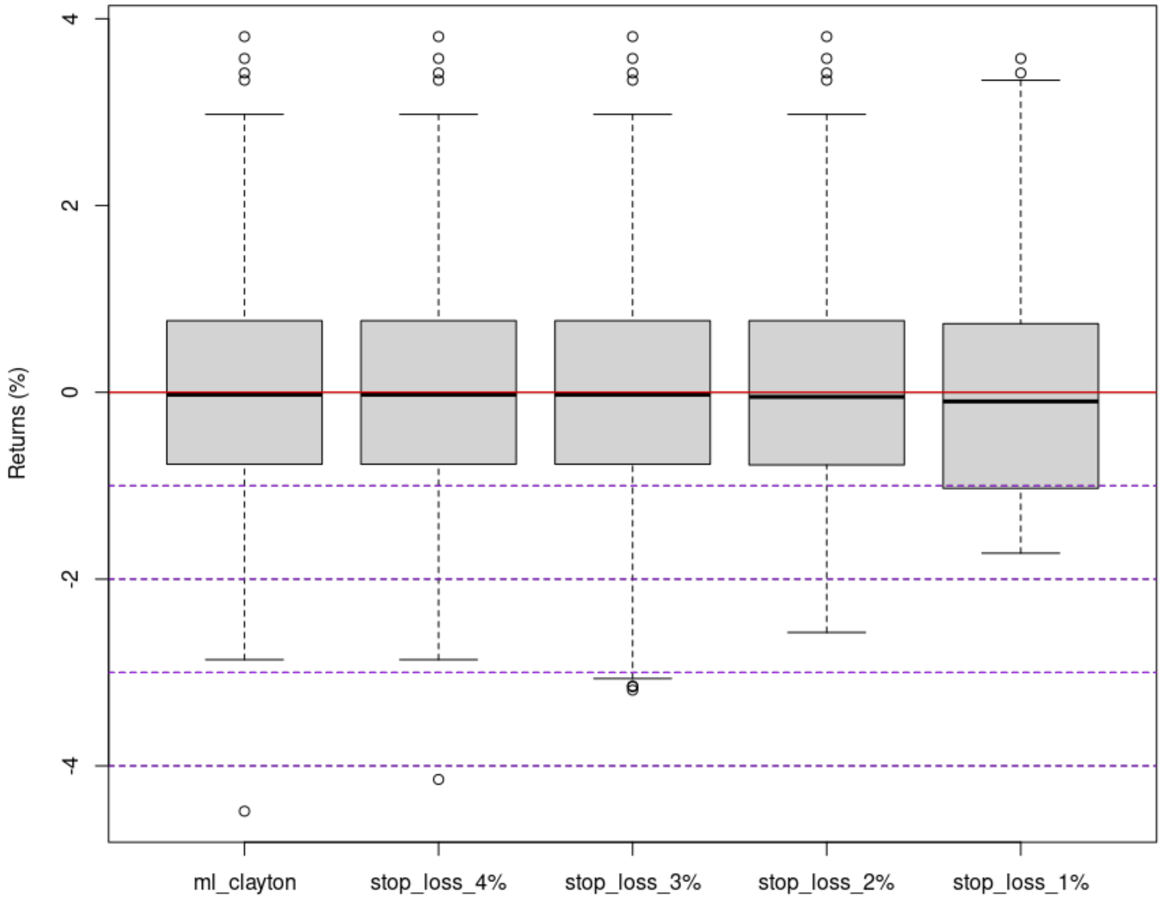


Figure 5.34: Comparison of return distributions for different stop loss thresholds for the entire period analyzed, days 1-353

Figure 5.35 illustrates the application of the technique with a threshold of −4% throughout the day 17/11/2021. As depicted in the Figure, the first negative return below the established limit is the largest compared to subsequent returns. Consequently, we observe that the strategy implementation did, in fact, ensure a less negative return, thereby avoiding a more extreme negative value.

