



Dynamic Return Scenario Generation Approach for Large-Scale Portfolio Optimisation Framework

David Neděla¹ · Sergio Ortobelli Lozza^{1,2} · Tomáš Tichý¹

Accepted: 16 December 2023
© The Author(s) 2024

Abstract

In this paper, we propose a complex return scenario generation process that can be incorporated into portfolio selection problems. In particular, we assume that returns follow the ARMA–GARCH model with stable-distributed and skewed t -copula dependent residuals. Since the portfolio selection problem is large-scale, we apply the multifactor model with a parametric regression and a nonparametric regression approaches to reduce the complexity of the problem. To do this, the recently proposed trend-dependent correlation matrix is used to obtain the main factors of the asset dependency structure by applying principal component analysis (PCA). However, when a few main factors are assumed, the obtained residuals of the returns still explain a non-negligible part of the portfolio variability. Therefore, we propose the application of a novel approach involving a second PCA to the Pearson correlation to obtain additional factors of residual components leading to the refinement of the final prediction. Future return scenarios are predicted using Monte Carlo simulations. Finally, the impact of the proposed approaches on the portfolio selection problem is evaluated in an empirical analysis of the application of a classical mean–variance model to a dynamic dataset of stock returns from the US market. The results show that the proposed scenario generation approach with nonparametric regression outperforms the traditional approach for out-of-sample portfolios.

Keywords ARMA–GARCH model · Conditional expectations · Large-scale portfolio optimisation · Principal component analysis · Trend analysis

✉ David Neděla
david.nedela@vsb.cz

Sergio Ortobelli Lozza
sergio.ortobelli@unibg.it

Tomáš Tichý
tomas.tichy@vsb.cz

¹ Department of Finance, VSB–Technical University of Ostrava, Sokolská třída 33, 70200 Ostrava, Czech Republic

² Department of Management, University of Bergamo, Via dei Caniana 2, 24127 Bergamo, Italy

1 Introduction

Since the introduction of the mean–variance portfolio theory by Markowitz (1952), many researchers have published papers that aim to perform optimal asset allocation according to the different preferences and risk attitudes of investors, see, among others, Konno and Yamazaki (1991), Rockafellar and Uryasev (2002), Biglova et al. (2004), Rachev et al. (2005), Farinelli et al. (2008), Sharma et al. (2017), Wei et al. (2021), and Bodnar et al. (2022). Recently, an emphasis has been placed on the dynamics and nonlinear relationships on which these models are built, see Ledoit and Wolf (2017), Kouaissah et al. (2022), and Wei et al. (2021). The foundation of Markowitz’s theory is derived from the fact that the investor always strives for the best compromise between the expected returns of the portfolio and the corresponding level of risk, expressed as the variance of historical observations. In the subsequent literature, various types of approaches originating from econometrics and operations research have been suggested that strive to surpass many models, see, among others, Rachev et al. (2008); Fabozzi et al. (2010), and Stádník (2022).

In empirical analyses of optimal portfolio selection (Kondor et al., 2007; Ortobelli and Tichý, 2015; Ortobelli et al., 2019), the solutions of portfolio optimisation models are highly dependent on the precise estimation of the input dataset (statistics), which typically captures the dependency and interconnectedness between series of historical returns. According to findings in the previous literature, the return series of a financial asset is characterised as a random variable that follows a non-normal distribution with an emphasis on heavy tails, see Fama (1965), Mandelbrot and Taylor (1967), Rachev et al. (2005), and Ortobelli et al. (2017). However, a considerable number of the financial models presented include a simplifying assumption that financial returns are normally distributed, see Markowitz (1952), Black and Scholes (1973), or Merton (1973). After the implementation of an α -stable (Lévy) type of distribution function to capture the distribution of financial data by Mandelbrot and Taylor (1967), this distribution became a widely used distribution in financial modelling, see Mittnik and Rachev (1989), Rachev and Mittnik (2000), or Xu et al. (2011).

In addition, reducing the portfolio selection problem’s dimensionality becomes more important with an increasing number of traded assets. In other words, for large-scale data analysis, the considerable complexity of estimating parameters increases as the number of variables increases, see Fan and Shi (2015). To alleviate this problem, naive diversification, resampling methods, shrinkage estimators, and other similar approaches have been introduced in the literature, see Ledoit and Wolf (2003), DeMiguel et al. (2009), Pflug et al. (2012), and Pun and Wong (2019). An approach that is widely used by financial analysts and researchers is the multifactor model (Ledoit & Wolf, 2003; Fan et al., 2008; Kouaissah & Hocine, 2021; Yang & Ling, 2023). As stated by Georgiev et al. (2015), approximating the returns using a multifactor model is an appropriate option in financial modelling, especially in the procedure of portfolio construction. Since several factors allow us to perfectly obscure the cross-sectional risks, we can reduce

the number of parameters for covariance matrix estimation Fan et al. (2008). The application of single-factor and multifactor models, the number of factors being based on copula dependency estimation, was done by Oh and Patton (2017). In particular, they proposed new models for the dependence structure with respect to the joint possibility of crashes and different dependence structures during market declines and rises. Such a model is useful for high-dimensional financial variables.¹ For normally distributed data, the ordinary least squares (OLS) estimator is an appropriate tool for estimating the coefficients of a multifactor model, see, among others, Ortobelli and Tichý (2015) and Ortobelli et al. (2017). However, given that returns do not follow a normal distribution, an alternative nonparametric approach based on a conditional expectations estimator was proposed by Ortobelli et al. (2019); this approach has also become a useful technique for performance evaluation.

Given the irregularity and volatility of return time series, in this paper, we propose and investigate a novel return scenario generation approach, concentrating on factors obtained from various principal component analyses (PCAs), which are modelled by the univariate ARMA(1,1)–GARCH(1,1) model with stable-distributed and skewed t -copula dependent residuals (Ross, 1978; Chamberlain & Rothschild, 1983; Ortobelli et al., 2019; Pegler, 2019).² This approach is subsequently integrated into dynamic large-scale portfolio selection strategies. Obviously, modelling portfolios and individual assets is a challenging problem, especially using high-dimensional datasets. In several analyses, it is noted that for modelling financial series, the ARMA–GARCH model works well, see Rachev and Mittnik (2000) and Georgiev et al. (2015). Recently, machine learning has become a more popular and suitable technique for predicting random series, see Rather et al. (2015), Ramezani et al. (2021), and Ma et al. (2021). A broad comparison of time-varying models for predicting returns is provided by Cenesizoglu and Timmermann (2012). In addition, we analyse return generation strategies that contain a parametric approximation and a nonparametric approximation of return series based on main factors (principal components) that are acquired employing PCA. Parametric regression analysis uses the OLS estimator as a simple tool for determining the coefficients of the linear regression between return and factor series, see Ross (1978) and Fan et al. (2008). On the other hand, the nonparametric regression is based on the Rupert and Wand (RW) estimator, see Ruppert and Wand (1994). In contrast to the recent literature (see Ortobelli and Tichý, 2015; Ortobelli et al., 2019), where the factors are determined from the Pearson linear correlation matrix or stable correlation matrix, we consider an alternative trend-dependent correlation measure inspired by the work of Ruttiens (2013).

In general, after applying PCA to the Pearson correlation, the first 5% of factors with the highest explanatory power capture approximately one-half of the original variability of a large-scale dataset. However, we observed that using the trend correlation allows

¹ Additional literature focusing on high-dimensional dependency estimation is discussed, for example, by Fan et al. (2016) and the references therein.

² We abbreviate autoregressive moving average as ARMA and generalized autoregressive conditional heteroscedasticity as GARCH.

us to improve the approximation process because of the higher explanatory power that results from this method. To capture the meaningful factors of returns, we develop a double-PCA approach, with a further PCA applied to the obtained residuals of returns. In particular, we assume that these residuals behave as another factor that expresses a significant portion of the variability. To do this, we apply the second PCA to the Pearson correlation matrix of the residuals to generate additional factors for the error parts. This process makes it possible to capture additional types and aspects of variability and improves the approximation as a whole.

Finally, we provide an ex-ante and ex-post empirical analysis of the application of different return approximation methods in return scenario generation strategies used for a subsequent mean–variance portfolio optimisation. In particular, we analyse the impact on numerous portfolio strategies that form an efficient frontier at individual moments of the selected time period. Especially, these strategies differ according to various levels of the minimal expected return that are equidistant from each other. Therefore, we show the performance statistics and diversification measures for individual strategies. Our approach enhances the existing literature by being the first to use double PCA for approximating returns, considering trend-dependent as well as Pearson correlations. In addition, the proposed approach, which includes the approximation and generation processes, outperforms both the classical mean–variance model and the strategy with approximated returns.

The rest of this paper is structured as follows. In Sect. 2, we discuss a multifactor model and different approaches for approximating returns suitable for a large-scale portfolio selection problem. Section 3 describes the properties of the ARMA–GARCH prediction model with stable-distributed residuals and a skewed t -copula dependency structure. In Sect. 4, we provide an empirical analysis of portfolio selection strategies that employ various methods to approximate the returns and a scenario generation model in the US equity market. Finally, a summary of the results is provided in Sect. 5.

2 Approaches to Approximating Returns

The purpose of this theoretical section is to characterise parametric and nonparametric techniques for approximating the scenario of asset returns in detail. Furthermore, we consider the decomposition of the dependency structure of assets expressed by the Pearson correlation measure and the trend-dependent measures (inspired by the “accrued returns variability” of Ruttiens, 2013) for factors obtained using PCA. The parametric approximation approach employs the classical linear OLS estimator. Additionally, a nonparametric technique is used that is based on nonparametric RW locally weighted least squares regression using a kernel estimator.

