LONG-TERM ASSET ALLOCATION BASED ON STOCHASTIC MULTISTAGE
MULTI-OBJECTIVE PORTFOLIO OPTIMIZATION

Guido Chagas  
EESP – FGV and Banco Itaú Unibanco

Pedro L. Valls Pereira  
EESP – FGV and CEQEF –FGV

ABSTRACT

Multi-Period Stochastic Programming (MSO) offers an appealing approach to identity optimal portfolios, particularly over longer investment horizons, because it is inherently suited to handle uncertainty. Moreover, it provides flexibility to accommodate coherent risk measures, market frictions, and most importantly, major stylized facts as volatility clustering, heavy tails, leverage effects and tail co-dependence. However, to achieve satisfactory results a MSO model relies on representative and arbitrage-free scenarios of the pertaining multivariate financial series. Only after we have constructed such scenarios, we can exploit it using suitable risk measures to achieve robust portfolio allocations. In this thesis, we discuss a comprehensive framework to accomplish that. First, we construct joint scenarios based on a combined GJR-GARCH + EVT-GPD + t-Copula approach. Then, we reduce the original scenario tree and remove arbitrage opportunities using a method based on Optimal Discretization and Process Distances. optimization taking into account market frictions such as transaction costs and liquidity restrictions. The proposed framework is particularly valuable to real applications because it handles various key features of real markets that are often dismissed by more common optimization approaches. Lastly, using the approximated scenario tree we perform a multi-period Mean-Variance-CVaR optimization taking into account market frictions such as transaction costs and liquidity restrictions. The proposed framework is particularly valuable to real applications because it handles various key features of real markets that are often dismissed by more common optimization approaches.

Keywords: Multiperiod Asset Allocation, Scenario Tree Generation and Reduction, Optimal Discretization, Mean-Variance-CVaR Optimization

Jel Code: G10, G11, G17
1 Introduction

Most, if not all, activities in financial markets explore risk-adjusted returns. They look for investment opportunities that offer attractive returns within acceptable risk limits. However, this seemingly simple question entails great challenges. First, we need to choose a proper way to weigh risks. But what is a good risk measure? Should we consider just one measure when comparing opportunities? Secondly, we seldom have exact information regarding the future performance of potential investments. In fact, we must take into account data uncertainty to make optimal decisions. Lastly, real markets are not perfect. They have restrictions, costs and limitations, known as market frictions, that often affect our decisions.

Furthermore, when considering a portfolio of such investment opportunities, we face an additional challenge: their performance is usually mutually dependent. Therefore, to find an optimal portfolio of investment opportunities (or, at least, one that is not dominated by any other), we also need to understand their dependence structure and use it to our advantage.

These insights suggest that, to perform a realistic portfolio optimization, we need a model that considers the (actual) individual and joint behaviour of the opportunities’ expected returns, takes into account suitable risk measures and evaluates the impacts of market frictions. This thesis proposes a integrated framework to accomplish that.

We start by modelling the individual and joint behaviour of the investment opportunities. These opportunities usually involve instruments whose series of returns (i) usually have particular statistical properties, (ii) are seldom independent and identically distributed and (iii) often display a nonlinear jointly behaviour. The more accurate our ability to explain these empirical features, the better we can appraise and allocate such opportunities.

These features, which are persistent across different times, markets and instruments, are known in the literature as stylized facts. Because of their importance, many empirical financial studies have suggested models to explain and, more interestingly, simulate multivariate financial time series that are capable of accommodating such key (individual and joint) empirical statistical facts. In fact, some
essential financial activities as risk management and portfolio optimization depend on elaborated models to account properly for these stylized facts in order to provide reliable conclusions.

Unfortunately, as we discuss later there are many stylized facts and, to the best of our knowledge, most models are capable of explaining simultaneously, at most, three of them. For instance, McNeil and Frey [213] suggested a combined approach using GARCH models and EVT to model extreme stock returns. Rockinger and Jondeau [268] proposed a different approach employing GARCH models and copulas to fit multivariate stock returns. Recent works, like Huang, et al. [149] and Wang et al. [310], combine these previous studies into a GARCH-EVT-Copula approach to study exchange rate markets.

In this thesis, we follow a similar framework and propose an unified approach to model four of the most significant facts: heavy tails, volatility clustering, leverage effects and tail co-dependence, which are particularly relevant to daily financial series (e.g., daily log returns).

Heavy tailed distributions are distributions whose tails exhibit seemingly polynomial decay. That means that extreme (absolute) returns are more common that those predicted by thin-tailed distributions, like the popular Gaussian distribution. For obvious reasons, the probability of occurrence of extreme observations is essential information for reliable risk management. Unfortunately, conventional econometric approaches are not suited to handle extreme returns because they are relatively rare. To do so, we need an alternative such as EVT discussed in the later sections.

Volatility clustering is another key property. Many statistical tools and techniques, like EVT for instance, rely on the assumption that series are independent and identically distribution (i.i.d.). Although daily raw returns do not show evidence of significant serial correlation, daily absolute or squared returns clearly exhibit strong serial correlation. In other words, large (absolute) returns are often followed by large returns while small (absolute) returns are typically followed by small returns, creating clusters of volatility. This empirical fact plainly invalidates the assumption of i.i.d. series. Conveniently, GARCH or SV models allow us to handle these volatility clusters and use (approximately) i.i.d filtered standardized residuals to perform subsequent analyses and modelling.
Leverage effects are also particularly relevant because they explain how “bad news” (i.e., negative shocks) have greater impact in volatility than “good news” (i.e., positive shocks). This behaviour can be accommodated using, for instance, asymmetric extensions of the GARCH models such as EGARCH, GJR-GARCH or TGARCH. These models are rather popular because they can reasonably explain the volatility clustering, the leverage effects and, to some extent, the heavy tails observed in financial daily series.

Lastly, tail co-dependence is likely one of the most important stylized facts because it describes how dependence among instruments and markets are stronger (in absolute terms) during turbulent periods. It indicates how cross-correlations change during these periods leading to increasing risk (or, similarly, decreasing diversification). Although some multivariate GARCH models like DCC and TVC are able to describe partially the joint behaviour of financial series, copulas offer a more flexible alternative because they detach the marginal modelling from the dependence structure modelling.

Once we have modelled these stylized facts, we use the calibrated model to simulate a sufficiently large number of paths and construct a discrete scenario tree that appropriately describes the expected behaviour of the daily log returns. Compared to other common scenario construction approaches, the method suggested in this thesis provides a better fit to the individual and joint dynamics of the financial series. However, the generated scenario tree is neither free of arbitrage opportunities nor tractable in real multi-period stochastic portfolio optimizations. Therefore, our next step is to construct a smaller but sufficiently similar (in some proper sense discussed later) scenario tree that is both manageable and arbitrage-free. To accomplish that, we extend the optimal discretization method based on process distances, proposed by Pflug [243] and further developed by Pflug and Pichler [245], [246] and Kovacevic and Pichler [186].

The process distance is a generalization of the Wasserstein distance. While the latter is a distance function between probability distributions, the former is a distance functional between stochastic processes. Recalling that scenario trees are simply filtered discrete probability spaces that represent the stochastic process model of the simulated data, we can use process distances to evaluate how similar (i.e., how close) two scenario trees are. Moreover, we can combine them with discretization
methods to approximate the original scenario trees by smaller, more manageable scenario trees.

These distances and their variants, which are inherently related to Optimal Transport Theory, have been vastly used in engineering applications, as pointed out by Villani [307]. Conversely, finance applications have been scarce because usual discretization methods based on process distances are not arbitrage-free. Therefore, we describe in the next sections how we can extend the optimal discretization approach to exclude arbitrage opportunities in order to construct a smaller but reasonably similar scenario tree. We also briefly discuss other possibilities to construct scenario trees for stochastic optimization such as conditional sampling methods, moment-matching approaches and scenario clustering techniques, and point out why we should exercise caution when using them in financial applications.

Lastly, once we have constructed the approximated scenario tree, we perform our concluding step: a multi-period multi-objective stochastic portfolio optimization. It is particularly important because there is wide literature regarding Multi-Period Stochastic Optimization (MSO) and Multi-Objective Optimization (MOO) but very little research combining them.

MSO has been extensively applied to capacity planning, transportation, logistics, and long-term asset allocation over the last decades, as pointed out by Wallace and Ziemba [309]. In these applications, future decisions depend on both preceding decisions and observations. Moreover, any optimal solution (i.e., optimal sequence of decisions) must take into account uncertainty. For financial applications in particular, Multi-Period Stochastic Programming (MSP) provides a sound alternative to handle such uncertainty, provided we can estimate the expected joint probability distribution of the returns using scenario trees. Besides, MSP also offers flexibility to accommodate markets frictions and elaborated risk constraints.

MOO, on the other hand, has been a little less popular. Although intuitively appealing, applications with multiple objectives have a formidable challenge: there is no unique optimal solution. Instead, there are many (possibly infinite) Pareto optimal solutions. The more complex and numerous the objective functions, the more difficult it is to identify the non-dominated solutions. As discussed by Marler and Arora [210], Ehrgott [87], and Abraham, Jain and Goldberg [1], there are various techniques to perform a MOO such as weighted sum approaches, epsilon-constraint methods, goal
attainment techniques and genetic algorithms, which we discuss in the literature review. Among them, the epsilon-constraint method is one of the few that can be efficiently combined with MSP, provided we have a small number of objectives.

Thus, in this work we combined a MSP approach with the epsilon-constraint method to perform our multi-period multi-objective stochastic portfolio optimization. The objectives are simple: (i) maximize expected return, (ii) minimize variance and (iii) minimize conditional value-at-risk. While the first two objectives are nearly mandatory in most practical portfolio optimization problems, the last one is chosen to manage market risk more accurately in real applications. Although, the objectives’ selection is arbitrary, it considers some of the most common objectives observed in real asset allocation applications, as remarked by Roman, Darby-Dowman and Mitra [269].

This thesis’ contributions to the literature are twofold: (i) it constructs arbitrage-free scenario trees that, differently from engineering applications (e.g. supply chain management, inventory control, energy production), are a fundamental cornerstone of financial applications and (ii) it provides an alternative framework to multi-period multi-objective stochastic portfolio optimization. Although quite useful for real financial applications, both subjects have been mildly explored by the literature.

The remainder of this work is organized as follows. Section 2 presents a brief review of the related literature. Section 3 discusses the theoretical background related to the scenario tree generation (i.e., GARCH models, EVT and copulas), the scenario tree approximation (i.e., optimal discretization using process distances) and finally the multi-period multi-objective portfolio optimization (i.e., Mean-Variance-CVaR optimization). Section 4 describes how to integrate these methods and theories into a unified framework. Finally, Section 5 presents some empirical results and Section 6 concludes, discussing possible further researches.
2 Literature Review

Many peculiar statistical features of financial time series have been comprehensively discussed in the literature. As stressed by Campbell, Lo, and MacKinlay [195], returns are hardly random walks. This belief dates back to Mandelbrot [202] [203], who was among the first to suggest that financial series are seldom Gaussian, often displaying higher kurtosis and more peaked distributions. In fact, financial returns exhibit a set of important empirical properties, known as stylized facts, that are observed in different markets and instruments and that influences various financial decision problems.

More precisely, Rydberg [274], Cont [58] and McNeil, Frey and Embrechts [214] point out that, among the most studied stylized facts regarding daily financial time series, we can underline:

(i) Presence of heavy tails
(ii) Presence of volatility clustering
(iii) Asymmetric distribution of raw returns
(iv) Lack of significant serial correlation in raw returns
(v) Strong serial correlation in squared and absolute returns
(vi) Negative correlation between volatility and returns (known as leverage effect)
(vii) Serial correlations of powers of absolute returns are strongest at power one (known as Taylor effect)
(viii) Considerable evidence of cross-correlation in squared and absolute returns
(ix) Presence of tail co-dependence between series of raw returns
(x) Time-varying cross-correlations
(xi) Long-range dependence

As stressed by Cont [58], it is somewhat difficult, if not impossible, model all these features together. Keeping that in mind, in this brief review we focus mostly on
four key stylized facts: heavy-tails, volatility clustering, leverage effect and tail co-
dependence; that we will accommodate in the framework proposed later. A more
detailed review of stylized facts is offered by Campbell, Lo and MacKinlay [42], Taylor
[295] [296], and Tsay [301].

Mandelbrot [202] [203] was one of the first who claimed that commodity and
stock extreme returns are rather more common than suggested by Gaussian
distributions. As an alternative, the author suggested stable Paretian distributions (i.e.,
distributions with tail index $\alpha < 2$) to accommodate series with such extreme
observations and preserve the assumption of independent and identically distributed
returns. Fama [97] [98] also supported this hypothesis pointing out that further studies
would be required to use it properly.

Although Mandelbrot and Fama correctly suggested that the Gaussian
distribution was inadequate to model financial returns, the proposed stable distribution
presented some disadvantages such as infinite variance (since $\alpha < 2$). Although there
are techniques to manage this family of distributions as discussed for instance by
Mittnik, Paolella and Rachev [222] [223] and Taleb [292], many studies like Hsu et al
[147], Hagerman [122], Perry [239], Ghose and Kroner [110] and Campbell, Lo and
MacKinlay [42] advocate that there are more suitable approaches, with less
counterfactual implications (e.g., variance does not appear to increase as sample size
increases, long-horizon financial returns seem to follow Gaussian distributions) to
accomplish this objective.

Consequently, alternative approaches were proposed to handle extreme
returns properly. Blattberg and Gonedes [42] suggested the Student's t-distribution as
a better alternative to capture the heavy-tail behaviour. Kon [180] questioned the
assumption that financial daily returns are stationary and pointed out that mixtures of
normal distributions are better suited than Student t distributions to describe such
leptokurtic distributions. Barndorff-Nielsen [16] introduced the flexible Generalized
Hyperbolic (GH) distributions whose particular cases were later applied to model
heavy-tails and, sometimes, asymmetric gains and losses. For example, Barndorff-
Nielsen [17] [18] recommended the Normal Inverse Gaussian (NIG) distribution.
Eberlein and Keller [85], and Eberlein, Keller and Prause [86] proposed a solution
based on hyperbolic distributions that was further generalized by Prause [250]. Hansen
[125] suggested a less common special case of the GH distribution: a skewed Student
t-distribution. These later studies appear to support Cont [58], who inferred that any
distribution capable of properly modelling the distribution of financial returns require at
least four parameters: location, scale, shape and asymmetry.

All approaches mentioned previously focused on modelling the whole
distribution of returns. Other interesting methods available to handle extreme
observations, which have been extensively applied in hydrology, meteorology and
insurance, comprise modelling the distribution tail’s behaviour separately. These
methods are based on Extreme Value Theory (EVT). Some of the first applications of
EVT in finance were proposed by Jansen and Vries [154], Koedijk et al. [178][179],
Longin [197] [198], Tsay [300], and Dacorogna at al. [59]. As underlined by Cont [58],
various works such as Longin [197], Lux [200] and Hauksson et al. [130] indicate that
a Fréchet distribution with shape parameter $0.2 \leq \xi < 0.4$ (i.e., with tail index $2.5 < \alpha \leq 5$) seems suitable to model the distribution tails of daily financial series. Cont [58] also
points out that such shape parameter range suggests that the variance is indeed finite
and tails are not as heavy as those achieved by stable Paretian distributions. In the
next section we briefly discuss these methods and we refer to Embrechts, Klüppelberg
and Mikosch [90], and Coles [53] while McNeil, Frey and Embrechts [214] and
Finkenstädt and Holger [102], for comprehensive reviews and applications of EVT.

Another major stylized fact in financial series is volatility clustering. Although
Mandelbrot [202] first remarked that large (small) returns tend to be followed by large
(small) returns, was Engle [93] who introduced the Autoregressive Conditional
Heteroscedastic (ARCH) model, one of the first serious approaches trying to explain
this empirical property. Soon after, Bollerslev [34] proposed a more parsimonious
extension: the Generalized Autoregressive Conditional Heteroscedastic (GARCH)
model, which became a keystone for real-world volatility modelling problems.

GARCH models provided an appealing alternative to earlier approaches
because they were able to better explain relevant properties of empirical daily returns.
First, daily returns do not appear to be independent and identically distributed.
Secondly, squared returns appear to be time varying and exhibit strong evidence of
serial correlation. Lastly, extreme returns appear to cluster together, independently of
their sign.
Furthermore, many sophisticated extensions of the original GARCH have been proposed over the last decades. For instance, there are various nonlinear specifications such as the Nonlinear ARCH (NARCH) model developed by Engle and Ng [96], the Asymmetric Power GARCH (APGARCH) model introduced by Ding et al. [77], and some variations of the Smooth Transition GARCH (STGARCH) model suggested by Hagerud [123], Gonzalez-Rivera [115] and Lubrano [199].

Other extensions focus on substantial evidence of long memory (i.e., very slow decay) in the empirical serial correlations of absolute or squared daily returns as discussed by Taylor [296], Ding et al. [77] and Ding and Granger [76]. The Fractionally Integrated GARCH (FIGARCH) proposed by Baillie, Bollerslev and Mikkelsen [12] and the Long Memory GARCH (LMGARCH) suggested by Karanasos et al. [167] are examples.

Additionally, fitting traditional GARCH models to series with structural changes may lead to potential misspecifications, as discussed by Hamilton and Susmel [124], Cai [41], and Mikosch and Ståricà [220]. Consequently, regime-switching models, such as the Markov-Switching GARCH (MS-GARCH) models suggested by Gray [117], Dueker [80] and Klaassen [172], were developed to take into account potential structural breaks.

Besides GARCH approaches, there are alternatives based on stochastic volatility such as Taylor [294] and Harvey and Shephard [128], and more recently, on realized volatility like Andersen et al. [1] [6], and Barndorff-Nielsen and Shephard [19] that also provide invaluable tools to properly explain some important stylized facts discussed in the literature. We do not discuss them here and refer to Bauwens, Hafner and Laurent [20], Andersen et al. [1] and Tsay [301] for further details regarding these alternatives.

Regardless of these various enhancements and alternatives, the basic univariate formulation GARCH(1,1) is still the most used according Bauwens, Hafner and Laurent [20], and McNeil, Frey and Embrechts [214]. The reasons are twofold: they are easier to fit and they provide reasonable estimates, particularly when combined with complementary techniques like extreme value theory and copulas. We will discuss some details of this particular specification in the next section and refer to Andersen et al. [7], Tsay [301], and Bauwens, Hafner and Laurent [20] for a comprehensive review.
Moreover, straightforward extensions of the GARCH(1,1) formulation are also commonly used to address other additional statistical features of daily financial series. In particular, some of these modifications lead to asymmetric models that can accommodate another key stylized fact: the leverage effect.

This feature was first discussed by Black [31] and Christie [52] who noticed that declining asset prices (i.e., negative log returns) are often follow by increasing volatility. Further research about this asymmetry of innovations shock led, for example, to the Exponential GARCH (EGARCH) model by Nelson [232], the GJR-GARCH model by Glosten, Jagannathan and Runkle [111] and the Threshold GARCH (TGARCH) by Rabemananjara and Zakoian [252] and Zakoian [312].

There are some differences among these extensions. For instance, moment conditions for the EGARCH are less strict than those for GJR-GARCH and TGARCH, and GJR-GARCH models the conditional variance while TGARCH focus on the conditional standard deviation. Nonetheless, to the best of our knowledge, there are not many studies comparing their results. Engle and Ng [96] is one exception that suggests that the GJR-GARCH model appears more appropriate to explain extreme shocks than the EGARCH model. Another exception can be found in Tsay [301] who provides a few examples suggesting that no model is statistically superior. Regarding practical applications, parsimonious EGARCH(1,1) and GJR-GARCH(1,1) specifications are often preferred according Andersen et al. [7], and Bauwens, Hafner and Laurent [20].