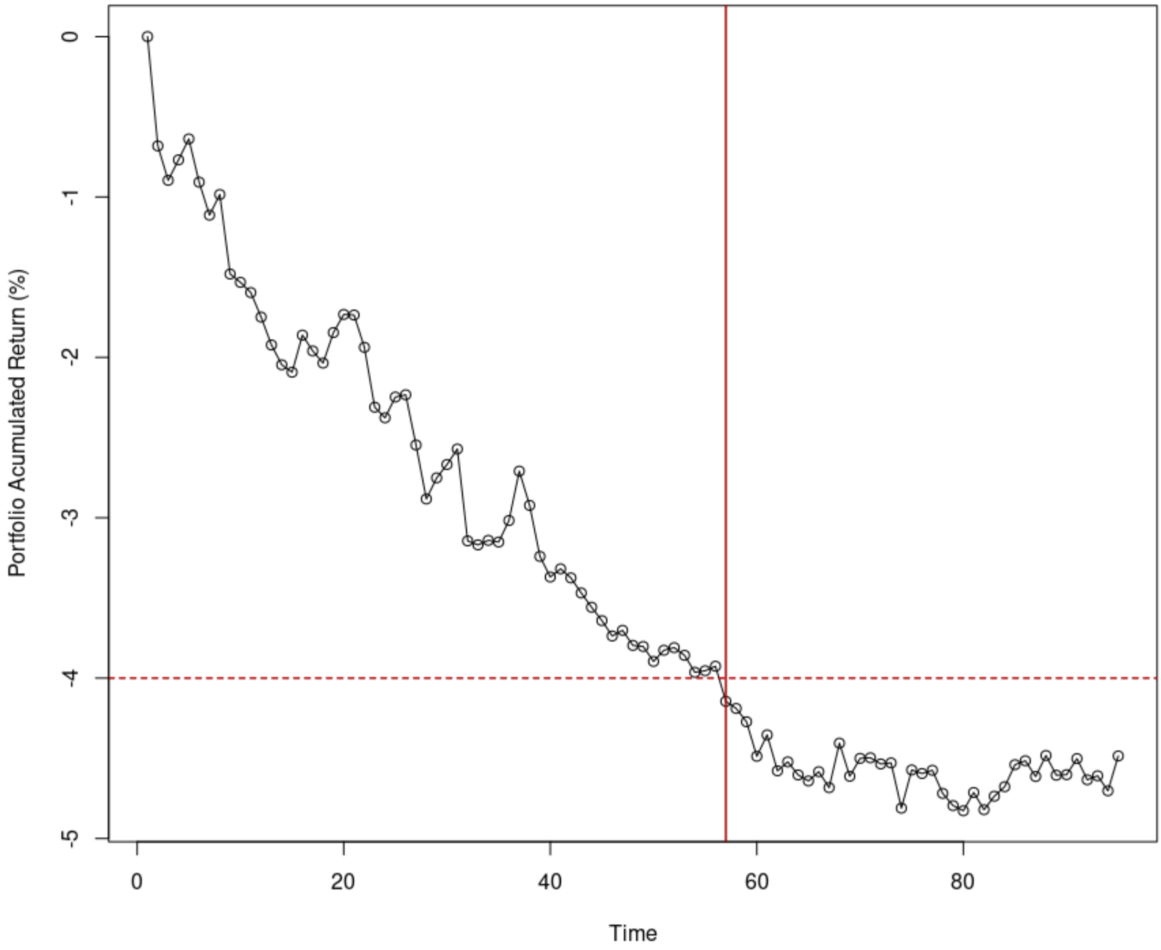


Figure 5.35: Example of applied stop loss of -4% for the day 17/11/2021. The horizontal line shows the stop loss threshold while the vertical line shows the return obtained by the use of the technique

However, using stop-loss does not guarantee that we will always incur more minor losses, as it is challenging to predict the optimal limit for each day and foresee whether the upcoming returns will continue the downward trend. Figure 5.36 illustrates the application of the technique with a threshold of -2% on day 11/10/2021. It is observed in the Figure that the first negative return below the established limit is not the most minor compared to subsequent returns. In this instance, we note an example where the technique adversely impacted the portfolio's return.

