September 2026

WORKING PAPER SERIES

2026-EQM-07

Multi-Moment and Multi-Time Nonparametric Frontier-Based Mutual Fund Rating: A Buy-and-Hold Backtesting Strategy with Robust Moment Statistics and a Baseline Comparison

Tiantian REN
Corresponding author: School of Business, Xiangtan University, Xiangtan 411105, China, tiantianren@xtu.edu.cn

Kristiaan KERSTENS
Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Economie Management, Lille F-59000, France, k.kerstens@ieseg.fr

Pieter Jan KERSTENS
Data & Analytics Competence Center, Flemish Institute for Technological Research, 2400 Mol, Belgium, pieterjan.kerstens@vito.be

Albane Christine TARNAUD
IESEG School of Management, Lille F-59000, France, a.tarnaud@ieseg.fr

IÉSEG School of Management Lille Catholic University 3, rue de la Digue F-59000 Lille Tel: 33(0)3 20 54 58 92
www.ieseg.fr

Multi-Moment and Multi-Time Nonparametric Frontier-Based Mutual Fund Rating: A Buy-and-Hold Backtesting Strategy with Robust Moment Statistics and a Baseline Comparison

Tiantian Ren[*], Kristiaan Kerstens[†], Pieter Jan Kerstens[‡], Albane Christine Tarnaud[§]

September 14, 2026

Abstract

This contribution introduces a comprehensive multi-moment and multi-time nonparametric frontier framework for mutual fund rating and selection. Without limiting ourselves to traditional central moments, we systematically integrate central moments, L-moments, and truncated L-moments into the multi-moment and multi-time frontier rating models. To validate out-of-sample performance of the proposed models, we design a novel baseline approach within a buy-and-hold backtesting framework ensuring that any outperformance stems from the rating capability rather than pure luck. Applying this framework to a large database of 713 European equity mutual funds yields three key findings. First, the robust L-moments provides a better balance between total returns and maximum drawdown compared to the central moments and truncated L-moments. Second, incorporating higher-order moments improves both total returns and maximum drawdown. Finally, the nonconvex rating models demonstrate the strongest statistical significance in providing superior maximum drawdown, whereas convex rating models tend to favour aggressive, higher-risk mutual funds.

JEL codes: D24, G11

Keywords: Data Envelopment Analysis; Free Disposal Hull; Shortage Function; Robust Moments; Frontier; Fund Rating.

[*]Corresponding author: School of Business, Xiangtan University, Xiangtan 411105, China, tiantianren@xtu.edu.cn

[†]Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, Lille F-59000, France, k.kerstens@ieseg.fr

[‡]Data & Analytics Competence Center, Flemish Institute for Technological Research, 2400 Mol, Belgium, pieterjan.kerstens@vito.be

[§]IESEG School of Management, Lille F-59000, France, a.tarnaud@ieseg.fr

1 Introduction

The seminal article of Markowitz (1952) on modern portfolio theory has convinced the scientific community that portfolio performance must be measured by exploring the trade-off between portfolio return and risk using a mean–variance (MV) portfolio optimization problem that simultaneously maximizes return and minimizes risk. There has been a continuous development of variations on these bi-objective MV optimization problems focusing on alternative risk measures: expected shortfall, entropy, semi-variance and other partial moment measures, mean absolute deviation, Value-at-Risk (VaR) in all its variations, etc. (see Feibel (2003) or Bacon (2008) for surveys).

The first two moments of a random variable's distribution are only consistent with the von Neumann-Morgenstern axioms of choice underlying expected utility theory under two conditions. First, asset processes must obey a normal distribution. Second, investors must have quadratic utility functions. The limitations of this MV approach are well-known. First, there exists a large empirical literature that shows that the normality of asset returns can be rejected for various financial asset classes in emerging and developed financial markets alike (for instance, Jondeau and Rockinger (2003)). Second, investors are often assigned the following preferences for higher moments: a positive preference for skewness, and a negative preference for kurtosis (see Scott and Horvath (1980)). The properties of the utility functions for risk-aversion preferences have been studied in a more general setting by Eeckhoudt and Schlesinger (2006). These authors discuss the broad class of mixed risk-aversion utility functions that represent a preference for odd moments and an aversion for even moments.

Handling higher-order moment portfolios requires special methods. Thus far, the literature is missing a single widely accepted criterion for handling these higher-order moment portfolios. One can differentiate between primal and dual approaches to compute higher-order moment portfolio frontiers. A rather widely used example of the primal approach builds upon the seminal article of Lai (1991): he determines mean-variance-skewness (MVS) optimal portfolios using Polynomial Goal Programming methods. An example of the dual approach that requires the specification of some indirect higher-moment utility function is found in Jondeau and Rockinger (2006): these authors obtain optimal portfolios after specifying parameters reflecting higher-moment preferences.

Another primal approach for calculating higher-order moment portfolio frontiers employs efficiency measures transposed from production theory to measure the efficiency of a portfolio or fund by estimating the distance between an observed portfolio or fund and its reference portfolio on a nonparametric estimation of a multidimensional portfolio frontier along a given projection direction. In this framework, the nonparametric portfolio frontier constitutes the benchmark to measure the portfolio efficiency, and the projection direction is compatible with investors' preferences for return, risk and higher-order moments (see Ren et al. (2024b) for a survey). The seminal article of Sengupta (1989) is the first to introduce an efficiency measure into a diversified MV portfolio model.

This introduction of an efficiency measure into portfolio models allows gauging performance over multiple dimensions and opens up entirely new perspectives. On the one hand, Briec et al. (2004)

establish duality between a general efficiency measure (named shortage function) and MV utility functions. Thereafter, Briec et al. (2007) use an extended shortage function to improve efficiency in MVS space by looking for increases in mean return and positive skewness and contractions in risk. These authors also develop a duality result with a MVS utility function. Kerstens et al. (2011a) develop perfectly general geometric representations of these MVS portfolio frontiers. Finally, for the general class of mixed risk-aversion utility functions, Briec and Kerstens (2010) evaluate portfolio performance for the general moments case by improving odd moments and reducing even moments, and establish a duality result with general moment utility functions. This general framework allows, among others, to determine optimal mean-variance-skewness-kurtosis (MVSK) portfolios. On the other hand, Morey and Morey (1999) construct a multiple time horizon assessment within a standard MV framework whereby one computes efficiency measures to evaluate MV performance over three time horizons simultaneously. Briec and Kerstens (2009) generalize this contribution by using a shortage function. While both these articles assume time separability, some authors develop dynamic optimization approaches over time: an example is provided by Lin et al. (2017). Finally, it is possible to simultaneously optimize performance over several moments and several time horizons as developed in Kerstens et al. (2022) and in this contribution.

An alternative approach to accommodate comparisons of entire probability distributions is to use efficiency measures and nonparametric portfolio frontiers in applying various stochastic dominance criteria. Examples of developments in this direction include Basso and Funari (2001), Branda and Kopa (2014), Kuosmanen (2007), Lozano and Guttiérez (2008a), and Lozano and Guttiérez (2008b), among others.

This multi-moment approach has been empirically applied in finance by Adam and Branda (2020), Boudt et al. (2020a), Boudt et al. (2020b), Branda (2013), Branda (2015), Joro and Na (2006), Jurczenko et al. (2006), Jurczenko and Yanou (2010), Krüger (2021), Lamb and Tee (2012), Massol and Banal-Estañol (2014), among others. Multi-time horizon empirical applications are available in Ren et al. (2021), Lin and Li (2024), among others.

The above multi-moment diversified nonparametric portfolio models are computationally very demanding (e.g., Briec et al. (2007)). Therefore, a lot of authors have been resorting to undiversified nonparametric production models in order to approximate these multi-moment diversified nonparametric portfolio models. Examples include Basso and Funari (2003), Galagedera and Silvapulle (2002), Haslem and Scheraga (2003), Kerstens et al. (2022), Tsolas (2014), among others (see Sarkar (2022) for a recent survey). Ren et al. (2024a) are the only contribution known to us that employs undiversified nonparametric production models to assess the role of risk-loving rather than risk-aversion preferences in a buy-and-hold backtesting framework. Liu et al. (2015) develop an approximation result stating that a diversified nonparametric convex (e.g., MV) portfolio model can be approximated by a convex undiversified nonparametric production model. Xiao et al. (2022) extend this result by proving that a diversified nonparametric nonconvex (e.g., MVS) portfolio model can be approximated by a specific nonconvex undiversified nonparametric production model.

Various attempts have been made to make the estimation of efficiency measures using deterministic nonparametric portfolio frontiers more robust. One direction is to use stochastic frontier analysis, but the disadvantage is that this requires parametric specifications and assumptions on efficiency distributions. Examples are Annaert et al. (2003) and Lamb and Tee (2024), among others. Another approach is to employ robust versions of these deterministic nonparametric portfolio frontiers based on partial frontiers: see, e.g., Daraio and Simar (2006) and Matallín-Sáez et al. (2014). Another avenue is to use shrinkage estimators in conjunction with efficiency measures and nonparametric portfolio frontiers (see Lamb and Tee (2026)).

In this contribution, following the article of Kerstens et al. (2011b) using undiversified nonparametric production models, we focus on robustifying the underlying central moment (CM) estimates by using L-moments (LM) (see Hosking (1990)) and truncated L-moments (TLM) (see Elamir and Seheult (2003)). These robust moments are rather sparsely used in finance (e.g., Darolles et al. (2009) or Tolikas and Gettinby (2009)), especially in a backtesting context (see Qin (2012) for an exception). Furthermore, we propose a novel baseline buy-and-hold backtesting framework to assess our deterministic nonparametric production or rating models. Instead of summarising absolute performance metrics, this backtesting procedure evaluates the statistical significance (p-values) of out-of-sample performance of our rating models against a baseline of randomly selected mutual funds ensuring that the observed results are driven by the rating models' inherent capabilities rather than pure luck.

Our contribution has three main goals. First, we want to test in our buy-and-hold backtesting strategy which of the classical (CM) or robust (LM or TLM) moments yields the best performance. Second, we want clarification on the role of including higher-order moments (i.e., MVS and MVSK) compared to traditional MV models. Earlier buy-and-hold backtesting results in Kerstens et al. (2022) confirmed the positive role of higher-order moments, but there we used the shortage function itself as a performance measure, while here we stick to two traditional financial performance measures: total returns and maximum drawdown. Third, we aim to assess to which extent convex and nonconvex undiversified nonparametric rating models perform better in convex and nonconvex portfolio settings in line with the above theoretical approximation results. Earlier buy-and-hold backtesting results in Kerstens et al. (2022) confirmed that nonconvex models perform slightly better, but it remains an open question whether this is due to skill or due to pure luck.