2.1 Trend-Dependent Correlation PCA

In general, measuring the strength of the dependence between random series (such as asset returns) is an essential aspect of most optimisation problems. An incorrect determination of the dependency function leads to errors in many decision-making

tasks, such as the development of the prediction framework or the optimisation framework. For this reason, researchers have studied and developed variations of dependency measures to capture the relations between random variables well, see Cherubini et al. (2004), Szegö (2004), Ruttiens (2013), Ortobelli and Tichý, (2015), and the references therein.

In the present paper, we assume that the portfolio contains z risky assets with a vector of log-returns $r = [r_1, r_2, \dots, r_z]$ and a vector of asset weights $x = [x_1, x_2, \dots, x_z]$. Moreover, we consider a situation in which short sales are not allowed (i.e. $x_i \geq 0$); the vector of portfolio weights x belongs to the simplex $S = \{x \in R^z \mid \sum x_i = 1; x_i \geq 0; \forall i = 1, \dots, z\}$. The log-return of assets between time $t - 1$ and time t of the i -th asset is defined as follows:

$$r_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right) \tag{1}$$

where $P_{i,t}$ represents the price of the i -th asset at time t .

To capture the dependency structure between return series that are typical for a heavy-tailed distribution, the classical correlation measures (e.g. Pearson, Kendall, or Spearman) are generally used. However, these correlation measures may lead to shortcomings in the results (Joe, 2014).

A commonly used linear dependency measurement is the Pearson coefficient of correlation, which is given by

$$\rho_{r_i, r_j}^{Pearson} = \frac{E[(r_i - \mu_i)(r_j - \mu_j)]}{\sigma_{r_i} \sigma_{r_j}} \tag{2}$$

where μ_i is the mean value of r_i , E is the operator of the expected value, and $\sigma_{r_i} = \sqrt{E(r_i - \mu_i)^2}$.

In recent years, several papers have analysed the suitability of these measures for the portfolio selection framework and concluded that the Pearson linear correlation coefficient is not relevant for heavy-tailed data, see Ortobelli and Tichý (2015) and Ortobelli et al. (2019). An alternative that overcomes this imperfection is the frequently used copula-based approach, see, among others, Cherubini et al. (2004) and Kouaissah et al. (2022). Pearson’s linear correlation represents one possibility, while concordance (rank) measures can also be applied to express the dependency between financial variables, these measures include, for example, Kendall’s tau, Spearman’s rho, Gini’s gamma, and Blomqvist’s beta. Ortobelli and Tichý (2015) provided a discussion of the properties of linear correlation measures and their applicability to the portfolio selection framework. Moreover, they noted that to express the dependency between random variables, any type of correlation can be used, but to reduce the dimensionality of the problem, only some of the semidefinite positive correlations are appropriate. Generally, classical correlation statistics ignore the factor of time, which is important for financial management and analysis.

Recently, Ruttiens (2013) suggested a different perspective on measuring the dependency while incorporating the time (trend) factor. In particular, his concept is based on the cumulative return $c_{i,t} = [c_{i,1}, c_{i,2}, \dots, c_{i,T}]$, calculated as

$c_{i,t} = c_{i,t-1} \exp(r_{i,t})$, and the linear non-volatile trend line $e_{i,t}$, which is called the ‘equally accrued return’ and leads to an identical final cumulative value, where $t = 1, 2, \dots, T$. Additionally, e_i is simply computed as a linearly weighted return $e_{i,t} = c_{i,0} + \frac{t}{T}(c_{i,T} - c_{i,0})$, where $c_{i,0}$ represents the initial investment. Thus, the spread series $c_{i,t} - e_{i,t}$ allows us to compare the historical path of investment in the i -th asset and its non-volatile alternative. Based on the work of Ruttians, we can formulate the following correlation measure ρ_{r_i, r_j}^{Rutt} between the i -th and j -th return series:

$$\rho_{r_i, r_j}^{Rutt} = \frac{E\{[(c_i - e_i) - m_i][(c_j - e_j) - m_j]\}}{\sigma_{(c_i - e_i)}\sigma_{(c_j - e_j)}} \quad (3)$$

where $m_i = E(c_i - e_i)$ is the mean spread of the i -th returns and $\sigma_{(c_i - e_i)} = \sqrt{E[(c_i - e_i) - m_i]^2}$ is the standard deviation of the spread series of the i -th returns. However, this formula can be reformulated more precisely by excluding the mean component m_i due to the fact that usually $\forall i m_i \neq e_i$. Essentially, we are proposing the use of a second moment of deviations $c_i - e_i$. We consider that the component m_i negatively affects the final correlation values, and it is not crucial to the correlation formula. The rationale for removing this component is discussed in more detail in Nedela et al. (2023). Therefore, we consider a modified version of the Ruttians correlation measure $\rho_{r_i, r_j}^{modRutt}$, which is formulated as follows:

$$\rho_{r_i, r_j}^{modRutt} = \frac{E[(c_i - e_i)(c_j - e_j)]}{\sigma_{(c_i - e_i)}\sigma_{(c_j - e_j)}} \quad (4)$$

where the standard deviation of the spreads is calculated as $\sigma_{(c_i - e_i)} = \sqrt{E(c_i - e_i)^2}$. By doing this, we can work precisely with deviations from the risk-free optimal alternative.

In order to reduce the dimensionality of a large-scale dataset while maintaining the asset dependency structure, we apply an exponentially weighted PCA to various correlation matrices to obtain particular factors (components) explaining the required level of variability of the dataset, see Ortobelli and Tichý (2015) or Ortobelli et al. (2019). In other words, we describe the joint behaviour of z random variables using several factors f_j , where we consider the following general linear multifactor model:

$$y_i = \alpha_i + \sum_{j=1}^k \beta_{i,j} f_j + \varepsilon_i \quad \text{for } i = 1, \dots, z \quad (5)$$

where y_i is a vector of the original variable with a variance–covariance matrix Σ that is decomposed into the constant α_i , a linear combination of the k uncorrelated factors f_j and coefficients $\beta_{i,j}$, and an uncorrelated error part ε_i .

To do this, the PCA introduced by Pearson (1901) and Hotelling (1933) can be used, as it is a permissible mathematical method that provides us with the identification of the main factors (components) characterised by non-zero variance.

The principle of the PCA method is to linearly transform the initial variable into a matrix of uncorrelated variables.

The factors are ordered, which means that the first factor obtained helps to explain the largest part of the variability of the original data, the second factor explains the second-largest part of the variability, and so on. Since we obtain z factors, for a large-scale dataset, we select only the first k factors that characterise a sufficient proportion of the variability. Note that several methods of determination have been proposed to find the optimal value of k , such as the Kaiser rule and cumulative variation (Kaiser, 1960).

2.2 Parametric and Nonparametric Regression Models for Approximating Returns

In portfolio selection, using the parametric regression model, we are able to replace the original z dependent return series $\{r_i\}_{i=1}^z$ with z new uncorrelated time series denoted as $\{w_i\}_{i=1}^z$ by employing the PCA method. In doing this, each r_i is derived as a linear function of the w_i series. For the general multifactor model (5), we can replace a random vector y_i with the return series $r_{i,t}$ that is derived by k -factors given by

$$\hat{r}_i = \alpha_i + \sum_{j=1}^k \beta_{i,j} f_j + \varepsilon_i, \quad \text{for } i = 1, \dots, z \quad (6)$$

where α_i is the fixed intercept of the i -th asset, $\beta_{i,j}$ is the coefficient related to the factor f_j , and ε_i is the error part of the i -th asset estimation. Note that ε_i is a composition of unused uncorrelated factors, and for this reason, ε_i is uncorrelated with the first k principal components (factors). Moreover, we can say that if the joint distribution of the original series is a multivariate Gaussian distribution, then both the factors f_j and the error part ε_i follow the Gaussian distribution and ε_i is independent of the factors.

In order to estimate the regression parameters, the OLS estimator is a simple and preferable method to use, especially when the regression is linear and the original series is normally distributed (Ortobelli et al., 2017; Kouaissah et al., 2022). However, it is evident that the returns of financial assets typically follow a fat-tailed distribution and a nonlinear dependency structure, see Fama (1965), Mandelbrot (1963), Rachev and Mittnik (2000), or Ortobelli et al. (2017).

Therefore, we should consider an alternative that fits the properties of financial returns, as stated by Ortobelli et al. (2017) or Ortobelli et al. (2019), such as a nonparametric regression method. This approach uses conditional expectations and a multivariate kernel estimator. Assuming that the factors are determined by performing PCA on the linear correlation matrix, the formulation of the nonparametric approach is as follows:

$$r = E(r | F = f) + \varepsilon = m(f) + \varepsilon \quad (7)$$

where F represents the matrix with $f = (f_1, \dots, f_k)$ vectors of k uncorrelated factors and ε is the error of estimation. Given the unknown general form of the function

$m(f)$, we should use a nonparametric method, as proposed by Nadaraya (1964) and Watson (1964), such as the Gasser–Müller kernel estimator (Gasser & Müller, 1984) or the locally weighted least squares method introduced by Ruppert and Wand (1994), which is employed in this analysis. The nonparametric estimation of conditional expectations using the Nadaraya–Watson predictor is presented in Herwartz (2017).

The general problem associated with the estimation of the regression function $m(f)$ is to find the value of the parameter a : this can be done, according to Ruppert and Wand (1994), by solving the following optimisation task:

$$\min_{a,b} \sum_{i=1}^T [r_i - a - b^T(f_i - f)]^2 K_H(f_i - f) \quad (8)$$

where $K_H(\cdot)$ is a multivariate kernel estimator with an $s \times s$ symmetric positive-definite matrix H that depends on the sample size T , and f_i is the i -th observation of the vector of factor f .