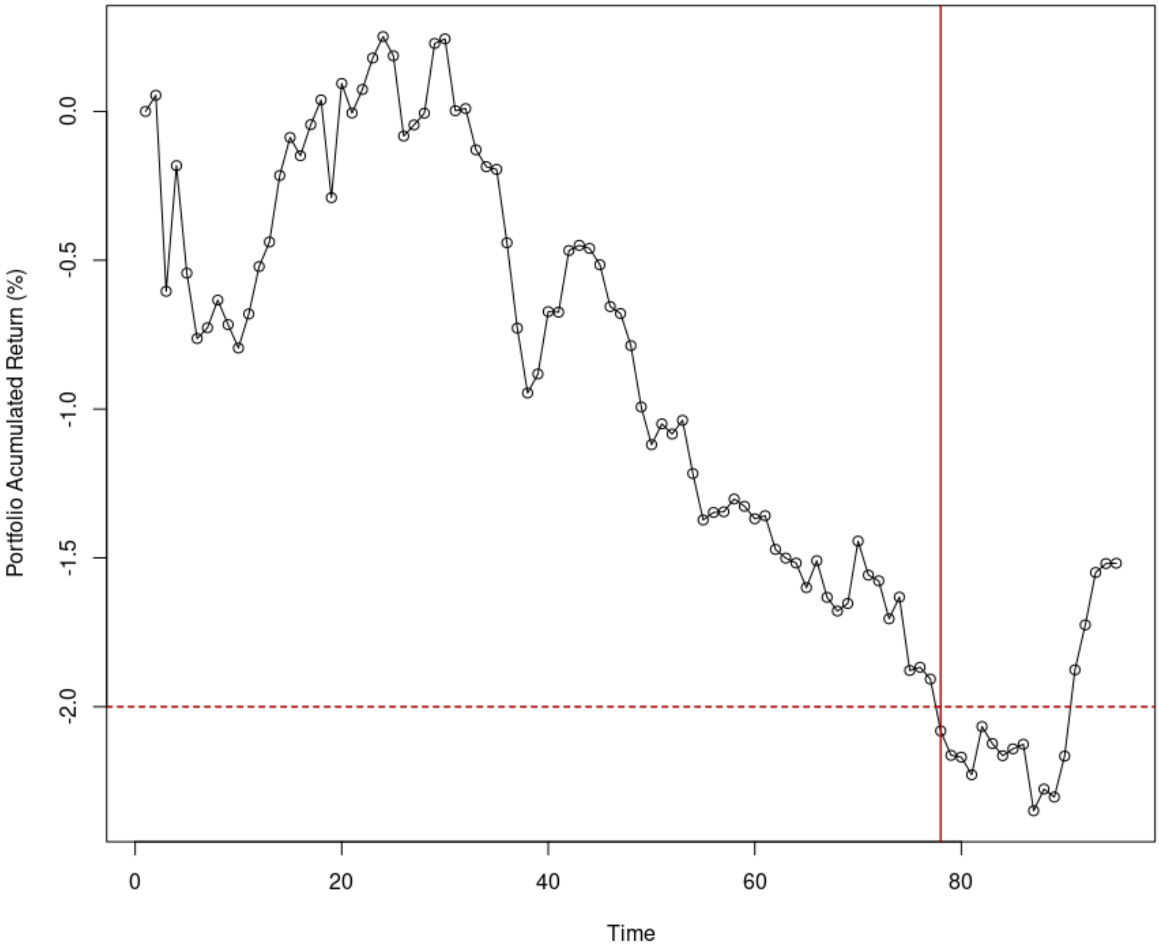


Figure 5.36: Example of applied stop loss of -2% for the day 11/10/2021. The horizontal line shows the stop loss threshold while the vertical line shows the return obtained by the use of the technique

Table 5.16 more precisely illustrates the challenge of selecting ideal stop loss values. By examining the frequency with which the technique enhanced returns and the total gain generated, it becomes evident that it was worthwhile only in the case of a stop loss set at -4%. For other stop loss values, the total of improved returns is less than 50% of the cases, resulting in negative gain values. Figure 5.37 compares using or not using a stop loss at -3%. In only 1 out of 5 cases, there was indeed an improvement in the outcome obtained.

Table 5.16: Stop loss analysis

	StopLoss -4%	StopLoss -3%	StopLoss -2%	StopLoss -1%
Number of Stops	1	5	30	127
Number of improvements	1	1	12	55
Gain	0.34%	-2.97%	-6.20%	-13.49%

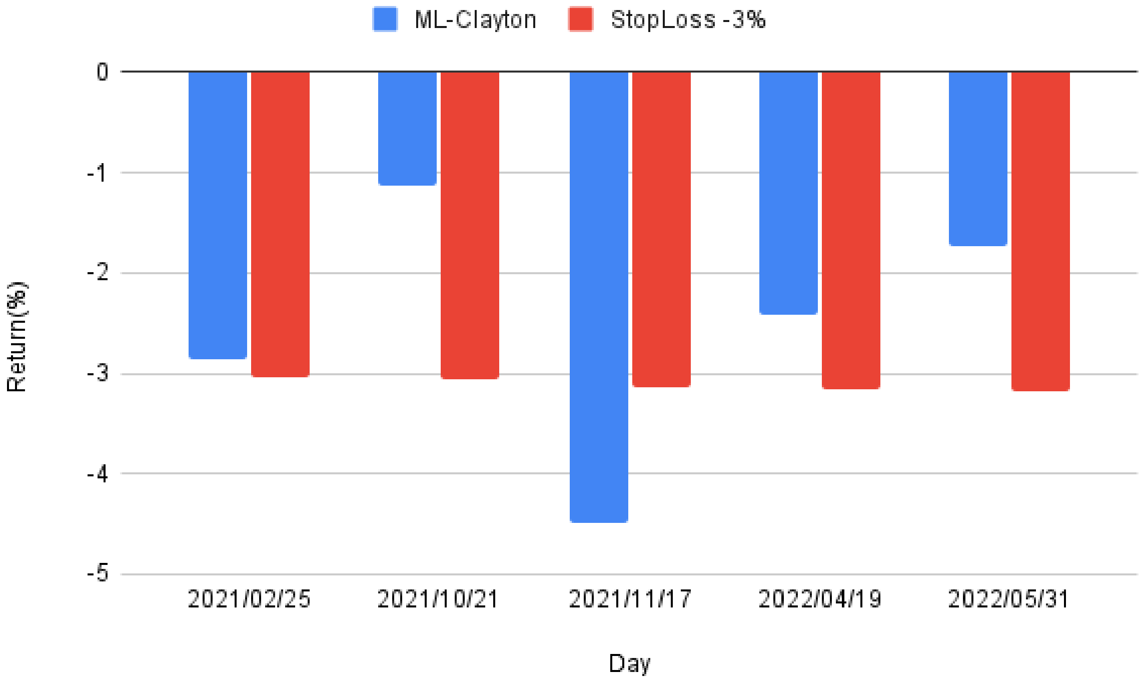


Figure 5.37: Portfolio return comparison using stop loss -3% against the pure strategie for different days

In Table 5.17, we compare the performance metrics of the pure ML-Clayton strategy with the same strategy combined with stop-loss. Upon examining the obtained averages, we notice that only in the case of a stop-loss of -4% did the average increase, while for the other stop-loss values, the average decreased. There are days when stop-loss yields a worse return than without its use. However, when analyzing the extreme negative values, the use of the technique tends to improve, as evidenced by the obtained Conditional Value at Risk (CVaR) values. As we prioritize a trade-off between risk and return, the best values were obtained for a stop-loss of -4%, yielding improved Sharpe ratio, Sortino ratio, Starr ratio, and Omega ratio values.