To this point we have discussed only univariate approaches but, unsurprisingly, various multivariate specifications were proposed over the last decades. Focusing on multivariate GARCH models for instance, Bollerslev et al. [37] introduced the Diagonal VECH (DVECH) model. Although quite general, it has limited practical applications because it requires too many parameters to be estimated. Baba, Engle, Kraft and Kroner [11] and Engle and Kroner [95] suggested the BEKK model, which further restricted the DVECH specification but still faced limitations with large dimensions. To address this curse of dimensionality, Bollerslev [36] proposed the Constant Conditional Correlation (CCC) Model whose constant correlation strong assumption was later relaxed in the Time-Varying Correlation (TVC) model of Tse and Tsui [302] and the Dynamic Conditional Correlation (DCC) model suggested by Engle [94]. More recent specifications of these dynamic correlation models, such as
Cappiello, Engle and Sheppard [44] and Hafner and Franses [121], also address additional features of financial series like leverage effects and also explain elaborate joint behaviours of financial markets such as contagion and volatility spillover effects. For further details, Bauwens, Laurent and Rombouts [21], and Silvennoinen and Teräsvirta [281] provide an inclusive review of multivariate GARCH models.

Despite of the appealing properties of multivariate GARCH models, some works such as McNeil, Frey, and Embrechts [214] and Patton [238] suggest that copulas, which were introduced by Sklar [282] and later promoted by Joe [160] and Nelsen [231], may offer a compelling alternative because they separate the dependence structure from the univariate marginal distributions allowing more flexibility (a key aspect for many practical applications). However, depending on the required degree of flexibility, accuracy and tractability, copula models may range from straightforward static combinations of different univariate distributions, such as those presented by Patton [235], to much more elaborated applications comprising dynamic multivariate processes like those discussed in Patton [236].

Among the most common copula types we have the (i) elliptical copulas like the Gaussian and t copulas, which are readily extended to multivariate settings but can explain only simple dependence structures; (ii) Archimedean copulas such as Frank, Gumbel and Clayton copulas that, conversely, are difficult to apply to high dimensional problems but offer greater flexibility to explain the joint dependence; (iii) vine copulas, which handle high dimensional dependence by combining flexible bivariate copulas families and; (iv) mixture copulas that are able to accommodate quite sophisticated dependence structures. We recommend Embrechts, Lindskog and McNeil [91] for a good review of elliptical and Archimedean families, Kurowicka and Joe [189] for a detailed discussion of vine copulas, and Joe [161] for a comprehensive review, including some mixture copula models.

Even though elaborated models like vine and mixture copulas provide greater flexibility, practical financial applications (i.e., applications that comprise many instruments) often favour more tractable approaches like elliptical copulas. Moreover, tail co-dependence is often the most pressing concern in these applications. For that reason, t copulas and their extensions are usually the preferred choice. See for example Demarta and McNeil [64].
Additionally, many recent financial applications combine copulas with other methods models to achieve better results. Important examples are GARCH-Copula models, discussed by Rockinger and Jondeau [162],[268], Huang et al. [150], and Riccetti [263],[264]. These models can be fitted using a two-step estimation approach, conceptually similar to that used by McNeil and Frey [213] to combine EVT and GARCH models.

As we elaborate in the next sections, we can integrate these two-step estimation approaches into a single three-step GJR-GARCH + EVT (GPD) + t-Copula framework to simulate multivariate trajectories (also known as paths) that have a common starting point. This simulated multivariate scenario fan, which is a particular case of a scenario tree, not only captures the uncertainty of the daily log returns but also accommodates reasonably well the four key stylized facts mentioned previously: heavy-tails, volatility clustering, leverage effect and tail co-dependence.

Similar works applied to exchange rates and commodities can be found Wang, Jin and Zhou [310] and Tang et al. [293]. Besides, other studies that jointly explain at least three stylized facts can be found in Xiong and Idzorek [311], which accounts for heavy tails, skewness, and asymmetric dependence, and Hu and Kercheval [148], which proposes a GHD approach to explain asymmetric dependence, volatility clustering and (partially) heavy tails. There are also a few elaborated works that capture four stylized facts. Patton [236], and Viebig and Poddig [306] are examples.

So far, we reviewed combined methods based exclusively on historical daily log returns. Nevertheless, there are several alternative methods available to simulate multivariate scenarios that use additional information such as economic variables or forward-looking indicators. Boender [33], Cariño et al. [45], and Cariño, Myers and Ziemba [46], for instance, construct scenarios using Vector Autoregressive (VAR) models. Mulvey and Vladimirou [228], on the other hand, generate scenarios using decomposition methods based on Principal Component Analysis (PCA). Mulvey [225], and Mulvey and Thorlacius [227] employ dynamic stochastic simulation to forecast both economic series and asset prices. For additional information regarding these different scenario generation methods, we refer to Ziemba and Mulvey [314], and Kouwenberg and Zenios [185].

These alternative methods based on economic variables or forward-looking indicators are vastly used in practical applications. However, they are not particularly
tailored to accommodate the stylized facts mentioned earlier, which are particularly relevant to portfolio optimization and risk management applications. For that reason, in this thesis we simulate the multivariate scenarios using the suggested GJR-GARCH + EVT (GPD) + t-Copula three-step approach, which can properly describe the multivariate underlying stochastic process of the daily log returns (at the expense of additional information that could prove useful in some cases).

This combined approach, however, exhibits two major limitations: the generated multivariate scenario tree (i) is often too large (i.e., computationally demanding) to be applied in practical finance applications and (ii) is not free of arbitrage opportunities. Therefore, we require additional steps to handle these limitations.

Regarding the first shortcoming, numerous techniques have been proposed to approximate the original scenario trees by more tractable ones, as highlighted by Dupačová, Consigli and Wallace [82] and Kouwenberg [184]. Dupačová, Consigli and Wallace [82], Høyland and Wallace [144] and Prékopa [251] also point out that the features considered by each particular stochastic optimization problem define which properties should be contemplated by a good scenario tree approximation (e.g., a traditional Mean-Variance optimization only entails the approximated first two moments of the original multivariate distribution). Therefore, true convergence towards the original scenario tree (based on the true underlying stochastic process) may be desired but is not mandatory to achieve optimal solutions for Stochastic Programming problems.

These approximation techniques seek to generate tractable scenario trees and guarantee that the different approximated scenario trees lead to similar, stable and optimal objective values, with small approximation errors. As reviewed in Dupačová, Consigli and Wallace [82], and Heitsch and Römisch [133], we discuss some common techniques used in finance, energy and logistics problems, which are commonly grouped in four categories: sampling algorithms, moment-matching approaches, clustering techniques or optimal discretization methods.

Sampling algorithms are likely the first approaches used to approximate stochastic processes. They comprise different techniques that basically draw possible values from a given distribution or stochastic process for each node of a predefined scenario tree. Among these sampling techniques, bootstrapping sampling, random
sampling, importance sampling, rejection sampling, conditional and Markov Chain Monte Carlo (MCMC) sampling are among the most used in practical applications.

Traditional or adjusted random sampling, also known as Monte Carlo sampling, are the simplest, and often the least efficient, of these methods. Essentially, they draw values randomly and associate them to probabilities using the predefined distribution. Similarly, bootstrapping sampling methods are just random sampling methods, with replacement, that draw values from the sample itself. Cariño et al. [45], Consigli [54], Consigli and Dempster [55], and Consigli, Ianquita and Moriggia [56] for instance, use these methods to construct their scenario trees.

Despite their popularity, random sampling techniques usually require too many draws to be reliable. Therefore, we often consider other, more efficient, techniques such as (sequential) importance sampling methods. These methods restrict (or give more importance to) the random draws to relevant regions of alternative distributions that are easier to handle. Dupačová, Consigli and Wallace [82], Dempster and Thompson [71], and Dempster [67] offer interesting applications of these methods. Moreover, Dupačová, Consigli and Wallace [82] also combines the sequential importance sampling with the moment matching and clustering approaches (described later) to achieve more reliable scenario approximations. Likewise, rejection sampling follows a similar approach to importance sampling but trades simplicity for efficiency, as discussed by Pflug and Pichler [247].

One important drawback of most of the previous sampling techniques is that current draws do not depend on previous draws. Consequently, they preclude any type of temporal dependence, which is quite relevant for numerous financial applications. Conditional sampling methods take away this shortcoming by keeping track of previously drawn values. The popular Sample Average Approximation method is an example of conditional sampling extensively discussed by Kleywegt, Shapiro and Homem-de-Mello [176] and Shapiro [278]. More recently, MCMC algorithms (such as Gibbs sampling and Metropolis-Hastings-Green) have been mixed with some of the previously mentioned methods like importance sampling, leading to advances in multidimensional optimization problems. We refer to Ekin, Polson and Soyer [88], and Parpas et al. [234] for recent examples.

Unfortunately, although sampling methods exhibit very nice asymptotic properties, large scenario trees are often intractable computationally, particularly in
high dimensional spaces (except for some very sophisticated MCMC sampling models), hindering most practical applications. Likewise, small scenario trees based on sampling methods usually lead to poor accuracy (i.e., large approximation errors and instability). For those reasons, alternative approximation techniques are often considered in practice.

Moment-matching (or, more generally, property-matching) approaches are among these alternatives. As highlighted by Dupačová, Consigli and Wallace [82], these methods do not require full knowledge of the true underlying stochastic process and, consequently, are particularly convenient when we have limited information about its probability distribution. Instead, property-matching approaches need only a few statistical properties of the distribution (e.g., moments, co-moments, percentiles, principal components, serial correlations) to simulate an approximated scenario tree, i.e., a discrete probability distribution that is consistent with the given properties.

Besides requiring less information (i.e., just some statistical properties of the underlying stochastic process), moment-matching approaches can match most properties using relatively few scenarios when compared to traditional sampling methods. They can also easily accommodate more elaborated dependency structures such as serial correlations.

However, they also present relevant shortcomings. First, as discussed by Dupačová, Consigli and Wallace [82] all relevant statistical properties for the optimization problem are assumed to be known before the optimization problem is solved, but that might not be always true. Second, Hochreiter and Pflug [141], and Pflug and Pichler [246] show that completely distinct approximated scenario trees (i.e., quite different stochastic processes) might lead to the desired statistical properties, possibly leading to unreliable solutions. Lastly, Kaut and Wallace [168] emphasises that, differently from sampling methods, property-matching approaches do not guarantee convergence by simply increasing the number of simulated scenarios, which may produce instability and large optimality gaps.

We suggest Höyland and Wallace [144], Höyland, Kaut and Wallace [145] and Dupačová, Consigli and Wallace [82] as well-known references regarding moment-matching methods and, more recently, Date, Mamon and Jalen [62], Consiglio, Carollo and Zenios [57], and Beraldi, Simone and Violi [26] as more elaborated applications of
such approximation methods. We also recommend King and Wallace [171] for a comprehensive review about the subject.

Scenario clustering techniques have also been employed as alternatives to traditional sampling methods. They use different methods of cluster analysis to bundle similar scenarios together to approximate the original scenario trees by smaller ones. However, as pointed out in Aggarwal [1], clustering techniques for time-series are more challenging because they fall within the class of contextual representation, i.e., techniques that should consider both behavioural (e.g., price) and contextual (e.g., time) attributes. In that regard, Liao [190], and Rani and Sikka [258] offer interesting surveys about time-series clustering techniques. They underline that among the many clustering categories available, three have been used (directly or indirectly) to handle time-series clustering: partitioning methods, hierarchical methods, and model-based methods.

Partitioning clustering methods constructs (crisp or fuzzy) partitions of the data, where each partition represents a cluster containing at least one element. As a popular example in this category, we can point out the k-means algorithms, which are characterized by the mean value of the elements in that cluster. Pranevičius and Šutienė [290][291], and Šutienė, Makackas, and Pranevičius [289] applied elaborated variations of the k-means algorithm to generate approximated scenario trees that preserve important statistical features of the underlying stochastic process such as intertemporal dependency structures. In addition, Beraldi and Bruni [25], and Gulpinar, Rustem and Settergren [120] discuss interesting variations of the k-means algorithm applied to stochastic optimization problems. However, as explained by Liao [190], basic partitioning methods perform well with spherical-shaped clusters and small or medium data sets but do not handle properly clusters with non-spherical or complex shapes, which are quite common in financial applications.

Hierarchical clustering methods constructs a multilevel cluster tree by grouping the data using agglomerative or divisive algorithms. Agglomerative (bottom-up) clustering puts each element in its own cluster and then merge clusters into larger clusters, until predefined conditions (e.g., desired number of clusters) are achieved. Divisive (top-down) clustering, on the other hand, start at the top with all elements in a single cluster and then splits it recursively until the predefined conditions are met. Unfortunately, plain hierarchical methods are also poorly equipped to handle complex
scenarios as discussed by Liao [190], and Pranevičius and Šutienė [290] because they cannot make adjustments once a merge or split action has been performed, which may lead to poor scenario structures. On the other hand, as demonstrated by Musmeci, Aste and Di Matteo [229], and Tumminello, Lillo and Mantegna [303], some elaborate extensions of hierarchical methods can achieve good results, at the expense of tractability.

Model-based clustering methods assume a model for each of the clusters and attempt to find the best fit for the data. Statistical-based methods falls into this category. Although fairly robust due to the use of more sophisticated frameworks (e.g., Hidden Markov Models) and similarity measures (e.g., Dynamic Time Warping, Longest Common Subsequence), these methods are usually rather complex and rare in the financial literature. As examples, we refer to Knab et al. [177], and Dias, Vermunt and Ramos [73][74] for sophisticated applications involving Hidden Markov Models (HMM) to construct clusters of financial time-series.

Lastly, optimal discretization methods have also been used as alternative approaches to approximated scenario trees, according Pflug [241][243], Römisch [270], Pflug and Römisch [248], and Heitsch and Römisch [132][133][134][135][137]. As stressed by Pflug [241][243], although more elaborated these methods offer a sound theoretical basis for approximating the true (and often intractable) underlying probability distribution by a manageable and suitable probability measure with finite support. In simply terms, such approximation is an optimization that seeks to minimize an appropriate probability distance metric between the original set of scenarios and the reduced set of scenarios.

In the literature, this approach is also known Monge-Kantorovich mass transport problem, which in the scenario tree context corresponds to the optimal redistribution of the approximated scenario probabilities in order to minimize a probability distance between the original and the approximated scenario tree. We refer to Villani [307] and Rachev and Rüschendorf [255][256] for detailed discussions about mass transportation theory and to Gröwe-Kuska, Heitsch and Römisch [118], Hochreiter, Pflug and Wozabal [142], and Heitsch and Römisch [133] for examples of applications outside financial markets.

Optimal discretization theory relies deeply on the choice of a suitable probability distance metric. As discuss by Rachev [253], Rachev, Stoyanov and
Fabozzi [257], and Gibbs and Su [112], numerous probability metrics are available but most, such as the Kolmogorov metric or the total variation distance, are poorly equipped to properly measure the distance between probability distributions. Among the appropriate metrics to that accomplish that, there are even fewer that suited multi-period stochastic optimization problems, i.e., that can be extended from probability distribution distances to stochastic process distances. The most common candidates in the literature are the Fortet-Mourier and Wasserstein distances.

Römisch [270], Heitsch and Römisch [132] and Dupačová, Gröwe-Kuska, and Römisch [83] are renowned examples of scenario reduction using the Fortet-Mourier metric. Likewise, Heitsch and Römisch [133]; Hochreiter and Pflug [141], Pflug [241][243], Pflug and Pichler [244][245][246][247], and Kovacevic and Pichler [186] are popular examples of scenario approximation using the Wasserstein metric. Moreover, Dempster, Medova and Sook [68] provides an interesting discussion about the statistical properties of the Wasserstein distance and its application in optimal discretization for stochastic programming.

In particular, Pflug and Pichler [244][245][246], and Kovacevic and Pichler [186] also explore in detail the Process (or Nested distance), which is an extension of the Wasserstein distance for multi-period problems (i.e., for problems where we should use some type of filtration distance, as explained by Heitsch and Römisch [132], and Heitsch, Römisch and Strugarek [138]). They show that, in such problems, we need to consider both the probability measures and the value-and-information structure (represented by a proper filtration or scenario tree) during the discretization process of a stochastic optimization problem. In this thesis, we follow this approach as well and explain it in detail in later sections.

Concerning the second limitation of the GJR-GARCH + EVT (GPD) + t-Copula framework to generate scenario trees, arbitrage-free scenarios are mandatory to real finance applications to avoid biased solutions, as stressed by Klaassen [173][174][175], Sodhi [286], Geyer, Hanke and Weissensteiner [107][108] and Consigli, Carollo, and Zenios [57]. They also underline that most of the scenario approximation methods do not take that into account. Accordingly, literature regarding arbitrage-free approximation methods is, at best, limited.

Klaassen [173], for instance, shows how arbitrage opportunities can lead to biased solutions (i.e., biased investment strategies) in stochastic programming
optimization problems. Klaassen [174] goes further and discusses a two-step process: one to construct an arbitrage-free scenario and a second to reduce the tree size by state and time aggregation while preserving the absence of arbitrage opportunities. Unfortunately, such approach does not guarantee that other statistical features (e.g., kurtosis, covariance or serial correlation) are preserved. Klaassen [175], on the other hand, extends Høyland and Wallace [144] moment-matching method, which preserves some predetermined statistical features of the underlying stochastic process but does not preclude arbitrage opportunities. Klaassen [175] describes how we can identify, ex-post, such opportunities and suggests how we can avoid them, ex-ante, by adding no-arbitrage constraints to Høyland and Wallace [144]’s moment-matching approximation method.

Klaassen [175] also suggests that we can rerun the scenario approximation method until an arbitrage-free scenario tree is achieved. However, Geyer, Hanke and Weissensteiner [107][108] shows that simply rerunning or even increasing randomly the branching scheme of the scenario tree do not guarantee an arbitrage-free scenario tree, but it does make the optimization problem less tractable. Adding constraints to exclude arbitrage opportunities lead to non-convex optimization problems, which usually require a global optimization approach. Consigli, Carollo, and Zenios [57], for instance, propose a refinement of the moment-matching method with no arbitrage-constraints that uses such an approach (based on lower bounding techniques).

Geyer, Hanke and Weissensteiner [107][108][109] also points out that the absence of arbitrage implies a lower bound on the branching scheme of the scenario tree. The number of possible children scenarios at each node (i.e., each state of nature, characterized by a particular realization of prices or returns) of the scenario tree usually must be greater than the number of non-redundant instruments considered in the optimization problem so that some statistical properties of the underlying stochastic process can be properly preserved. Note that this condition is more stringent than Harrison and Kreps [127]’s no-arbitrage condition.

Additionally, Jobst and Zenios [159], Fersti and Weissensteiner [100], and Geyer, Hanke and Weissensteiner [107] stress that multi-period portfolio optimization problems should use objective (i.e., real-world) probability measures rather than risk-
neutral probability measures to account properly for the (estimated) market price of risk.

Finally, assuming that tractability and arbitrage conditions have been achieved, we still need to evaluate the quality of the optimal solutions provided by these techniques. As we mentioned earlier, such quality is typically evaluated in the literature based on two key characteristics: optimality gap and stability.

The optimality gap, also known as the approximation error, between the approximated and the original scenario tree, which represents the true underlying stochastic process, is a major aspect of the approximation quality. Since the optimal objective values or policies for true underlying stochastic problem are hardly available (otherwise we would not need approximations in first place), we must find indirect methods to estimate such gap. Many studies like Kaut and Wallace [168], Römisch [270], Heitsch and Römisch [136], and Pflug [241] debate possible methods to control it as well as their implications to the discretization process. Pflug [241], in particular, provides a good review about the subject and proposes an important lemma that establishes an upper bound on the optimality gap, which we discuss in later sections. Such result is paramount to our optimal discretization approach based on the Wasserstein and Process distances, as we will see later.