The essential aspect of the performance of a kernel function is the precise selection of the bandwidth rather than the type of function, see Hall and Kang (2005). For general multivariate kernel estimators $K_H(\cdot)$, Scott (2015) suggests the application of an s -dimensional multivariate Gaussian density that incorporates the variance–covariance bandwidth $H = \text{diag}(h_1, \dots, h_s)$ according to the following rule:

$$\text{Scott's rule in } \mathbb{R}^s : \hat{h}_i = \hat{\sigma}_i T^{-1/(s+4)}, i = 1, \dots, s \quad (9)$$

where $\hat{\sigma}_i$ represents the estimated standard deviation of the i -th factor f_i , and the parameter T is the length of the sample of observations. The task of this approach is to minimise the mean integrated squared error (MISE) obtained from the estimation. Additional discussions concerning the bandwidth selection procedure are presented by, e.g., Mugdadi and Ahmad (2004) and Borrajo et al. (2017). According to existing evidence presented by Ortobelli et al. (2019) and Kouaissah et al. (2022), Scott's bandwidth selection is chosen to be used in the empirical analysis in this paper.

3 ARMA–GARCH Model with Stable Distribution and Skewed t Copula Dependency Structure

In this theoretical section, we briefly introduce the method for the simulation of return scenarios, which is motivated by techniques from Rachev et al. (2008) and Biglova et al. (2009, 2014). To do this, we take into consideration that many of the random variables used in the financial environment follow the ARMA(1,1)–GARCH(1,1) model with stable-distributed and skewed t -copula dependent residuals, see Sun et al. (2008), Kim et al. (2011), and Georgiev et al. (2015). Thus, we can simply model individual return series or principal components (factors) scenarios for the future. The ARMA part of the model is used for capturing the autocorrelations, and by incorporating the GARCH part of the model, we are able to reflect several essential features of financial

time series, such as leptokurticity, conditional heteroscedasticity, and volatility clustering, see Ha and Lee (2011) and Georgiev et al. (2015).

Since the dimensionality of a large-scale portfolio problem is significant and the multifactor model with PCA is used to eliminate this property, we are able to apply the univariate ARMA(1,1)–GARCH(1,1) model to the t -th observation of the j -th obtained factor $f_{j,t}$ for $j = 1, \dots, k$, see Biglova et al. (2014) and Georgiev et al. (2015). In this situation, the model is formulated as follows:

$$\begin{aligned} f_{j,t} &= c_j + a_j f_{j,t-1} + b_j \epsilon_{j,t-1} + \epsilon_{j,t} \\ \epsilon_{j,t} &= \sigma_{j,t} u_{j,t} \\ \sigma_{j,t}^2 &= \alpha_{j,0} + \alpha_{j,1} \epsilon_{j,t-1}^2 + \beta_{j,1} \sigma_{j,t-1}^2 \end{aligned} \tag{10}$$

where $\alpha_{j,0} \geq 0, \alpha_{j,1} \geq 0, \beta_{j,1} \geq 0$, and $\alpha_{j,0} + \alpha_{j,1} + \beta_{j,1} \leq 1$, for $j = 1, \dots, k, t = 1, \dots, T$, and $u_{j,t}$ usually represents independent identically distributed (i.i.d.) random variables with a mean equal to zero and a variance of 1, which is typical for distributions with skewness and heavy tails. Note that the parameters of the models are usually estimated using the maximum likelihood method or quasi-maximum likelihood method, see Biglova et al. (2014), Ha and Lee (2011), and Georgiev et al. (2015).

From the multifactor model, considering only the first k factors, we still obtain the error part ϵ_i for each return r_i , which represents unexplained residuals that are independent of each other, as well as the factors f_j . Since we wish to reduce the problem’s dimensionality, it is not desirable to consider a vast number of factors. According to the empirical findings, the error part ϵ_i for each vector of returns r_i (in a large-scale problem) could have a higher explanatory power than the most significant factor f_j . For this reason, the significance of these errors cannot be overlooked and neglected. To improve the whole simulation process, we apply a second PCA to the Pearson correlation matrix of the vectors of errors ϵ_i for $i = 1, \dots, z$, which are obtained from the previous PCA. By performing the second PCA, we reduce the dimensionality of the unexplained residuals to obtain other significant factors for the final approximation. For this reason, the identical ARMA(1,1)–GARCH(1,1) model (10) is applied in order to generate the scenarios of an additional l factors $f_{\epsilon_{ij}}$, for $j = 1, \dots, l$, from the i -th return residuals.

Moreover, although we extend the matrix of factors used with the second PCA, we still obtain the remaining residual part ϵ_i , which is naturally smaller than it would be if only a single PCA was used. According to Georgiev et al. (2015), to model these residuals, we consider the identical ARMA(1,1)–GARCH(1,1) model with a mathematical formulation given by

$$\begin{aligned} \epsilon_{i,t} &= c + a \epsilon_{i,t-1} + b \epsilon_{i,t-1} + \epsilon_{i,t} \\ \epsilon_{i,t} &= \sigma_{i,t} u_{i,t} \\ \sigma_{i,t}^2 &= \alpha_{i,0} + \alpha_{i,1} \epsilon_{i,t-1}^2 + \beta_{i,1} \sigma_{i,t-1}^2 \end{aligned} \tag{11}$$

where $\alpha_{i,0} \geq 0, \alpha_{i,1} \geq 0, \beta_{i,1} \geq 0, \alpha_{i,0} + \alpha_{i,1} + \beta_{i,1} \leq 1, i = 1, \dots, n$, and $t = 1, \dots, T$. Simultaneously, we determine the model parameters with the help of the classical maximum likelihood method. Again, we assume that the interdependencies of the

unexplained residuals $\epsilon_{i,t}$ are expressed by a skewed t copula dependency with stable-distributed marginals.

For the previous models (10) and (11), we assume that the residuals $\epsilon_{j,t}$ follow the α_j -stable distribution and skewed t -copula interdependence. The distribution of financial return series is generally characterised by skewness, a higher peak, and two heavier tails compared to a normal distribution (Mittnik & Rachev, 1989; Xu et al., 2011). Therefore, we proceed by performing the following steps. First, we have to approximate the empirical standardised residuals $\hat{u}_{j,t} = \frac{\hat{\epsilon}_{j,t}}{\sigma_{j,t}}$ using the α_j -stable distribution $S_{\alpha_j}(\sigma_j, \beta_j, \mu_j)$, where $\alpha \in (0, 2]$, $\sigma_j \geq 0$, $\beta_j \in [-1, 1]$, and $\mu_j \in \mathbb{R}$, see Rachev and Mittnik (2000). The parameter α is a stability index that determines the behaviour of the tails, σ_j is the scale parameter, μ_j is the shift parameter, and β_j is the skewness parameter. The residuals are simply computed, e.g. from the model (10), by $\hat{\epsilon}_{j,t} = f_{j,t} - a_{j,0} - a_{j,1}f_{j,t-1} - b_{j,1}\epsilon_{j,t-1}$ for $j = 1, \dots, k$. To estimate the parameters of the α_j -stable distribution, we use the maximum likelihood method. Then, we simulate S stable-distributed future scenarios for each of the standardized residual series $\hat{u}_{j,t}$ and determine the sample distribution for the same series as $F_{\hat{u}_{j,T+1}}^{Stable}(x) = \frac{1}{S} \sum_{s=1}^S I_{\{\hat{u}_{j,T+1}^{(s)} \leq x\}}$, $x \in \mathbb{R}$, $j = 1, \dots, k$, where $\hat{u}_{j,t}^{(s)}$ is the s -th future standardised residual.

We use the maximum likelihood method to fit the parameters P and γ of a skewed t distribution (ν degrees of freedom) from $\hat{u}_{j,t}$ using the following formulation:

$$V = \mu + \gamma g(W) + \sqrt{g(W)}Z \quad (12)$$

where $g : [0, \infty) \rightarrow [0, \infty)$, γ is a vector of skewness parameters, $Z \sim N(0, \Sigma)$ with a covariance matrix $\Sigma = [\sigma_{i,j}]$, and $W \sim IG(\frac{\nu}{2}, \frac{\nu}{2})$, while IG represents an inverse Gaussian distribution. In accordance with the literature, we consider that $\nu = 5$, and then $\hat{\Sigma} = [cov(x) - \frac{(2\nu^2)}{(\nu-2)^2(\nu-4)}\hat{\gamma}\hat{\gamma}'] \frac{(\nu-2)}{\nu}$, see Zugravu et al. (2013). After the estimation, we simulate Q scenarios S^q using (12), where $S^q = (S_1^q, \dots, S_k^q)$, k is the number of factors, and $q = 1, \dots, Q$. If we denote the marginal distributions by $F_{S_j}(x)$, then we can formulate $((S_1^q, \dots, F_k(S_k^q)) = (P_1(X_1 \leq S_1^q), \dots, P_k(X_k \leq S_k^q))$, and we obtain S scenarios from a uniform random vector $V = (V_1, \dots, V_k)$ with a joint distribution given by the copula $C_{v,P,\gamma}^t$. Finally, the vector V can be transformed into dependent standardised residuals $\hat{u}_{j,t+1} = (F_j^{Stable})^{-1}(V_j)$, considering the marginal α_j -stable distribution. We are able to generate $\hat{\epsilon}_{j,T+1} = \hat{\sigma}_{j,T+1}\hat{u}_{j,T+1}$, where $\hat{\sigma}_{j,T+1}$ is the volatility forecast obtained using the ARMA(1,1)–GARCH(1,1) model.