Table 5.17: Comparison metrics of different stop loss thresholds for the entire period analyzed, days 1-353

	ML Clayton	**StopLoss -4%**	**StopLoss -3%**	**StopLoss -2%**	**StopLoss -1%**
Mean	5.41E-04	5.50E-04	4.56E-04	3.64E-04	1.57E-04
Std	1.25E-02	1.24E-02	1.26E-02	1.25E-02	1.14E-02
Downside	8.09E-03	8.03E-03	8.26E-03	8.21E-03	6.89E-03
CVaR	2.36E-02	2.32E-02	2.38E-02	2.28E-02	1.93E-02
Sharpe	4.34E-02	4.43E-02	3.63E-02	2.90E-02	1.37E-02
Sortino	6.69E-02	6.85E-02	5.52E-02	4.43E-02	2.28E-02
Starr	2.29E-02	2.38E-02	1.91E-02	1.59E-02	8.16E-03
Omega	1.12E+00	1.12E+00	1.10E+00	1.08E+00	1.03E+00

Chapter 6

Conclusion

In this dissertation, we have developed a novel methodology that integrates machine learning algorithms, copula functions, intraday financial asset data, and extreme value theory distributions for the future generation of the multivariate probability distribution of financial assets. The fundamental hypothesis of this work posits that the utilization of a trained machine learning algorithm in conjunction with intraday asset data can produce more accurate predictions of asset volatility and correlation compared to classical approaches in the literature. Consequently, improved predictions can result in more precise future distributions, potentially leading to enhanced financial outcomes in terms of risk and return when employing portfolio optimization models.

In our experiments, we utilized historical real data from 2008 to 2022, with a 5-minute frequency, encompassing 29 stocks from the Brazilian stock exchange. We proposed a methodology to assess the results by comparing the predictions of the proposed algorithms with classical models from the volatility and correlation forecasting literature. This was achieved through the analysis of prediction errors using hypothesis tests. It was observed that, in the majority of cases, our approach yielded predictions with smaller errors. Furthermore, through our analyses, it was possible to validate and address the hypotheses raised in the study:

1. Asset volatility (relized variance) exhibits dynamic characteristics.

2. The correlation among assets is also dynamic, as evident from the analysis of the intraday Kendall correlation time series of assets. This observation suggests that adjusting the copula function daily based on Kendall correlation predictions is ideal for better modeling the correlation structure among assets.

3. Training a machine learning model properly, as described in the methodology along with the utilized features, can lead to more accurate predictions for both one-day-ahead realized variance and intraday Kendall correlation compared with the baselines.

4. The more accurate predictions of volatility and correlation had a positive impact on the portfolio optimization model, resulting in improved returns and risk for the trading strategy during the analyzed period,

Further, we conclude with the experiments that it is possible to combine the proposed methodology with market trading strategies, such as Stop Loss, to attempt to enhance financial gains. However, selecting the optimal threshold for this technique is challenging, and in our experiments, gains were only realized when setting the threshold to extremely high values. Small threshold values ended up hindering the achieved gains.

6.1 Future work

In future work, we aim to apply the same methodology with a larger dataset, drawn from previously published works in the literature, to obtain more robust results and validate whether the positive findings persist. Exploring additional data sources, including stocks from different markets, foreign exchange rates, commodities, and even cryptoassets, holds potential interest for further analysis. Additionally, we plan to extend the methodology to generate multivariate distributions of assets over a longer time horizon than one day ahead. It would be insightful to assess the feasibility of making predictions further into the future.

From a machine learning perspective, we intend to apply the state-of-the-art algorithms such as N-Beats [Oreshkin et al., 2020], N-Hits [Challu et al., 2023], and transformers [Wolf et al., 2020], as they represent the latest advancements in the field, capable of achieving remarkable results. Also proposing new features to enhance predictions is intriguing. Feature engineering techniques would be relevant in this context. Another promising analysis involves understanding which features or market periods contribute to better predictions. An approach in this regard would be to attempt using the SHAP (SHapley Additive exPlanations) method proposed by Lundberg and Lee [2017]. By better understanding the predictions, it may be possible to train better or even propose new algorithms to predict volatility and Kendall correlation between assets and evaluate how each feature influences the predictions.

Lastly, it is also compelling to extend the study from the perspective of copula functions. The Gaussian, t-student, and Clayton functions were analyzed in the current work using Kendall correlation to fit the functions. As mentioned in the paper, some parameters of certain copula functions may be time-dependent and cannot be calibrated via Kendall's correlation (for example the degrees of freedom from the Student t copula). In future work, we also intend to address this problem by trying to predict in some way these parameters. Also, it would be interesting to evaluate whether using Spearman correlation could yield better results. Testing other copula functions beyond those already mentioned would also be valuable.

Bibliography

C. Alexander. *Market risk analysis. Volume 2, Practical financial econometrics*. The Wiley Finance Series ; v.2. Wiley, Chichester, England, 1st edition edition, 2008. ISBN 1-282-34997-X.

C. Almeida and C. Czado. Efficient bayesian inference for stochastic time-varying copula models. *Computational Statistics & Data Analysis*, 56(6):1511–1527, 2012. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.08.015. URL https://www.sciencedirect.com/science/article/pii/S0167947311003148.

T. G. Andersen and T. Bollerslev. Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998. doi: 10.2307/2527343.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. The Distribution of Realized Exchange Rate Volatility. *Journal of the American Statistical Association*, 96():42–55, 2001. doi: 10.1198/016214501750332965.

T. G. Andersen, T. Bollerslev, and F. X. Diebold. Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4):701–720, 2007. doi: 10.1162/rest.89.4.701.