Stability is another major aspect related to the quality of the approximated scenario tree. According Kaut and Wallace [168], King and Wallace [171], and Dempster, Medova and Sook [69], stability analyses usually focus on both in-sample and out-of-sample stability of the objective function rather than the stability of the solutions, which may vary significantly due to the highly nonlinearity and non-convexity of the optimization problem.

Definitions of in-sample and out-of-sample stability vary in the related literature. In this thesis, we follow King and Wallace [171]'s definitions. In-sample stability implies that optimal solutions do not vary significantly if we generate new scenario trees with similar cardinality, using the same method. The concept of in-sample stability implicitly assumes that the entertained scenario approximation method does not produce the exactly same scenario tree for the same input data. Out-of-sample stability, on the other hand, guarantees that the generated optimal solutions lead to optimal objective function values that are close to the true objective value. Clearly, out-of-sample stability is achieved if the optimality gap is sufficiently small.
Unfortunately, since the true objective function value is typically unknown in practical applications, out-of-sample stability must be evaluated approximately by alternative tests, as we discuss later.

Kaut and Wallace [168], and King and Wallace [171] also highlight that in-sample stability does not imply out-of-sample stability and vice-versa, but both help to establish a lower bound on the cardinality of the scenario tree. Various works in the literature evaluate the stability of different methods to approximate scenario trees. For a more comprehensive review of the subject, we refer to Dupačová, Gröwe-Kuska and Römisch [83], Rachev and Römisch [254], Heitsch and Römisch [136], and Heitsch, Römisch and Strugarek [138].

The various concepts and techniques mentioned previously suggest that, in general, no particular scenario construction approach is superior in all circumstances. They often depend heavily on the context and problem at hand. Dempster, Medova and Sook [68], and Löhndorf [196] for example, compare sampling algorithms, moment-matching approaches, clustering techniques or optimal discretization methods, and conclude that differences are minimal, except for simple sampling approaches that are outperformed.

It is also worth mentioning that some methods perform the scenario tree generation and reduction steps simultaneously, rather than generating a large scenario tree and then approximating it by a tractable one. In this thesis, we focus on the latter, which seems to be more common in the recent literature. For studies and applications regarding the former, we suggest Cariño et al. [45], Mulvey [225], Dempster and Thorlacius [72], and Gulpinar, Rustem and Settergren [120].

Lastly, provided a manageable, sufficiently accurate and arbitrage-free scenario tree is available, we proceed to the portfolio optimization step. Since Markowitz’s seminal work [208], portfolio optimization has been a major subject in financial markets. Although elegant, his single-period Mean-Variance approach entails various simplifications that are hardly satisfied in real markets. For example, variance is not a coherent risk measure; actual financial problems commonly require decisions involving time varying (i.e., uncertain) variables over multiple periods (e.g., weeks, months or years); portfolio rebalances incur in transaction costs and, sometimes, liquidity issues and; taxes and regulatory restrictions often influence portfolio performance.
A myopic (i.e., single-period) decision is optimal only in very particular circumstances. Typically, as pointed out by Mulvey, Pauling and Madey [226] and Brandt [40], (i) it fails to benefit or hedge against the presence of dynamic patterns and relationships among financial instruments, (ii) it does not account for market liquidity and transaction costs and (iii) it deals poorly with uncertainty. To accommodate such elaborated market dynamics and frictions, various alternatives and extensions were suggested. We refer to Guastaroba [119], Jarraya [155], Mansini, Ogryczak, and Speranza [205] [206], and Kolm, Tütüncü and Fabozzi [166] for recent surveys. Among these various alternatives, Multi-Period Stochastic Optimization (MSO) approaches were particularly appealing because of their flexibility to handle key features of real markets.

One key difference of multi-period approaches from single-period ones is that their optimal solutions do not comprehend a single allocation decision. Instead, they comprise a whole sequence of successive allocation decisions (i.e., rebalances), known as allocation policy or strategy. Although an optimal policy rarely offers the best outcome for any particular scenario, it does suggest a noninferior course of action (in the Pareto sense) that considers the impacts and probabilities of all possible scenarios, i.e., it suggests an allocation strategy that takes into account uncertainty. Shapiro, Dentcheva and Ruszczyński [279], Dupačová, Hurt and Štěpán [84], Kall and Wallace [165], and King and Wallace [171] provide numerous examples that discuss this difference and emphasize the advantages of the multi-period stochastic approach.

There are many approaches to multi-period optimization such as dynamic stochastic control, multi-period stochastic programming, sampling-based methods (e.g., acceptance-rejection sampling, and importance sampling) and metaheuristic methods (e.g., genetic algorithms, simulated annealing, and particle swarm optimization). Early models like Mossin [224], Samuelson [276] and Merton [215], [216] focused on manageable approaches that could be solved analytically, often based on continuous-time frameworks with unrealistic simplifying assumptions. Conversely, recent advances in processing power have greatly improved the tractability of discrete-time Multi-Period Stochastic Programming (MSP) models, which offer greater flexibility in financial applications as demonstrated, for instance, by Mulvey and Vladimirov [228], Dantzig and Infanger [60], and Dupačová, Hurt and Štěpán [84]. Ziemba [313] goes further and states that scenario-based discrete multi-period stochastic programming
usually offers a superior alternative to simulation, control theory and continuous-time approaches.

Such advantage, however, relies on a critical and rather challenging step: the construction and approximation of (often) multivariate scenarios. Stability, tractability and accuracy of MSP’s allocation policies or strategies depend heavily on the quality and number of the simulated scenarios, commonly represented as scenario trees. Therefore, the steps mentioned earlier, and explained in detail in the next sections, are essential to provide a reliable and manageable scenario tree that can be used by a proper MSP approach. For further details regarding discrete-time stochastic decision models, we suggest Uryasev and Pardalos [305], and Dupačová, Hurt and Štěpán [84].

In this thesis, a proper MSP approach refers to a stochastic programming setup that handles satisfactorily, in the Artzner et al. [9] and Föllmer and Schied [105]’s sense, the risk described by the approximated scenario tree, as discussed, for instance, by Ruszczyński and Shapiro [273] and Rockafellar [265]. Markowitz’s original Mean-Variance approach is certainly tractable and elegant but it does not account properly for actual market risk. To accomplish that, various risk measures have been proposed in the literature but few have achieved wide acceptance among academics and practitioners, as discussed by Nawrocki [230].

The Lower Partial Moment (LPM) family of risk measures (or risk functionals), proposed and explored by Bawa [20], Bawa and Lindenberg [21] and Fishburn [95], included some of these measures. Such LPM measures are often called safety risk measures in contrast to the more traditional dispersion risk measures, such as variance or absolute deviation. As discussed by Nawrocki [230], the advantages of LPM risk measures are twofold: they can handle non-normal returns and accommodate a myriad of different investors’ utility functions (i.e., risk seeking, risk-neutral or risk averse behaviours).

These properties are particularly relevant for practical financial applications because: (i) non-normality of financial returns are commonly observed in real markets, as pointed out by Campbell, Lo and MacKinlay [42], and Taylor [295] [296]; (ii) derivation of a precise (and unique) utility function is challenging, if not impossible at all, as stressed by Roy [271], and Kahneman and Tversky [163].
Therefore, LPM provides more realistic, albeit not perfect (e.g., usually difficult to optimize or extend to standard portfolio theory), risk measures that are consistent with stochastic dominance (for positive exponents) and focus on what typically matters most for investors: the downside risk. We refer to Roy [271], Markowitz [209], Mao [207], and Fishburn [103] for the original arguments regarding these characteristics and to Unser [304], Bertsimas, Lauprete and Samarov [28], and Pflug and Römisch [248] for recent discussions about LPM risk measures.

Among the most renowned downside risk measures, we underline the Semivariance (SV), suggested by Markowitz [209], the Semi-Absolute Deviation (SAD), proposed by Speranza [288], the Conditional Value-at-Risk (CVaR), proposed by Rockafellar and Uryasev [266][267] and the Conditional Drawdown-at-Risk (CDaR), suggested by Chekhlov, Uryasev and Zabarankin [49][50]. In particular, the CVaR, also known as Average Value-at-Risk (AVaR) or Expected Shortfall (ES), has been quite popular in recent practical applications according Xiong and Idzorek [311], Krokhmal, Palmquist and Uryasev [187], and Krokhmal, Zabarankin and Uryasev [188], because it captures tail risks reasonably well and it is computationally easy to optimize.

A comprehensive review of these various risk measures, their strengths and limitations can be found in Dowd [79], Pflug and Römisch [248], Dempster, Pflug and Mitra [70], and Rachev, Stoyanov, and Fabozzi [257]. Pflug and Römisch [248], in particular, describes in detail the optimization of two types of risk measures: acceptability type and deviation type functionals, which are especially suited for MSP. As we discuss in later sections, we use the same approach to accommodate the risk measures in the multi-period portfolio optimization step.

Unfortunately, a single risk measure is rarely enough to accommodate adequately all investor’s goals in most practical portfolio optimization problems, as pointed out by Roman, Darby-Dowman and Mitra [269]. As an example, variance (or, correspondingly, standard deviation) is not a coherent risk measure but it is usually a compulsory objective, from most investors’ perspectives. In such cases, a second (and coherent) risk measure could be considered in order to manage risk properly. These real-world cases, where multiple objectives must be evaluated simultaneously, often lead to Multi-Objective Optimization (MOO) problems.

However, differently from MSP literature MOO literature is somewhat more recent. There are a few early examples like Jean [156] and Konno, Shirakawa and
Yamazaki [182] but, only recently, improvements in processing power made it possible to evaluate and explore new venues and ideas in practical applications. As examples of promising applications for portfolio optimization, we can mention the Multi-Objective Ant Colony Algorithm (MOACO), applied by Doerner et al. [78], the Non Dominated Sorting Genetic Algorithm (NSGA-II), used by Lin, Wang and Yan [191], and the Multiple Objective Evolutionary Algorithms (MOEA), entertained by Fieldsend, Matatko and Peng [101]. We suggest Abraham, Jain and Goldberg [1] and Ehrgott [87] for a comprehensive review on multicriteria optimization and Jarraya [155] for a recent survey of MOO applications in asset allocation.

The recent advances in MOO are very appealing but most of them lack a key feature required for many practical financial applications: they cannot handle multi-period settings. In fact, the financial literature has barely explored approaches that can solve stochastic multistage multi-objective problems. Roman, Darby-Dowman and Mitra [269] and Abdelaziz [1] are among the few examples. Most studies combining multi-period multi-objective optimization such as Fazlollahi, and Maréchal [99] are found in Engineering. To contribute on that front, in this thesis we discuss an additional, but narrowly explored, advantage of MSP approaches: the flexibility to accommodate multi-objective functions in multi-period portfolio optimization problems. In the next section, we discuss in detail the theoretical background required to accomplish that.
3 Theoretical Background

In this section we discuss the theoretical framework required to:

(i) Estimate a multivariate probability distribution of daily log returns based on a GJR-GARCH + EVT (GPD) + t-Copula combined approach that is able to explain four important stylized facts: heavy-tails, volatility clustering, leverage effect and tail co-dependence.

(ii) Simulate a sufficiently large tree of multivariate scenarios using the combined approach, which is assumed to be close to the true underlying stochastic process governing the multivariate price evolution of the respective financial instruments.

(iii) Approximate the original large scenario tree by a smaller one with tractable cardinality and no arbitrage opportunities.

(iv) Optimize a Mean-Variance-CVaR portfolio using Stochastic Programming together with the approximated scenario, which is assumed to describe properly the parameters uncertainty.

Figure 1 – Stochastic Portfolio Optimization Process Overview

Following McNeil, Frey and Embrechts [214], King and Wallace [171], and Pflug and Pichler [246], before we elaborate on each one of these steps, we start with some fundamental concepts and definitions.
Definition 1 A σ-algebra (sigma-algebra) $\mathcal{F}$ on a scenario set $\Omega$ is a collection of subsets of $\Omega$ such that:

(i) $\Omega \in \mathcal{F}$

(ii) $\mathcal{F}$ is closed under complementation (i.e., if $X \in \mathcal{F}$ then $X^C \in \mathcal{F}$)

(iii) $\mathcal{F}$ is closed under countable unions (i.e., if $X_1, X_2, \cdots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} X_i \in \mathcal{F}$)

(iv) $\mathcal{F}$ is closed under countable intersections (i.e., if $X_1, X_2, \cdots \in \mathcal{F}$ then $\bigcap_{i=1}^{\infty} X_i \in \mathcal{F}$)

The elements of $\mathcal{F}$ are also called events of $\mathcal{F}$. Two important σ-algebras on $\Omega$ are the trivial σ-algebra $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and the complete σ-algebra $\mathcal{F}_\infty = P(\Omega)$, which comprises the power set (i.e., the collection of all subsets) of $\Omega$. Additionally, given any collection $\mathcal{F}$ of subsets of $\Omega$, the smallest σ-algebra containing all subsets of $\mathcal{F}$, denoted $\sigma(\mathcal{F})$, is called the σ-algebra generated by $\mathcal{F}$.

Definition 2 Given a scenario set $\Omega$ and a σ-algebra $\mathcal{F}$ on $\Omega$, a measure is a mapping $\mu$ from $\mathcal{F}$ to the extended real line $\mathbb{R} = [-\infty, +\infty]$ such that:

(i) $\mu(\emptyset) = 0$

(ii) $\mu(X) \geq 0, \forall X \in \mathcal{F}$

(iii) $\mu$ is countably additive (i.e., if $X_1, X_2, \cdots \in \mathcal{F}$ are pairwise disjoint sets in $\mathcal{F}$ then $\mu(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} \mu(X_i)$)

Definition 3 Given a scenario set $\Omega$ and a σ-algebra $\mathcal{F}$ on $\Omega$, the pair $(\Omega, \mathcal{F})$ forms a measurable space. A measurable space is simply a set $\Omega$ equipped with a σ-algebra $\mathcal{F}$, whose elements are called measurable sets.

Definition 4 Given a scenario set $\Omega$, a σ-algebra $\mathcal{F}$ on $\Omega$ and a measure $\mu$, the triple $(\Omega, \mathcal{F}, \mu)$ forms a measure space.

Definition 5 Given a measure space $(\Omega, \mathcal{F}, \mu)$, the measure $\mu$ is a probability measure if:

(i) $\mu(X) \in [0,1], \forall X \in \mathcal{F}$

(ii) $\mu(\Omega) = 1$

Definition 6 A probability space is a measure space $(\Omega, \mathcal{F}, \mu)$ such that the measure $\mu$ is a probability measure.
**Definition 7** Given a probability space \((\Omega, \mathcal{F}, P)\) such that \(\Omega\) is a scenario space, \(\mathcal{F}\) is a \(\sigma\)-algebra on \(\Omega\) and \(P\) is a probability measure, a (real-valued) random variable \(X\) and a (real-valued) random vector \(Y = (Y_1, \ldots, Y_n)\) correspond, respectively, to the mappings:

\[
X: \Omega \to \mathbb{R} \quad \quad \quad \quad (1)
\]
\[
Y: \Omega \to \mathbb{R}^n \quad \quad \quad \quad (2)
\]

Such that \(X\) and \(Y\) are \(\mathcal{F}\)-measurable.

**Definition 8** Given a random variable \(X\), \(F\) and \(f\) are, respectively, the cumulative distribution function and the probability density function of \(X\), provided the derivative of \(F\) exists, such that:

\[
F: \mathbb{R} \to [0,1] \quad \quad \quad \quad (3)
\]
\[
F(x) = \int_{-\infty}^{x} f(v)dv = P(X \leq x) \quad \quad \quad \quad (4)
\]

**Definition 9** Given a random vector \(X = (X_1, \ldots, X_n)\), the multivariate distribution function \(F\) of \(X_1, \ldots, X_n\) corresponds to:

\[
F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \quad \quad \quad \quad (5)
\]

Such that the marginal distributions of \(F\) are

\[
F_i(x_i) = \lim_{x_{i+1}, \ldots, x_n \to \infty} F(x_1, \ldots, x_n), \quad i = 1, \ldots, n \quad \quad \quad \quad (6)
\]

### 3.1 Scenario Tree Generation

As we discuss later in detail, an optimization based on stochastic programming assumes that the underlying stochastic process or probability distribution of the parameters is known. However, except for very unrealistic cases (i.e., cases that rely on very strong simplifying assumptions), we cannot easily solve multiperiod stochastic optimization problems using continuous processes or distributions because they are functional optimization problems (i.e., solutions are functions, not values).
In practice, we often need to approximate the continuous stochastic process or probability distribution by a similar (in some statistical sense) process or distribution that has finitely many discrete scenarios like a scenario tree structure. Scenario trees offer an intuitive way to represent the underlying stochastic process or probability distribution, providing an easy structure to accommodate the uncertainty of the relevant parameters and, in the former case, the way that information is revealed over time. In other words, we can represent a stochastic process \( \{\xi_1, \xi_2, \ldots, \xi_T\} \) or a probability distribution associated to \( \{\xi_T\} \) with a discrete scenario process represented by a tree structure.

**Definition 10** Scenario trees are acyclic directed graphs comprising nodes and leafs (i.e., nodes without descendants) that represent discrete conditional realizations of a particular random \( n \)-vector at a given decision stage \( t \). Nodes can have neither siblings nor more than one parent. Consequently, scenario trees have only a single root node. Each node is connected to its descendants by arcs (or branches) and each arc is associated to a conditional probability of occurrence. Each possible scenario path (or trajectory) from the root node to a leaf node is also associated to a (unconditional) probability, which is given by the product of the arc probabilities pertaining to that path. The number of stages and the branching scheme determines the topology of the scenario tree. The \( n \)-vector values of the root node, which defines the beginning of the optimization horizon, are assumed to be known.

For instance, following Pflug and Pichler [246] notation, let’s assume a particular scenario tree with \( N + 1 = 11 \) nodes labelled \( \{0, \ldots, N\} \) and \( T=2 \) stages \( \{\mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2\} \) such that:
Figure 2 – Scenario Tree Structure Example

Note that each scenario corresponds to a specific scenario path (or trajectory):

\[ \omega_1 \rightarrow \text{Path 0, 1 and 4} \]
\[ \omega_2 \rightarrow \text{Path 0, 1 and 5} \]
\[ \omega_3 \rightarrow \text{Path 0, 2 and 6} \]
\[ \omega_4 \rightarrow \text{Path 0, 2 and 7} \]
\[ \omega_5 \rightarrow \text{Path 0, 2 and 8} \]
\[ \omega_6 \rightarrow \text{Path 0, 3 and 9} \]
\[ \omega_7 \rightarrow \text{Path 0, 3 and 10} \]

In this example, besides the root node stage \( N_0 = \{0\} \) we have two other stages: \( N_1 = \{1,2,3\} \) and \( N_2 = \{4,5,6,7,8,9,10\} \). Moreover, for a given node \( n \in \{0, \ldots, N\} \), we define that:

\[ n \in N_t \Rightarrow \text{pred}(n) = n^- \in N_{t-1} \text{ (parent node)} \quad (7) \]
\[ n \in N_t, m \in N_{t+1}, \text{pred}(m) = n \Rightarrow m \in n^+ \text{ (children nodes)} \quad (8) \]
\[ n \in \mathcal{N}_t, m \in \mathcal{N}_s, s < t \Rightarrow \text{pred}_s(n) = m \Rightarrow n \succ m \text{ (all descendant nodes)} \quad (9) \]

Each leaf node \( n \in \mathcal{N}_f \) is associated with a possible scenario \( \omega_i \in \mathbb{R}^k, i = 1, \ldots, 7 \), which has a unconditional probability \( P(n) = P(\omega_i) \). Consequently, each non-leaf node \( m \in \mathcal{N}_t, t < T \) has probability:

\[
P(m) = \sum_{n > m} P(n) \quad (10)
\]

Moreover, given the conditional probability \( P(n|n^-) \), between nodes \( n \) and \( n^- \) (its parent node), we have:

\[
P(n|n^-) = \frac{P(n, n^-)}{P(n^-)} = \frac{P(n)}{P(n^-)} \quad (11)
\]

\[
P(m) = P(n|n^-) \prod_{n > m} P(m|m^-), n \in \mathcal{N}_r, n \succ m \quad (12)
\]

\[
P(m = 0) = 1 \text{ (root node is known)} \quad (13)
\]

Many simulating approaches applied in financial applications, as the one used in this thesis for instance, often generate a scenario fan, which is a particular case of a general scenario tree, rather than a traditional scenario tree. They are usually easier to generate but less efficient (in the sense that they require too many nodes). Once a scenario fan is simulated, we can approximate it by a smaller scenario tree using optimal discretization, as we discuss in the next section.