Anticipating our main results, we can conclude the following. First, our backtesting results seem to indicate that using robust LM offers a better balance between total returns and maximum drawdown compared to classical CM and the TLM. Second, the backtesting reveals that extending the evaluation space to include higher-order moments yields consistent improvements in both total returns and maximum drawdown compared to the MV setting. Third, the convex models occasionally capture higher absolute returns at the cost of exposure to severe maximum drawdown, while the nonconvex models yield lower absolute returns but experience better maximum drawdown.

Our contribution is structured as follows. Section 2 presents the methodology in detail by first developing the diversified multi-moment and multi-time horizon portfolio models followed by their convex and nonconvex approximations using undiversified multi-moment and multi-time horizon production models. This is followed by a succinct introduction to the definition of LM and TLM in comparison with traditional CM. The next section develops the basic buy-and-hold backtesting framework. The new comparison against a wide series of baseline buy-and-hold backtesting strategies provides a statistical test to determine whether our deterministic production models perform well due to skill or due to luck. Section 4 describes the extensive sample of European mutual funds that is analysed in our backtesting framework. The next section 5 first defines the thirty convex and nonconvex undiversified multi-moment and multi-time horizon production models with CM, LM and TLM and with moment selection: MV, MVS and MVSK. Thereafter, we gauge the out-of-sample performance of the thirty buy-and-hold strategies using two traditional performance measures: (i) the realized total returns, and (ii) the maximum drawdown. Finally, we make a comparison with the baseline buy-and-hold backtesting strategies and we verify the statistical significance of our rating models by the p-values of our backtesting. A final Section 6 concludes.

2 Methodology

2.1 Diversified Frontier Rating Models in a Multi-Moment and Multi-Time Framework

To introduce some basic notation and definitions, consider the problem of selecting a portfolio (i.e., a fund of funds) from n mutual funds (MFs) at time t. Let $R_{1,t}, \ldots, R_{n,t}$ be the random returns of MFs $1, \ldots, n$. These MFs can be characterized by a set of moments of an arbitrary order. A portfolio of MFs $\omega_t = (\omega_{1,t}, \ldots, \omega_{n,t})$ is represented by a vector of proportions invested in each of the n MFs with $\sum_{j=1}^{n} \omega_{j,t} = 1$. If shorting is excluded, then these investment weights must satisfy non-negativity conditions ($\omega_{j,t} \geq 0$). Thus, the set of admissible portfolios at time t is written as follows:

$$\Im_t = \left\{ \omega_t \in \mathbb{R}^n : \sum_{j=1}^{n} \omega_{j,t} = 1, \omega_{j,t} \geq 0, j = 1, \ldots, n \right\}. \tag{1}$$

To simplify the notation and facilitate the analysis of higher-order moments, we adopt the following general formula. The return of a MF portfolio at the give time t is expressed as $R(\omega_t) = \sum_{j=1}^{n} \omega_{j,t} R_{j,t}$. Since the portfolio return $R(\omega_t)$ is a random variable depending on the returns of observed MFs, its distributional characteristics are captured by its moment sequence. Consistent with the utility expansion of Scott and Horvath (1980), we assume that rational investors exhibit a preference for odd moments (positive marginal utility) and an aversion to even moments (negative marginal utility). Accordingly, within the nonparametric efficiency framework, we classify the m even moments as input-like variables (to be minimized) and the s odd moments as output-like variables (to be maximized). Let $x[R(\omega_t)] \in \mathbb{R}^m$ denote the vector of input-like variables, such that $x[R(\omega_t)] = (x_1[R(\omega_t)], \ldots, x_m[R(\omega_t)])$, and let $y[R(\omega_t)] \in \mathbb{R}^s$ denote the vector of output-like variables, such that $y[R(\omega_t)] = (y_1[R(\omega_t)], \ldots, y_s[R(\omega_t)])$.

Then, we define the function $\Phi : (\omega_t \in \Im_t) \to \mathbb{R}^m \times \mathbb{R}^s$ as $\Phi(\omega_t) = (x[R(\omega_t)], y[R(\omega_t)])$ to represent the odd moments and even moments for a given portfolio of MFs. In the remainder, the multi-moment image of $\Im_t$ can be represented as $\Phi(\omega_t)$.

Based on the above notations and definitions, it is useful to define the multi-moment disposal representation at time t set through:

$$\begin{aligned}\mathcal{DR}^t = \big\{(x^t, y^t) \in \mathbb{R}^m \times \mathbb{R}^s \mid \forall i \in \{1, \ldots, m\} : x_i^t \geq x_i[R(\omega_t)],\\ \forall r \in \{1, \ldots, s\} : y_r^t \leq y_r[R(\omega_t)], \omega_t \in \Im_t\big\}.\end{aligned} \tag{2}$$

The boundary of the multi-moment disposal representation set at time t is called the diversified frontier in the single-time framework.

In practical investment scenarios, investors do not rely solely on the performance of financial assets within a single period. Instead, they incorporate performance across multiple distinct time horizons —such as returns over the past 1, 3, 5, or 10 years— into their asset selection decisions. For potential investors, financial performance data spanning these various time horizons offers substantially more information than single-period metrics. This rationale serves as one of the primary reasons why numerous stock and fund rating agencies currently provide multi-period performance results.

Following the fundamental idea of Briec and Kerstens (2009) to construct a multi-time model, we define a multi-time path $\mathscr{W} = (\omega^t)_{t=1}^T = \mathop{\times}\limits_{t=1}^{T} \omega^t$ in a multi-moment framework. Similarly, we define a time-dependent direction vector as $\mathscr{G} = (g^t)_{t=1}^T = \mathop{\times}\limits_{t=1}^{T} g^t$, where $g^t \in \mathbb{R}_-^m \times \mathbb{R}_+^s$. Additionally, we denote $\Delta = (\delta^1, \ldots, \delta^T) \in \mathbb{R}^T$ and employ the convention $\Delta\mathscr{G} = (\delta_1 g^1, \ldots, \delta_T g^T) \in (\mathbb{R}^m \times \mathbb{R}^s)^T \cong \mathbb{R}^{m\times T} \times \mathbb{R}^{s\times T}$. In a discrete time framework, for a give time horizon T the multi-moment space has the dimension $\mathbb{R}^{(m+s)\times T}$ and consists of all sequences of dated even-order and odd-order moments of the form:

$$\mathscr{X}(\mathscr{W}) = \mathop{\times}\limits_{t=1}^{T} (x[R(\omega_t)], y[R(\omega_t)]). \tag{3}$$

Then, we can define a multi-time and multi-moment representation for a given time horizon T as $\mathscr{DR} = \mathop{\times}\limits_{t=1}^{T} \mathcal{DR}^t$, which represents the Cartesian product of the disposal representation at each time t (see (2)). Similarly, we can define a multi-horizon or temporal, discounted shortage function in a multi-moment framework as follows.

Definition 2.1. For any multi-time and multi-moment path $\mathscr{W}$, the time discounted multi-time shortage function $\mathcal{S}^T$ in the direction of G is defined as:

$$\mathcal{S}^T(\mathscr{W}, \mathscr{G}) = \sup\Big\{\frac{1}{T}\sum_{t=1}^{T} \xi^{T-t}\delta_t \mid \mathscr{X}(\mathscr{W}) + \Delta\mathscr{G} \in \mathscr{DR}\Big\},$$

where ξ is the time discount factor.

For a given time horizon T, this definition involves the arithmetic mean of the simultaneous decreases in even-order moments and increases in odd-order moments along the path of direction $\mathscr{G}$, thereby projecting the observed multi-moment path $\mathscr{W}$ onto the boundary of the representation set $\mathscr{DR}$.

This shortage function for a MF o ($o \in 1, \ldots, n$) under evaluation can be computed by solving the following program:

$$
\begin{aligned}
\max \quad & \frac{1}{T}\sum_{t=1}^{T} \xi^{T-t}\delta_t \\
\text{s.t.} \quad & x_i\big(\sum_{j=1}^{n} \omega_{j,t} R_{j,t}\big) \leq x_{io}^t - \delta_t g_{io}^t, \quad i = 1, \ldots, m, \\
& y_r\big(\sum_{j=1}^{n} \omega_{j,t} R_{j,t}\big) \geq y_{ro}^t + \delta_t g_{ro}^t, \quad r = 1, \ldots, s, \\
& \sum_{j=1}^{n} w_{j,t} = 1, \quad \delta_t \geq 0, \quad t = 1, \ldots, T.
\end{aligned}
\tag{4}
$$

This multi-time and multi-moment diversified evaluation model provides a concrete evaluation of MF o. However, this diversified model requires nonlinear programming in most cases, its potential complexity and computational burden make this model suffer from the dilemma of being unsuitable for large-scale evaluations. This dilemma may be especially evident when higher-order moments are included in the practical use of this diversified evaluation method.

2.2 Nondiversified Frontier Rating Models in a Multi-Moment and Multi-Time Framework

In this subsection, we develop the multi-time and multi-moment MF performance measure based on the convex and nonconvex efficient variable returns to scale frontier methods instead of the diversified efficient frontier method in subsection 2.1. These nondiversified frontier-based efficiency measures are all simply computed by linear (or binary mixed integer) programming, which are much easier and time-saving when applied in the large-scale and multi-dimensional MF rating issue.

Nonparametric frontier rating models are rooted in the concept of the efficient frontier derived from production theory. Consider a sample of n MFs evaluated over a specific time horizon. At any specific time instance t, the j-th MF (where $j = 1, \ldots, n$) is defined by a vector of input-like variables x_{ij}^t ($i = 1, \ldots, m$) and output-like variables y_{rj}^t ($r = 1, \ldots, s$), aiming to minimize input-like variables while maximizing output-like variables.

We define a unified representation of convex and nonconvex production possibility sets (PPS), denoted as P_Λ^t. This PPS encompasses all feasible input-output combinations for the sample of n

MFs at time t:

$$P^t_\Lambda = \Big\{(x^t, y^t) \in \mathbb{R}^m \times \mathbb{R}^s \mid \forall i \in \{1,\dots,m\} : x^t_i \geq \sum_{j=1}^{n} \lambda_j x^t_{ij},$$

$$\forall r \in \{1,\dots,s\} : y^t_r \leq \sum_{j=1}^{n} \lambda_j y^t_{rj}, \lambda \in \Lambda\Big\}, \tag{5}$$

where:
$\Lambda \equiv \Lambda^C = \{\lambda \in \mathbb{R}^n \mid \sum_{j=1}^{n} \lambda_j = 1 \text{ and } \forall j \in \{1,\dots,n\} : \lambda_j \geq 0\}$ if convexity is assumed, and
$\Lambda \equiv \Lambda^{NC} = \{\lambda \in \mathbb{R}^n \mid \sum_{j=1}^{n} \lambda_j = 1 \text{ and } \forall j \in \{1,\dots,n\} : \lambda_j \in \{0,1\}\}$ if nonconvexity is assumed.