J. Arnerić, T. Poklepović, and J. W. Teai. Neural network approach in forecasting realized variance using high-frequency data. *Business Systems Research*, 9(2):18–34, 2018. doi: 10.2478/bsrj-2018-0016.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. doi: 10.1111/1467-9965.00068.

M. Bai and L. Sun. Application of copula and copula-CVaR in the multivariate portfolio optimization. In *Combinatorics, Algorithms, Probabilistic and Experimental Methodologies*, pages 231–242. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-74450-4_21.

T. B. Baillie, R.T. and H. Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74():3–30, 1996. doi: 10.1016/S0304-4076(95)01749-6.

A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974. ISSN 00911798. doi: 10.1214/aop/1176996548. URL http://www.jstor.org/stable/2959306.

G. H. Bauer and K. Vorkink. Forecasting multivariate realized stock market volatility. *Journal of Econometrics*, 160():93–101, 2011. doi: 10.1016/j.jeconom.2010.03.021.

L. Bauwens, S. Laurent, and J. V. K. Rombouts. Multivariate garch models: a survey. *Journal of Applied Econometrics*, 21(1):79–109, 2006. doi: https://doi.org/10.1002/jae.842. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.842.

F. Björnsjö. CAN DEEP LEARNING BEAT TRADITIONAL ECONOMETRICS IN FORECASTING OF REALIZED VOLATILITY?,. *Uppsala University, Department of Statistics*, ():, 2020.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. doi: 10.1016/0304-4076(86)90063-1.

T. Bollerslev, A. J. Patton, and R. Quaedvlieg. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18, 2016. doi: 10.1016/j.jeconom.2015.10.007.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

A. Bucci. Realized Volatility Forecasting with Neural Networks. *Journal of Financial Econometrics*, 18(3):502–531, 2020a. doi: 10.1093/jjfinec/nbaa008.

A. Bucci. Realized Volatility Forecasting with Neural Networks. *Journal of Financial Econometrics*, 18(3):502–531, 2020b. doi: 10.1093/jjfinec/nbaa008.

J. F. Caldeira, G. V. Moura, M. S. Perlin, and A. A. Santos. Portfolio management using realized covariances: Evidence from Brazil. *EconomiA*, 18():328–343, 2017. doi: 10.1016/j.econ.2017.04.002.

M. C. Cario and B. L. Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Citeseer, 1997.

L. Chakkalakal, U. Hommel, and W. Li. Transport infrastructure equities in mixed-asset portfolios: estimating risk with a Garch-Copula CVaR model. *Journal of Property Research*, 35(2):117–138, 2018. doi: 10.1080/09599916.2018.1461126.

C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza Ramirez, M. Mergenthaler Canseco, and A. Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, Jun.

2023. doi: 10.1609/aaai.v37i6.25854. URL https://ojs.aaai.org/index.php/AAAI/article/view/25854.

C.-L. Chang, J.-Á. Jiménez-Martín, E. Maasoumi, M. McAleer, and T. Pérez-Amaral. Choosing expected shortfall over VaR in Basel III using stochastic dominance. *International Review of Economics & Finance*, 60:95–113, 2019. doi: doi.org/10.2991/icefs-17.2017.11.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

X. Chen and Y. Fan. Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics*, 135(1):125–154, 2006. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2005.07.027. URL https://www.sciencedirect.com/science/article/pii/S0304407605001776.

R. Chiriac and V. Voev. Modelling and forecasting multivariate realized volatility. *Journal of Applied Econometrics*, 26():922–947, 2011. doi: 10.1002/jae.1152.

K. Christensen, M. Siggaard, and B. Veliyev. A Machine Learning Approach to Volatility Forecasting*. *Journal of Financial Econometrics*, 21(5):1680–1727, 06 2022. ISSN 1479-8409. doi: 10.1093/jjfinec/nbac020. URL https://doi.org/10.1093/jjfinec/nbac020.

G. A. Christodoulakis and S. E. Satchell. Correlated arch (corrarch): Modelling the time-varying conditional correlation between financial asset returns. *European Journal of Operational Research*, 139(2):351–370, 2002. ISSN 0377-2217. doi: 10.1016/S0377-2217(01)00361-7. EURO XVI: O.R. for Innovation and Quality of Life.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001. doi: 10.1080/713665670.

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009. doi: 10.1093/jjfinec/nbp001.

D. Creal, S. J. Koopman, and A. Lucas. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795, 2013. doi: https://doi.org/10.1002/jae.1279. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1279.

P. Date, R. Mamon, and L. Jalen. A new moment matching algorithm for sampling from partially specified symmetric distributions. *Operations Research Letters*, 36(6):669–672, 2008. doi: 10.1016/j.orl.2008.07.004.

I. De Lira Salvatierra and A. J. Patton. Dynamic copula models and high frequency data. *Journal of Empirical Finance*, 30:120–135, 2015. ISSN 0927-5398. doi: https://doi.org/10.1016/j.jempfin.2014.11.008. URL https://www.sciencedirect.com/science/article/pii/S0927539814001169.

S. Demarta and A. J. McNeil. The t Copula and Related Copulas. *International Statistical Review*, 73(1):111–129, 2005. doi: 10.1111/j.1751-5823.2005.tb00254.x.

L. Deng, C. Ma, and W. Yang. Portfolio optimization via pair copula-garch-evt-cvar model. *Systems Engineering Procedia*, 2:171–181, 2011. ISSN 2211-3819. doi: https://doi.org/10.1016/j.sepro.2011.10.020. URL https://www.sciencedirect.com/science/article/pii/S2211381911001093. Complexity System and Engineering Management.