**Definition 11** A scenario fan is a scenario tree whose nodes (except for the leaf nodes) have a single descendant node.
A key aspect of any scenario tree structure is that its topology governs how information is revealed at each stage. Consequently, the tree is implicitly linked to the natural filtration associated to the underlying stochastic process and, as pointed out by Pflug and Pichler [246], it can be assumed to be equivalent to the pertaining filtered probability space.

**Definition 12** Given a probability space $(\Omega, \mathcal{F}, P)$, an increasing sequence of $\sigma$-algebras $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_T)$ on $\Omega$ is a natural filtration if $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$ for $t = 0, 1, \ldots, T - 1$. A filtration represents the evolution of the available information up to time $t$. Additionally, the probability space $(\Omega, \mathcal{F}, P)$ endowed with a filtration $\mathcal{F}$ is called a filtered probability space.

**Definition 13** A stochastic process $X = (X_1, X_2, \ldots, X_T)$ on $(\Omega, \mathcal{F}, P)$ is said to be adapted to the filtration $\mathcal{F}$ if $X_t$ is $\mathcal{F}_t$-measurable for $t \geq 0$. It means that the stochastic process at time $t$ depends only on the past information imposed by $\mathcal{F}_t$. This property is known as a nonanticipativity constraint. Moreover, any stochastic process $X$ is adapted to its natural filtration (i.e., its history process).
For instance, in the previous example suppose that the σ-algebras $\mathcal{F}_\ell$ are generated by the possible scenarios (i.e., atoms) associated with the nodes pertaining to each stage of the tree. Consequently, the filtration $\mathcal{F} = (\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2)$ comprises:

$$\mathcal{F}_0 = \sigma(\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7\}) = \sigma(\Omega)$$

$$\mathcal{F}_1 = \sigma(\{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5\}, \{\omega_6, \omega_7\})$$

$$\mathcal{F}_2 = \sigma(\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, \{\omega_6\}, \{\omega_7\}) = \mathcal{P}(\Omega)$$

As $t$ increases, additional information is revealed according the structure imposed by the scenario tree and the natural filtration of the underlying stochastic process. There is one-to-one relationship between them.

For multistage problems in particular, this information structure is because crucial. Although two stochastic process may exhibit the same discrete distribution at the final stage (i.e., the same scenarios and unconditional probabilities), they might not reveal information in the same way, which may lead to different optimal solutions. To
see that, let us consider another example. Suppose we have the following scenario
trees A and B:

![Scenario Tree Comparison - Tree A](image1)

![Scenario Tree Comparison - Tree B](image2)

**Figure 5 – Scenario Tree Comparison – Tree A**

**Figure 6 – Scenario Tree Comparison – Tree B**
Both scenario trees lead to the same multivariate discrete distribution (i.e., same realizations with identical unconditional probabilities) at the final stage $T = 2$. However, the underlying stochastic processes related to trees A and B are different. To see that, just consider the node $n = 1$ in both scenario trees. On scenario tree A, node 1 leads to one single outcome and, consequently, there is no uncertainty regarding the expected scenario at the stage $T$ (i.e., the realized scenario is acknowledged before stage $T$). Conversely, on scenario tree B, node 1 leads to three possible scenarios. Therefore, the realized scenario is not known until an event (conditional on node 1) at stage $T$ materializes. In that case, the optimal solutions for the underlying stochastic processes represented by scenario trees A and B are likely to be different (depending on values sitting on each node). Hence, any suitable optimal discretization process should take into account the (dis)similarity between the information structures (i.e., the natural filtrations) pertaining to the original and approximated scenarios (i.e., discrete stochastic processes).

Another famous example in the literature highlighting the importance of the information structure (i.e., the way information is revealed as the process evolves) involves a fair coin tossed three times. In this example, one dollar is paid if (A) head appears at least two times or if (B) head appears at the last toss. The following scenario trees represent these payoffs.
Figure 7 – Scenario Tree for Coin Bet A

Figure 8 – Scenario Tree for Coin Bet B
After the last toss, the payoff distributions for both bets are equal but that the payoff for bet A can be anticipated (i.e., can be revealed before the last toss) for some scenarios. Therefore, earlier tosses might provide additional (and relevant) information for bet A. Accordingly, the information structures of these bets are not the same and any optimal decision based on them should consider this dissimilarity.

These examples show that the evolution of information disclosure is paramount to any decision process. Obviously, we also require the content (i.e., the values) of the information disclosed since decisions are based on both the possible values and their (conditional) probabilities.

Fortunately, a scenario tree can describe both the information process and the value process. The information (or tree) process $Y$, related to the filtration $\mathcal{F}$, corresponds to the tree structure that establishes how information is revealed. The discrete value process $X$ corresponds to the value or vectors sitting in the tree nodes that comprise the revealed content. The pair $(X,Y)$ is called (scenario) value-and-information pair.

**Definition 14** A tree process is a stochastic process $Y = (Y_1,Y_2,\cdots,Y_T)$ with values on some (Polish) state space $\mathcal{N}_t$, $t = 0,\cdots,T$, such that:

(i) $\mathcal{N}_0$ is a singleton

(ii) $\mathcal{N}_t$ are pairwise disjoint

(iii) $\sigma(Y_t) = \sigma(Y_0,\cdots,Y_t), t = 0,\cdots,T$

(iv) $\mathcal{F} = \sigma(Y) = (\sigma(Y_0),\sigma(Y_1),\cdots,\sigma(Y_T))$ is a filtration

A tree process $Y = (Y_1,Y_2,\cdots,Y_T)$ can be constructed using the history process of any stochastic process $X = (X_1,X_2,\cdots,X_T)$:

$Y_0 = X_0$

$Y_1 = (X_0,X_1)$

$\vdots$

$Y_T = (X_0,X_1,\cdots,X_T)$

Assuming that all state spaces $\mathcal{N}_t$ of the tree process $Y$ are Polish and given general Borel sets $\mathcal{A}_t$, $t = 1,\cdots,T$, Shiryaev [283] demonstrates that we can compute
the joint distribution $P$ of the tree process $Y = (Y_1, Y_2, \cdots, Y_T)$ using its conditional probabilities:

$$P_1(A_1) = P(Y_1 \in A_1)$$

$$P_2(A_2|B_1) = P(Y_2 \in A_2|Y_1 = B_1)$$

$$\vdots$$

$$P_t(A_t|B_1, \cdots, B_{t-1}) = P(Y_t \in A_t|Y_1 = B_1, Y_2 = B_2, \cdots, Y_{t-1} = B_{t-1})$$

In addition, Shiryaev [283] also shows that given a filtration $\mathcal{F} = \sigma(Y)$, a stochastic process $X$ adapted to $\mathcal{F}$ and a measurable function $f_t$, we have:

$$X_t = f_t(Y_t)$$

(14)

Finally, a tree process induces a probability distribution $P$ on $\mathcal{N}_t$. Therefore, if we choose $\Omega = \mathcal{N}_T$, we get a filtration $\mathcal{F} = \sigma(Y) = (\sigma(Y_0), \cdots, \sigma(Y_T)) = (\mathcal{F}_0, \cdots, \mathcal{F}_T)$ on $\Omega$, which leads to a filtered probability space $(\Omega, \mathcal{F}, P)$.

**Definition 15** Given two tree processes $Y$ and $\tilde{Y}$ defined on possibly different filtered probability spaces $(\mathcal{N}, \mathcal{F}, P)$ and $(\tilde{\mathcal{N}}, \tilde{\mathcal{F}}, \tilde{P})$, Borel sets $A_t, t = 1, \cdots, T$ and conditional probabilities such that:

$$P_t(A_t|B_1, \cdots, B_{t-1}) = P(Y_t \in A_t|Y_1 = B_1, Y_2 = B_2, \cdots, Y_{t-1} = B_{t-1}), t = 1, \cdots, T$$

(15)

$$\tilde{P}_t(A_t|B_1, \cdots, B_{t-1}) = \tilde{P}(Y_t \in A_t|Y_1 = B_1, Y_2 = B_2, \cdots, Y_{t-1} = B_{t-1}), t = 1, \cdots, T$$

(16)

The tree processes are equivalent if:

(i) There are bijective functions $k_1, \cdots, k_T$ mapping the state spaces $\mathcal{N}_t$ of $Y_t$ to the state spaces $\tilde{\mathcal{N}}_t$ of $\tilde{Y}_t$ such that the processes $(Y_1, \cdots, Y_T)$ and $(k_1^{-1}(\tilde{Y}_1), \cdots, k_T^{-1}(\tilde{Y}_T))$ have the same distribution.

(ii) $\tilde{P}$ is the pushforward measure of $P$ for the measurable functions $k_t$.

$$\tilde{P}_t(A_t) = P_{1}(k_1^{-1}(\tilde{A}_1))$$

$$\tilde{P}_t(A_t|B_1, \cdots, B_{t-1}) = P_{t}(k_t^{-1}(\tilde{A}_t)|k_1^{-1}(\tilde{B}_1), \cdots, k_{t-1}^{-1}(\tilde{B}_{t-1})), t = 2, \cdots, T$$
Finally,

**Definition 16** Given and tree process $Y$ and a value process $X$ such that $X$ is adapted to the filtration $\mathcal{F}$ generated by $\sigma(Y)$, the pair $(X,Y)$ is called scenario value-and-information pair.

**Definition 17** Given two scenario value-and-information pairs $(X,Y)$ and $(\tilde{X},\tilde{Y})$ such that $X_t = f_t(Y_t)$ and $\tilde{X}_t = \tilde{f}_t(\tilde{Y}_t)$, the pairs are equivalent if:

(i) Tree processes $Y$ and $\tilde{Y}$ are equivalent (in the sense of Definition 15).

(ii) $\tilde{f}_t = f_t \circ k_t^{-1}$ a.s. for the equivalent mappings $k_t$.

These definitions indicate that scenario value-and-information pairs and scenario trees are interchangeable. Consequently, given a multi-period stochastic optimization problem and two equivalent scenario value-and-information pairs, we should achieve similar solutions.

In addition, as we discuss later, a scenario value-and-information pair can also be described by a nested distribution, a concept proposed by Pflug [243] and further discussed by Pflug and Pichler [245]. The nested distribution and the related process distance concepts are particularly useful when dealing with approximations of stochastic processes rather than probability measures (i.e., approximations in multi-period settings rather single-period settings) because they do not require two separate
measures of dissimilarity (i.e., one for distributions and another for filtrations) to govern the quality of the approximation.

However, before we delve into how we can perform such multi-period approximations using process distances, in the next subsections we cover the theoretical background required to simulate a (usually large) multivariate scenario tree that properly describes real financial series. They discuss how to (i) elaborate and calibrate a combined econometric model capable of simulating the four major financial stylized facts discussed earlier: heavy-tails, volatility clustering, leverage effect and tail co-dependence, and (ii) generate a large set of multivariate trajectories of log returns, represented by a scenario fan, based on this combined econometric model.

3.1.1 GARCH Models

As we discussed in the previous section, Generalized Autoregressive Conditional Heteroscedastic (GARCH) models are particularly appealing because they can explain reasonably well key stylized facts of financial series like volatility clustering, leverage effects and, to some extent, heavy-tails.

Despite of some quite interesting nonlinear and regime-switching extensions proposed recently, in this study we focus on simpler asymmetric GARCH models that are more tractable and still can properly accommodate these key statistical features.

A simple GARCH (p,q) process entails:

\[
\varepsilon_t = r_t - E_{t-1}[r_t] \\
\varepsilon_t = z_t \sigma_t \tag{17}
\]

\[
E[\varepsilon_t] = 0, E[\varepsilon_t \varepsilon_{t-s}] = 0 \text{ for } s \neq 0
\]

\[
\sigma_t^2 = a_0 + \sum_{j=1}^{q} a_j \varepsilon_{t-j}^2 + \sum_{i=1}^{p} b_i \sigma_{t-i}^2 \tag{18}
\]

\[
a_0 > 0, a_j \geq 0, j = 1 \cdots q, b_i \geq 0, i = 1 \cdots p
\]

\[
\sum_{j=1}^{q} a_j + \sum_{i=1}^{p} b_i < 1
\]
Where \( \{r_t\} \) is a series of daily log returns, \( \{z_t\} \) is a white noise sequence and \( \sigma_t^2 \) is the conditional variance at time \( t \). The restrictions in equation (19) ensure that the conditional variance is positive and that the unconditional variance is finite:

\[
Var(\varepsilon_t) = E[\varepsilon_t^2] = \frac{a_0}{1 - (\sum_{j=1}^{q} a_j + \sum_{i=1}^{p} b_i)}
\]

(20)

The conditional mean \( E_{t-1}[y_t] \) is usually specified as a low order ARMA model. For daily log returns in particular, empirical evidence suggests that serial correlation in raw returns is often statistically insignificant leading to no relevant autoregressive term in the conditional mean specification.

Using a basic GARCH (1,1) formulation as a reference:

\[
\sigma_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2
\]

(21)

\[
a_0 > 0, \quad a_1 > 0, \quad b_1 > 0
\]

\[
a_1 + b_1 < 1
\]

We can see how volatility clustering and, to some extent, heavy-tails can be explained. According to Bauwens, Hafner and Laurent [20], most estimates of the coefficients \( a_1 \) and \( b_1 \) lie, respectively, in the ranges [0.02 0.05] and [0.75, 0.98]. We notice that for these values of \( a_1 \) and \( b_1 \), a large value \( \sigma_{t-1}^2 \) prompts a large value \( \sigma_t^2 \) and a small value of \( \sigma_{t-1}^2 \) induces a small value \( \sigma_t^2 \), provided \( \varepsilon_{t-1}^2 \) is not too large. Consequently, the volatility clustering observed in daily financial returns can be reasonably explained by estimating appropriate values for \( a_1 \) and \( b_1 \). Besides, provided

\[
1 - 2a_1^2 - (a_1 + b_1)^2 > 0
\]

and assuming that \( z_t \) follows a Gaussian distribution, He [131] shows that

\[
\kappa_4 = \frac{E[\varepsilon_t^4]}{(E[\varepsilon_t^2])^2} = 3 \left\{ \frac{1 - (a_1 + b_1)^2}{1 - (a_1 + b_1)^2 - 2a_1^2} \right\} > 3
\]

(22)

That means that a GARCH(1,1) process also induces a leptokurtic unconditional distribution, i.e., an unconditional distribution that exhibits heavier tails than those suggested by a Gaussian distribution. Unfortunately, assuming Gaussian innovations, the induced unconditional kurtosis is not large enough to fully explain the tail behaviour of daily financial returns. There are, however, alternatives: (i) we can assume innovations with Student t-distributions, as recommended by Bollerslev [35], or Generalized Error Distributions (GED), as suggested by Nelson [232] or; (ii) we can
combine GARCH models with Extreme Value Theory (EVT) as proposed by McNeil and Frey [213]. We shall explore the latter alternative in the next section.

Parameters of GARCH models are often estimated using Maximum-Likelihood Estimation (MLE) or Quasi-Maximum-Likelihood Estimation (QMLE), depending on the choice of the innovation distribution. Unfortunately, as stressed in Andersen et al. [7], GARCH log-likelihood functions are often ill-behaved and poor choices of starting values can lead to convergence problems. Moreover, constraints such as positive variance can compromise optimization performance. As a consequence, simpler specifications are often preferred in practical applications.

These simpler specifications comprise, for example, early formulations of asymmetric GARCH models like the EGARCH and GJR-GARCH models. They are also widely used because they can explain a third important stylized fact: the leverage effect.

The EGARCH (p,q) process substitutes equation (19) for

$$\ln \sigma_t^2 = a_0 + \sum_{j=1}^{q} \left( a_j \left( \frac{\varepsilon_{t-j}}{\sigma_{t-j}} \right) - E \left( \frac{\varepsilon_{t-j}}{\sigma_{t-j}} \right) \right) + \sum_{i=1}^{p} b_i \ln \sigma_{t-i}^2 \quad (23)$$

The additional standardized terms can capture potential asymmetric impacts of innovations. For instance, if $d_i$ is negative, bad news has a larger impact on volatility. Besides, the logarithmic transformation guarantees positive variance and relaxes the positive constraint requirements. On the other hand, EGARCH processes might lead to infinite variance under non-Gaussian innovations (for example, with Student-t innovations) as underlined by Nelson [232]. In addition, conditional variance’s forecasts are biased according Jensen’s inequality

$$E[\sigma_t^2] \geq e^{E[\ln \sigma_t^2]}$$

On the other hand, the GJR-GARCH formulation substitutes equation (19) for

$$\sigma_t^2 = a_0 + \sum_{j=1}^{q} a_j \varepsilon_{t-j}^2 + \sum_{i=1}^{p} b_i \sigma_{t-i}^2 + \sum_{j=1}^{q} d_j I(\varepsilon_{t-j} < 0) \varepsilon_{t-j}^2 \quad (24)$$

$$a_0 > 0, \; a_j \geq 0, \; d_j \geq 0, \; j = 1 \ldots q, \; b_i \geq 0, \; i = 1 \ldots p$$

Where $I(.)$ is an indicator function such that
\[ I(\varepsilon_{t-j} < 0) = \begin{cases} 1 & \text{if } \varepsilon_{t-j} < 0 \\ 0 & \text{if } \varepsilon_{t-j} \geq 0 \end{cases} \quad (25) \]

Notice that potential asymmetric impacts on the conditional variance are explained by the last term. In this specification, we need to impose restrictions on the coefficients to ensure a positive conditional variance. Moreover, GJR-GARCH processes have more stringent stationarity conditions than EGARCH process, as discussed by Ling and McAleer [192].

For example, a GJR-GARCH (1,1) process is stationary and has finite variance provided

\[ a_1 + b_1 + \frac{d_1}{2} < 1 \quad (26) \]

Moreover, the fourth moment exists if

\[ b_1^2 + 2a_1b_1 + \kappa_2a_2^2 + b_1d_1 + \kappa_2 \left( a_1d_1 + \frac{d_1^2}{2} \right) < 1 \quad (27) \]

\[ \kappa_2 = \begin{cases} 3 & \text{if } z_t \sim N(0,1) \\ \frac{\nu - 2}{\nu - 4} & \text{if } z_t \sim t(0,1, \nu) \end{cases} \]

To the best of our knowledge, there are few studies comparing performance of EGARCH and GJR-GARCH models. Engle and Ng [96], and Peters [240] recommend the GJR-GARCH over the EGARCH specification while Alberg, Shalit and Yosef [1] suggests otherwise. Without loss of generality, we follow the former and focus on the GJR-GARCH formulation in our analysis.

### 3.1.2 Extreme Value Theory

Although some models of the GARCH family can adequately emulate important key stylized facts such as volatility clustering and leverage effects, they are often still unable to fully model heavy-tailed distributions. To accomplish that, an additional technique such as Extreme Value Theory (EVT) may be used to handle tails properly, where very few observations are usually available.
EVT encompasses two major approaches to model extreme returns: Block Maxima (BM) and Peaks over Threshold (POT) that use, respectively, two fundamental results from EVT. We briefly discuss both approaches next.