Following the logic of Kerstens et al. (2022), we assume that T represents the number of consecutive time periods in the evaluation horizon. On this basis, we define the multi-time path of input-like variables and output-like variables for the j-th MF ($j = 1,\dots,n$) as a sequence $Z_j = (x^t_j, y^t_j)^T_{t=1}$, where x^t_j and y^t_j represent the vectors of m inputs and s outputs at time t, respectively. The multi-time PPS is defined as the Cartesian product of the single-period PPSs:

$$\mathbf{P}^T_\Lambda = P^1_\Lambda \times \cdots \times P^T_\Lambda \subset (\mathbb{R}^m \times \mathbb{R}^s)^T \cong \mathbb{R}^{m\times T} \times \mathbb{R}^{s\times T}, \tag{6}$$

where P^t_Λ, ($t = 1,\dots,T$) is the PPS at time t mentioned previously in (5).

Additionally, we further define a time dependent direction vector as $G = (g^1,\dots,g^T) \in (\mathbb{R}^m_- \times \mathbb{R}^s_+)^T \cong \mathbb{R}^{m\times T}_- \times \mathbb{R}^{s\times T}_+$, representing a given multi-time direction path, where $g^t = (-g^t_x, g^t_y) \in \mathbb{R}^m_- \times \mathbb{R}^s_+$ denotes the direction vector at time $t \in \{1,\dots,T\}$. Then, we denote $\Theta = (\beta_1,\dots,\beta_T) \in \mathbb{R}^T$ and $\Theta \cdot G = (\beta_1 g^1,\dots,\beta_T g^T) \in (\mathbb{R}^m \times \mathbb{R}^s)^T \cong \mathbb{R}^{m\times T} \times \mathbb{R}^{s\times T}$.

Typically, the rational economic agents possess a time preference favouring immediate yields over future proceeds. In the context of retrospective assessment, this preference implies that the results of recent economic behaviors carry more weight than those of earlier periods. In light of the time preference of an investor in a portfolio context, the multi-period evaluation method proposed herein posits that the time dimension is non-neutral. Specifically, during retrospective benchmarking, efficiency gains achieved in the distant past are assigned lower weights than those realized in the nearby past. Accordingly, we introduce a time discounting factor denoted ξ ($0 < \xi < 1$) to weigh the efficiency measures over the time horizon. The time-discounted multi-time shortage function is formally defined as:

Definition 2.2. For any observation $Z \in (\mathbb{R}^m \times \mathbb{R}^s)^T \cong \mathbb{R}^{m\times T} \times \mathbb{R}^{s\times T}$, the time discounted multi-time shortage function S^T_Λ in the direction of G is defined as:

$$S^T_\Lambda(Z; G) = \sup\Big\{\frac{1}{T}\sum_{t=1}^{T} \xi^{T-t}\beta_t \mid Z + \Theta \cdot G \in \mathbf{P}^T_\Lambda\Big\}.$$

In Definition 2.2, the input-like variables to be reduced and the output-like variables to be expanded can, e.g., be a vector of even and odd moment characteristics of the MF return. This multi-time shortage function allows for simultaneously seeking increases in output-like variables and reductions in input-like variables over the whole time horizon.

To compute the multi-time shortage function, one needs to solve the following mathematical program for MF unit $o \in 1, \ldots, n$ in need of assessment:

$$
\begin{aligned}
\max \quad & \frac{1}{T}\sum_{t=1}^{T} \xi^{T-t}\beta_t \\
\text{s.t.} \quad & \sum_{j=1}^{n} \lambda_j^t x_{ij}^t \leq x_{io}^t - \beta_t g_{io}^t, \quad i = 1, \ldots, m, \\
& \sum_{j=1}^{n} \lambda_j^t y_{rj}^t \geq y_{ro}^t + \beta_t g_{ro}^t, \quad r = 1, \ldots, s, \\
& \sum_{j=1}^{n} \lambda_j^t = 1, \quad \beta_t \geq 0, \quad t = 1, \ldots, T, \\
& \forall j = 1, \ldots, n: \begin{cases} \lambda_j^t \geq 0, \quad t = 1, \ldots, T, & \text{under convexity,} \\ \lambda_j^t \in \{0, 1\}, \quad t = 1, \ldots, T, & \text{under nonconvexity.} \end{cases}
\end{aligned}
\tag{7}
$$

It is apparent that model (7) exhibits a block diagonal structure, formulated as a linear programming (LP) problem under the convexity assumption or a binary mixed-integer linear programming (BMILP) problem under the nonconvexity assumption. This means that the MF assessment in each time period is independent of the other time periods and that one can solve each time period separately with the objective function replaced by $\max \beta_t$ and calculate $\frac{1}{T}\sum_{t=1}^{T} \xi^{T-t}\beta_t$ afterwards from those T sub-problem results.

If the multi-time improvement direction path for the MF o under evaluation is defined as $G = (|x_o^t|, |y_o^t|)_{t=1}^T$, where $x_o^t = (x_{1o}^t, \ldots, x_{mo}^t)$ and $y_o^t = (y_{1o}^t, \ldots, y_{so}^t)$, then the efficiency score derived from the multi-time shortage function represents the time-discounted weighted average of the maximum proportional simultaneous improvement of initial inputs and outputs across all time periods. The optimal value for models (7) greater than zero implies that the input-like variables and output-like variables of the evaluated MF o can be improved in at least one or more time periods, thereby indicating that this MF is inefficient within the multi-time evaluation framework. Conversely, an optimal value of zero signifies that the MF's input-like variables and output-like variables are efficient in each respective time period. Evidently, the criterion for defining efficient units under the multi-time framework is significantly more stringent than the identification of efficient ones within a single-period framework.

2.3 L-Moments and Truncated L-Moments

The conventional central moments (CM) have faced several criticism in the financial literature due to their extreme sensitivity to outliers and poor sampling properties, particularly when MF or asset return distributions deviate from normality. Fabozzi et al. (2007) indicate that the practical unreliability of quantitative portfolio models commonly stems from these statistical instabilities. Recent research in operations research and financial engineering has shifted a lot of attention towards the robust portfolio management, aiming to explicitly incorporate statistical precision into the portfolio allocation and evaluation process.

Therefore, we opt to duplicate the analysis of CM with the more robust univariate L-moments, denoted LM, (see Hosking (1990)) and truncated L-moments, denoted TLM, (see Elamir and Seheult (2003)) in the proposed multi-time and multi-moment rating methods. Given the need to handle heavy-tailed distributions, these LM are defined as strategically selected linear functions of the expected values of ordinal statistics. Consequently, LM provide estimators that are not only more robust to extreme observations, but are also capable of characterizing a wider range of heavy-tailed distributions. Indeed, LM exist for any distribution with a finite mean, a property that has facilitated their adoption in fields such as meteorology and hydrology, and, more recently, in empirical finance (e.g., Tolikas and Gettinby (2009), Kerstens et al. (2011b)).

While LM provide a robust alternative to CM, their existence is not guaranteed for distributions characterized by extremely heavy tails and undefined means, such as the Cauchy distribution. An alternative robust version of LM is to censor or truncate a predetermined percentage of the extreme values from the sample before estimating the mean and the standard deviation from the untruncated sample values. In these TLM estimators, the expectations of order statistics from a conceptual sample—used in the standard definition of population LM — are replaced by the expectations of order statistics from a larger conceptual sample. The increase in this conceptual sample size is exactly equal to the total magnitude of truncation.

2.3.1 L-Moments: Linear Combinations of Order Statistics

LM are expectations of certain linear combinations of order statistics. Unlike conventional moments, LM are less sensitive to the effects of outliers and provide more reliable inferences from samples following non-normal distributions. Following the seminal work of Sillito (1969) and Hosking (1990), we define population LM λ_r as robust alternatives to the CM μ_r. Let X be a random variable representing MF returns with cumulative distribution function $F(\cdot)$, and let $X_{1:n} \leq X_{2:N} \leq \cdots \leq X_{N:N}$ be the order statistics of a random sample of size N from $F(\cdot)$. In general the LM λ_r of order r can be defined as:

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E[X_{r-k:r}] \tag{8}$$

where $\binom{r-1}{k}$ is the binomial coefficient weighting each $E[X_{r-k:r}]$. Being linear functions (not powers) of the data, LM are more robust than CM to outliers, particularly in small samples, and enable more secure inference about the underlying distribution of X. Furthermore, if the random variable X has a finite expectation, then all LM exist. A key consequence is that LM of any order can be computed even when CM of the corresponding order do not exist.

In particular, we estimate the first four LM to capture the essential characteristics of the MF return distribution and these are respectively expressed as follows:

$$\lambda_1 = E[X_{1:1}] \tag{9}$$

$$\lambda_2 = \frac{1}{2}E[X_{2:2} - X_{1:2}] \tag{10}$$

$$\lambda_3 = \frac{1}{3}E[(X_{3:3} - X_{2:3}) - (X_{2:3} - X_{1:3})] \tag{11}$$

$$\lambda_4 = \frac{1}{4}E[(X_{4:4} - X_{1:4}) - 3(X_{3:4} - X_{2:4})] \tag{12}$$

The first four LM are estimated to capture the essential characteristics of the MF return distribution. As we can see that the first LM $\lambda_1 = E[X_{1:1}]$ is the same as the population mean $\mu_1 = E(X)$, representing the expected return of the MF. The second LM λ_2, also known as L-scale, measures the dispersion of the MF's returns. Intuitively, λ_2 can be explained as follows: consider an investor who invested in MF X in two periods; the expected difference between the higher return ($X_{2:2}$) and the lower return ($X_{1:2}$) of the two realized returns characterizes the dispersion of X's returns. If the returns are clustered close to each other, then this expected difference is small (and it is zero if and only if X yields one return for certain). If the returns are scattered far apart, then this expected difference is large.