A. Dias, P. Embrechts, et al. Dynamic copula models for multivariate high-frequency data in finance. *Manuscript, ETH Zurich*, 81:1–42, 2004.

F. Diebold and R. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63, 1995. doi: https://doi.org/10.1080/07350015.1995.10524599. URL https://EconPapers.repec.org/RePEc:bes:jnlbes:v:13:y:1995:i:3:p:253-63.

R. Donaldson and M. Kamstra. An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance*, 4():17–46, 1997. doi: 10.1016/S0927-5398(96)00011-4.

P. Embrechts, A. J. McNeil, and D. Straumann. *Correlation and Dependence in Risk Management: Properties and Pitfalls*, page 176–223. Cambridge University Press, 2002. doi: 10.1017/CBO9780511615337.008.

R. Engle. Dynamic conditional correlation. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002. doi: 10.1198/073500102288618487.

R. F. Engle. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987–1007, 1982. doi: 10.2307/1912773.

W. H. Enzo Giacomini and V. Spokoiny. Inhomogeneous dependence modeling with time-varying copulae. *Journal of Business & Economic Statistics*, 27(2):224–234, 2009. doi: 10.1198/jbes.2009.0016. URL https://doi.org/10.1198/jbes.2009.0016.

M. R. Fengler and O. Okhrin. Managing risk with a realized copula parameter. *Computational Statistics & Data Analysis*, 100:131–152, 2016. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2014.07.011. URL https://www.sciencedirect.com/science/article/pii/S0167947314002151.

J. Fleming, K. C., and O. B. The Economic Value of Volatility Timing. *Journal of Finance*, 56():329–352, 1999. doi: 10.1111/0022-1082.00327.

J. Fleming, K. C., and O. B. The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, 67():473–509, 2003. doi: 10.1016/S0304-405X(02)00259-3.

A. Goel, A. Sharma, and A. Mehra. Robust optimization of mixed cvar starr ratio using copulas. *Journal of Computational and Applied Mathematics*, 347:62–83, 2019. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2018.08.001. URL https://www.sciencedirect.com/science/article/pii/S0377042718304746.

G. Guastaroba, R. Mansini, and M. G. Speranza. On the effectiveness of scenario generation techniques in single-period portfolio optimization. *European Journal of Operational Research*, 192(2):500–511, 2009. doi: https://doi.org/10.1016/j.ejor.2007.09.042.

T. Guldimann. The story of RiskMetrics. *Risk*, 13(1):56–58, 2000.

E. S. Gunnarsson, H. R. Isern, A. Kaloudis, M. Risstad, B. Vigdel, and S. Westgaard. Prediction of realized volatility and implied volatility indices using ai and machine learning: A review. *International Review of Financial Analysis*, 93:103221, 2024. ISSN 1057-5219. doi: https://doi.org/10.1016/j.irfa.2024.103221. URL https://www.sciencedirect.com/science/article/pii/S1057521924001534.

C. M. Hafner and H. Manner. Dynamic stochastic copula models: estimation, inference and applications. *Journal of Applied Econometrics*, 27(2):269–295, 2012. doi: https://doi.org/10.1002/jae.1197. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1197.

K. Høyland and S. W. Wallace. Generating scenario trees for multistage decision problems. *Management science*, 47(2):295–307, 2001. doi: 10.1287/mnsc.47.2.295.9834.

K. Høyland, M. Kaut, and S. W. Wallace. A heuristic for moment-matching scenario generation. *Computational Optimization and Applications*, 24(2):169–185, 2003. doi: 10.1023/A:1021853807313.

C.-C. Hsu, C.-P. Tseng, and Y.-H. Wang. Dynamic hedging with futures: A copula-based garch model. *Journal of Futures Markets*, 28(11):1095–1116, 2008. doi:

https://doi.org/10.1002/fut.20345. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/fut.20345.

J.-J. Huang, K.-J. Lee, H. Liang, and W.-F. Lin. Estimating value at risk of portfolio by conditional copula-garch method. *Insurance: Mathematics and Economics*, 45(3):315–324, 2009. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2009.09.009. URL https://www.sciencedirect.com/science/article/pii/S0167668709001267.

W. K. Härdle, O. Okhrin, and Y. Okhrin. Dynamic structured copula models. *Statistics & Risk Modeling*, 30(4):361–388, 2013. doi: doi:10.1524/strm.2013.2004. URL https://doi.org/10.1524/strm.2013.2004.

M. Izzeldin, M. K. Hassan, V. Pappas, and M. Tsionas. Forecasting realised volatility using ARFIMA and HAR models. *Quantitative Finance*, 19(10):1627–1638, 2019. doi: 10.1080/14697688.2019.1600713.

X. Jin. Large portfolio risk management with dynamic copulas. *SSRN Electronic Journal*, 12 2009. doi: 10.2139/ssrn.1483296.

H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. 1997. doi: https://doi.org/10.1201/9780367803896.

E. Jondeau and M. Rockinger. The copula-garch model of conditional dependencies: An international stock market application. *Journal of International Money and Finance*, 25(5):827–853, 2006. ISSN 0261-5606. doi: https://doi.org/10.1016/j.jimonfin.2006.04.007. URL https://www.sciencedirect.com/science/article/pii/S0261560606000350.