**Block Maxima Method**

The central result underlying block maxima methods is the Fisher-Tippett-Gnedenko theorem, initially proposed by Fisher and Tippett [104], and fully proved by Gnedenko [114]. It shows that, under certain conditions, the asymptotic distribution for the series of maxima (minima) converges to the Weibull, Gumbel or Fréchet distributions, independently of the distribution of the series.

**Theorem 1** *(Fisher-Tippett-Gnedenko theorem)* Let \( \{r_1, \ldots, r_n\} \) be a sequence of independent and identically distributed random variables with common distribution function \( F \), \( M_n = \max_{i \leq n} r_i \) the maximum order statistic and \( P(M_n \leq x) = P(r_1 \leq x, \ldots, r_n \leq x) = [F(x)]^n \). If there exist constants \( \sigma_n, \mu_n \in \mathbb{R}, \sigma_n > 0 \) and a non-degenerate distribution function \( H \) such that:

\[
\lim_{n \to \infty} P\left( \frac{M_n - \mu_n}{\sigma_n} \leq x \right) = \lim_{n \to \infty} \left[ F(\sigma_n x + \mu_n) \right]^n = H(x) \tag{28}
\]

Then \( H \) converges in distribution to one of the three limiting families:

**Gumbel (or Type I):** \( H(x) = e^{-e^{-x}}, -\infty < x < +\infty \) \( \tag{29} \)

**Fréchet (or Type II):** \( H(x) = \begin{cases} e^{-\left(1 + \xi x\right)^{-\frac{1}{\xi}}} & \text{if } x > -\frac{1}{\xi} \\ 0 & \text{otherwise} \end{cases} \) \( \tag{30} \)

**Weibull (or Type III):** \( H(x) = \begin{cases} e^{-\left(1 + \xi x\right)^{-\frac{1}{\xi}}} & \text{if } x < -\frac{1}{\xi} \\ 1 & \text{otherwise} \end{cases} \) \( \tag{31} \)

In other words, the Fisher-Tippett-Gnedenko theorem shows that if \( F \) belongs to the maximum domain of attraction (MDA) of \( H \) for some non-degenerate distribution function \( H \) then \( H \) must converge weakly to one of these distributions of properly normalized maxima. Moreover, these limiting distributions are particular cases of the Generalized Extreme Value (GEV) distribution (Figure 10), introduced by von Mises [221] and Jenkinson [157].
The shape parameter $\xi$ determines the corresponding limiting distribution for the tail and, therefore, its thickness. It is also common to refer to the tail behaviour using the *tail index* $\alpha = \xi^{-1}$. If $\xi < 0$, $H_\xi$ converges to the Weibull family that includes distributions with finite support like the beta distribution. If $\xi = 0$, $H_\xi$ converges to the Gumbel family that encompasses distributions like the normal, log-normal and hyperbolic whose right tails exhibit an exponential decay (i.e., thin-tailed behaviour). Finally, if $\xi > 0$, $H_\xi$ converges to the Fréchet family that comprehends, for example, Student t, Pareto and stable distributions whose right tails follow a polynomial decay (i.e., heavy-tailed behaviour).

Given that daily financial series of returns typically display heavy tails as empirical research suggests, the Fréchet distribution seems a suitable candidate to model the asymptotic behaviour of their tails, even if we do not know, as is usually the case, the actual distributions of the financial series.

As we do not know *a priori* the normalized maxima $x$ required by $H_\xi(x)$, we need to identify, besides the shape parameter $\xi$, the scale $\sigma$ and location $\mu$ parameters. Differently from the former, the latter two often depend on the distribution of the financial series. Therefore, we use an alternative definition for the GEV distribution:

$$H_\xi(x) = \begin{cases} e^{-(1+\xi x)^{-\frac{1}{\xi}}} & \text{if } \xi \neq 0 \text{ such that } 1 + \xi x > 0 \\ e^{-e^{-x}} & \text{if } \xi = 0 \end{cases}$$
\[ H_\xi(y) = H_{\xi, \sigma, \mu}(x) = \begin{cases} e^{-\left(1 + \frac{\xi x - \mu}{\sigma}\right)^{-\frac{1}{\xi}}} & \text{if } \xi \neq 0 \text{ such that } 1 + \frac{\xi x - \mu}{\sigma} > 0 \\ e^{-\frac{x - \mu}{\sigma}} & \text{if } \xi = 0 \end{cases} \]  

(33)

Then, we can apply a parametric approach like maximum-likelihood to estimate these parameters as suggested by Embrechts, Klüppelberg and Mikosch [90] and Tsay [301]. First, we divide the sample in \( m \) non-overlapping subperiods of length \( k \). Then we select the maximum return in each subperiod and construct a sequence of maximal returns \( \{M_i\}_{i=1}^m \), assumed to have a distribution. Supposing that the sequence of maximal returns is independent and identically distributed, the likelihood function is:

\[ l(M_1, \ldots, M_m | \xi, \sigma, \mu) = \prod_{i=1}^m h_{\xi, \sigma, \mu}(M_i) \]  

(34)

\[ h_{\xi, \sigma, \mu}(M_i) = \begin{cases} \frac{1}{\sigma} \left[ 1 + \frac{\xi - \frac{M_i - \mu}{\sigma}}{\sigma} \right]^{-\frac{1+\xi}{\xi}} e^{-\left(1 + \frac{\xi}{\sigma}\right)^{-\frac{1}{\xi}}} & \text{if } \xi \neq 0 \\ \frac{1}{\sigma} e^{-\frac{M_i - \mu}{\sigma}} & \text{if } \xi = 0 \end{cases} \]  

(35)

And, consequently, the log-likelihood function is:

\[ L(M_1, \ldots, M_m | \xi, \sigma, \mu) = \sum_{i=1}^m \ln[h_{\xi, \sigma, \mu}(M_i)] \]  

(36)

\[ h_{\xi, \sigma, \mu}(M_i) = \begin{cases} -\ln \sigma - \frac{1 + \xi}{\xi} \ln \left[ 1 + \frac{\xi (M_i - \mu)}{\sigma} \right] - \left[ 1 + \frac{\xi (M_i - \mu)}{\sigma} \right]^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ -\ln \sigma - \frac{M_i - \mu}{\sigma} - e^{-\frac{M_i - \mu}{\sigma}} & \text{if } \xi = 0 \end{cases} \]  

(37)

Embrechts, Klüppelberg and Mikosch [90] and Hosking [143] discuss potential numerical procedures to find the maximum-likelihood estimates for the shape, scale and location parameters of the GEV distribution. Based on Smith [284], we can show that these estimates are consistency and asymptotic efficiency as long as \( \xi > -1/2 \).

Although straightforward, the previous parametric approach relies on the strong simplifying hypothesis that extreme returns follow precisely a GEV distribution. A more realistic approach, for example, would be to assume that extreme returns follow approximately a GEV distribution. Unfortunately, this alternative semi-parametric
method is more involved. We refer to Embrechts, Klüppelberg and Mikosch [90] for further details regarding this approach.

The traditional extreme value approach based on the limiting GEV distribution is rather simple and extensively used but it presents some shortcomings, as pointed out by Embrechts [89], Tsay [301] and Gilli and Këllezi [113]. First, it does not use all available data efficiently, i.e., it uses only maxima from (possibly sizable) subsamples. Second, estimation results might vary greatly depending on the number of subperiods \( m \) and the subperiod sample size \( k \) and it is not clear how to identify, supposing it exists, a suitable trade-off between \( m \) and \( k \). Lastly, the Fisher-Tippett-Gnedenko theorem holds if financial series are (almost) independent and identically distributed or if the sample size \( k \) is excessively large so that the block maxima sequence can be assumed to be independent. So we need to handle these limitations before we can apply this approach to financial series. The technique discussed next avoids the first and second problems while the last one is handled in a later section, where we combine GARCH, EVT and copula methods to properly simulate daily multivariate return series.

**Peaks over Thresholds Method**

As an alternative to block maxima approach, Davison and Smith [63] and Smith [285] introduced an alternative approach focused on the exceedances beyond a given (large) threshold, also known as Peaks over Threshold (POT), that does not exhibit some of the shortcomings of the traditional approach.

Rather than focusing on the distribution function \( F(x) \) of the independent and identically distributed returns, this method focuses on the conditional excess distribution \( F_u(y) \) over a sufficiently large threshold \( u \) (i.e., on the exceedances distribution) where \( y = x - u, x > u \). \( F_u(y) \) is given by:

\[
F_u(y) = P(X \leq y + u | X > u) = \frac{P(X \leq y + u, X > u)}{P(X > u)} \Rightarrow \\
F_u(y) = \frac{P(u < X \leq y + u)}{1 - P(X \leq u)} \\
F_u(y) = \frac{F(y + u) - F(u)}{1 - F(u)} 
\]

(38)

Or, in terms of the tail’s distribution \( \bar{F} = 1 - F \):
\[ P(X > y + u | X > u) = 1 - F_u(y) = 1 - \frac{F(y + u) - F(u)}{1 - F(u)} \]

\[ 1 - F_u(y) = \frac{1 - F(u) - F(y + u) + F(u)}{1 - F(u)} \Rightarrow \]

\[ 1 - F_u(y) = \frac{1 - F(y + u)}{1 - F(u)} \Rightarrow \]

\[ \bar{F}_u(y) = \frac{\bar{F}(y + u)}{\bar{F}(u)} \Rightarrow \]

\[ \bar{F}(y + u) = \bar{F}(u) \bar{F}_u(y) \quad (39) \]

So, we can estimate the tail distribution \( \bar{F}(y + u) \) by individually estimating \( \bar{F}(u) \) and \( \bar{F}_u(y) \). The fundamental result underlying the estimation of \( \bar{F}_u(y) \) was proposed by Pickands [249], and Balkema and de Haan [14]. Basically, it shows that under certain conditions the conditional distribution \( F_u(x) \) converges to a Generalized Pareto Distribution (GPD) (Figure 11) where

\[
G_{\xi, \sigma(u)}(y) = \begin{cases} 
1 - \left(1 + \xi \frac{y}{\sigma(u)}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\
1 - e^{-\frac{y}{\sigma(u)}} & \text{if } \xi = 0
\end{cases} \quad \text{where } \beta > 0 \quad (40)
\]

Figure 11 – Generalized Pareto Distributions

With support \( x \geq \mu \) when \( \xi \geq 0 \) and \( 0 < x \leq -\frac{\sigma(u)}{\xi} \) when \( \xi < 0 \). Similarly to GEV distribution, the GPD can be defined by a shape parameter \( \xi \), a scale parameter \( \sigma \) and a location parameter \( \mu \) (here assumed to be zero).
Theorem 2 (Pickands-Balkema-de Haan theorem) Let \( \{r_1, \ldots, r_n\} \) be a sequence of independent and identically distributed random variables with common distribution function \( F \) and common conditional distribution \( F_u \) over a sufficiently large threshold \( u \). If \( F \in MDA(H_\xi) \) there is a positive measurable function \( \sigma(u) \) such that:

\[
\lim_{u \to x_F} \sup_{0 \leq x < x_F - u} |F_u(x) - G(x)\sigma(u)| = 0 \quad (41)
\]

Where \( x_F \) is the right endpoint (finite or infinite) of the distribution \( F \).

We observe that, while the GEV distribution \( H_{\xi, \sigma, \mu} \) describes the limiting distributions of normalized maxima of a distribution \( F \) that belongs to the \( MDA(H_\xi) \), the GPD \( G_{\xi, \sigma(u)} \) describes the respective limiting distribution of exceedances beyond a given sufficiently large threshold. Incidentally, the shape parameter \( \xi \) is the same for both distributions.

If \( \xi < 0 \), \( G_{\xi, \sigma(u)} \) converges to distributions whose right tails are finite like the beta distribution. If \( \xi = 0 \), \( G_{\xi, \sigma(u)} \) converges to distributions whose right tails exhibit an exponential decay like the normal. Lastly, if \( \xi > 0 \), \( G_{\xi, \sigma(u)} \) converges to distributions whose right tails follow a polynomial decay like the Student t or Pareto distributions. Consequently, a GPD with positive shape parameter \( \xi \) seems a likely candidate to model the asymptotic behaviour of the conditional tails of daily return series.

To find \( \tilde{F}_u(y) \), we start by applying Theorem 1:

\[
\lim_{n \to \infty} [F(\sigma_n x + \mu_n)]^n = H_\xi(x) \Rightarrow \\
\lim_{n \to \infty} [1 - \tilde{F}_n(\sigma_n x + \mu_n)]^n = H_\xi(x) \Rightarrow \\
\lim_{n \to \infty} n \ln [1 - \tilde{F}_n(\sigma_n x + \mu_n)] = \ln H_\xi(x) \quad (42)
\]

Defining a sequence of thresholds \( u_n(x) = \sigma_n x + \mu_n \) and assuming, for large values of \( u_n(x) \) such that \( \tilde{F}_n(u_n(x)) \approx 1 \), a first-order Taylor series approximation:

\[
\ln [1 - \tilde{F}_n(u_n(x))] \approx -\tilde{F}_n(u_n(x)) \Rightarrow \\
\lim_{n \to \infty} n \ln [1 - \tilde{F}_n(\sigma_n x + \mu_n)] \approx \lim_{n \to \infty} n \tilde{F}_n(u_n(x)) \approx \ln H_\xi(x) \Rightarrow \\
\lim_{n \to \infty} n \tilde{F}_n(u_n(x)) \approx -\ln H_\xi(x) \quad (43)
\]

Finally, for \( n \) sufficiently large:
\[ n\bar{F}(u) \approx -\ln H_\xi \left( \frac{u - \mu}{\sigma} \right) \Rightarrow \]
\[ \bar{F}(u) \approx -\frac{1}{n} \ln H_\xi \left( \frac{u - \mu}{\sigma} \right) \Rightarrow \]
\[ \bar{F}(u) \approx \begin{cases} \frac{-1}{n} \ln e^{-\left[1 + \xi \frac{u - \mu}{\sigma}\right]} & \text{if } \xi \neq 0 \\ \frac{-1}{n} \ln e^{-\frac{u - \mu}{\sigma}} & \text{if } \xi = 0 \end{cases} \]
\[ \bar{F}(u) \approx \begin{cases} \frac{1}{n} \left(1 + \xi \frac{u - \mu}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ \frac{1}{n} e^{-\frac{u - \mu}{\sigma}} & \text{if } \xi = 0 \end{cases} (44) \]
\[ \bar{F}(y + u) \approx \begin{cases} \frac{1}{n} \left(1 + \xi \frac{y + u - \mu}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ \frac{1}{n} e^{-\frac{y + u - \mu}{\sigma}} & \text{if } \xi = 0 \end{cases} (45) \]

Using these results, we can estimate \( \bar{F}_u(y) \) by recalling that:
\[ \bar{F}_u(y) = \frac{\bar{F}(y + u)}{\bar{F}(u)} \Rightarrow \]
\[ \bar{F}_u(y) = \frac{\frac{1}{n} \left(1 + \xi \frac{y + u - \mu}{\sigma}\right)^{-\frac{1}{\xi}}}{\frac{1}{n} \left(1 + \xi \frac{u - \mu}{\sigma}\right)^{-\frac{1}{\xi}}} \Rightarrow \]
\[ \bar{F}_u(y) = \left(1 + \xi \frac{y + u - \mu}{\sigma}\right)^{-\frac{1}{\xi}} \Rightarrow \]
\[ \bar{F}_u(y) = \left(\frac{\sigma + \xi(y + u - \mu)}{\sigma + \xi(u - \mu)}\right)^{-\frac{1}{\xi}} = \left(1 + \frac{\xi y}{\sigma + \xi(u - \mu)}\right)^{-\frac{1}{\xi}} \] (46)

Alternatively, we could find \( \bar{F}_u(x) \) based on Theorem 2:
\[ \lim_{u \to x_F} \sup_{0 \leq x < x_F - u} |F_u(x) - G_{\xi, \sigma(u)}(x)| = 0 \Rightarrow \]
\[ \lim_{u \to x_F} \sup_{0 \leq x < x_F - u} |1 - F_u(x) - 1 + G_{\xi, \sigma(u)}(x)| = 0 \Rightarrow \]
\[
\lim_{u \to \infty} \sup_{0 \leq x < x_F - u} |\bar{F}_u(x) - \bar{g}_{\xi, \sigma(u)}(x)| = 0
\]

As a result, for \( u \) sufficiently large, we can assume that:

\[
\bar{F}_u(x) \approx \bar{g}_{\xi, \sigma(u)}(x)
\]

Lastly, supposing that we have a sufficiently large number \( N_u \) of observations greater than the threshold \( u \), as proposed by Embrechts, Klüppelberg and Mikosch [90], we can consider the empirical approximation:

\[
\bar{F}(x) \approx \bar{F}_n(x) = \frac{N_u}{n}
\]

Using these results, we can estimate \( \bar{F}(y + u) \) by:

\[
\bar{F}(y + u) = \bar{F}_u(y)\bar{F}_u(y) \Rightarrow \\
\bar{F}(y + u) = \frac{N_u}{n} \left(1 + \xi \frac{y - u}{\sigma}\right)^{-\frac{1}{\xi}}
\]

Then, we can use a maximum-likelihood approach to estimate the shape and scale parameters \( \xi \) and \( \sigma \) of the tail distribution \( \bar{F}(y + u) \), as discussed by Embrechts, Klüppelberg and Mikosch [90] and Hosking and Wallis [146].

As pointed out by McNeil, Frey and Embrechts [92], the POT approach uses the data more efficiently than the BM approach but it still presents two major challenges: (i) appropriate choice of thresholds and (ii) unreliable results when return series exhibit serial correlation.

The first challenge refers to the trade-off between variance and bias as discussed by McNeil [211], and McNeil and Saladin [212]. A large threshold \( u \) means low bias because fewer observations outside the tail are selected to estimate the GPD parameters. Conversely, it also means higher variance because the tail itself contains fewer observations, which increases the estimates' variance. Castillo and Hadi [48] further emphasizes that the threshold \( u \) must select significantly large quantiles (preferably above 0.99) so we can apply the Pickands-Balkema-de Haan theorem.

The most two common approaches to choose suitable thresholds are the Hill estimator plot discussed by Hill [139], Resnick and Stărică [261] and the empirical mean excess function plot recommended by Embrechts, Klüppelberg and Mikosch [41].
Following McNeil, Frey and Embrechts [214], using the order statistics $X_{n,n} \leq \cdots \leq X_{1,n}$ and assuming a large sample size $n$ and sufficiently small number of extremes $k$, beyond the threshold $u$, we can employ the standard form of the Hill estimator:

$$\hat{\alpha}_{k,n} = \left( \frac{1}{k} \sum_{i=1}^{k} \ln(X_{i,n}) - \ln(X_{k,n}) \right)^{-1}, \ 2 \leq k \leq n \quad (51)$$

To plot estimates for different sets of order statistics (i.e., different values of $k$) and look for a stable region in the Hill estimator plot, where we can assume that $\hat{\alpha}^{-1} \sim \hat{\xi}$.

Similarly, we can also use the empirical mean excess function plot as described by McNeil, Frey and Embrechts [214] to identify appropriate thresholds. Embrechts, Klüppelberg and Mikosch [90] shows that a key property of the GPD is that the mean excess function is linear in the threshold $u$, provided $u$ is sufficiently large. Thus, we can use the empirical estimator of the mean excess function:

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u) I_{X_i > u}}{\sum_{i=1}^{n} I_{X_i > u}} \quad (52)$$

$$I_{X_i > u} = \begin{cases} 1, & \text{if } X_i > u \\ 0, & \text{if } X_i \leq u \end{cases}$$

To plot estimates of the empirical mean excess function for different values of $u$. If the distribution tail follows a GPD with positive shape parameter $\xi$, the plot should exhibit a linear upward trend for satisfactorily higher threshold values.