The third LM λ_3, L-skewness, measures the skewness of the returns of MF. This can be explained by considering an investor who invested in MF X in three periods; the expected difference between the mean and the median return of the three realized returns characterizes the skewness of the return distribution. For a symmetric distribution, this expected difference is zero, whereas for a positively (negatively) skewed distribution, this expected difference is positive (negative). The fourth LM λ_4, L-kurtosis, provides a measure of kurtosis that allows for measuring the peakedness and tail heaviness of the MF's return series. λ_4 characterizes the expected difference between the extreme returns ($X_{4:4}$ and $X_{1:4}$) relative to the spread of the central returns ($X_{3:4}$ and $X_{2:4}$). A high λ_4 indicates heavier tails and higher exposure to tail risk.

In the context of MF performance evaluation, the odd-order LM (λ_1 and λ_3) are treated as output-like variables to be maximized, as they represent higher average returns and favourable positive skewness. Conversely, the even-order moments (λ_2 and λ_4) are treated as input-like variables to be minimized, as they represent volatility and tail risk—undesirable attributes that signify the cost of achieving a given level of return of the MF under evaluation.

2.3.2 Truncated L-Moments and Robust Variable Construction

The conceptual foundation of a TLM $\lambda_r^{(t_1,t_2)}$ lies in its modification of the sample size and the selection of order statistics. Specifically, for an r-th order moment, we consider a conceptual sample of size $m = r+t_1+t_2$. By truncating the t_1 smallest and t_2 largest observations, the TLM is defined as the expected value of a linear combination of the remaining r intermediate order statistics. This truncation process effectively assigns zero weight to the extreme tails of the distribution, ensuring that the resulting measures are governed by the more stable, central portion of the return series.

Mathematically, the TLM $\lambda_r^{(t_1,t_2)}$ of order r is expressed as:

$$\lambda_r^{(t_1,t_2)} = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E[X_{r-k+t_1:r+t_1+t_2}]. \tag{13}$$

It is clear that TLM reduce to LM when $t_1 = t_2 = 0$. While there will be applications when unequal amounts of truncation ($t_1 \neq t_2$) may be appropriate, we focus here on the symmetric case $t_1 = t_2 = t$, and rewrite (13) as:

$$\lambda_r^{(t)} = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E[X_{r-k+t:r+2t}]. \tag{14}$$

Since we adopt symmetric truncation ($t_1 = t_2 = 1$) to obtain robust estimators, the conceptual sample size increases by 2 for each moment (i.e., size becomes $r + 2$). The first four TLM are then defined as follows:

$$\lambda_1^{(1)} = E[X_{2:3}] \tag{15}$$

$$\lambda_2^{(1)} = \frac{1}{2}E[X_{3:4} - X_{2:4}] \tag{16}$$

$$\lambda_3^{(1)} = \frac{1}{3}E[X_{4:5} - 2X_{3:5} + X_{2:5}] \tag{17}$$

$$\lambda_4^{(1)} = \frac{1}{4}E[X_{5:6} - 3X_{4:6} + 3X_{3:6} - X_{2:6}] \tag{18}$$

Note that $\lambda_1^{(1)}$, $\lambda_2^{(1)}$, $\lambda_3^{(1)}$, $\lambda_4^{(1)}$ are population measures of location, scale, skewness and kurtosis analogous to the corresponding standard LM, but with enhanced robustness.

3 Backtesting Framework

3.1 Backtesting Strategy: Buy-and-hold

The backtesting strategy is a buy-and-hold strategy where the timing effect is mitigated by simulating the buy-and-hold strategy at different time periods. We have $T = 180$ months of daily data. We use a selection of historical MF return data to compute statistical moments of the return distribution and determine the portfolio weights. This selection of historical MF return data is made with a sliding window of width $W \in \{12, 36, 60\}$ (in months). Let $W_{max} = \max\{12, 36, 60\} = 60$ be the maximum window width in months. We use the first W_{max} months as training data and start the buy-and-hold strategy for month $t = W_{max} + 1$ to determine the portfolio weights. We compute the statistical moments over the time period $[t - W, t]$ for time window W and do an efficiency analysis where the uneven moments are used as outputs and the even moments are used as inputs. Let $\beta_{k,t}^W$ be the inefficiency score for MF k at time t for window width W. We repeat the previous step for different window widths W and record the corresponding efficiency scores $\{\beta_{k,t}^W\}_{k=1}^K$ $\forall W \in \{12, 36, 60\}$. Next, we aggregate these inefficiency scores for different W using the time-discounted formula (with $\xi = 0.95$) $\bar{\beta}_{k,t} = \frac{1}{3}(0.95\beta_{k,t}^{12} + 0.95^3\beta_{k,t}^{36} + 0.95^5\beta_{k,t}^{60})$ for all MFs $k = 1, \ldots, K$. In this way model (7) is implemented by exploiting the time separability property. Finally, the uniform portfolio weights at time t are assigned to the J most efficient MFs as determined by $\bar{\beta}_{k,t}$. This portfolio is held until the last period of data. In order to mitigate a possible timing effect of when the portfolio is constructed, we repeat this analysis for a portfolio constructed at time $t = W_{max} + 2$ using historical data shifted by one month in our time window. In total we repeat the above $T - 2 \cdot W_{max} = 60$.

The entire strategy is summarized below in pseudocode:

---

for $t = W_{max} + 1, W_{max} + 2, \ldots, T - W_{max} - 1$ **do**
    for $W \in \{12, 36, 60\}$ **do**
        Calculate statistical moments (CM, LM or TLM) over the time window $[t - W, t]$.
        Perform $S_{NC}^T(Z;G)$ or $S_C^T(Z;G)$ efficiency analysis using uneven moments as outputs and even moments as inputs.
        Let $\{\beta_{k,t}^W\}_{k=1}^K$ be the inefficiency scores.
    end for
    Calculate weighted mean of the inefficiency scores $\bar{\beta}_{k,t} = \frac{1}{3}(0.95\beta_{k,t}^{12} + 0.95^3\beta_{k,t}^{36} + 0.95^5\beta_{k,t}^{60})$ for all MFs $k = 1, \ldots, K$.
    Rank $\{\bar{\beta}_{k,t}\}_{k=1}^K$ and assign uniform portfolio weights to top J most efficient assets.
    Hold the constructed portfolio until the last period of data.
end for

We repeat the above for different values of $J \in \{10, 20, 30\}$.

3.2 Baseline: A New Proposal

The performance of the constructed portfolios are compared to a set of baseline portfolios. This determines whether the performance of the constructed portfolios is simply attributable to luck or not. One of the simplest baseline portfolios is a deterministic buy-and-hold strategy of 1 MF in our dataset simulated at different time periods. This buy-and-hold strategy is simulated separately for each MF in the dataset. This yields 713 baseline portfolios for each rebalancing period $t = W_{max} + 1, W_{max} + 2, \ldots, T - W_{max} - 1$. Obviously, alternative baseline portfolios could also be adopted: e.g., randomly selecting multiple funds to construct equally or randomly weighted portfolios. To be in line with our buy-and-hold backtesting strategy, we adopt the current single fund approach as our baseline that follows the same philosophy.

We can then evaluate how likely the performance of a constructed portfolio is compared to all baseline portfolios. Let a_t be a performance metric (e.g., total return) of a constructed portfolio at rebalancing period t and $X_t \in \{\bar{a}_{k,t}\}_{k=1}^K$ a random variable representing the same performance metric for a baseline portfolio. The probability that a constructed portfolio is outperformed by the baseline portfolios on the selected performance metric is then:

$$\mathbf{P}[X_t > a_t] = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{>a_t}(\bar{a}_{k,t})$$

where $\mathbf{1}_{>a_t}(x)$ is the indicator function that equals 1 when $x > a_t$ and 0 otherwise. A low probability (p-value) indicates that the outperformance of the constructed portfolio compared to the baseline is unlikely due to simple luck. These p-values are calculated for the following selection of performance metrics: total returns, and maximum drawdown. The definitions of the buy-and-hold backtesting strategy and the baseline strategy are available at a GitHub link: `https://github.com/kepiej/fundrating`.

4 Sample Description

The data used in this study are obtained from the Thomson Reuters Mutual Funds Holdings database and the Refinitiv Lipper Funds Lipper database, which provide comprehensive information on MF's various risk and performance measures, as well as a wide range of fund-specific characteristics. We applied several filters to an initial sample of MFs distributed in Europe.

First, we retain only MFs reporting Net Asset Values (NAV) on a daily basis, with pricing conducted on trading days from Monday to Friday, as differences in valuation frequency can affect the reliability of risk and return estimates used in portfolio construction and analysis. When NAVs are not reported daily, then return series exhibit irregular observation intervals and missing values, which require interpolation or other forms of reconstruction. Such procedures may introduce measurement error and generate artificial smoothing of returns that will appear mechanically less volatile. This issue is often amplified when the MFs' holdings are illiquid or valued using model-based approaches rather than observable market quotes. By focusing on daily-valued MFs, we reduce the risk of biased estimates of volatility, correlations, and performance measures.

Also, heterogeneous pricing frequencies may induce stale pricing effects. When valuation is not performed daily, then NAV updates may reflect delayed adjustments to the MFs' portfolio value, leading to spurious autocorrelation and distorted cross-sectional correlations across MFs. Imposing a daily pricing requirement ensures temporal alignment of return observations and reduces biases stemming from asynchronous price updates.

Thus, by restricting the sample to daily NAV MFs addresses these issues by ensuring a uniform high-frequency return structure and full alignment of observations across MFs.

Second, we restrict the sample to equity MFs only in order to ensure comparability across MFs and to maintain a homogeneous investment universe. We exclude MFs primarily investing in other asset classes such as bonds, money market instruments, real estate, or multi-asset portfolios, as these MFs exhibit fundamentally different return-generating processes, risk characteristics, and pricing dynamics. Equity MFs are typically invested in highly liquid, exchange-traded securities and evaluated against equity market indices, whereas other MF categories may involve different valuation conventions and less synchronous pricing. Overall, this restriction ensures a coherent and comparable sample.

Third, we restrict the sample to MFs domiciled in on-shore European countries[1] in order to

[1] Austria, Belgium, Bulgaria, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary,

ensure homogeneous regulatory frameworks, disclosure requirements, and tax regimes, all of which may affect MF behavior, reporting practices, and the comparability of performance measures.

Fourth, we exclude MFs explicitly targeting institutional investors in order to improve sample homogeneity in terms of investor clientele and investment constraints. In line with this objective, we retain only MFs with a minimum investment threshold below the equivalent of EUR 100,000, thereby ensuring relative homogeneity in fee structures and investor accessibility. Institutional share classes and institution-only MFs often differ systematically from retail-oriented MFs in terms of fee structures, minimum investment requirements, liquidity conditions, and reporting practices. These differences can affect observed performance, risk-taking behavior, and return dynamics, thereby limiting comparability across MFs within the sample.