4 Empirical Analysis of the US Stock Market

In this section, we characterise the dataset and present an analysis, investigation, and discussion of the approximation of the returns and the approaches for predicting returns introduced in Sects. 2 and 3; these approaches are incorporated into the complex mean–variance portfolio selection strategies.

Table 1 Comparison of the average of parameters computed using all daily returns for the considered dataset

Mean (%)	SD (%)	Skew	Kurt	Min	Max
0.0348	2.3440	-0.1849	11.6921	-0.2347	0.1757
α	β	σ	μ	JB(95%)	KS(95%)
1.6953	0.1046	0.0129	-8×10^{-05}	0.0078	0.0456

4.1 Characterisation of the Dataset and Empirical Procedure

To perform the empirical analysis, we consider a dynamic dataset of daily historical return observations of 942 stocks traded on the US market that formed the components of the S &P500 index. The data period, which begins on 2 January 2002 and runs until 31 December 2021, has a total of 5037 daily observations. The dynamic nature of the dataset lies in the fact that we captured all the changes that occurred in the composition of the index during the analysed period at regular 3-month intervals. This allows us to provide a more precise and realistic analysis and eliminate the influence of survivorship bias (see, e.g., Brown et al., 1995). The dataset was gathered from the Thomson Reuters Datastream.

Since we assume in this analysis that individual stock returns follow a stable distribution, we estimate the particular parameters (α , σ , β , μ) using the maximum likelihood method; the results are reported in Table 1. Furthermore, we show the average results for the mean (Mean), standard deviation (SD), skewness (Skew), kurtosis (Kurt), minimum (Min), and maximum (Max). Finally, we compute the percentage of rejections for the Jarque–Bera (JB) and Kolmogorov–Smirnov (KS) statistics to test the normality of the data with a 95% confidence level. We emphasise that when considering the whole period of return observations, we observe that the returns are not stationary due to the occurrence of financial crises. However, if we split the dataset into crisis-affected and non-crisis periods, we can observe that the data in the non-crisis periods are stationary.³

According to the average values, the normal distribution is rejected for almost all return series by both statistical tests. Moreover, stock returns present a negative skewness and high kurtosis (on average), which indicate the presence of heavy tails. This conclusion is also confirmed by the β parameter and the α parameter, which has an average value below the threshold of 2. Since the time period includes several crises, the absolute value of the average minimum is lower than the absolute value of the average maximum.

Due to the sensitivity of portfolio optimisation problems to the frequency of re-calibration, we consider a monthly re-calibration interval (meaning 21 trading days), with a half-year rolling window of historical observations (126 trading days), to determine all necessary parameters of the approximation and generation models. Moreover, short sales of stocks are not allowed during investment, and the limit of

³ Although the GARCH model is conditioned by the stationarity of the data, this assumption is usually met since we are performing an empirical analysis with short periods of observations.

the investment in an individual asset is not defined, i.e. $x_i \leq 1$. The initial amount invested in each portfolio is $W_0 = 1$.

The following steps are performed to optimise the various portfolios and then compute the performance statistics and the final wealth paths:

Step 1: PCA is applied to the trend-dependent correlation matrix $\Theta = (\rho_{r_i, r_j}^{modRutt})_{z \times z}$ defined in Eq. (4) to obtain z factors (principal components) from among all available asset returns to reduce the dimensionality using Eq. (5). By doing this, the selected $k = 6$ factors (we observe that six factors are sufficient to explain more than 95% of the total trend-variance in all cases) of a large-scale portfolio are obtained. Furthermore, we compute the residual part $\hat{\varepsilon}_i$ for each return r_i and create a Pearson correlation matrix to consider the correlation between these residual series. Then, we apply a second PCA to the Pearson correlation matrix in order to find an additional $l = 25$ factors that capture approximately 40%–60% of the residual variability. Thus, the total number of factors is 31. Moreover, a small error component is still obtained.

Remark 1 The approximation procedure based on two separate PCAs performed on different types of correlation matrices takes into account different aspects of the variability. In particular, using this approach, we capture the market risk using the Pearson correlation and the trend-dependent risk using the modified Ruttens correlation. Furthermore, according to the empirical results, we can observe the benefit of applying PCA to the trend-dependent correlation in the strong explanatory power of a small number of principal factors. For example, to explain approximately 85% of the total portfolio variability, using the Pearson correlation PCA, we need to employ more than three times as many factors compared to the trend-dependent correlation PCA. Therefore, this approach allows us to more effectively capture the variability of the large-scale portfolio. Recall that we put an emphasis on the trend-dependent variability, which we consider an essential aspect of the portfolio variability.

In addition, the second PCA applied to the Pearson correlation provides insight into the market portfolio variability. To capture important factors, we apply PCA to the dependency matrix of the residuals of the returns obtained from the initial PCA. The residuals are essentially used to mitigate the double-capture impact of the main factors.

Step 2: In the portfolio selection scenario including the return generation process (hereinafter referred to as S-generation), future scenarios of all $j = 1, \dots, 31$ factors f_j and a small error part ε_i are generated while supposing that the factors obtained from the previous step follow the ARMA(1,1)–GARCH(1,1) models (10) and (11) with stable-distributed residuals, which are interdependent according to the skewed t -copula function explained in Sect. 3. Using the maximum likelihood method, models for each factor and error series are fitted individually. In addition, a Monte Carlo simulation is performed to simulate $S = 1000$ future scenarios of individual series.⁴ In contrast, for the second portfolio selection scenario (hereinafter referred to as

⁴ Based on our proprietary analysis, we did not observe a significant difference in the final portfolio performance for a higher number of scenarios, e.g. $S = 2000$ or $S = 5000$. However, using a higher number of scenarios is significantly more computationally expensive and time-consuming.

S-approximation), the scenario generation process is not considered, and only historical approximate returns obtained using the OLS or RW techniques are used in later steps.

Step 3: Approximated return series r_i are computed by employing the linear parametric OLS estimator or nonparametric RW estimator presented in Sect. 2.2 to the simple factors (principal components) or the generated factors from *Step 2* depending on the selected portfolio selection strategy.

Step 4: The optimal portfolio x_{opt} is found by solving the minimisation of the global variance framework (for the last strategy, the maximisation of the expected return) for admissible levels of the expected return M , for the generated asset returns (using the S-generation strategy) or approximated returns (using the S-approximation strategy) from *Step 2* as follows:

$$\begin{aligned}
 \min (x_{opt}' \Sigma x_{opt}) \\
 E(x_{opt}' r) \geq M \\
 x_{opt}' \varphi = 1 \\
 0 \leq x_i \leq 1; i = 1, \dots, z
 \end{aligned}
 \tag{13}$$

where Σ is the variance–covariance matrix and $\varphi' = (1, 1, \dots, 1)$, it is a vector of ones. These optimisations problems are solved using MATLAB solvers.

Step 5: In the final step, we compute the ex-post final wealth of portfolios as follows:

$$W_{t_{k+1}} = W_{t_k} (x_{opt}' r_{t_k}^{ex-post})
 \tag{14}$$

where $r_{t_k}^{ex-post}$ is the vector of gross returns for the time period between t_k and t_{k+1} . The time $t_{k+1} = t_k + \zeta$, where $\zeta = 21$. The algorithm, from *Step 1* to *Step 5*, is then repeated until no more observations are available.

4.2 Ex-post and Ex-Ante Results and Discussion

In this subsection, we present the results obtained from the ex-post empirical analysis based on the empirical process above. These results are provided in Tables 2 and 3 and Figs. 1, 2 and 3. The portfolio statistics that we consider are the mean (Mean(%)), standard deviation (SD(%)), VaR5%, CVaR5%, and final wealth (Final W).⁵ We compute two performance measures,⁶ i.e. the Sharpe ratio (SR) and the Rachev ratio (RR). To compute the excess return in performance ratios, we consider a 3-month US Treasury bill return as the risk-free rate. Note that in all tables, we

⁵ For a detailed explanation and mathematical formulation of the Value at Risk (VaR) and Conditional Value at Risk (CVaR) indicators, see e.g. Rachev et al. (2008) or Biglova et al. (2004).

⁶ Performance measures usually compare the excess return to unit of particular risk measure. The most used performance measure is the Sharpe ratio, in which the standard deviation represents the risk (Sharpe, 1994) In contrast, the Rachev ratio considers only the CVaR, which indicates both the expected return and the expected loss (Biglova et al., 2004).