The second challenge requires that we handle dependence before employing the POT approach. Resnick and Stărică [262], for instance, suggests using standardized residuals from ARMA processes to mitigate this problem. McNeil and Frey [213] extends the idea by using the standardized residuals from GARCH processes. In the next section, we develop the latter to account for leverage effects and tail co-dependence.

Lastly, we can also extend the POT approach previously described to a point process framework, i.e., rather than handling the peaks over threshold and the occurrence times as separate processes, we can combine them into a point process of exceedances. We do not explore this alternative in this work and refer to Resnick [260] for further details.
This brief review suggests that EVT provides an elegant way to deal with univariate extreme returns. However, it is not easily extended to a multivariate framework because there is no longer a unique definition for an extreme events (i.e., we may entertain different concepts of order). Fortunately, as we discuss next copulas provide an interesting alternative to handle multivariate heavy-tailed distributions.

3.1.3 Copulas

Although Pearson’s correlation is likely one of the most applied measure of dependence in practical finance models, it is not always able to describe properly how close assets move together. Financial series often display higher (in absolute terms) dependence during turbulent periods than in calmer times. In other words, linear measures such as Pearson’s correlation cannot deal appropriately with the nonlinear dependence structures exhibited by financial markets, particularly in the tails of the multivariate distribution.

As we mentioned earlier some elaborated multivariate GARCH models, like the DCC proposed by Engle [94] and TVC suggested by Tse and Tsui [302], are capable of accounting for time-varying dependence. However, they often impose restrictions that may reduce the flexibility to accommodate stylized facts of individual series.

An alternative to explain and, more interestingly, simulate the dependence structure observed among financial series are copulas, proposed by Sklar [282]. In basic terms, copula are links between multivariate distributions and their marginal distributions. They provide greater flexibility to (i) model particular properties of each univariate series and (ii) select a suitable dependence structure for their joint behaviour. Besides, they also offer a straightforward method to simulate draws from joint distributions, once we estimate the marginal distributions and copulas’ parameters.

According Nelsen [231] and McNeil, Frey and Embrechts [214], an n-dimensional copula is simply a multivariate uniform distribution $C(u_1,\ldots,u_n)$ such that $C: [0,1]^n \rightarrow [0,1]$, which exhibits the following properties:

(i) $C(u_1,\ldots,u_n)$ is grounded and n-increasing for all $u_i \in [0,1], i = 1,\ldots,n$. 
(ii) \( C(1, \ldots, u_i, \ldots, 1) = u_i \) for all \( u_i \in [0,1] \), \( i = 1, \ldots, n \).

(iii) \( C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_n) = 0 \) for all \( u_i \in [0,1] \), \( i = 1, \ldots, n \).

In addition, the key result underlying the copula approach is the Sklar theorem, which let us decouple the modelling of the marginal distributions from the dependence structure. It states that we can either extract a copula from a multivariate distribution function or combine marginal distributions to a copula to construct a multivariate distribution function.

**Theorem 3** (Sklar theorem) Let \( F(F_1, \ldots, F_n) \) be an \( n \)-dimensional distribution function with continuous margins \( F_1, \ldots, F_n \). Then there exists a unique copula \( C: [0,1]^n \rightarrow [0,1] \) with uniform margins such that

\[
F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))
\]

Another important result is the Fréchet–Hoeffding theorem, which defines the lower and upper limits of a particular copula. This is important because some copulas can describe only a certain range of dependence.

**Theorem 4** (Fréchet–Hoeffding theorem) Let \( C \) be an \( n \)-dimensional copula. Then for any \( (u_1, \ldots, u_n) \in [0,1]^d \) we have

\[
\max \left( 1 - n + \sum_{i=1}^{n} u_i ; 0 \right) \leq C(u_1, \ldots, u_n) \leq \min(u_1, \ldots, u_n)
\]

Given these results, the next step involves selecting an appropriate copula to describe the dependence structure. As we pointed out previously, in this study we use a Student t copula, from the elliptical family, to accommodate tail dependence. Although elliptical copulas impose tail symmetry and do not offer a closed form, they are tractable in higher dimensions. Conversely, other families like Archimedean and vine copulas usually offer more flexibility at the cost of tractability. Even t copulas can be extended to more elaborated models such as the skewed t or grouped t copulas, as discussed by Demarta and McNeil [64], but these extensions are less tractable as well.

The Student t copula \( C_{\rho,\nu} \) is easily defined in terms of an \( n \)-dimensional multivariate standardized Student t distribution function \( T_{\rho,\nu} \) and a univariate standardized Student t distribution function \( t_{\nu} \). Given the correlation matrix \( \rho \), the degrees of freedom \( \nu \) and the gamma function \( \Gamma(\cdot) \), we have:
\[ T_{\rho,v}(x_1, \ldots, x_n) = T_{\rho,v}(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \frac{\Gamma(\frac{\nu + n}{2}) |\rho|^{-\frac{1}{2}}} {\left(\frac{\nu}{2}\right)^n \frac{\nu + n}{2}} \left(1 + \frac{x^T \rho^{-1} x}{\nu}\right)^{-\frac{\nu + n}{2}} dx \] (55)

\[ t_v(x_i) = u_i, \; u_i \in [0,1] \; \text{and} \; i = 1, \ldots, n \] (56)

\[ C^v_{\rho,v}(u_1, \ldots, u_n; \rho, v) = C^v_{\rho,v}(u; \rho, v) = T_{\rho,v}(t^{-1}_v(u_1), \ldots, t^{-1}_v(u_n)) = T_{\rho,v}(y) \] (57)

Additionally, we can express the corresponding density function \( c^v_{\rho,v} \) as:

\[ c^v_{\rho,v}(u_1, \ldots, u_n; \rho, v) = |\rho|^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu + n}{2}) \left(\frac{\nu}{2}\right)^n} {\left[\Gamma(\frac{\nu + n}{2})\right]^n \frac{\nu}{2}} \left(1 + \frac{\nu^T \rho^{-1} \nu}{\nu}\right)^{-\frac{\nu + n}{2}} \prod_{i=1}^{n} \left(1 + \frac{[t^{-1}_v(u_i)]^2}{\nu}\right)^{-\frac{\nu + 1}{2}} \] (58)

Following Bouyé et al. [39] and Cherubini, Luciano and Vecchiato [51], for a vector of parameters \( \theta \), \( K \) observations and a log-likelihood function \( l(\theta) \) such that:

\[ l(\theta) = \sum_{k=1}^{K} \ln \left( c^\rho_v(F_1(x_1^k), \ldots, F_n(x_n^k); \theta) \right) + \sum_{k=1}^{K} \sum_{i=1}^{N} \ln f_i(x_i^k; \theta_i) \] (59)

\[ l(\theta) = \sum_{k=1}^{K} \ln \left( c^\rho_v(u_1^k, \ldots, u_n^k; \theta) \right) \] (60)

\[ l(\theta) \propto -\frac{K}{2} \ln \rho - \frac{v + n}{2} \sum_{k=1}^{K} \ln \left(1 + \frac{\nu^T \rho^{-1} \nu}{\nu}\right) + \frac{v + 1}{2} \sum_{k=1}^{K} \sum_{i=1}^{N} \ln \left(1 + \frac{[t^{-1}_v(u_i)]^2}{\nu}\right) \] (61)

We can use a parametric method such as the Exact Maximum Likelihood (EML) to estimate jointly both the dependence structure and the marginal parameters. Unfortunately, the EML method is computationally intensive for high dimensional distributions.

Alternatively, we can use a two-step approach proposed by Joe [160] and Genest, Ghoudi and Rivest [106] that first estimates the individual parameters of the marginal distributions and then estimates the copula parameters. Joe [160] method, called Inference Functions for Margins (IFM), assumes that the univariate marginal distributions are known and that we can start by separately estimating their parameters \( \hat{\theta}_i \):
\[ l(\theta) = \sum_{k=1}^{K} \ln c^t_{\theta, \omega}(F_1(x^k_1; \theta_1), \ldots, F_n(x^k_n; \theta_n); \omega) + \sum_{k=1}^{K} \sum_{i=1}^{N} \ln f_i(x^k_i; \theta_i) \quad (62) \]

\[ \hat{\theta}_i = \arg \max_{\theta_i} \sum_{k=1}^{K} \ln f_i(x^k_i; \theta_i) \quad (63) \]

Then, we use these values to estimate the parameters of the dependence structure:

\[ \hat{\omega} = \arg \max_{\omega} \sum_{k=1}^{K} \ln c^t_{\theta, \omega}(F_1(x^k_1; \theta_1), \ldots, F_n(x^k_n; \theta_n); \omega) \quad (64) \]

Conversely, Genest, Ghoudi and Rivest [106] method, known as Canonical Maximum Likelihood (CML) or Pseudo Maximum Likelihood, does not assume parametric distributions for the margins. Instead, it relies on the univariate empirical cumulative distribution functions \( \hat{F}_i \) to generate uniform variates \( \hat{u}_i^k \), which are subsequently used to estimate the copula parameters:

\[ \hat{\omega} = \arg \max_{\omega} \sum_{k=1}^{K} \ln c^t_{\theta, \omega}(\hat{u}_1^k, \ldots, \hat{u}_n^k; \omega) \quad (65) \]

For large samples, both methods yield similar results but for a small number of observations Kim, Silvapulle and Silvapulle [170] strongly recommends the CML approach. Their study suggests that IFM method is not robust against misspecifications of the marginal distributions and that, consequently, the CML method generally performs better in practical applications, where margins are often unknown.

### 3.1.4 Simulating a Multivariate Scenario Tree

Once we have estimated the copula parameters and the marginal distributions, we have a powerful tool to jointly simulate the daily financial series. Based on Sklar theorem, given an n-dimensional copula \( C^t_{\theta, \omega}(u_1, \ldots, u_n) \) and a multivariate distribution \( F(x_1, \ldots, x_n) \) with margins \( F_1(x_1), \ldots, F_n(x_n) \), such that:

\[ F(x_1, \ldots, x_n) = C^t_{\theta, \omega}(F_1(x_1), \ldots, F_n(x_n)) \quad (66) \]

We have:
Therefore, we simply need to generate \( n \) dependent uniform variates. According Cherubini, Luciano and Vecchiato [51], the general iterative approach to accomplish that involves the conditional copula distribution function:

\[
C(u_i|u_1, \ldots, u_{i-1}) = \int_0^{u_i} c(v|u_1, \ldots, u_{i-1}) dv 
\]

\[
c(v|u_1, \ldots, u_{i-1}) = \frac{c(u_1, \ldots, u_i)}{c(u_1, \ldots, u_{i-1}, 1)} = \frac{\partial i C(u_1, \ldots, u_i)}{\partial u_i} \frac{\partial i-1 C(u_1, \ldots, u_{i-1}, 1)}{\partial u_{i-1}} \]

\[
C(u_i|u_1, \ldots, u_{i-1}) = \frac{\partial i-1 C(u_1, \ldots, u_{i-1}, u_i)}{\partial u_i \partial u_{i-1}} \frac{\partial i-1 C(u_1, \ldots, u_{i-1}, 1)}{\partial u_{i-1}} \]  

(68)

We begin by drawing \( u_1 \) from a standard uniform distribution and then uses the inverse of the conditional copula distribution to compute the subsequent dependent uniform variates \( u_2, \ldots, u_n \):

\[
u_i = C^{-1}(v_i|u_1, \ldots, u_{i-1}) \]

\( i = 1, \ldots, n \)

Where each \( v_i \) is also drawn from a standard uniform distribution.

Fortunately, Cherubini, Luciano and Vecchiato [51] also points out that there is an easier method to generate dependent uniform variates from a \( \tau \) copula. Given a correlation matrix \( \rho \) and the degrees of freedom \( \nu \), we can:

(i) Find the Cholesky decomposition \( A \) of \( \rho \)

(ii) Draw \( n \) random variates \( z = (z_1, \ldots, z_n)' \) from the standard normal distribution

(iii) Draw a random variate \( s \) from the chi-squared distribution with \( \nu \) degrees of freedom

(iv) Set \( y = Az \) and \( x = y\sqrt{\nu/s} \)

Set \( u_i = t_\nu(x_i), i = 1, \ldots, n \), where \( t_\nu \) is the univariate Student \( \tau \) distribution with \( \nu \) degrees of freedom
We use the method above to generate dependent multivariate random residuals. Then, we combine each series of residuals with its previously fitted GJR-GARCH + EVT (GPD) + t-Copula model to simulate fully dependent univariate trajectories for all daily financial series. For example, suppose we simulate $m$ multivariate scenarios (i.e., paths) for $n$ instruments over $T$ periods. We are basically generating the following scenario tree:

![Figure 12 – Multivariate Scenario Fan Graph](image)

Each multivariate scenario $i$ is simulated independently from other scenarios and has probability $P(\omega_i) = P_i = 1/m$, $i = 1, \cdots, m$. Moreover, following Pflug and Pichler [247], we can also represent the scenario tree as a nested distribution (further explained in later subsections):

![Figure 13 – Multivariate Scenario Fan Structure](image)
Note that in the structural scenario representation, we omit the root node because it is not stochastic (i.e., it is assumed to be known \textit{a priori}). As a numerical example, suppose we have $m = 4$ scenarios, $n = 3$ instruments and $T = 2$ periods. Suppose also that the node values are:

$$y^{0} = (1.0, 1.0, 1.0)$$

$$y^{1,1} = (2.0, 1.3, 1.7), \ y^{1,2} = (0.9, 1.4, 1.2), \ y^{1,3} = (0.9, 1.4, 1.2), \ y^{1,4} = (0.7, 1.1, 0.8)$$

$$y^{2,1} = (2.2, 1.7, 1.6), \ y^{2,2} = (1.0, 1.3, 1.0), \ y^{2,3} = (0.8, 1.2, 1.0), \ y^{2,4} = (0.6, 1.0, 0.9)$$

The respective scenario tree can be represented as:

\[ 
\begin{array}{c}
\begin{array}{c}
\text{Figure 14 – Multivariate Scenario Fan Graph Example (m=4, n=3 and T=2)}
\end{array}
\end{array}
\]

As we discussed previously, this simulated multivariate scenario tree offers a reasonably accurate way to represent a (continuous) underlying stochastic process by a discrete approximation. However, in practical financial applications it faces two
important shortcomings: it is often too large (i.e., too computationally demanding to be optimized within a reasonable amount of time) and it is not free of arbitrage opportunities. Therefore, in the next section we introduce an additional step to address these limitations.

3.2 Scenario Tree Approximation

Although scenario trees based on the combined GJR-GARCH + EVT (GPD) + t-Copula approach provide an interesting framework for representing daily log returns, they lead to two relevant shortcomings in the optimization process that must be addressed before they can be used. First, stochastic portfolio optimizations based on simulated scenarios are computationally demanding, particularly in multi-period problems, because a huge number of scenarios are often required to achieve proper accuracy. Second, there is no guarantee that the simulated scenarios preclude arbitrage opportunities, which may induce unreliable optimization solutions.

An alternative to eliminate the first shortcoming is to approximate the original set of simulated scenarios to a much smaller set, with adjusted probabilities, using some metric distance to keep both scenarios reasonably similar. This approach is often known as optimal discretization. As we saw in the literature review, several authors have explored this approach to approximate the original scenario tree by smaller (and, consequently, more tractable) tree.

Optimal discretization does not require knowledge of the corresponding stochastic optimization model as, for example, moment-matching methods. This is mostly interesting for general portfolio optimization applications where the structure of the constraints and objective functions (and, therefore, the relevant statistical properties of the optimization problem) differ from one investor to another. On the other hand, alike most conditional sampling and clustering techniques, optimal discretization must be particularly careful to avoid biased and unstable (in-sample and out-of-sample) solutions.

In-sample stability is related to internal consistency. It requires that if we simulate many approximated scenarios using the same discretization method, all of them must lead to similar objective function values. If the optimal values are not similar
for slightly different scenarios generated using the same process and data, the discretization method is not stable. Following Kaut and Wallace [168], most in-sample stability tests focus on the objective function values rather than on the solution values because objective functions are often flat, which means that an optimal objective function value may be attained by quite different solutions, making solution stability very hard to achieve.

Hence, given an objective function $f(x, \xi)$ where $x$ is a solution and $\xi$ is the parameter uncertainty embodied by the approximated scenario tree, in-sample stability means that, for two approximated scenario trees generated by the same discretization method leading to $\xi_i$ and $\xi_j$, we should have:

$$\min_{x \in X} f(x, \xi_i) \approx \min_{x \in X} f(x, \xi_j) \quad \text{(70)}$$

In addition to the potential instability issues, we also need to keep the optimality gap small. This optimality gap is simply the approximation error (or bias) between the optimal objective function values achieved using the true underlying stochastic process and the approximated scenario tree. Therefore, given a stochastic optimization problem such that $f(x, u)$ is the objective function, $G$ is the true underlying distribution function and $\tilde{G}$ is the approximated discrete distribution, we can define the approximation error $e(F, \tilde{F})$:

$$\min_{x \in \mathbb{X}} F(x) = \min_{x \in \mathbb{X}} \int f(x, u) dG(u) \quad \text{(71)}$$

$$\min_{x \in \mathbb{X}} \tilde{F}(x) = \min_{x \in \mathbb{X}} \int f(x, u) d\tilde{G}(u) \quad \text{(72)}$$

$$e(F, \tilde{F}) = F\left(\arg\min_{x} \tilde{F}(x)\right) - F\left(\arg\min_{x} F(x)\right) \geq 0 \quad \text{(73)}$$

$$e(F, \tilde{F}) = F(\tilde{x}^*) - F(x^*) \geq 0 \quad \text{(74)}$$

Pflug [241] shows that, although the true optimal objective value $F(x^*)$ is unavailable for most practical applications, we can define an upper bound to the approximation error $e(F, \tilde{F})$ based on the following lemma:

**Lemma 1**

$$e(F, \tilde{F}) \leq 2\sup_{x} |F(x) - \tilde{F}(x)| \geq 0 \quad \text{(75)}$$
Proof. Let \( x^* \in \text{argmin} \, F \), \( \tilde{x}^* \in \text{argmin} \, \tilde{F} \) and \( \epsilon = \sup_x |F(x) - \tilde{F}(x)| \). Let also \( M = \{x : F(x) \leq F(x^*) + 2\epsilon\} \) and suppose that \( \tilde{x}^* \notin M \). Then:

\[
F(x^*) + 2\epsilon < F(\tilde{x}^*) \leq \tilde{F}(\tilde{x}^*) + \epsilon \leq \tilde{F}(x^*) + \epsilon \leq F(x^*) + 2\epsilon \tag{76}
\]

This contradiction shows that \( \tilde{x}^* \in M \) and, consequently:

\[
e(F, \tilde{F}) = F(\tilde{x}^*) - F(x^*) \leq 2\epsilon \tag{77}
\]

The quality of stochastic optimization solution rely heavily on the level of stability and optimality gap. These levels, in turn, depend greatly on the chosen metric distance for the approximated scenario. Among the many available metrics, the Wasserstein or Kantorovich distance and its extensions seem particularly suited according various works such as Hochreiter and Pflug [141], Pflug [241][243], Pflug and Pichler [245][246] and Heitsch, Römisch [270], and Römisch and Strugarek [138]. The Wasserstein distance seems to be particular popular because it exhibits some interesting properties when applied to stochastic programming problems. For instance, (i) it metricizes convergence in distribution, (ii) it is intuitively extended to stochastic processes and (iii) it is also computationally straightforward to handle.

Unfortunately, most traditional discretization methods do not handle the second shortcoming. In fact, despite of the recent advances of in optimal discretization, there is very limited research concerning the prevention of arbitrage opportunities. To address that, we need to adjust the discretization approach to avoid scenarios that lead to these opportunities. Arbitrage-free scenarios might not be particularly essential for energy generation and logistics applications as we pointed out in the literature review but they are critical for financial applications. For that reason, additional constraints must be imposed in the optimal discretization approach to prevent scenario tree nodes that lead could to these opportunities.