M. Karmakar. Dependence structure and portfolio risk in indian foreign exchange market: A garch-evt-copula approach. *The Quarterly Review of Economics and Finance*, 64:275–291, 2017. ISSN 1062-9769. doi: https://doi.org/10.1016/j.qref.2017.01.007. URL https://www.sciencedirect.com/science/article/pii/S106297691730025X.

M. Kaut and S. W. Wallace. *Evaluation of scenario-generation methods for stochastic programming*. Humboldt-Universität zu Berlin, Mathematisch - Naturwissenschaftliche Fakultät II, Institut für Mathematik, 2003. doi: 10.18452/8296.

H. Y. Kim and C. H. Won. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type modelss. *Expert Systems with Applications*, 103():25–37, 2018. doi: 10.1016/j.eswa.2018.03.002.

I. Kojadinovic and J. Yan. Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and*

Economics, 47(1):52–63, 2010. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2010.03.008. URL https://www.sciencedirect.com/science/article/pii/S0167668710000363.

L. Koliai. Extreme risk modeling: An evt–pair-copulas approach for financial stress tests. *Journal of Banking & Finance*, 70:1–22, 2016. ISSN 0378-4266. doi: https://doi.org/10.1016/j.jbankfin.2016.02.004. URL https://www.sciencedirect.com/science/article/pii/S037842661600042X.

H. Konno and H. Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. *Management Science*, 37(5):519–531, 1991. doi: 10.1287/mnsc.37.5.519.

R. Kouwenberg. Scenario generation and stochastic programming models for asset liability management. *European Journal of operational research*, 134(2):279–292, 2001. doi: 10.1016/S0377-2217(00)00261-7.

W. Kristjanpoller and M. C. Minutolo. Volatility forecast using hybrid Neural Network models. *Expert Systems with Applications*, 41():2437–2442, 2014. doi: 10.1016/j.eswa.2013.09.043.

W. Kristjanpoller and M. C. Minutolo. A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. *Expert Systems with Applications*, 109():1–11, 2018. doi: 10.1016/j.eswa.2018.05.011.

E. LeDell and S. Poirier. H2o automl: Scalable automatic machine learning. 2020. URL https://api.semanticscholar.org/CorpusID:221338558.

Y. Liu. Novel volatility forecasting using deep learning–Long Short Term Memory Recurrent Neural Networks. *'Expert Systems with Applications*, 132():99–109, 2019. doi: 10.1016/j.eswa.2019.04.038.

Y. Liu and R. Luger. Efficient estimation of copula-garch models. *Computational Statistics & Data Analysis*, 53(6):2284–2297, 2009. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2008.01.018. URL https://www.sciencedirect.com/science/article/pii/S0167947308000182. The Fourth Special Issue on Computational Econometrics.

F. M. Longin. *Extreme Value Theory: An Introductory Overview*, page 53–53. Wiley, 2017. ISBN 9781118650196. doi: 10.1002/9781118650318.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing*

Systems, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

P. M. Lurie and M. S. Goldberg. An approximate method for sampling correlated random variables from partially-specified distributions. *Management science*, 44(2):203–218, 1998. doi: 10.1287/mnsc.44.2.203.

L. Maciel, R. Ballini, and F. Gomide. Evolving Possibilistic Fuzzy Modeling for Realized Volatility Forecasting With Jumps. *IEEE Transactions on Fuzzy Systems*, 25():302–314, 2017. doi: 10.1109/TFUZZ.2016.2578338.

R. Mansini, W. Ogryczak, and M. G. Speranza. Twenty years of linear programming based portfolio optimization. *European Journal of Operational Research*, 234(2):518–535, 2014. doi: 10.1016/j.ejor.2013.08.035.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi: 10.1111/j.1540-6261.1952.tb01525.x.

H. M. Markowitz. *Portfolio selection: efficient diversification of investments*. Yale University Press, 1959. ISBN 9780300013726.

R. D. Martin, S. Z. Rachev, and F. Siboulet. *Phi-alpha optimal portfolios and extreme risk management*, chapter 17, pages 223–248. Wiley, 2003. ISBN 9780470023518.

F. J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. doi: 10.1080/01621459.1951.10500769.

M. P. Massimo Guidolin. *Essentials of Time Series for Financial Applications*. Academic Press, 1st edition edition, 2018. ISBN 9780128134092.

C. M. Mesquita, C. A. Valle, and A. C. M. Pereira. Dynamic portfolio optimization using a hybrid mlp-har approach. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1075–1082, 2020. doi: 10.1109/SSCI47803.2020.9308249.

C. M. Mesquita, C. A. Valle, and A. C. M. Pereira. Scenario generation for financial data with a machine learning approach based on realized volatility and copulas. *Computational Economics*, 2023. ISSN 1572-9974. doi: https://doi.org/10.1007/s10614-023-10387-2. URL https://link.springer.com/article/10.1007/s10614-023-10387-2#citeas.

S. B. Messaoud and C. Aloui. Measuring risk of portfolio : GARCH-Copula model. *Journal of Economic Integration*, 30(1):172–205, 2015. doi: 10.11130/jei.2015.30.1.172.

E. Messina and D. Toscani. Hidden Markov models for scenario generation. *IMA Journal of Management Mathematics*, 19(4):379–401, 2008. doi: 10.1093/imaman/dpm026.