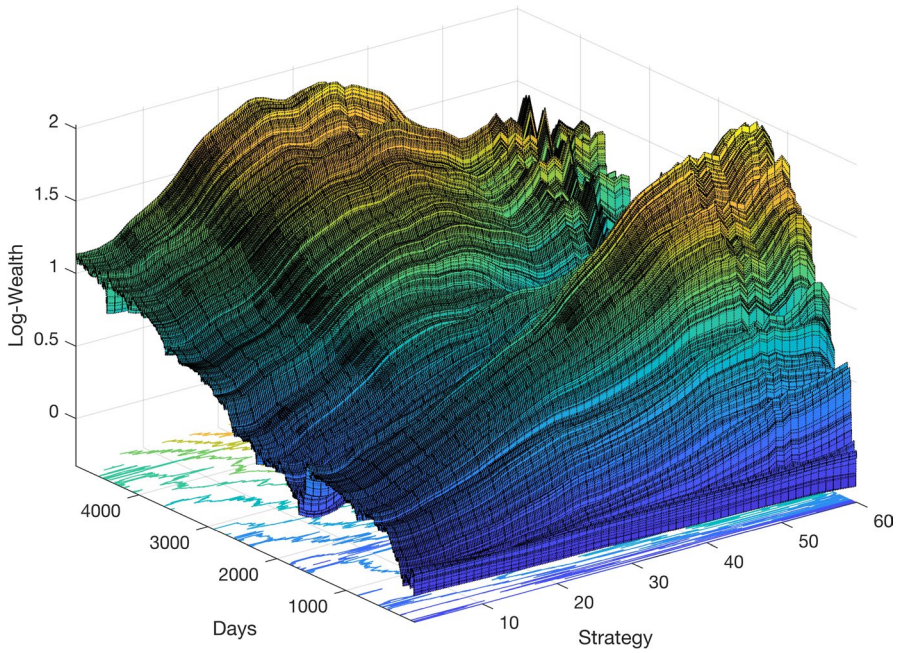


Fig. 1 Ex-post Log-Wealth for mean-variance strategies using S-approximated scenario with OLS estimator

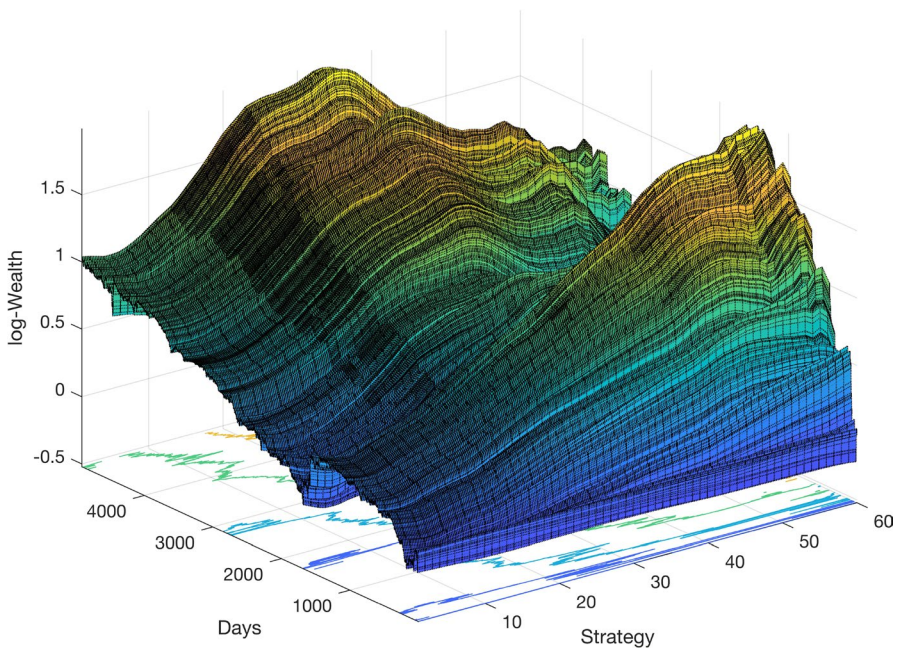


Fig. 2 Ex-post Log-Wealth for mean-variance strategies using S-approximated scenario with RW estimator

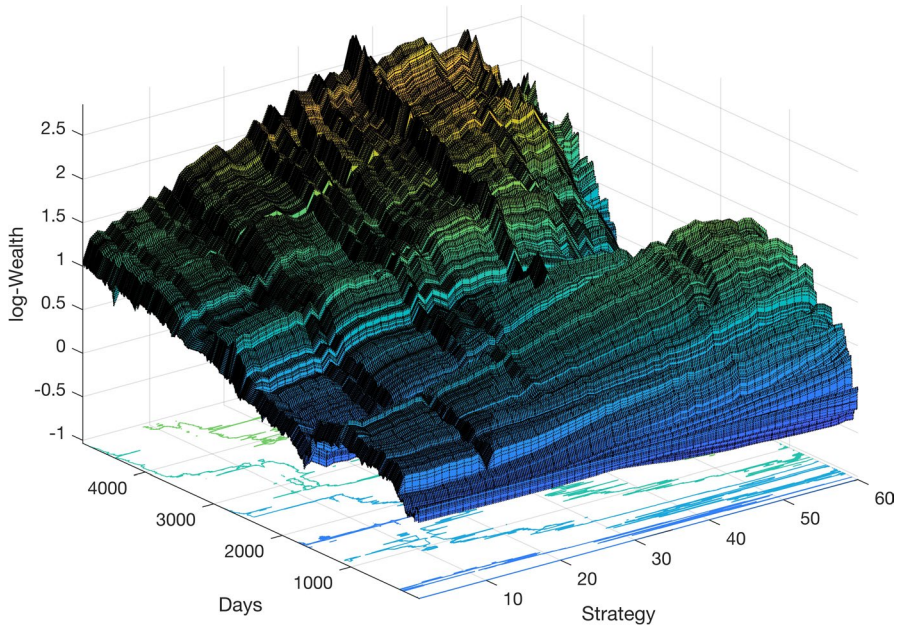


Fig. 3 Ex-post Log-Wealth for mean-variance strategies using S-generated scenario with RW estimator

show results only for selected portfolio strategies to illustrate the significant differences between the portfolio statistics.

In Figs. 1 and 2, we show the ex-post paths of the logarithm of wealth (Log-Wealth) obtained for various portfolio strategies using the OLS and RW approximations without the scenario generation process, respectively. For this comparison, we provide a performance and risk analysis of parametric and nonparametric approximation techniques applied to a dynamic dataset for the first time. From the paths in Figs. 1 and 2, we can see that the RW approach effectively smooths the wealth path for particular portfolio strategies. In particular, the trend-correlation PCA with RW approximation leads to a slightly lower variability, especially after the impact of the financial crisis (2008–2009). Generally, due to the dynamic dataset of stocks, portfolios have been dealing with the effects of the crisis for a long time, and this is usually observable in the real world. Interestingly, strategies that require the highest risk generate the best profitability until the beginning of the crisis, but after that point, the returns are not sufficient to achieve the corresponding level before the COVID-19 market downturn. Furthermore, for risk-preferred strategies (from the interval 50–60), the highest differences in performance occur; these differences are expressed as peaks. Overall, the RW approach outperforms the OLS approach for middle portfolio strategies, and it also obtains the highest Final W. Therefore, we use the RW approximation approach for the further analysis of the predictive power.⁷

⁷ The benefits of the RW estimator with respect to the estimation error are provided by Ortobelli et al. (2017, 2019).

Table 2 Portfolio statistics computed for daily ex-post returns obtained for selected mean–variance strategies considering S-approximated scenario with OLS and RW estimators

Strategy	S-approximated OLS						S-approximated RW						
	Mean (%)	SD (%)	VaR5%	CVaR5%	Final W	RR	Mean (%)	SD (%)	VaR5%	CVaR5%	Final W	SR	RR
1	0.0231	0.7745	0.0102	0.0188	3.1094	0.0287	0.0211	0.7549	0.0101	0.0185	2.8236	0.0268	0.8784
2	0.0229	0.7693	0.0103	0.0187	3.0868	0.0287	0.0209	0.7521	0.0102	0.0185	2.7873	0.0266	0.8740
3	0.0230	0.7658	0.0102	0.0186	3.0931	0.0289	0.0206	0.7507	0.0103	0.0185	2.7478	0.0262	0.8691
4	0.0230	0.7639	0.0103	0.0186	3.0953	0.0290	0.0202	0.7508	0.0104	0.0185	2.7012	0.0258	0.8657
5	0.0233	0.7635	0.0104	0.0186	3.1471	0.0294	0.0202	0.7517	0.0105	0.0185	2.7026	0.0258	0.8622
11	0.0274	0.7965	0.0118	0.0194	3.8358	0.0333	0.0237	0.7849	0.0117	0.0193	3.2044	0.0291	0.8635
12	0.0280	0.8072	0.0121	0.0197	3.9627	0.0336	0.0244	0.7947	0.0119	0.0195	3.3217	0.0296	0.8643
13	0.0289	0.8190	0.0123	0.0200	4.1399	0.0342	0.0252	0.8057	0.0124	0.0198	3.4420	0.0301	0.8649
14	0.0299	0.8321	0.0127	0.0203	4.3431	0.0349	0.0262	0.8175	0.0128	0.0201	3.6163	0.0309	0.8665
15	0.0311	0.8460	0.0130	0.0206	4.5965	0.0357	0.0274	0.8305	0.0131	0.0205	3.8403	0.0319	0.8685
21	0.0365	0.9469	0.0147	0.0232	5.9940	0.0376	0.0351	0.9313	0.0146	0.0230	5.5951	0.0367	0.8861
22	0.0369	0.9680	0.0151	0.0237	6.1256	0.0372	0.0360	0.9518	0.0152	0.0234	5.8517	0.0369	0.8888
23	0.0375	0.9903	0.0155	0.0242	6.3027	0.0370	0.0371	0.9742	0.0156	0.0240	6.1833	0.0372	0.8910
24	0.0380	1.0138	0.0159	0.0248	6.4757	0.0366	0.0374	0.9961	0.0159	0.0245	6.2822	0.0367	0.8925
25	0.0381	1.0379	0.0164	0.0254	6.5082	0.0359	0.0377	1.0199	0.0164	0.0251	6.3811	0.0361	0.8939
31	0.0373	1.2063	0.0191	0.0298	6.2506	0.0302	0.0390	1.1883	0.0193	0.0293	6.8016	0.0321	0.8964
32	0.0375	1.2402	0.0197	0.0307	6.2943	0.0295	0.0395	1.2205	0.0197	0.0301	6.9663	0.0317	0.8970
33	0.0376	1.2746	0.0201	0.0315	6.3343	0.0288	0.0397	1.2542	0.0202	0.0309	7.0229	0.0309	0.8960
34	0.0372	1.3109	0.0206	0.0325	6.2125	0.0277	0.0393	1.2895	0.0206	0.0318	6.9076	0.0298	0.8950
35	0.0357	1.3479	0.0211	0.0333	5.7870	0.0259	0.0387	1.3265	0.0211	0.0327	6.6952	0.0285	0.8939
41	0.0314	1.6087	0.0245	0.0397	4.6793	0.0190	0.0301	1.5892	0.0243	0.0392	4.3955	0.0184	0.8794
42	0.0297	1.6624	0.0256	0.0411	4.3087	0.0174	0.0288	1.6411	0.0252	0.0405	4.1062	0.0170	0.8776