The application of these filters results in a final sample of 713 MFs: comprising both capitalization and distribution MFs. These MFs differ in their income distribution policies, ranging from full reinvestment of income to partial or full distribution at varying frequencies. To ensure consistency and comparability of performance measures across MFs, we collect both income distributions and daily NAV, with all series denominated in euros, in order to accurately compute total returns accruing to a representative euro-based investor over time. The analysis covers a 15-year period from 13 May 2010 to 12 May 2025.

5 Empirical Backtesting Results

5.1 Nondiversified Multi-Moment and Multi-Time Rating Models: Classification and Selection

The 30 buy-and-hold backtesting strategies differ primarily in their MF selection processes, since the best-performing MFs are identified using various nonparametric nondiversified rating models within a multi-time framework. We opt for nondiversified rather than diversified rating models within a multi-time framework because the latter are computationally prohibitive and the former lend themselves naturally to a buy-and-hold backtesting strategy.

These 30 scenarios are summarized in Table 1. Each scenario adopts either a multi-time convex (denoted 'c') or a multi-time nonconvex (denoted 'nc') nonparametric frontier specification. To systematically investigate the impact of statistical robust moments on the MF performance evaluation, these frontier specifications focus on three distinct statistical frameworks: traditional central moments (CM), L-moments (LM), and truncated L-moments (TLM). Furthermore, to thoroughly account for varying degrees of heavy tails and extreme outliers in the return distributions, the TLM are differentiated into three symmetric truncation levels: 1-99, 5-95, and 10-90. Each of these statistical configurations is applied within three multidimensional spaces depending on the moment order (r): the first two moments (MV, $r \in \{1, 2\}$) to capture the basic mean-variance trade-off, the first three moments (MVS, $r \in \{1, 2, 3\}$) to also account for skewness, and the first four moments

Italy, Norway, Portugal, Slovenia, Spain, Sweden, Switzerland, the Netherlands, and the United Kingdom.

(MVSK, $r \in \{1, 2, 3, 4\}$) to further incorporate kurtosis. For instance, the abbreviation LM-MVSKc designates the convex frontier model with the first four LM selected.

Table 1: Multi-time rating models in different frameworks

Convexity assumption	Statistical framework	Moment order (r)	Abbreviation
Model (1) convex	CM	$r \in \{1, 2\}$ within MV framework	CM-MVc
		$r \in \{1, 2, 3\}$ within MVS framework	CM-MVSc
		$r \in \{1, 2, 3, 4\}$ within MVSK framework	CM-MVSKc
	LM	$r \in \{1, 2\}$ within MV framework	LM-MVc
		$r \in \{1, 2, 3\}$ within MVS framework	LM-MVSc
		$r \in \{1, 2, 3, 4\}$ within MVSK framework	LM-MVSKc
	TLM	$r \in \{1, 2\}$ within MV framework	TLM-MVc
		$r \in \{1, 2, 3\}$ within MVS framework	TLM-MVSc
		$r \in \{1, 2, 3, 4\}$ within MVSK framework	TLM-MVSKc
Model (1) nonconvex	CM	$r \in \{1, 2\}$ within MV framework	CM-MVnc
		$r \in \{1, 2, 3\}$ within MVS framework	CM-MVSnc
		$r \in \{1, 2, 3, 4\}$ within MVSK framework	CM-MVSKnc
	LM	$r \in \{1, 2\}$ within MV framework	LM-MVnc
		$r \in \{1, 2, 3\}$ within MVS framework	LM-MVSnc
		$r \in \{1, 2, 3, 4\}$ within MVSK framework	LM-MVSKnc
	TLM	$r \in \{1, 2\}$ within MV framework	TLM-MVnc
		$r \in \{1, 2, 3\}$ within MVS framework	TLM-MVSnc
		$r \in \{1, 2, 3, 4\}$ within MVSK framework	TLM-MVSKnc

Based on the 30 multi-time rating models presented in Table 1, the buy-and-hold backtesting exercises are executed to empirically evaluate their out-of-sample performance in practical MF selection. As introduced in Section 3, we implement a buy-and-hold strategy where one selects the 10, 20, or 30 best performing MFs according to the rankings derived from these rating models and hold these selections to the end of sample period. Our backtesting analysis is iteratively performed using a sliding time window. In total, this simulation is repeated 60 times across the sample period. Consequently, it serves as a objective mechanism to benchmark the out-of-sample performance of various nonparametric rating models – ranging from traditional CM assessments to robust LM and TLM frameworks– in a realistic MF selection context.

5.2 Traditional Backtesting Performance Indicators

In our empirical backtesting exercises, the out-of-sample performance of the 30 buy-and-hold strategies is gauged using two primary indicators: (i) the realized total returns, and (ii) the maximum drawdown. The former quantifies the absolute profitability realized by each strategy, while the latter serves as a critical measure of its downside risk exposure. As stated previously, the buy-and-hold backtesting is executed iteratively over 60 sliding time windows. Therefore, each strategy generates a sequence of 60 out-of-sample measurements for both total returns and maximum drawdown. To comprehensively summarize the results across these 60 backtesting iterations, Table 2 reports the average values of the total returns and the maximum drawdown for each buy-and-hold strategy across 60 time windows of backtesting.

Table 2: Performance results for buy-and-hold strategies: Average values of the total returns and maximum drawdown across 60 backtesting windows

Methods	Total returns			Maximum drawdown		
	MF(10)	MF(20)	MF(30)	MF(10)	MF(20)	MF(30)
CM-MVc	0.5029	0.5586	0.5859	-0.2758	-0.2840	-0.2889
CM-MVnc	0.4103	0.5236	0.5531	-0.2625	-0.2751	-0.2835
LM-MVc	0.5187	0.5833	0.5988	-0.2864	-0.2880	-0.2905
LM-MVnc	0.4396	0.4898	0.5143	-0.2716	-0.2776	-0.2845
TLM-MVc (1-99)	0.4637	0.4805	0.4791	-0.3329	-0.3238	-0.3207
TLM-MVnc (1-99)	0.3190	0.3092	0.3160	-0.2733	-0.2810	-0.2870
TLM-MVc (5-95)	0.5106	0.5069	0.4930	-0.3358	-0.3252	-0.3251
TLM-MVnc (5-95)	0.2656	0.2648	0.3030	-0.2668	-0.2800	-0.2839
TLM-MVc (10-90)	0.5402	0.5300	0.5295	-0.3351	-0.3270	-0.3262
TLM-MVnc (10-90)	0.2507	0.2586	0.3162	-0.2703	-0.2766	-0.2807
CM-MVSc	0.5380	0.5561	0.5847	-0.2608	-0.2718	-0.2796
CM-MVSnc	0.5072	0.5200	0.5291	-0.2375	-0.2626	-0.2730
LM-MVSc	0.5479	0.5823	0.6037	-0.2855	-0.2886	-0.2904
LM-MVSnc	0.5027	0.5234	0.5651	-0.2661	-0.2810	-0.2861
TLM-MVSc (1-99)	0.4500	0.4624	0.4890	-0.3161	-0.3092	-0.3100
TLM-MVSnc (1-99)	0.4124	0.4188	0.4398	-0.2679	-0.2856	-0.2896
TLM-MVSc (5-95)	0.5743	0.5410	0.5291	-0.3221	-0.3221	-0.3192
TLM-MVSnc (5-95)	0.4582	0.3988	0.3903	-0.2738	-0.2784	-0.2868
TLM-MVSc (10-90)	0.6153	0.5789	0.5546	-0.3211	-0.3243	-0.3228
TLM-MVSnc (10-90)	0.4284	0.4564	0.4147	-0.2881	-0.2865	-0.2890
CM-MVSKc	0.5468	0.5666	0.5874	-0.2513	-0.2723	-0.2793
CM-MVSKnc	0.4631	0.4945	0.4901	-0.2350	-0.2610	-0.2715
LM-MVSKc	0.5658	0.5860	0.5924	-0.2639	-0.2803	-0.2858
LM-MVSKnc	0.4562	0.4889	0.5078	-0.2427	-0.2660	-0.2774
TLM-MVSKc (1-99)	0.5046	0.5043	0.5208	-0.3111	-0.3078	-0.3063
TLM-MVSKnc (1-99)	0.4823	0.4461	0.4430	-0.2703	-0.2748	-0.2803
TLM-MVSKc (5-95)	0.5444	0.5358	0.5210	-0.3206	-0.3165	-0.3147
TLM-MVSKnc (5-95)	0.4363	0.4187	0.3922	-0.2794	-0.2810	-0.2849
TLM-MVSKc (10-90)	0.5721	0.5366	0.5047	-0.3124	-0.3134	-0.3151
TLM-MVSKnc (10-90)	0.3984	0.4753	0.4623	-0.2958	-0.2986	-0.2974

The first key observation is that the robust LM provide a more balanced enhancement in profitability without severe risk degradation compared to both CM and TLM. By utilizing linear combinations of order statistics rather than the higher-power deviations from the mean used in CM, the robust LM is able to effectively smooth out noise within the sample. As observed in columns 2–4 of Table 2, the frontier-based rating models with LM consistently generate higher average total returns than their CM counterparts across various frameworks in most cases (mainly c). For instance, the buy-and-hold strategy deviated from LM-MVSKc rating model achieves an average return of 0.5658, outperforming the 0.5468 of the strategy based on CM-MVSKc when buying the 10 best MFs. Importantly, because the robust LM framework avoids mechanically truncating extreme tails, it preserves critical sensitivity to downside vulnerabilities during MF efficiency evaluations. On occasion, TLM performs better than CM and LM, but such results are rare. In terms of maximum drawdown, CM performs better than LM in that it has smaller absolute values on average. However, by effectively penalizing MFs with fragile downside profiles, the LM-based strategies exhibit smaller out-of-sample maximum drawdown (in absolute values) compared to the TLM-based strategies in most scenarios, as shown in columns 5–7 of Table 2.