S. A. Monfared and D. Enke. Volatility Forecasting Using a Hybrid GJR-GARCH Neural Network Model. *Procedia Computer Science*, 36():246–253, 2014. doi: 10.1016/j.procs.2014.09.087.

R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, NY, USA, second edition, 2006.

D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991. doi: 10.2307/2938260.

B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting, 2020.

A. J. Patton. On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation. *Journal of Financial Econometrics*, 2(1):130–168, 01 2004. ISSN 1479-8409. doi: 10.1093/jjfinec/nbh006. URL https://doi.org/10.1093/jjfinec/nbh006.

A. J. Patton. Modelling asymmetric exchange rate dependence. *International economic review*, 47(2):527–556, 2006. doi: https://doi.org/10.1111/j.1468-2354.2006.00387.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2354.2006.00387.x.

J. Pickands. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119 – 131, 1975. doi: 10.1214/aos/1176343003. URL https://doi.org/10.1214/aos/1176343003.

S. Pong, M. B. Shackleton, S. J. Taylor, and X. Xu. Forecasting currency volatility: A comparison of implied volatilities and AR(FI)MA models. *Journal of Banking & Finance*, 28():2541–2563, 2004. doi: 10.1016/j.jbankfin.2003.10.015.

K. Ponomareva, D. Roman, and P. Date. An algorithm for moment-matching scenario generation with application to financial portfolio optimisation. *European Journal of Operational Research*, 240(3):678–687, 2015. doi: 10.1016/j.ejor.2014.07.049.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038.

T. H. Roh. Forecasting the volatility of stock price index. *Expert Systems with Applications*, 33():916–922, 2007. doi: 10.1016/j.eswa.2006.08.001.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL http://dx.doi.org/10.1037/h0042519.

P. J. Rousseeuw and G. Molenberghs. Transformation of non positive semidefinite correlation matrices. *Communications in Statistics - Theory and Methods*, 22(4):965–984, 1993. doi: 10.1080/03610928308831068.

M. Sahamkhadam, A. Stephan, and R. Östermark. Portfolio optimization based on GARCH-EVT-Copula forecasting models. *International Journal of Forecasting*, 34(3): 497–506, 2018. doi: 10.1016/j.ijforecast.2018.02.004.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *Revstat Statistical Journal*, 10:33–60, 03 2012. doi: 10.57805/revstat.v10i1.110.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

J. E. Smith. Moment methods for decision analysis. *Management science*, 39(3):340–358, 1993. doi: 10.1287/mnsc.39.3.340.

M. K. So and C. Y. Yeung. Vine-copula garch model with dynamic conditional dependence. *Computational Statistics & Data Analysis*, 76:655–671, 2014. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2013.08.008. URL https://www.sciencedirect.com/science/article/pii/S0167947313002958. CFEnetwork: The Annals of Computational and Financial Econometrics.

H. G. Souto and A. Moradi. Introducing nbeatsx to realized volatility forecasting. *Expert Systems with Applications*, 242:122802, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.122802. URL https://www.sciencedirect.com/science/article/pii/S0957417423033043.

B. T., P. A. J., and Q. R. Modeling and forecasting (un)reliable realized covariances for more reliable financial decisions. *Journal of Econometrics*, 207():71–91, 2018. doi: 10.1016/j.jeconom.2018.05.004.

S. J. Taylor. *Asset Price Dynamics, Volatility, and Prediction.* . Princeton University Press, , edition, 2007. ISBN 9780691134796.

Y. K. Tse and A. K. Tsui. A multivariate garch model with time-varying correlations. *Journal of Business and Economic Statistics*, 20:351–362, 2002. doi: 10.2139/ssrn.250228.

C. D. Vale and V. A. Maurelli. Simulating multivariate nonnormal distributions. *Psychometrika*, 48(3):465–471, 1983. doi: 10.1007/BF02293687.

D. I. Vortelinos. Forecasting realized volatility: HAR against Principal Components Combining, neural networks and GARCH. *Research in International Business and Finance*, 39:824–839, 2017. doi: 10.1016/j.ribaf.2015.01.004.

Z.-R. Wang, X.-H. Chen, Y.-B. Jin, and Y.-J. Zhou. Estimating risk of foreign exchange portfolio: Using VaR and CVaR based on GARCH–EVT-Copula model. *Physica A: Statistical Mechanics and its Applications*, 389(21):4918–4928, 2010. doi: 10.1016/j.physa.2010.07.012.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

S. Yitzhaki. Stochastic dominance, mean variance, and Gini's mean difference. *American Economic Review*, 72(1):178–185, 1982. doi: 10.2307/1808584.

C. Zhang, Y. Zhang, M. Cucuringu, and Z. Qian. Volatility Forecasting with Machine Learning and Intraday Commonality*. *Journal of Financial Econometrics*, 22(2):492–530, 03 2023. ISSN 1479-8409. doi: 10.1093/jjfinec/nbad005. URL https://doi.org/10.1093/jjfinec/nbad005.

O. Çepni, R. Gupta, D. Pienaar, and C. Pierdzioch. Forecasting the realized variance of oil-price returns using machine learning: Is there a role for u.s. state-level uncertainty? *Energy Economics*, 114:106229, 2022. ISSN 0140-9883. doi: https://doi.org/10.1016/j.eneco.2022.106229. URL https://www.sciencedirect.com/science/article/pii/S0140988322003723.