Table 2 (continued)

Strategy	S-approximated OLS										S-approximated RW																		
	Mean (%)	SD (%)	VaR5%	CVaR5%	Final W	SR	RR	Mean (%)	SD (%)	VaR5%	CVaR5%	Final W	SR	RR	Mean (%)	SD (%)	VaR5%	CVaR5%	Final W	SR	RR								
43	0.0279	1.7164	0.0266	0.0424	3.9336	0.0157	0.8834	0.0273	1.6974	0.0262	0.0420	3.8152	0.0155	0.8757	0.0256	1.8353	0.0282	0.0454	3.5222	0.0135	0.8799	0.0248	1.8166	0.0277	0.0450	3.3862	0.0132	0.8746	
44	0.0266	1.7739	0.0275	0.0439	3.6968	0.0145	0.8809	0.0262	1.7550	0.0267	0.0434	3.6143	0.0144	0.8749	0.0165	2.2680	0.0340	0.0560	2.2462	0.0069	0.8743	0.0156	2.2529	0.0343	0.0557	2.1466	0.0065	0.8687	
45	0.0256	1.8353	0.0282	0.0454	3.5222	0.0135	0.8799	0.0248	1.8166	0.0277	0.0450	3.3862	0.0132	0.8746	0.0193	2.3216	0.0354	0.0570	2.5801	0.0079	0.8828	0.0122	2.3371	0.0351	0.0578	1.8249	0.0049	0.8672	
51	0.0165	2.2680	0.0340	0.0560	2.2462	0.0069	0.8743	0.0156	2.2529	0.0343	0.0557	2.1466	0.0065	0.8687	0.0124	2.4405	0.0368	0.0601	1.8372	0.0047	0.8758	0.0096	2.4277	0.0369	0.0601	1.6039	0.0036	0.8684	
52	0.0193	2.3216	0.0354	0.0570	2.5801	0.0079	0.8828	0.0122	2.3371	0.0351	0.0578	1.8249	0.0049	0.8672	0.0094	2.5405	0.0384	0.0625	1.5904	0.0034	0.8795	0.0065	2.5261	0.0381	0.0624	1.3728	0.0022	0.8719	
53	0.0124	2.4405	0.0368	0.0601	1.8372	0.0047	0.8758	0.0096	2.4277	0.0369	0.0601	1.6039	0.0036	0.8684	0.0071	2.6392	0.0403	0.0651	1.4199	0.0024	0.8821	0.0046	2.6270	0.0405	0.0650	1.2565	0.0014	0.8780	
54	0.0094	2.5405	0.0384	0.0625	1.5904	0.0034	0.8795	0.0065	2.5261	0.0381	0.0624	1.3728	0.0022	0.8719	0.0050	3.0218	0.0450	0.0746	1.2813	0.0014	0.8998	0.0007	3.0073	0.0451	0.0744	1.0340	-0.0001	0.8956	
55	0.0071	2.6392	0.0403	0.0651	1.4199	0.0024	0.8821	0.0046	2.6270	0.0405	0.0650	1.2565	0.0014	0.8780	0.0073	3.1916	0.0480	0.0784	1.4306	0.0020	0.9068	0.0025	3.1814	0.0482	0.0784	1.1287	0.0005	0.9016	
58	0.0050	3.0218	0.0450	0.0746	1.2813	0.0014	0.8998	0.0007	3.0073	0.0451	0.0744	1.0340	-0.0001	0.8956	0.0058	3.3883	0.0500	0.0831	1.3267	0.0014	0.9071	0.0007	3.3827	0.0506	0.0832	1.0325	-0.0001	0.9031	
59	0.0073	3.1916	0.0480	0.0784	1.4306	0.0020	0.9068	0.0025	3.1814	0.0482	0.0784	1.1287	0.0005	0.9016	0.0058	3.3883	0.0500	0.0831	1.3267	0.0014	0.9071	0.0007	3.3827	0.0506	0.0832	1.0325	-0.0001	0.9031	
60	0.0058	3.3883	0.0500	0.0831	1.3267	0.0014	0.9071	0.0007	3.3827	0.0506	0.0832	1.0325	-0.0001	0.9031															

Table 3 Portfolio statistics computed for daily ex-post returns obtained for selected mean–variance strategies using S-generation scenario with RW estimators

Strategy	Mean (%)	SD (%)	VaR5%	CVaR5%	Final W	SR	RR
1	0.0215	0.7739	0.0103	0.0194	2.8752	0.0266	0.8555
2	0.0264	0.8783	0.0116	0.0224	3.6638	0.0291	0.8888
3	0.0277	0.8908	0.0118	0.0226	3.9072	0.0302	0.9055
4	0.0269	0.8908	0.0118	0.0226	3.7555	0.0292	0.9038
5	0.0288	0.8947	0.0122	0.0226	4.1223	0.0312	0.9100
11	0.0315	0.9373	0.0132	0.0237	4.6952	0.0327	0.8908
12	0.0339	0.9269	0.0130	0.0231	5.2908	0.0356	0.9075
13	0.0319	0.9210	0.0133	0.0232	4.7956	0.0337	0.8783
14	0.0357	0.9433	0.0135	0.0234	5.7762	0.0369	0.9124
15	0.0372	0.9573	0.0138	0.0237	6.2089	0.0379	0.9114
21	0.0382	1.0427	0.0154	0.0256	6.5178	0.0358	0.9208
22	0.0393	1.0702	0.0159	0.0261	6.9002	0.0359	0.9359
23	0.0402	1.0703	0.0159	0.0261	7.2154	0.0368	0.9321
24	0.0436	1.1062	0.0169	0.0268	8.5155	0.0386	0.9560
25	0.0418	1.1195	0.0171	0.0271	7.7972	0.0366	0.9496
31	0.0464	1.2446	0.0198	0.0299	9.7916	0.0366	0.9614
32	0.0500	1.2734	0.0199	0.0303	11.6492	0.0386	0.9721
33	0.0471	1.2947	0.0200	0.0310	10.0892	0.0357	0.9621
34	0.0493	1.3231	0.0205	0.0315	11.2668	0.0366	0.9669
35	0.0476	1.3525	0.0211	0.0323	10.3713	0.0346	0.9632
36	0.0491	1.3850	0.0216	0.0329	11.1447	0.0348	0.9681
37	0.0564	1.4371	0.0220	0.0337	15.9319	0.0386	0.9972
38	0.0486	1.4513	0.0224	0.0345	10.9050	0.0329	0.9724
41	0.0492	1.5460	0.0241	0.0366	11.1920	0.0312	0.9792
42	0.0492	1.5832	0.0248	0.0374	11.2256	0.0305	0.9828
43	0.0507	1.6233	0.0252	0.0383	12.0663	0.0307	0.9880
44	0.0509	1.6688	0.0260	0.0393	12.2096	0.0300	0.9924
45	0.0506	1.7216	0.0266	0.0403	12.0096	0.0289	1.0024
51	0.0462	2.0444	0.0322	0.0484	9.6701	0.0222	1.0045
52	0.0496	2.1273	0.0334	0.0501	11.4292	0.0229	1.0120
53	0.0415	2.1841	0.0341	0.0520	7.6842	0.0186	0.9984
54	0.0392	2.2741	0.0356	0.0542	6.8502	0.0168	1.0011
55	0.0380	2.3745	0.0372	0.0564	6.4541	0.0156	1.0129
58	0.0291	2.6958	0.0415	0.0650	4.1733	0.0105	0.9968
59	0.0276	2.8348	0.0435	0.0685	3.8761	0.0094	0.9906
60	0.0329	3.0117	0.0453	0.0728	5.0387	0.0106	0.9836

In order to compare particular ex-post portfolio statistics of selected strategies, we present Table 2, which includes the results of both approximation techniques. Generally, the results indicate that the best performance (based on the Mean and Final W) is achieved for the middle strategies using the RW estimator. In other words, the most advantageous attitude of the investor is that in which the risk and reward are considered similarly important. Usually, the riskiness of portfolios (SD, VaR5%, or CVaR5%) is similar or slightly lower in the RW approximation scenario compared to the OLS approach. On the contrary, the worst performance is achieved when the expected return of the portfolio is maximised regardless of the risk.

Moreover, we use the portfolio selection strategy with the simulated Q scenarios of factors obtained using the ARMA–GARCH model, which should provide additional benefits. By doing this, we are able to generate scenarios of the future returns of assets that are used for optimisation. In Table 3, we present the portfolio statistics obtained when the S-generation scenario with the RW regression estimator is used, and Fig. 3 shows the final Log-Wealth paths of all strategies.

As can be seen in Table 3, the performance results of the portfolio strategies based on the predicted return series significantly outperform the portfolios that are obtained when only the approximate return series are considered. In particular, the final wealth and mean return of the best strategy are approximately doubled compared to those of the S-approximation strategy with the RW estimator, as shown in Table 2. However, when this approach is used, slightly higher values of the risk indicators (SD, VaR5%, and CVaR5%) are obtained, and they do not differ significantly from the S-approximation scenario. Overall, the obtained performance ratios (SR and RR) are also at a higher level, which means that this strategy is better suited to a real application.