Second, the TLM involve a relatively pronounced trade-off between profit gain and tail-risk blindness in MF performance appraisal. Comparing the average total returns of the TLM-based strategies with those of the LM-based and CM-based strategies in Table 2, it is evident that while moderate truncation can yield higher average total returns in a few specific scenarios, it generally fails to maintain superior profitability across the majority of cases. For instance, although the truncated TLM-MVSc (10-90) model achieved an average total return of 0.6153 under the MF(10) scenario — exceeding the 0.5479 of its LM-MVSc counterpart — this advantage generally diminishes as the selected MF size increases. In the MF(30) scenario, its average total return yields 0.5546, lagging behind the 0.6037 of the LM-based model. However, in terms of the downside risk, the out-of-sample average maximum drawdown of TLM-based strategies are generally larger than those of the CM-based and LM-based strategies. For example, the average maximum drawdown of the TLM-MVc (1-99) strategy deepens to -0.3329, whereas the corresponding CM-MVc strategy is contained at -0.2758. This to some extent indicates that systematically eliminating downside extreme tail observations (e.g., historical severe losses) during the in-sample efficiency assessment may inadvertently omit some information about the structural downside risk of the MFs. Particularly under the convex evaluation frameworks where models actively seek hypothetical high-yield combinations, this tail-risk blindness ultimately leads to a higher risk exposure for the out-of-sample MF portfolios.

Third, incorporating higher-order moments (such as skewness and kurtosis) yields a mild improvement in both the out-of-sample average total returns and maximum drawdown. In terms of the total returns, the advantage of MVSK frameworks over traditional MV models is relatively pronounced when selecting a smaller number of MFs. For instance, the CM-MVSKc strategy yields an average total return of 0.5468 in the MF(10) scenario, outperforming the 0.5029 of CM-MVc. As the number of selected MFs increases, this advantage presented by backtesting strategies with higher-order moments tends to diminish on average. Similar results are apparent within the LM and TLM frameworks. In terms of the maximum drawdown, the buy-and-hold strategies incorporating higher-order moments perform slightly better than the basic MV framework across various selected MF sizes, especially for the CM and a bit less uniform for LM and TLM. By incorporating skewness and kurtosis into the in-sample MF efficiency assessment, these nonparametric evaluation models can potentially screen out MFs harboring skewed and heavy-tailed risks, resulting in general (not uniformly) in a reduction in the out-of-sample maximum drawdown during the backtesting exercises.

Fourth, comparing the c and nc frontier-based models reveals a clear trade-off: the nc models generally yield lower average total returns, but these provide consistently superior downside protection. Under the c assumption, a MF is evaluated against an idealized linear combination of best-performing peers. This stringent criterion forces the c models to sometimes select extreme MF (e.g., high-yield and high-risk MFs) with extreme upward momentum, thereby capturing higher out-of-sample returns. Conversely, the nc models form a step-wise frontier without virtual combinations, which selects less extreme MF and focuses on a broader, more conservative set of locally dominant MF. Furthermore, the gap between c and nc models is particularly pronounced under the TLM framework for both total returns and maximum drawdown.

While Table 2 provides a comprehensive overview of the average performance indicators for the 30 strategies, it is equally crucial to examine the cross-sectional distribution and out-of-sample stability of the selected MFs. To this end, Figures 1 and 2 present the boxplots for the out-of-sample total returns and maximum drawdown per strategy, respectively. In these figures, the vertical layers correspond to the different selected MF sizes (i.e., the best 10, 20, and 30 MFs selected), while the structural annotations (e.g., MV, MVS, MVSK) denote the specific higher-order moment frameworks. Additionally, the triangles within the boxes represent the mean values. For the total returns shown in Figure 1, the boxes positioned higher along the vertical axis are preferable, indicating superior out-of-sample return for the corresponding strategies. In addition, since maximum drawdown are recorded as negative values in Figure 2, the boxes positioned closer to the horizontal zero-axis are better, signifying smaller absolute maximum drawdown.

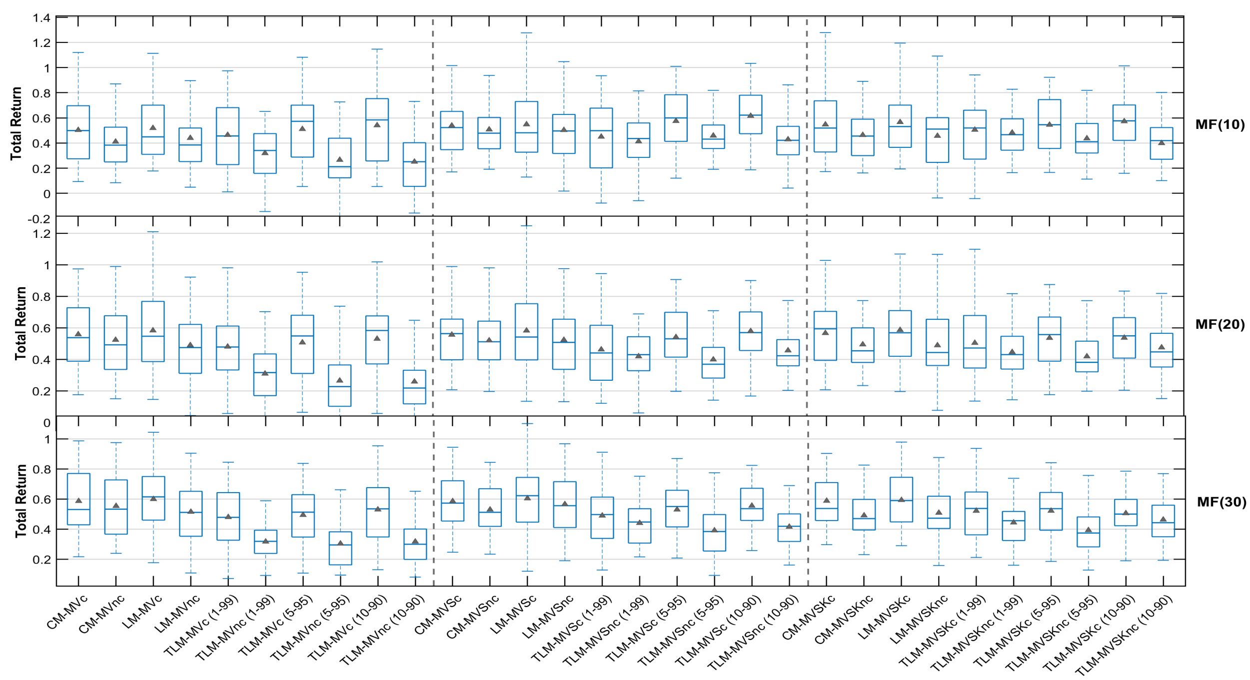

Figure 1: Distributions of total returns for the 30 buy-and-hold strategies across 60 backtesting windows

The visual patterns in Figures 1 and 2 are consistent with the aforementioned analysis from

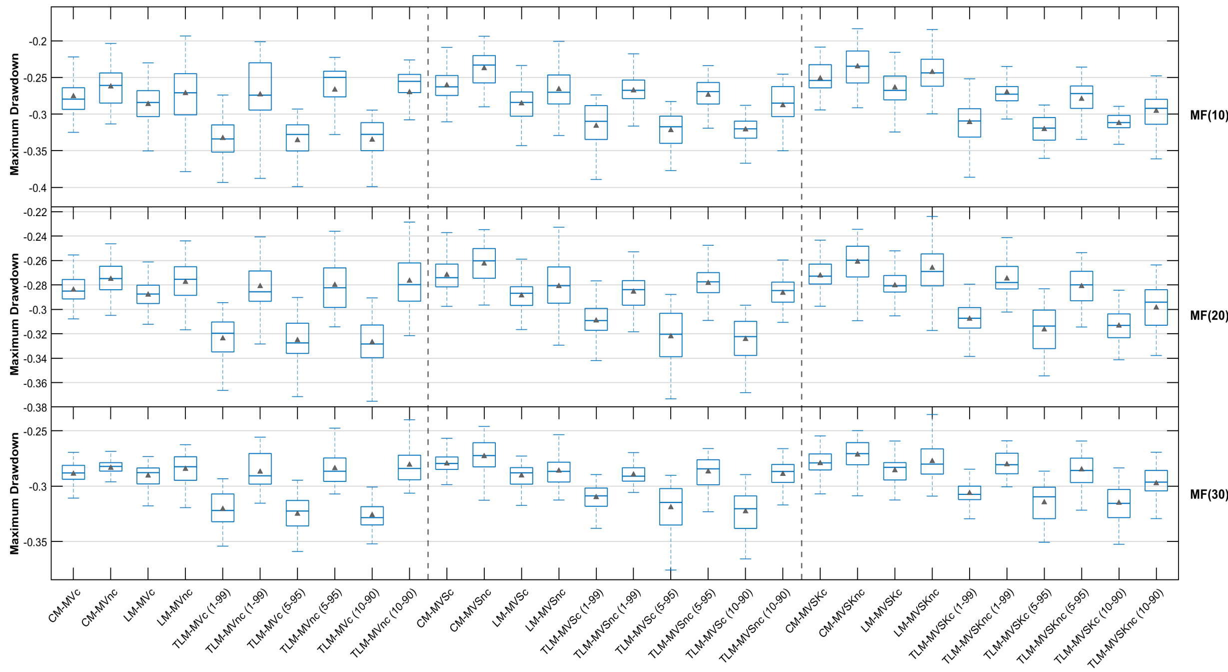

Figure 2: Distributions of maximum drawdown for the 30 buy-and-hold strategies across 60 backtesting windows

Table 2. Beyond validating the average values of the two indicators, the boxplots provide two additional observations regarding the distributions of both indicators. First, as the selected MF size expands from MF(10) to MF(30) in Figure 1, the interquartile ranges noticeably compress across almost all models. This visual convergence effectively illustrates the diversification effect, demonstrating how holding more MFs reduces the cross-sectional variance of out-of-sample returns. Second, regarding structural tail risks, Figure 2 reveals that the frontier-based strategies under the c assumption frequently generate long tailed lower whiskers and severe negative outliers. By contrast, the strategies under the nc assumption are closer to the horizontal zero-axis.

5.3 Comparison with the baseline: p-values of backtesting

In this section, we evaluate the statistical significance of the 30 buy-and-hold strategies by computing the p-values as outlined in the backtesting framework (see subsection 3.2). The primary purpose of this process is to verify whether these strategies exhibit robust out-of-sample performance in terms of total returns and maximum drawdown by ensuring that the observed performance stems from the model's inherent MF selection capability rather than being attributable to cheer luck. Table 3 reports the average p-values derived from 60 backtesting windows across 30 strategies. As previously defined, the reported p-values represent the probability that a constructed strategy is outperformed by a naive baseline defined as a randomly selected single MF from the dataset. Therefore, a lower p-value indicates a reduced likelihood that the proposed strategy is being defeated by random selection. In other words, smaller p-values signify stronger statistical evidence that the strategy's effectiveness regarding the corresponding performance metric (i.e., total returns and maximum drawdown) is robust and beyond mere chance. These p-values are reported in Table 3 which is structured completely analogous to Table 2 except that total returns and maximum drawdown values are now replaced by p-values.