In the following subsections, we discuss in detail the Wasserstein and Process distance concepts, and describe a method to approximate a large scenario tree by a tractable and arbitrage-free one. First, we describe how we can use the Wasserstein and Processes distances to compute, respectively, the distance between distributions and stochastic processes. Then, we describe the approach proposed by Pflug and Pichler [246], which performs optimal discretizations using the Process distance, and extend it to prevent arbitrage opportunities.
3.2.1 Distributions Distance

Optimal discretization relies on the concept of probability metrics to evaluate the dissimilarity between different probability distributions. Such discretization approaches minimize an appropriately chosen probability metric to approximate the original and often continuous distribution by a smaller but similar (in some statistical sense) discrete one, which can be used in practical optimization problems.

Definition 18 Given a set of probability measures $\mathcal{P}$ on $\mathbb{R}^m$ and $P_1, P_2, P_3 \in \mathcal{P}$, a probability metric $d$ on $\mathcal{P} \times \mathcal{P}$ is a distance function between probability distributions such that:

(i) $d(P_1, P_2) = 0 \iff P_1 = P_2$ (78)
(ii) $d(P_1, P_2) \geq 0$ (79)
(iii) $d(P_1, P_2) = d(P_2, P_1)$ (80)
(iv) $d(P_1, P_2) \leq d(P_1, P_3) + d(P_2, P_3)$ (81)

Definition 19 Given a set $\Xi$ and a probability metric $d: \Xi \times \Xi \to \mathbb{R}$, the pair $(\Xi, d)$ is called metric space. A metric space $(\Xi, d)$ is said to be separable and complete if it has a countable dense subset and if every Cauchy sequence in $(\Xi, d)$ converges. A complete separable metric space is called Polish space.

Various metrics or distances (e.g., Kolmogorov metric, Lévy-Prokhorov metric, total variation distance) can be defined on a given set but few are suited to measure distances between discrete probability distributions. Two metrics in particular, the Fortet-Mourier and the Wasserstein distances, have been favoured in the optimal discretization literature because of their appealing computational and statistical properties, particularly in Polish spaces. As pointed out by Pflug [241], Pflug and Pichler [244], and Kovacevic and Pichler [186], they are especially relevant for scenario approximation because they metricize weak convergence and because discrete probability measures are dense in the corresponding spaces of probability measures.

Definition 20 Given all continuous bounded functions $h$, a sequence of probability measures $P_n$ on $\mathbb{R}^M$ converge weakly as $n \to \infty$ if:
\[
\int h dP_n \to \int h dP \tag{82}
\]

In other words, assuming a Polish space a proper distance \(d\) metricizes weak convergence from a discrete approximated distribution \(P_n\) to the true underlying distribution \(P\), i.e. \(d(P_n, P) \to 0\), if and only if \(P_n \Rightarrow P\). This property, exhibited by the Wasserstein distance, is paramount to the approximation scenario technique used in the thesis.

The Wasserstein distance is generalization of the Kantorovich distance (also known as Wasserstein distance of order \(r = 1\)) that can also be generalized from probability measures to stochastic processes, as we discuss in the next subsection. It is directly associated to the concept of optimal transportation cost between distributions, which is extensively applied in engineering applications.

**Definition 21** Given two probability spaces \((\Omega, \mathcal{F}, P)\) and \((\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})\) with elements \(\omega \in \Omega\) and \(\tilde{\omega} \in \tilde{\Omega}\), a joint probability measure \(\pi(\omega, \tilde{\omega})\) on \(\Omega \times \tilde{\Omega}\), and a (transportation) cost function \(c(\omega, \tilde{\omega})\) between them the optimal transportation cost is given by:

\[
\inf_{\tilde{\pi}} \int_{\tilde{\Omega} \times \tilde{\tilde{\Omega}}} c(\omega, \tilde{\omega}) \pi(d\omega, d\tilde{\omega}) \tag{83}
\]

Such that, for all measurable sets \(A \in \mathcal{F}\) and \(B \in \tilde{\mathcal{F}}\):

\[
\pi(A \times \tilde{\Omega}) = P(A) \tag{84}
\]

\[
\pi(\Omega \times \tilde{B}) = \tilde{P}(B) \tag{85}
\]

**Definition 22** Given two probability spaces \((\Omega, \mathcal{F}, P)\) and \((\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})\) and two \(\mathbb{R}^m\)-valued random variables \(\xi: \Omega \to \mathbb{R}^m\) and \(\tilde{\xi}: \tilde{\Omega} \to \mathbb{R}^m\), the inherited distance (from \(\xi\) and \(\tilde{\xi}\) between \(\Omega\) and \(\tilde{\Omega}\) is defined either as:

\[
d(\omega, \tilde{\omega}) = \|\xi(\omega) - \tilde{\xi}(\tilde{\omega})\| \text{ or} \tag{86}
\]

\[
d(\xi(\omega), \tilde{\xi}(\tilde{\omega})) = \|\xi(\omega) - \tilde{\xi}(\tilde{\omega})\| \tag{87}
\]

For some norm \(\|\cdot\|\) in \(\mathbb{R}^m\).
Definition 23 Given an inherited distance \( d(\omega, \tilde{\omega}) \) between \( \Omega \) and \( \tilde{\Omega} \), the Wasserstein distance is defined either as:

\[
\begin{align*}
    d_r(P, \tilde{P}) &= \left( \inf_{\pi} \int_{\Omega \times \tilde{\Omega}} d(\omega, \tilde{\omega})^r \pi(d\omega, d\tilde{\omega}) \right)^{\frac{1}{r}} \\
    d_r(P, \tilde{P}) &= \left( \inf_{\pi} \int_{\mathbb{R}^m \times \mathbb{R}^m} d(u, v)^r \pi(du, dv) \right)^{\frac{1}{r}}
\end{align*}
\]

As we will discuss in detail later, multi-period stochastic programs depend on multivariate stochastic processes \( \{\xi_t\}_{t=1}^T \), which are defined on some probability space \( (\Omega, \mathcal{F}, P) \) such that \( \xi_t \in \mathbb{R}^d \). In addition, a decision \( x_t \) at time \( t \) cannot depend on future realizations (i.e., \( \xi_{t+k}, k \geq 1 \)) meaning that \( x_t \) is measurable with respect to \( \mathcal{F}_t(\xi) \). This restriction is known as nonanticipativity condition.

3.2.2 Stochastic Processes Distance

As we have underlined earlier, multi-period stochastic optimization problems often require a simpler (i.e., tractable) discrete stochastic process \( \tilde{\xi} \) with some filtration \( \tilde{\mathcal{F}} \) that approximates the original underlying stochastic process \( \xi \), with another filtration \( \mathcal{F} \). In other words, given two general multi-period stochastic optimization problems:

\[
\max \{ \mathcal{A}[H(x_0, \xi_1, \ldots, x_{T-1}, \xi_T)]: x \ll \mathcal{F} \} \tag{90}
\]

\[
\max \{ \mathcal{A}[H(x_0, \tilde{\xi}_1, \ldots, x_{T-1}, \tilde{\xi}_T)]: x \ll \tilde{\mathcal{F}} \} \tag{91}
\]

We need to identify \( \tilde{\xi} \) and \( \tilde{\mathcal{F}} \) such that the optimality gap (i.e., the approximation error) between the solutions of these optimization problems is stable and small. However, an important detail cannot be overlooked: in practical applications, these processes and filtrations are rarely defined in the same probability space \( (\Omega, \mathcal{F}, P) \). Therefore, the approximation metric used to measure the distance (i.e., the optimality gap) between these processes must take it into account. The process or nested distance, proposed by Pflug [243] and further detailed in Pflug and Pichler [244][245][246], is particularly interesting.
3.2.3 Scenario Tree Reduction

The optimal discretization approach applied in this section follows the techniques presented in Pflug [243], Pflug and Pichler [246], and Kovacevic and Pichler [186], which are based on process distances.

3.2.4 No-Arbitrage Scenario Tree Reduction

Informally, an arbitrage opportunity consists of a riskless investment with a positive return that, in real financial markets, does not persist for long (because investors quickly explore it). Unfortunately, as discussed by Klassen [173][174][175], and Geyer, Hanke and Weissensteiner [107], portfolio optimization models are ill equipped to handle arbitrage opportunities, because they often lead to unbounded or biased solutions. For that reason, scenario trees must preclude such opportunities to achieve reliable solutions. Formally, following Ingerson [151]:

**Definition 24** Given non-redundant instruments \( k = 1, \ldots, K \), their allocations \( \theta_{k,t} \) and prices \( p^n_{k,t} \) at stage \( t \) and state (i.e., scenario tree node) \( n \), and their prices \( p^m_{k,t+1} \) at stage \( t + 1 \) and states \( m \) such that \( m \in n^+ \) (i.e., scenario tree direct descendent nodes), an arbitrage opportunity exists if either:

(i) It is possible to construct zero net investment portfolio or investment strategy \( \theta_t = (\theta_{1,t}, \ldots, \theta_{K,t}) \) that is profitable in at least one future state of nature (i.e., that delivers a positive return in at least one scenario arc) and never produces a loss, whatever the future state of nature (i.e., that does not deliver negative returns in any scenario arc).

\[
\sum_{k=1}^{K} \theta_{k,t} p^n_{k,t} = 0 \quad (92)
\]

\[
\sum_{k=1}^{K} \theta_{k,t} p^m_{k,t+1} \geq 0, \text{ for all } m \in n^+ \quad (93)
\]
\[ \sum_{k=1}^{K} \theta_{k,t} P_{k,t+1}^m > 0, \text{for at least one } m \in n^+ \quad (94) \]

(i) It is possible to construct an investment portfolio or strategy \( \theta_t = (\theta_1, t, \cdots, \theta_K) \) that has a positive cash flow today and generates a nonnegative return, whatever the future state of nature.

\[ \sum_{k=1}^{K} \theta_{k,t} P_{k,t}^n < 0 \quad (95) \]

\[ \sum_{k=1}^{K} \theta_{k,t} P_{k,t+1}^m \geq 0, \text{for all } m \in n^+ \quad (96) \]

Putting it differently, as imposed by Harrison and Kreps theorem, given \( P_{k,t}^n \) and \( P_{k,t+1}^m, m \in n^+ \) the scenario tree does not exhibit arbitrage opportunities if and only if there is a state price vector \( \pi \) such that:

\[ P_{k,t+1}^m \pi = P_{k,t}^n, \text{for all } m \in n^+ \quad (97) \]

However, as explained by Jobst and Zenios [159], Fersti and Weissensteiner [100], and Geyer, Hanke and Weissensteiner [107], portfolio optimization must be carried on objective (i.e., real world) probability measures rather than risk-neutral probability measures.

3.3 Stochastic Portfolio Optimization

Once we have constructed our tractable scenario tree, we need to choose an appropriate optimization method that properly explores the impacts and outcomes of potential sequence of allocations in each scenario and, as we discuss earlier, the Stochastic Programming (SP) is a suited approach to accomplish that.

4 Methodology

4.1 Data Description
The combined GJR-GARCH + EVT (GPD) + t-Copula that we discuss next is applied to fit the joint behaviour of five global stock indexes: IBX, S&P 500, FTSE 100, DAX and Nikkei 225 from January 4, 2005 to August 15, 2015 comprising 2.673 trading days.

According Rydberg [274], most daily financial series analyzed in the literature are based on log returns of closing prices. We adopt the same approach and use Dollar closing prices to compute the multivariate financial series of $N = 5$ assets and $T = 2.672$ daily log returns:

$$r_{i,t} = \ln \frac{S_{i,t}}{S_{i,t-1}} \quad (98)$$

$$i = 1, \ldots, N; \ t = 1, \ldots, T$$

<table>
<thead>
<tr>
<th></th>
<th>IBX</th>
<th>S&amp;P 500</th>
<th>FTSE 100</th>
<th>DAX</th>
<th>Nikkei 225</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0262%</td>
<td>0.0207%</td>
<td>0.0039%</td>
<td>0.0277%</td>
<td>0.0151%</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.3230%</td>
<td>1.2639%</td>
<td>1.4560%</td>
<td>1.6460%</td>
<td>1.5018%</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.2630</td>
<td>-0.3327</td>
<td>-0.0310</td>
<td>-0.0317</td>
<td>-0.4776</td>
</tr>
<tr>
<td>Median</td>
<td>0.097%</td>
<td>0.074%</td>
<td>0.067%</td>
<td>0.055%</td>
<td>0.048%</td>
</tr>
<tr>
<td>Mean Absolute Deviation</td>
<td>1.631%</td>
<td>0.804%</td>
<td>0.978%</td>
<td>1.133%</td>
<td>1.073%</td>
</tr>
<tr>
<td>Lower Semi Standard Deviation</td>
<td>1.679%</td>
<td>0.920%</td>
<td>1.050%</td>
<td>1.173%</td>
<td>1.098%</td>
</tr>
<tr>
<td>Upper Semi Standard Deviation</td>
<td>1.605%</td>
<td>0.866%</td>
<td>1.099%</td>
<td>1.154%</td>
<td>1.024%</td>
</tr>
</tbody>
</table>

Table 1 – Univariate Statistics – Daily Log Returns

Unfortunately, there are only 2.379 trading days with no missing prices (Table 2 and Figure 16) and we need to handle the missing information before we can proceed. Standard techniques for missing data cannot be applied without distorting considerably the multivariate series. For example, row-wise removal decreases the sample size and creates irregular (i.e., non-daily) time series. Single imputation techniques, such as copying the last available price or interpolating the nearest prices, may introduce substantial biases.

Instead, we use a better alternative: an Expectation-Maximization (EM) algorithm for recovering multivariate normal missing data proposed by Meucci [217]. It still induces some distortions, but they do not appear to be severe (Figure 16 and Figure 17) as usually happens with the standard methods.
### Table 2 – Missing Return Patterns

<table>
<thead>
<tr>
<th>Occurrences</th>
<th>DAX</th>
<th>FTSE</th>
<th>SP</th>
<th>IBX</th>
<th>Nikkei</th>
<th>Missing Returns per Row</th>
</tr>
</thead>
<tbody>
<tr>
<td>2378</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>91</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>132</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DAX</th>
<th>FTSE</th>
<th>SP</th>
<th>IBX</th>
<th>Nikkei</th>
<th>Missing Returns per Row</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>79</td>
<td>91</td>
<td>139</td>
<td>156</td>
<td>527</td>
</tr>
</tbody>
</table>

#### Figure 16 – Missing Returns Histogram
4.2 Scenario Tree Generation

The framework proposed in this study combines the approaches proposed by Diebold, Schuermann, and Stroughair [75], McNeil and Frey [213], and Rockinger and
Jondeau [162][268]. We substitute an asymmetric GJR-GARCH model for the original GARCH model to encompass potential leverage effects and expand the univariate analysis to a multivariate setting using a t-Copula to account for the dependency structure among the financial series.

Our combined approach comprises four major steps:

(i) Fit a GJR-GARCH model to each of the univariate series to generate approximately i.i.d. standardized residuals
(ii) Fit GPD tails and a smoothed-kernel body to each series of standardized residuals
(iii) Estimate a t-copula using the transformed residuals of the previous margins
(iv) Simulate dependent paths using the combined models

Before the initial step, we examine the series looking for indication of the four stylized facts: heavy-tails, volatility clustering, leverage effect and tail co-dependence. Cumulative returns and daily log returns can help us identify signs of these features. Additionally, correlograms and tests such as the Augmented Dickey-Fuller test, Ljung-Box Q test and Engle test for ARCH effects can further support our analysis.

Once we confirm that these stylized facts seem to be present, we can model them. First, we take into account volatility clustering and leverage effects by fitting a GJR-GARCH process with Student t innovations to each series of log daily returns. This step constructs sequences of approximately i.i.d. standardized residuals with heavy tails that we use in the subsequent steps.

As pointed out earlier, we selected a parsimonious specification GJR-GARCH (1,1) to model the conditional variance because it is more tractable and still provides adequate estimates. Conversely, we do not assume a priori that the conditional mean does not have a significant autoregressive term. Instead we compare low order ARMA specifications to verify that conditional raw returns indeed exhibit little serial correlation are close to zero, as suggested by the literature.

Thus, we assume that:

\[ \varepsilon_t = r_t - E_{t-1}[r_t] \]  \hspace{1cm} (99)
\[ E_{t-1}[r_t] = c + \phi r_{t-1} + \theta r_{t-1} \]  \hspace{1cm} (100)
\[ \varepsilon_t = z_t \sigma_t \]  

\[ \sigma_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2 + d_1 I(\varepsilon_{t-1} < 0) \varepsilon_{t-1}^2 \]  

\[ a_0 > 0, \ a_1 \geq 0, \ b_1 \geq 0, \ d_1 \geq 0 \]  

Where \{r_t\} is the series of daily log returns, \( E_{t-1}[r_t] \) is the conditional return, whose best specification is likely to omit autoregressive and moving average terms, \{z_t\} is a i.i.d. sequence with Student t distribution and \( \sigma_t^2 \) is the conditional variance at time \( t \). We then model each univariate series of daily log returns using the formulation above and examine the residuals diagnostics.

We perform a visual analysis of the daily standardized residuals to check how reasonable is the hypothesis of (approximately) i.i.d. standardized residuals. We also check for remaining signs of volatility clustering, leverage effects and heavy-tails. Given the discussed literature, we expect only the latter to be present. Further checks employ correlograms, Ljung-Box Q tests and Engle tests for ARCH effects of the raw and squared standardized residuals to confirm that the former two stylized facts have been adequately modelled.

Once we have the i.i.d. standardized residual series, we can improve the fitting by separating each series’ distribution in three parts: the lower tail, the upper tail and the central part of distribution. First, we need to decide which particular distribution is suitable to each one of these parts. Based on the Peaks over Threshold (POT) method of Extreme Value Theory discussed earlier, it seems reasonable to assume that, given a sufficiently large threshold \( u \) and shape parameter \( \xi > 0 \), tails follow approximately a GPD

\[ G_{\xi, \sigma(y)}(y) = \begin{cases} 
1 - \left( 1 + \frac{y}{\sigma(u)} \right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\
1 - e^{-\frac{y}{\sigma(u)}} & \text{if } \xi = 0 
\end{cases} \]  

where \( \beta > 0 \)  

Note that for the upper tails (i.e., the positive extreme observations), we can fit a GPD directly to the daily standardized residuals while for the lower tails (i.e., the negative extreme observations), we need to fit a GPD to the negative standardized residuals.

The first step we need to estimate each GPD is to define proper thresholds, which separate the tails from the centre of the distributions. As we underlined in the
previous section, these thresholds are not easily identified. One alternative is to
analyze the quantile-quantile and empirical excess mean plots to obtain an
approximate idea of each tail’s threshold. Once the lower and upper thresholds are
selected, we proceed with the GPD estimation and identify, for each standardized
residual series, the shape parameter $\xi$ (or, equivalently, the tail index $\alpha = 1/\xi$) and the
scale parameter $\sigma$ of each tail.

Then, based on selected thresholds we use the central part of standardized
residuals to fit the body of each distribution to a Gaussian kernel using a Kernel Density
Estimator (KDE). This approach assumes that a large number of observations are
available to properly describe the central part of distribution.

After we have fitted the parametric tails and kernel body of each distribution of
standardized residuals, we evaluate the results. First, we compare the empirical upper
and lower tail distributions to the suggested GPD model. Secondly, we compare the
empirical cumulative distribution functions to the semi-parametric ones generated by
each combined GJR-GARCH + EVT (GPD) model. Lastly, we evaluate how properly
each proposed semi-parametric distribution fits the empirical distribution using
quantile-quantile and probability density function plots.