Similarly to the previous analysis, Fig. 3 shows the wealth paths obtained after optimisation was applied to the returns obtained using the S-generated scenario with the RW estimator. In particular, it is clearly evident that the surface plot of the paths is less smooth than that of the S-approximated strategy with the RW estimator shown in Fig. 2. The paths of adjoining strategies behave identically but at different levels, leading to sharp peaks. However, the main benefit of this scenario lies in its faster adaptability during and after the crisis periods.

For the final evaluation, the ex-ante diversification analysis is performed using several indicators to examine the changes in the weights between re-calibration moments. The turnover ρ expresses the impact of re-optimisation on the portfolio composition. The value of ρ_k after the k -th re-calibration is formulated as follows:

$$\rho_k = \sum_{i=1}^z |x_i^k - x_i^{k-1}| \tag{15}$$

where x_i^k is the weight assigned to the asset i at the re-calibration time k . Moreover, ρ_k is contained in the interval $[0, 2]$; a value of 0 indicates that all weights remained unchanged after the k -th re-calibration, and a value of 2 means that there has been a complete transformation of the portfolio composition.

In addition, we show a simple portfolio diversification measure: the number of assets that form the components of the portfolio at the k -th re-calibration time.

Using this measure, we are able to examine the changes in the portfolio composition with respect to particular strategies. Logically, the # (number) of assets is in the interval $[1, z]$. In a unique situation, when the investment is interrupted, the value of this measure is 0.

Finally, a diversification measure based on a concentration index is analysed. In the literature, we can find it as the Herfindahl–Hirschman index (HHI), which was originally applied to express the concentration of the market, see, e.g. Hirschman (1964). However, due to its explanatory power, it may be applied to portfolio analysis. It is calculated as the sum of the squares of the weights of the individual assets:

$$HHI = \sum_{i=1}^z x_i^2 \quad (16)$$

where $HHI \in [\frac{1}{z}, 1]$. If the concentration HHI goes to zero (one), the investment is divided into a large (small) set of assets, and vice versa.

In Table 4, the mean values of all the diversification measures mentioned above are reported for selected portfolio strategies.

The results in Table 4 confirm that portfolio strategies that minimise risk have more assets in the portfolio, and the number of assets decreases with increasing acceptance of risk (higher return requirement). This fact is supported by the literature related to the effect of diversification, see Woerheide and Persson (1992), Egozcue et al. (2011) and the references therein. The RW regression generally encompasses more components in the portfolio, which is relevant for the less risky strategies. However, an interesting finding is visible when the RW approximation and generation processes are combined. The generated portfolios are divided into a larger number of assets, even for high-risk strategies. Moreover, this scenario leads to a relatively high value of turnover compared to scenarios without the generation step, which indicates more significant changes in the vector of optimal weights. Note that turnover affects the corresponding transaction costs.

Overall, the proposed approach for generating return scenarios based on double PCA, where PCA is applied first to classical correlations and then to trend-dependent correlations, and the ARMA–GARCH process with modified residuals in the mean–variance portfolio selection strategy outperforms simple portfolio selection approaches that consider only approximate returns.

5 Conclusion

In this paper, we proposed a new approach for the generation of return scenarios using the approximation of the returns and the ARMA–GARCH model with stable-distributed residuals following the dependency defined by the skewed t -copula function. In particular, we employed PCA twice: the trend-dependent correlation matrix is considered in the first step. Then, the linear correlation matrix of the residuals is included in the second PCA. By doing this, we determined the factors that best describe a large proportion of the portfolio variability, which was

Table 4 Ex-ante portfolio composition statistics obtained for selected mean–variance strategies while applying different scenarios

Strategy	S-approximated OLS			S-approximated RW			S-generated RW		
	ρ	# assets	HHI	ρ	# assets	HHI	ρ	# assets	HHI
1	0.9363	14.7554	0.2821	0.8266	22.7897	0.2136	1.0731	37.8283	0.2009
2	0.9416	15.2790	0.2761	0.8322	23.2403	0.2103	1.1538	27.9871	0.2331
3	0.9528	15.6524	0.2683	0.8416	23.3433	0.2049	1.1742	31.6524	0.2252
4	0.9622	15.8069	0.2597	0.8526	23.4592	0.1987	1.1856	35.0687	0.2158
5	0.9726	15.8841	0.2507	0.8671	23.5451	0.1923	1.2051	34.0858	0.2146
11	1.0609	15.8412	0.1991	0.9695	23.2146	0.1534	1.3218	31.6438	0.1832
12	1.0770	15.6996	0.1925	0.9853	22.8627	0.1479	1.3455	29.4635	0.1834
13	1.0929	15.6567	0.1866	1.0009	22.5150	0.1429	1.3656	31.7167	0.1746
14	1.1060	15.5451	0.1813	1.0167	22.2275	0.1385	1.3753	27.5794	0.1733
15	1.1167	15.3777	0.1764	1.0308	21.8541	0.1345	1.3999	26.4464	0.1714
21	1.1627	14.2060	0.1621	1.0900	19.5880	0.1198	1.4799	26.8970	0.1583
22	1.1710	13.9099	0.1617	1.0990	19.2575	0.1195	1.5037	23.6094	0.1613
23	1.1798	13.6266	0.1616	1.1088	18.9227	0.1198	1.5048	22.2833	0.1589
24	1.1867	13.3133	0.1622	1.1157	18.5579	0.1208	1.5257	25.4592	0.1617
25	1.1932	13.2275	0.1631	1.1225	18.0601	0.1221	1.5281	20.6652	0.1643
31	1.2184	11.5665	0.1749	1.1551	15.0815	0.1358	1.5589	19.8155	0.1818
32	1.2218	11.2661	0.1776	1.1580	14.5923	0.1391	1.5614	19.4592	0.1840
33	1.2251	11.0429	0.1810	1.1611	14.1588	0.1427	1.5654	22.9485	0.1869
34	1.2281	10.7639	0.1849	1.1650	13.6695	0.1469	1.5667	18.7124	0.1927
35	1.2334	12.6524	0.1906	1.1692	13.2189	0.1512	1.5699	20.9399	0.1955
36	1.2359	10.1159	0.1945	1.1737	12.6781	0.1559	1.5742	19.9614	0.2024
37	1.2372	9.7897	0.1999	1.1789	12.2318	0.1613	1.5770	15.3004	0.2156
38	1.2385	9.4850	0.2059	1.1852	11.7296	0.1676	1.5720	19.0300	0.2156
39	1.2374	9.2275	0.2127	1.1904	11.2232	0.1749	1.5827	17.8584	0.2238
40	1.2352	8.9313	0.2197	1.1925	10.7210	0.1828	1.5778	18.8755	0.2316
41	1.2327	8.5923	0.2263	1.1927	10.2489	0.1910	1.5809	18.4034	0.2377
42	1.2391	10.2361	0.2351	1.1942	9.7554	0.1997	1.5777	15.3262	0.2469
43	1.2363	11.8798	0.2440	1.2004	11.3348	0.2107	1.5787	15.0515	0.2591
44	1.2368	11.4893	0.2537	1.2009	10.8283	0.2207	1.5858	16.6352	0.2706
45	1.2401	9.0773	0.2643	1.2018	10.3562	0.2319	1.5868	14.4936	0.2806
51	1.2257	7.1931	0.3532	1.2035	9.8541	0.3272	1.5715	20.8712	0.3716
52	1.2190	8.8584	0.3732	1.2022	9.2618	0.3487	1.5757	16.2403	0.3940
53	1.2182	6.2403	0.3965	1.2030	7.3391	0.3735	1.5812	18.9700	0.4136
54	1.2149	10.4807	0.4247	1.2023	8.9571	0.4027	1.5776	14.9700	0.4388
55	1.2107	10.2833	0.4555	1.1981	8.2403	0.4340	1.5745	15.0773	0.4613
58	1.1815	9.8155	0.5857	1.1763	8.8712	0.5711	1.5603	15.1845	0.5927
59	1.1752	6.4549	0.6595	1.1693	7.8712	0.6486	1.5700	12.4120	0.6639
60	1.1761	3.5150	0.7677	1.1790	2.4807	0.7603	1.5689	12.0815	0.7588

expressed by different types of measures. In addition, we suggested both parametric OLS and nonparametric RW regression estimators. This was followed by an empirical investigation focused on the benefits of including the return scenario generation process in the portfolio optimisation strategy for a dynamic dataset consisting of US stocks.

The results of the first comparison, in which the impact of the OLS and RW estimation techniques on various mean–variance portfolio strategies without the generation step was analysed, show that using the RW estimator the best portfolio performance is achieved. Additionally, the riskiness of these portfolios is lower than it is when the OLS estimator is used; this is reflected in the smoother wealth paths. Second, we compared the portfolio statistics obtained by incorporating the dynamic process of factor scenario generation (ARMA–GARCH model), while the RW estimator is utilised for different portfolio strategies with those obtained without the generation process. The ex-post results of the new proposed approach show that it significantly outperforms the scenarios without the generation part in terms of profitability, while the risk increased slightly. According to the ex-ante results of diversification indicators, if the RW approach is used for return approximation, the portfolio is usually spread over a larger number of assets.

In general, our empirical findings suggest that the complex scenario consisting of the double-PCA, nonparametric approximation based on the RW estimator and the ARMA–GARCH prediction model is significantly more profitable than the other considered scenarios. The benefits of this scenario lie in the favourable properties of the dynamic ARMA–GARCH prediction and, more importantly, in its cooperation with the double PCA, which leads to an increase in the portfolio performance. We observed that the proposed procedure is able to approximate future asset return scenarios well. This is apparent specifically for mean–variance portfolio strategies that express the attitude of the average risk-seeking investor.

Acknowledgements The authors would like to thank the Czech Science Foundation (GACR) under the Project 23-06280S and SGS Research Project SP2023/019 of VSB – Technical University of Ostrava for financial support of this work.