Table 3: Performance results for 30 buy-and-hold strategies: Average p-values of the total returns and maximum drawdown across 60 backtesting windows

Methods	p-values for total returns			p-values for maximum drawdown		
	MF(10)	MF(20)	MF(30)	MF(10)	MF(20)	MF(30)
CM-MVc	0.5022	0.4314	0.4001	0.1011	0.1118	0.1251
CM-MVnc	0.5799	0.4696	0.4372	0.0770	0.0904	0.1088
LM-MVc	0.4933	0.4208	0.3916	0.1388	0.1216	0.1281
LM-MVnc	0.5489	0.4949	0.4663	0.1239	0.1006	0.1142
TLM-MVc (1-99)	0.5251	0.4966	0.5016	0.3565	0.2956	0.2786
TLM-MVnc (1-99)	0.6743	0.6868	0.6848	0.1288	0.1097	0.1214
TLM-MVc (5-95)	0.4928	0.4897	0.4988	0.3698	0.3073	0.3053
TLM-MVnc (5-95)	0.7424	0.7361	0.7108	0.1145	0.1062	0.1110
TLM-MVc (10-90)	0.4678	0.4635	0.4643	0.3649	0.3179	0.3107
TLM-MVnc (10-90)	0.7516	0.7461	0.6983	0.1271	0.0999	0.1053
CM-MVSc	0.4725	0.4348	0.3966	0.0727	0.0857	0.1011
CM-MVSnc	0.4827	0.4608	0.4504	0.0406	0.0698	0.0892
LM-MVSc	0.4673	0.4266	0.3921	0.1317	0.1241	0.1281
LM-MVSnc	0.4927	0.4650	0.4207	0.0876	0.1117	0.1176
TLM-MVSc (1-99)	0.5319	0.5188	0.4918	0.2656	0.2131	0.2124
TLM-MVSnc (1-99)	0.5765	0.5688	0.5457	0.0895	0.1194	0.1240
TLM-MVSc (5-95)	0.4192	0.4461	0.4519	0.2925	0.2911	0.2724
TLM-MVSnc (5-95)	0.5317	0.6039	0.6122	0.1026	0.0975	0.1204
TLM-MVSc (10-90)	0.3821	0.4083	0.4245	0.2804	0.3011	0.2905
TLM-MVSnc (10-90)	0.5619	0.5379	0.5883	0.1508	0.1198	0.1241
CM-MVSKc	0.4588	0.4247	0.3928	0.0563	0.0866	0.1000
CM-MVSKnc	0.5159	0.4892	0.4920	0.0355	0.0660	0.0844
LM-MVSKc	0.4480	0.4118	0.3956	0.0780	0.1046	0.1171
LM-MVSKnc	0.5225	0.4929	0.4706	0.0488	0.0767	0.0975
TLM-MVSKc (1-99)	0.4928	0.4839	0.4623	0.2397	0.2056	0.1919
TLM-MVSKnc (1-99)	0.4963	0.5441	0.5457	0.0830	0.0924	0.1010
TLM-MVSKc (5-95)	0.4447	0.4447	0.4612	0.2809	0.2586	0.2447
TLM-MVSKnc (5-95)	0.5540	0.5839	0.6117	0.1179	0.1063	0.1138
TLM-MVSKc (10-90)	0.4145	0.4442	0.4760	0.2270	0.2332	0.2467
TLM-MVSKnc (10-90)	0.5906	0.5118	0.5299	0.1709	0.1711	0.1609

Focusing on the comparison of these 30 backtesting strategies, several findings can be observed by analysing Table 3. First, for total returns the LM framework exhibits lower p-values than both the CM and TLM models under the c assumption, although this pattern is observed in only a minority of cases under the nc assumption. One exception is LM-MVSKc that has higher p-value than CM-MVSKc, while this is exactly reversed for LM-MVSKnc compared to CM-MVSKnc.

Furthermore, an internal comparison of the TLM variants reveals a rather mixed pattern related to the trimming parameter regarding total returns. Sometimes the trimming decreases the p-values (seemingly mainly for the c case), while sometimes it increases (seemingly mainly for the nc case), and sometimes patterns are mixed. This same Table 3 reveals that the CM models yield lower p-values for maximum drawdown than their LM and TLM counterparts in all cases, except for TLM-MVnc (10-90) compared to CM-MVnc for MF(30). In addition, an internal comparison of the TLM variants reveals again rather mixed patterns related to the trimming parameter for maximum drawdown.

Second, incorporating higher-order moments on average improves the p-values for total returns and for maximum drawdown. But, there are exceptions to this general tendency. For instance, for total returns CM-MVnc has lower p-value than CM-MVSnc and CM-MVSKnc for MF(30). For example, for maximum drawdown TLM-MVnc (10-90) has lower p-value than TLM-MVSnc (10-90) and TLM-MVSKnc (10-90) for MF(10) to MF(30). However, we do not necessarily observe a monotonous improvement over the moments: adding skewness on average improves compared to MV, and adding skewness and kurtosis also on average improves compared to MV, but moving from MVS to MVSK does not guarantee an improvement. E.g., for MF(30) under total returns, we observe that moving from LM-MVnc to LM-MVSnc reduces the p-value, but then moving to LM-MVSKnc increases it again to a p-value above the initial one for LM-MVnc.

Finally, we contrast c versus nc rating models. Regarding the p-values of total returns, the c framework consistently demonstrates lower p-values than their nc counterparts. As previously discussed in section 5.2, by forming a step-wise frontier without virtual combinations, the nc models favour a more conservative set of locally dominant MF. Regarding the p-values of maximum drawdown, the nc framework generally yields lower p-values compared to the c framework, a pattern consistently observed across the MV, MVS, and MVSK models and the CM, LM, and TLM variations. This statistical evidence confirms that the nc framework is superior at mitigating severe drawdown and providing robust downside protection.

Figures 3 and 4 offer further graphical overviews of the p-value performance for total returns and maximum drawdown per buy-and-hold strategy by stacking a box plot per strategy: these Figures are similarly structured as Figures 1 and 2 above. Based on the distributional characteristics, two additional observations emerge. First, there is a clear difference in the stability of statistical significance between out-of-sample total returns and maximum drawdown. In Figure 3, the p-values for total returns exhibit wide interquartile ranges spanning much of the [0, 1] interval. This indicates that there is variation in the model's ability to outperform the baseline statistically in terms of total returns. By contrast, Figure 4 shows that p-values for maximum drawdown are more densely concentrated at lower levels, which suggests that the downside risk protection provided by the frontier-based models is relatively consistent. This result is not surprising; rather, it indirectly yet compellingly demonstrates the power of diversification. The buy-and-hold strategy of selecting funds via frontier-based models is a portfolio strategy whose total returns are averaged out to an ordinary level, but it successfully offsets the extreme downside risks faced by single funds. Second,

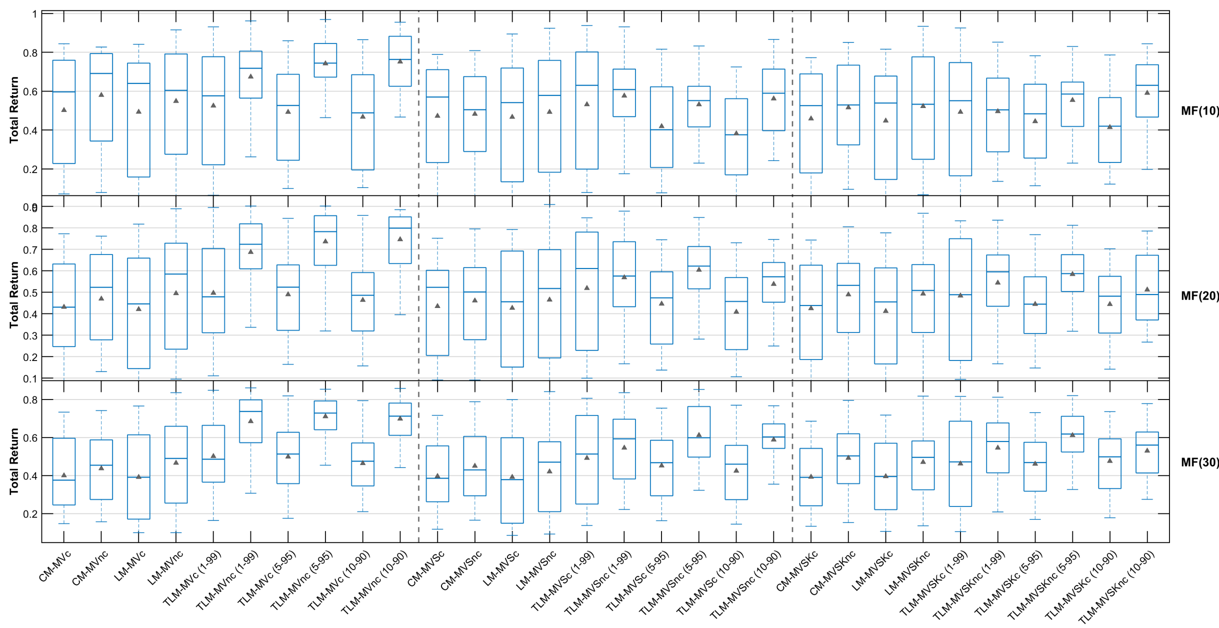

Figure 3: Distributions of the p-values for the 30 buy-and-hold strategies across 60 backtesting windows in terms of total returns