Provided the combined GJR-GARCH + EVT(GPD) models are adequate, we
can proceed and assess the joint dependence structure by using the standardized
residuals to fit the a t-copula. The approach is straightforward: employing the GJR-
GARCH + EVT (GPD) cumulative distribution functions, we transform the residuals into
uniform variates that are used to fit the t-copula using the CML method described
earlier. As pointed out in the literature, we expect to explain some of the tail co-
dependence between the margins observed in the bivariate scatter and density plots.

As the final step, we use the combined GJR-GARCH + EVT (GPD) + t Copula
model to simulate dependent samples (or trajectories) of the daily log returns. As
described in the previous section, we first generate samples of dependent uniform
variates using the fitted t copula. Then, we transform back these variates into
standardized residuals using the inverse of their distributions based on calibrated GPD
tails and kernel-smoothed bodies. Lastly, we filter the standardized residuals into the
ARMA + GJR-GARCH models to simulate the dependent daily log returns.
This combined approach generates multivariate samples that exhibit the four key stylized facts discussed previously: heavy-tails, volatility clustering, leverage effect and tail co-dependence. As practical applications, we perform a multi-objective Mean-Variance-CVaR portfolio optimization; using samples simulated from a multivariate Gaussian distribution and from the proposed GJR-GARCH + EVT (GPD) + t Copula approach. We present the results and analyses in the next section.

5 Empirical Results

In this section, we present our empirical findings and comment some of the results. All analysis were performed using Matlab R2015a (version: 8.5), except for the missing data analysis discussed in the previous section, which were partially implemented using R 3.2.3.

We begin by an initial inspection of the cumulative log returns of these financial series (Figure 18). It suggests some interesting features such as time-varying volatilities, erratic extreme variations and some degree of co-dependence.

![Figure 18 – Stock Indexes’ Cumulative Returns](image-url)
We first examine the univariate behaviour of each series. A closer look at the daily log returns (Figure 21) seem to confirm the presence of volatility clustering, heavy-tails and some leverage effects in all series.

Figure 19 – ACF and PACF of Raw Returns
Figure 20 – ACF and PACF of Squared Returns
Augmented Dickey-Fuller tests (Table 3) with significance level of 5% reject the hypothesis of unit roots in these series. Additionally, as suggested by previous researches, the sample Autocorrelation Functions (ACFs) (Figure 19) and sample Partial Autocorrelation Functions (PACFs) (Figure 20) as well as the Ljung-Box Q-tests (Table 4) show little evidence of serial correlation in raw returns. They do, however,
suggest strong evidence of serial correlation in squared returns (with apparently very slow decay) and consequently reject the assumption of i.i.d. series. Similar analyses of the absolute returns also corroborate these findings.

<table>
<thead>
<tr>
<th>ADF Test (Critical Value = -1.94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-Value</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>S&amp;P 500</td>
</tr>
<tr>
<td>FTSE 100</td>
</tr>
<tr>
<td>DAX</td>
</tr>
<tr>
<td>Nikkei 225</td>
</tr>
</tbody>
</table>

Table 3 – ADF Tests

<table>
<thead>
<tr>
<th>Ljung-Box Q-test on Squared Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>S&amp;P 500</td>
</tr>
<tr>
<td>FTSE 100</td>
</tr>
<tr>
<td>DAX</td>
</tr>
<tr>
<td>Nikkei 225</td>
</tr>
</tbody>
</table>

| Critical Values | 3.84 | 5.99 | 7.81 | 9.49 | 11.07 | 12.59 | 14.07 | 15.51 |

Following Tsay [301], we set the number of lags to $\ln T = \ln 2.672 \cong 7.89$, rounded up to the next integer, to increase the performance of the Ljung-Box Q test.

Table 4 – Ljung Box Q-test on Squared Residuals

<table>
<thead>
<tr>
<th>ARCH Engle test for Residual Heteroscedasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>S&amp;P 500</td>
</tr>
<tr>
<td>FTSE 100</td>
</tr>
<tr>
<td>DAX</td>
</tr>
<tr>
<td>Nikkei 225</td>
</tr>
</tbody>
</table>

| Critical Values | 3.84 | 5.99 | 7.81 | 9.49 | 11.07 | 12.59 | 14.07 | 15.51 |

Table 5 – ARCH Engle test for Residual Heteroscedasticity

Fitting a ARMA (p,q) + GJR-GARCH (1,1) model to each one of the univariate series, we observe that most of them are best explained by parsimonious ARMA (0,0) + GJR-GARCH (1,1) processes (Table 6), which suggests that the conditional means are constant and, in these cases, not statistically different from zero (Table 7). The exceptions are the IBX and S&P 500 series. However, the autoregressive term is not
statistically relevant in the former (Table 8) and neither is the moving average term in the latter (Table 9). Therefore, in this study we also assume ARMA (0,0) + GJR-GARCH (1,1) specifications for these series (Table 7).

<table>
<thead>
<tr>
<th>IBX</th>
<th>P</th>
<th>Q</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>-13,561.9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>-13,562.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>-13,554.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>-13,542.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S&amp;P 500</th>
<th>P</th>
<th>Q</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>-17,559.9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>-17,559.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>-17,554.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FTSE 100</th>
<th>P</th>
<th>Q</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>-16,407.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>-16,399.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>-16,391.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DAX</th>
<th>P</th>
<th>Q</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>-15,433.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>-15,425.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>-15,417.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nikkei 225</th>
<th>P</th>
<th>Q</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>-15,786.9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>-15,784.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>-15,779.3</td>
</tr>
</tbody>
</table>

Table 6 – BIC Comparison of ARMA (P,Q) + GJR-GARCH (1,1) Models
Table 7 – ARMA (0,0) + GJR-GARCH (1,1) Specifications

<table>
<thead>
<tr>
<th></th>
<th>IBX ARMA(0,0) Model</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>S&amp;P 500 ARMA(0,0) Model</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter</td>
<td>Value</td>
<td>Std. Error</td>
<td>t Stat</td>
<td></td>
<td>Parameter</td>
<td>Value</td>
<td>Std. Error</td>
<td>t Stat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.91</td>
<td></td>
<td>Constant</td>
<td>0.0004</td>
<td>0.0001</td>
<td>3.13</td>
<td></td>
</tr>
<tr>
<td>GJR-GARCH(1,1) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GJR-GARCH(1,1) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>5.66</td>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>3.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GARCH(1)</td>
<td>0.9194</td>
<td>0.0105</td>
<td>87.65</td>
<td></td>
<td>GARCH(1)</td>
<td>0.8864</td>
<td>0.0136</td>
<td>64.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ARCH(1)</td>
<td>0.0061</td>
<td>0.0108</td>
<td>0.57</td>
<td></td>
<td>ARCH(1)</td>
<td>0.0000</td>
<td>0.0168</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leverage(1)</td>
<td>0.1106</td>
<td>0.0156</td>
<td>7.09</td>
<td></td>
<td>Leverage(1)</td>
<td>0.1953</td>
<td>0.0256</td>
<td>7.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student t Innovations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student t Innovations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DoF</td>
<td>12.7322</td>
<td>2.8893</td>
<td>4.41</td>
<td></td>
<td>DoF</td>
<td>6.0985</td>
<td>0.8277</td>
<td>7.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FTSE 100 ARMA(0,0) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FTSE 100 ARMA(0,0) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter</td>
<td>Value</td>
<td>Std. Error</td>
<td>t Stat</td>
<td></td>
<td>Parameter</td>
<td>Value</td>
<td>Std. Error</td>
<td>t Stat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.88</td>
<td></td>
<td>Constant</td>
<td>0.0005</td>
<td>0.0002</td>
<td>2.35</td>
<td></td>
</tr>
<tr>
<td>GJR-GARCH(1,1) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GJR-GARCH(1,1) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>2.75</td>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>3.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GARCH(1)</td>
<td>0.9080</td>
<td>0.0108</td>
<td>83.90</td>
<td></td>
<td>GARCH(1)</td>
<td>0.9078</td>
<td>0.0122</td>
<td>74.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ARCH(1)</td>
<td>0.0092</td>
<td>0.0105</td>
<td>0.88</td>
<td></td>
<td>ARCH(1)</td>
<td>0.0019</td>
<td>0.0122</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leverage(1)</td>
<td>0.1444</td>
<td>0.0191</td>
<td>7.57</td>
<td></td>
<td>Leverage(1)</td>
<td>0.1493</td>
<td>0.0199</td>
<td>7.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student t Innovations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student t Innovations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DoF</td>
<td>9.4403</td>
<td>1.5830</td>
<td>5.96</td>
<td></td>
<td>DoF</td>
<td>8.3243</td>
<td>1.3512</td>
<td>6.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DAX ARMA(0,0) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DAX ARMA(0,0) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parameter</td>
<td>Value</td>
<td>Std. Error</td>
<td>t Stat</td>
<td></td>
<td>Parameter</td>
<td>Value</td>
<td>Std. Error</td>
<td>t Stat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.95</td>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>4.05</td>
<td></td>
</tr>
<tr>
<td>GJR-GARCH(1,1) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GJR-GARCH(1,1) Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>4.05</td>
<td></td>
<td>Constant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>4.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GARCH(1)</td>
<td>0.8802</td>
<td>0.0155</td>
<td>56.66</td>
<td></td>
<td>GARCH(1)</td>
<td>0.8802</td>
<td>0.0155</td>
<td>56.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ARCH(1)</td>
<td>0.0296</td>
<td>0.0159</td>
<td>1.86</td>
<td></td>
<td>ARCH(1)</td>
<td>0.0296</td>
<td>0.0159</td>
<td>1.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leverage(1)</td>
<td>0.1154</td>
<td>0.0224</td>
<td>5.15</td>
<td></td>
<td>Leverage(1)</td>
<td>0.1154</td>
<td>0.0224</td>
<td>5.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student t Innovations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student t Innovations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DoF</td>
<td>9.2877</td>
<td>1.4095</td>
<td>6.59</td>
<td></td>
<td>DoF</td>
<td>9.2877</td>
<td>1.4095</td>
<td>6.59</td>
<td></td>
</tr>
</tbody>
</table>
We notice that there is strong evidence of volatility clustering and leverage effects in all series. On the other hand, we also notice that ARCH terms are not statistically different from zero, which might have led to identification problems and compromise quasi-maximum likelihood estimates as underlined in Andersen et al. [7]. Fortunately, it seems not to be the case as residual diagnostics suggest.

A visual examination of the daily standardized residuals appears to support that they are nearly i.i.d. (Figure 23). Analyses of the ACFs and PACFs further corroborate this assumption (Figure 24 and Figure 25). Lastly, Quantile-Quantile (QQ) plots against Gaussian and Student t-distributions indicate that we can improve the

---

### Table 8 – ARMA (1,0) + GJR-GARCH (1,1) Specification – IBX

<table>
<thead>
<tr>
<th>IBX</th>
<th>ARMA(1,0) Model</th>
<th>GJR-GARCH(1,1) Model</th>
<th>Student t Innovations</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR{1}</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.66</td>
</tr>
<tr>
<td>GARCH{1}</td>
<td>0.9185</td>
<td>0.0105</td>
<td>87.86</td>
</tr>
<tr>
<td>ARCH{1}</td>
<td>0.0040</td>
<td>0.0108</td>
<td>0.37</td>
</tr>
</tbody>
</table>

### Table 9 – ARMA (0,1) + GJR-GARCH (1,1) Specification – S&P 500

<table>
<thead>
<tr>
<th>S&amp;P 500</th>
<th>ARMA(0,1) Model</th>
<th>GJR-GARCH(1,1) Model</th>
<th>Student t Innovations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA{1}</td>
<td>0.0005</td>
<td>0.0003</td>
<td>1.47</td>
</tr>
<tr>
<td>GARCH{1}</td>
<td>0.8899</td>
<td>0.0134</td>
<td>66.43</td>
</tr>
<tr>
<td>ARCH{1}</td>
<td>0.0000</td>
<td>0.0164</td>
<td>0.00</td>
</tr>
</tbody>
</table>
fitting of the tails (Figure 26), except maybe for FTSE 100 series. Although Gaussian distributions are too thin to accommodate the extreme observations, Student t distributions do not seem to achieve good results regarding tail behaviour, possibly because they are fitted to the whole distributions rather than to the tails individually.

Figure 22 – Filtered Residuals and Filtered Conditional Volatility
Figure 23 – Standardized Residuals
Figure 24 – ACF and PACF of Raw Standardized Residuals
Figure 25 – ACF and PACF of Squared Standardized Residuals
As we saw previously, an interesting alternative to accomplish that is to use the Peaks over Threshold (POT) method to fit each tail of standardized residuals individually. First, we use the QQ plots (Figure 26) and the Empirical Excess Mean (EEM) plots (Figure 27 and Figure 28) to select an appropriate threshold for each tail (Table 10).
Figure 27 – Empirical Mean Excess Plots – Lower Tails
Figure 28 – Empirical Mean Excess Plots – Upper Tails
Given the selected lower and upper thresholds, we estimate the GPD parameters of each tail, fit the kernel-smoothed bodies and evaluate the results (Figure 29 to Figure 33). We see that the lower tail of the IBX and DAX distributions and the upper tail of the FTSE 100 distribution have finite variance but not finite fourth moment (Table 11). We also notice that distribution tails do not seem to be symmetric. Nonetheless, an overall analysis of the cumulative function, density function and QQ plots suggest that proposed GJR-GARCH + EVT (GPD) approach provides a good fit to the heavy-tails, volatility clustering, leverage effect and tail co-dependence observed in the empirical distributions (Figure 31, Figure 32 and Figure 33).

<table>
<thead>
<tr>
<th>GDP Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lower Tail</strong></td>
</tr>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>S&amp;P 500</td>
</tr>
<tr>
<td>FTSE 100</td>
</tr>
<tr>
<td>DAX</td>
</tr>
<tr>
<td>Nikkei 225</td>
</tr>
</tbody>
</table>

Table 10 – GPD Thresholds

<table>
<thead>
<tr>
<th>GDP Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lower Tail</strong></td>
</tr>
<tr>
<td><strong>Tail Index</strong></td>
</tr>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>S&amp;P 500</td>
</tr>
<tr>
<td>FTSE 100</td>
</tr>
<tr>
<td>DAX</td>
</tr>
<tr>
<td>Nikkei 225</td>
</tr>
</tbody>
</table>

Table 11 – Tail Fitting – GPD Parameters
Figure 29 – Lower Tail Fitting – Standardized Residuals
Figure 30 – Upper Tail Fitting – Standardized Residuals
Figure 31 – Empirical Cumulative Distribution Functions
Figure 32 – QQ Plots of the Standardized Residuals vs GJR-GARCH + GPD Distributions
Therefore, assuming that the GJR-GARCH + EVT (GPD) models are suited to model the marginal distributions, we use them to fit a t-copula that captures their...
dependence structure. As we described previously, the approach is simple: we use the GJR-GARCH + EVT (GPD) cumulative distribution functions to transform the residuals into uniform variates. Then we fit the t-copula to the uniform univariates using the CML method. As we observe in the bivariate scatter and density plots (Figure 34 to Figure 37) all series, except for the Nikkei 225, appear to exhibit some degree of tail co-dependence. The t copula estimated parameters (Table 12) also indicate a high correlation between most of them series.

<table>
<thead>
<tr>
<th>t-Copula Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>IBX</td>
</tr>
<tr>
<td>S&amp;P 500</td>
</tr>
<tr>
<td>FTSE 100</td>
</tr>
<tr>
<td>DAX</td>
</tr>
<tr>
<td>Nikkei 225</td>
</tr>
</tbody>
</table>

Table 12 – t-Copula Parameters

Figure 34 – Dependence between Series – Part 1
Figure 35 – Dependence between Series – Part 2
Figure 36 – Bivariate Histograms – Part 1
Lastly, we use the complete combined approach GJR-GARCH + EVT (GPD) + \( t \) Copula to simulate 500,000 cross and serially dependent samples of the daily series. We begin by construing dependent uniform variates using the fitted \( t \) copula as explained in the previous section.

Figure 37 – Bivariate Histograms – Part 2

Figure 38 – Mean-CVaR Efficient Frontiers
Figure 39 – Mean-Variance Efficient Frontiers
6 Conclusions and Further Research

In the thesis, we discuss three key steps to perform a long-term multistage multi-objective portfolio optimization based on stochastic programming. First, we develop and estimate a combined GJR-GARCH + EVT-GPD + t-Copula econometric model to simulate a large number of discrete multivariate trajectories, which is represented by a large scenario tree. Secondly, we approximate the original scenario tree by a smaller one that is both tractable and arbitrage-free. Lastly, we use the approximated scenario tree to optimize a Mean-Variance-CVaR portfolio.

The first step combines different econometric models, extending approaches similar to those suggested by McNeil and Frey [213], and Jondeau and Rockinger [162], to simulate four major stylized facts: volatility clustering, heavy tails, leverage effects and tail co-dependence, which are typically indispensable to proper risk management and portfolio optimization in financial markets. The reason is straightforward: without unrealistic simplifying assumptions, proper optimization models (i.e., models that can handle real-world multivariate distributions) often achieve superior results because the simulated scenario tree (or, more precisely, scenario fan) describe better key empirical features of the financial series.

However, the original scenario tree is neither tractable nor arbitrage-free. Therefore, we need a second step that (i) approximates (i.e., reduce the size) the original tree by a smaller one and (ii) removes potential opportunities of arbitrage. Otherwise, practical portfolio optimization problems are likely to be either unsolvable (within a reasonable amount of time) or biased. To accomplish both objectives, this step extends the optimal discretization approach, proposed by Pflug [203] and developed by Pflug and Pichler [205][206], and Kovacevic and Pichler [155], to preclude arbitrage opportunities.

Finally, the last step exploits the approximated scenario tree to perform a multi-objective multi-period stochastic portfolio Mean-Variance-CVaR optimization that considers market frictions like transaction costs and liquidity restrictions. Maximizing expected return and minimizing variance are virtually mandatory objectives in most real-world portfolio optimization problems. On the other hand, minimizing the Conditional Value-at-Risk is incorporated to handle market risk more accurately in
practical asset allocation applications. This final step extends the approach discussed by Roman, Darby-Dowman and Mitra [269] to our multi-period setting.

Results discussed in the previous section suggest that these combined steps offer a thorough and reliable framework to optimize such portfolios. The absence of questionable assumptions (e.g., no market frictions or liquidity restrictions) and better uncertainty modelling (e.g., co-dependence and heavy tails) combined with a more realistic set of objectives seem better equipped to manage practical portfolios. Although some of these portfolio optimization enhancements have been proposed in the literature, to the best of our knowledge they have never been combined into a single framework.

Such improvements have only been possible because of the recent advances in computing power, which proved invaluable to implement approaches like multi-period stochastic programming. Such approaches are instrumental to real applications because they handle key features of real markets that are often dismissed by common optimization approaches.

Furthermore, the techniques and models used in the proposed framework can be replaced by newer extensions that have been recently explored and/or further developed. The Student t distribution, for instance, can be replaced by Hansen’s skewed-t distribution [125] to accommodate also potential gain and loss asymmetries in daily returns, a fifth and potentially relevant stylized fact. Bayesian Stochastic Volatility models, based on Bayesian Markov Chain Monte Carlo (MCMC) algorithms, can substitute the GJR-GARCH models to mitigate with potential model misspecifications, such as discussed by Jacquier, Polson and Rossi [152][153], and Bauwens, Hafner and Laurent [20]. The t copulas can be replaced by vine copulas, as suggested by Kurowicka and Joe [189], and Patton [237][238], which are able to explain more elaborated dependence structures and include some additional financial multivariate properties like tail dependence asymmetries. Lastly, we assumed that the true probability distribution was fully known. In practice, however, this is hardly true. In financial applications, the probability model is usually unknown and can only be estimated. This model “ambiguity”, as discussed by Pflug and Pichler [246] should also be accounted by the stochastic optimization problems to produce solutions that are more robust.
References