Funding Open access publishing supported by the National Technical Library in Prague. This research has been supported by the VSB – Technical University of Ostrava institutional grant SP2023/019 and by the Czech Science Foundation GACR 23-06280S. The financial support of the European Union under the REFRESH – Research Excellence For REgion Sustainability and Hightech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition is acknowledged as well.

Declarations

Conflict of interest No potential conflict of interest was reported by the author(s).

Consent for publication The authors have consented to the submission of the work to the journal.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Biglova, A., Ortobelli, S., & Fabozzi, F. (2014). Portfolio selection in the presence of systemic risk. *Journal of Asset Management*, 15(5), 285–299.
- Biglova, A., Ortobelli, S., Rachev, S. T., et al. (2004). Different approaches to risk estimation in portfolio theory. *The Journal of Portfolio Management*, 31(1), 103–112.
- Biglova, A., Ortobelli, S., Rachev, S., et al. (2009). Modeling, estimation, and optimization of equity portfolios with heavy-tailed distributions. In S. Satchell (Ed.), *Optimizing optimization: The next generation of optimization applications and theory*. Amsterdam: Academic Press.
- Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81(3), 637–654.
- Bodnar, T., Lindholm, M., Niklasson, V., et al. (2022). Bayesian portfolio selection using VaR and CVaR. *Applied Mathematics and Computation*, 427, 127120.
- Borrajó, M. I., González-Manteiga, W., & Martínez-Miranda, M. D. (2017). Bandwidth selection for kernel density estimation with length-biased data. *Journal of Nonparametric Statistics*, 29(3), 636–668.
- Brown, S. J., Goetzmann, W. N., & Ross, S. A. (1995). Survival. *Journal of Finance*, 50(3), 853–873.
- Cenesizoglu, T., & Timmermann, A. (2012). Do return prediction models add economic value? *Journal of Banking & Finance*, 36(11), 2974–2987.
- Chamberlain, G., & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(1), 1281–1304.
- Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula Methods in Finance*. Chichester: Wiley.
- DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus Naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5), 1915–1953.
- Egozcue, M., Fuentes García, L., Wong, W., et al. (2011). Do investors like to diversify? A study of Markowitz preferences. *European Journal of Operational Research*, 215(1), 188–193.
- Fabozzi, F. J., Dashan, H., & Guofu, Z. (2010). Robust portfolios: Contributions from operations research and finance. *Annals of Operations Research*, 176(1), 191–220.
- Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business*, 38(1), 34–105.
- Fan, J., Fan, Y., & Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1), 186–197.
- Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1), C1–C32.
- Fan, J., Liao, Y., & Sh, X. (2015). Risks of large portfolios. *Journal of Econometrics*, 186(2), 367–387.
- Farinelli, S., Ferreira, M., Rossello, D., et al. (2008). Beyond Sharpe ratio: Optimal asset allocation using different performance ratios. *Journal of Banking & Finance*, 32(10), 2057–2063.
- Gasser, T., & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11(3), 171–185.
- Georgiev, K., Kim, Y. S., & Stoyanov, S. (2015). Periodic portfolio revision with transaction costs. *Mathematical Methods of Operations Research*, 81(3), 337–359.
- Ha, J., & Lee, T. (2011). NM-QELE for ARMA-GARCH models with non-gaussian innovations. *Statistics & Probability Letters*, 81(6), 694–703.
- Hall, P., & Kang, K. H. (2005). Bandwidth choice for nonparametric classification. *The Annals of Statistics*, 33(1), 284–306.
- Herwartz, H. (2017). Stock return prediction under Garch—An empirical assessment. *International Journal of Forecasting*, 33(3), 569–580.
- Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review*, 54(5), 761–762.

- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(7), 498–520.
- Joe, H. (2014). *Dependence modeling with copulas* (1st ed.). Boca Raton: Chapman and Hall/CRC.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–151.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., et al. (2011). Time series analysis for financial market meltdowns. *Journal of Banking & Finance*, 35(8), 1879–1891.
- Kondor, I., Pafka, S., & Nagy, G. (2007). Noise sensitivity of portfolio selection under various risk measures. *Journal of Banking & Finance*, 31(5), 1545–1573.
- Konno, H., & Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. *Management Science*, 37(5), 519–531.
- Kouaissah, N., & Hocine, A. (2021). Forecasting systemic risk in portfolio selection: The role of technical trading rules. *Journal of Forecasting*, 40(4), 708–729.
- Kouaissah, N., Ortobelli, S., & Jebabli, I. (2022). Portfolio selection using multivariate semiparametric estimators and a copula PCA-based approach. *Computational Economics*, 60, 833–859.
- Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5), 603–621.
- Ledoit, O., & Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *Review of Financial Studies*, 30(12), 4349–4388.
- Ma, Y., Han, R., & Wang, W. (2021). Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, 165, 113973.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, 36(4), 394–419.
- Mandelbrot, B., & Taylor, H. M. (1967). On the distribution of stock price differences. *Operations Research*, 15(6), 1057–1062.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91.
- Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica*, 41(5), 867–887.
- Mittnik, S., & Rachev, S. T. (1989). Stable distributions for asset returns. *Applied Mathematics Letters*, 2(3), 301–304.
- Mugdadi, A. R., & Ahmad, I. A. (2004). A bandwidth selection for kernel density estimation of functions of random variables. *Computational Statistics & Data Analysis*, 47(1), 49–62.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1), 141–142.
- Neděla, D., Ortobelli, S., & Tichý, T. (2023). Mean-variance vs trend-risk portfolio selection. *Review of Managerial Science*. <https://doi.org/10.1007/s11846-023-00660-x>
- Oh, D. H., & Patton, A. J. (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics*, 35(1), 139–154.
- Ortobelli, S., Kouaissah, N., & Tichý, T. (2017). On the impact of conditional expectation estimators in portfolio theory. *Computational Management Science*, 14(4), 535–557.
- Ortobelli, S., Kouaissah, N., & Tichý, T. (2019). On the use of conditional expectation in portfolio selection problems. *Annals of Operations Research*, 274(1), 501–530.
- Ortobelli, S., & Tichý, T. (2015). On the impact of semidefinite positive correlation measures in portfolio theory. *Annals of Operations Research*, 235(1), 625–652.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6), 559–572.
- Pegler, M. (2019). Large-dimensional factor modeling based on high-frequency observations. *Journal of Econometrics*, 208(1), 23–42.
- Pflug, G., Pichler, A., & Wozabal, D. (2012). The 1/n investment strategy is optimal under high model ambiguity. *Journal of Banking & Finance*, 36(2), 410–417.
- Pun, C. S., & Wong, H. Y. (2019). A linear programming model for selection of sparse high-dimensional multiperiod portfolios. *European Journal of Operational Research*, 273(2), 754–771.
- Rachev, S. T., Menn, C., & Fabozzi, F. J. (2005). *Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing*. New York: Wiley.
- Rachev, S., & Mittnik, S. (2000). *Stable Paretian Models in Finance*. Chichester: Wiley.
- Rachev, S. T., Stoyanov, S. V., & Fabozzi, F. J. (2008). *Advanced Stochastic Models, Risk Assessment and Portfolio Optimization: The Ideal Risk, Uncertainty and Performance Measures*. New York: Wiley Finance.

- Ramezani, R., Peymanfar, A., & Ebrahimi, S. B. (2021). An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in Tehran stock exchange market. *Applied Soft Computing*, 82, 105551.
- Rather, A. M., Agarwal, A., & Sastry, V. N. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, 42(6), 3234–3241.
- Rockafellar, R., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26, 1443–1471.
- Ross, S. (1978). Mutual fund separation in financial theory: The separating distributions. *Journal of Economic Theory*, 17(2), 254–286.
- Ruppert, D., & Wand, M. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3), 1346–1370.
- Ruttiens, A. (2013). Portfolio risk measures: The time's arrow matters. *Computational Economics*, 41(3), 407–424.
- Scott, D. (2015). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.
- Sharma, A., Utz, S., & Mehra, A. (2017). Omega-CVaR portfolio optimization and its worst case analysis. *OR Spectrum*, 39(2), 505–539.
- Sharpe, W. F. (1994). The sharpe ratio. *Journal of Portfolio Management*, 21(1), 49–58.
- Stádník, B. (2022). Convexity arbitrage-the idea which does not work. *Cogent Economics & Finance*, 10(1), 2019361.
- Sun, W., Rachev, S., Stoyanov, S. V., et al. (2008). Multivariate skewed student's t copula in the analysis of nonlinear and asymmetric dependence in the German equity market. *Studies in Nonlinear Dynamics & Econometrics*. <https://doi.org/10.2202/1558-3708.1572>
- Szegö, G. (2004). *Risk measures for the 21st century*. Chichester: Wiley.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, 26(4), 359–372.
- Wei, J., Yang, Y., Jiang, M., et al. (2021). Dynamic multi-period sparse portfolio selection model with asymmetric investors' sentiments. *Expert Systems with Applications*, 177, 114945.
- Woerheide, W., & Persson, D. (1992). An index of portfolio diversification. *Financial Services Review*, 2(2), 73–85.
- Xu, W., Wu, C., Dong, Y., et al. (2011). Modeling Chinese stock returns with stable distribution. *Mathematical and Computer Modelling*, 54(1–2), 610–617.
- Yang, S., & Ling, N. (2023). Robust projected principal component analysis for large-dimensional semiparametric factor modeling. *Journal of Multivariate Analysis*, 195, 105155.
- Zugravu, B., Oanea, D. C., & Anghelache, V. G. (2013). Analysis based on the risk metrics model. *Romanian Statistical Review*, 61(2), 145–154.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.