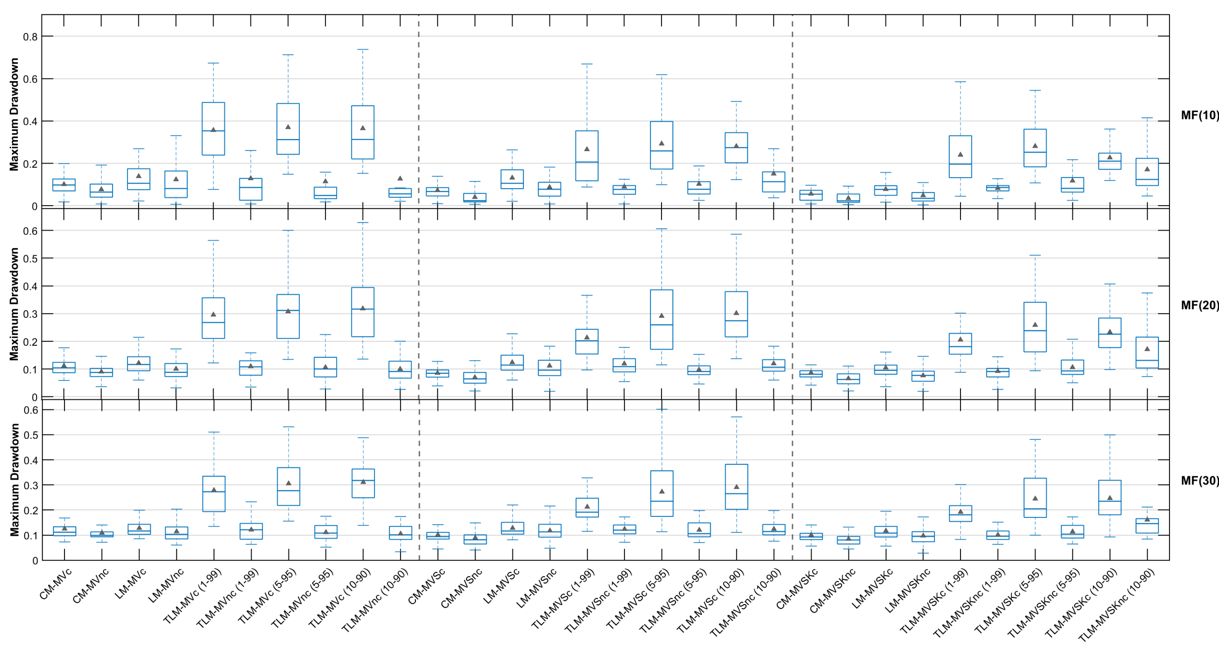

Figure 4: Distributions of the p-values for the 30 buy-and-hold strategies across 60 backtesting windows in terms of maximum drawdown

Figure 4 clearly illustrates the superiority of the nc models over their c counterparts in downside risk management. This contrast is particularly prominent within the TLM frameworks. The c models exhibit wider interquartile ranges and higher upper whiskers across all trimming levels, indicating significant volatility and instability in controlling out-of-sample maximum drawdown. By contrast, their corresponding nc counterparts maintain narrow distributions clustered near zero. In addition, as the selected MF number expands from MF(10) to MF(30), the vertical spread of distributions in Figure 4 generally narrows, implying that a larger MF pool contributes to a more stable risk management.

6 Conclusions

This contribution introduces a comprehensive multi-moment and multi-time nonparametric frontier framework for the rating and selection of MFs with the help of the shortage function. To address the computational difficulties inherent in diversified frontier rating models when incorporating higher-order moments, we adopt nondiversified c and nc frontier rating models transposed from production theory. Building upon established multi-moment and multi-time rating principles (see Kerstens et al. (2022)), the methodological contribution of this study is twofold. First, alongside the classical moments (CM), we systematically integrate the robust LM and TLM (which are still rarely used in finance) into the MF performance appraisal. To the best of our knowledge, this is the first study to formally incorporate CM, LM and TLM into a nonparametric frontier-based MF rating method within a buy-and-hold backtesting framework. Second, we propose a novel baseline approach for this backtesting procedure: rather than merely reporting absolute performance metrics, this procedure evaluates the statistical significance (p-values) of out-of-sample performance against a baseline of randomly selected MF ensuring that the observed results are driven by the models' inherent rating capabilities rather than sheer luck.

The empirical validity of the proposed framework is investigated based on a sample of 713 European equity MFs by the buy-and-hold backtesting strategy. By evaluating 30 multi-time frontier-based rating models across the MV, MVS, and MVSK settings over 60 backtesting windows, our buy-and-hold backtesting investigation has aimed to clarify the following three issues. First, our results reveal that the use of robust LM offers a superior balance between profitability and risk control compared to classical CM and the TLM. While TLM can occasionally yield higher total returns in specific scenarios, its systematic truncation of extreme tail observations inadvertently induces a tail-risk blindness during the in-sample efficiency evaluation, which generally translates into larger out-of-sample maximum drawdown. This finding questions the current preference in finance applications of TLM instead of LM (see Darolles et al. (2009), Qin (2012) and Tolikas and Gettinby (2009), among others). Second, the backtesting results indicate that extending the evaluation space to include higher-order moments (i.e., MVS and MVSK) under these classic and robust moment frameworks tends to provide consistent improvements in both total returns and maximum drawdown compared to the MV setting. Third, a clear trade-off emerges between total returns and maximum drawdown depending on the convexity assumption. By benchmarking against idealized virtual combinations, the c rating models tend to favour aggressive MFs by occasionally capturing higher absolute returns, but at the cost of exposing the out-of-sample portfolios to severe downside vulnerabilities. By contrast, the nc rating models restrict evaluations to locally dominant observable MFs resulting in more conservative MF selections that consistently provide superior downside protection. Crucially, the p-values derived from our baseline backtesting framework reveal that while consistently outperforming a naive random selection in terms of total returns remains statistically challenging across all models, the proposed nonparametric rating models under the nc assumption exhibit strong statistical significance in controlling maximum drawdown.

Obviously, the proposed methodology and backtesting results have certain limitations that offer avenues for future research. First, it would be good if the MF rating with robust moments can be extended to a more robust frontier framework (e.g., using partial frontiers). Second, another interesting future research direction could focus on extending our fundamental methodology to a metafrontier framework when several nonhomogeneous classes of MFs are considered (see Jin et al. (2024)). Third, further intensive backtesting could prove beneficial in at least the following directions: the effect of the length of validation period and/or holdout period, and the selection of a baseline strategy.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (No. 72401246), the Humanities and Social Sciences Fund of the Ministry of Education of China (No. 24YJC630173), and the Hunan Provincial Natural Science Foundation of China (No. 2026JJ60527).

References

Adam, L. and Branda, M. (2020). Risk-aversion in Data Envelopment Analysis models with diversification. *Omega*, 102:102338.

Annaert, J., van den Broeck, J., and Vander Vennet, R. (2003). Determinants of mutual fund underperformance: A Bayesian stochastic frontier approach. *European Journal of Operational Research*, 151(3):617–632.

Bacon, C. (2008). *Practical Portfolio Performance Measurement and Attribution*. Wiley, New York, 2nd edition.

Basso, A. and Funari, S. (2001). A Data Envelopment Analysis approach to measure the mutual fund performance. *European Journal of Operational Research*, 135(3):477–492.

Basso, A. and Funari, S. (2003). Measuring the performance of ethical mutual funds: A DEA approach. *Journal of the Operational Research Society*, 54(5):521–531.

Boudt, K., Cornilly, D., Van Holle, F., and Willems, J. (2020a). Algorithmic portfolio tilting to harvest higher moment gains. *Heliyon*, 6(3):e03516.

Boudt, K., Cornilly, D., and Verdonck, T. (2020b). A coskewness shrinkage approach for estimating the skewness of linear combinations of random variables. *Journal of Financial Econometrics*, 18(1):1–23.

Branda, M. (2013). Diversification-consistent data envelopment analysis with general deviation measures. *European Journal of Operational Research*, 226(3):626–635.

Branda, M. (2015). Diversification-consistent Data Envelopment Analysis based on directional-distance measures. *Omega*, 52:65–76.

Branda, M. and Kopa, M. (2014). On relations between DEA-risk models and stochastic dominance efficiency tests. *Central European Journal of Operations Research*, 22(1):13–35.

Briec, W. and Kerstens, K. (2009). Multi-horizon Markowitz portfolio performance appraisals: A general approach. *Omega: The International Journal of Management Science*, 37(1):50–62.

Briec, W. and Kerstens, K. (2010). Portfolio selection in multidimensional general and partial moment space. *Journal of Economic Dynamics and Control*, 34(4):636–656.

Briec, W., Kerstens, K., and Jokung, O. (2007). Mean-variance-skewness portfolio performance gauging: A general shortage function and dual approach. *Management Science*, 53(1):135–149.

Briec, W., Kerstens, K., and Lesourd, J. (2004). Single period Markowitz portfolio selection, performance gauging and duality: A variation on the Luenberger shortage function. *Journal of Optimization Theory and Applications*, 120(1):1–27.

Daraio, C. and Simar, L. (2006). A robust nonparametric approach to evaluate and explain the performance of mutual funds. *European Journal of Operational Research*, 175(1):516–542.

Darolles, S., Gourieroux, C., and Jasiak, J. (2009). L-performance with an application to hedge funds. *Journal of Empirical Finance*, 16(4):671–685.

Eeckhoudt, L. and Schlesinger, H. (2006). Putting risk in its proper place. *American Economic Review*, 96(1):280–289.

Elamir, E. A. and Seheult, A. H. (2003). Truncated l-moments. *Journal of Statistical Planning and Inference*, 115(2):439–458.

Fabozzi, F., Kolm, P., Pachamanova, D., and Focardi, S. (2007). *Robust Portfolio Optimization and Management*. Wiley, New York.

Feibel, B. (2003). *Investment Performance Measurement*. Wiley, New York.

Galagedera, D. and Silvapulle, P. (2002). Australian mutual fund performance appraisal using Data Envelopment Analysis. *Managerial Finance*, 28(9):60–73.

Haslem, J. and Scheraga, C. (2003). Data Envelopment Analysis of Morningstar's large-cap mutual funds. *Journal of Investing*, 12(4):41–48.

Hosking, J. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society*, 52B(1):105–124.

Jin, Q., Basso, A., Funari, S., Kerstens, K., and Van de Woestyne, I. (2024). Evaluating different groups of mutual funds using a metafrontier approach: Ethical vs. non-ethical funds. *European Journal of Operational Research*, 312(3):1134–1145.

Jondeau, E. and Rockinger, M. (2003). Conditional volatility, skewness, and kurtosis: Existence, persistence, and comovements. *Journal of Economic Dynamics and Control*, 27(10):1699–1737.

Jondeau, E. and Rockinger, M. (2006). Optimal portfolio allocation under higher moments. *European Financial Management*, 12(1):29–55.

Joro, T. and Na, P. (2006). Portfolio performance evaluation in mean-variance-skewness framework. *European Journal of Operational Research*, 175(1):446–461.

Jurczenko, E., Maillet, B., and Merlin, P. (2006). Hedge funds portfolio selection with higher-order moments: A nonparametric mean-variance-skewness-kurtosis efficient frontier. In Jurczenko, E. and Maillet, B., editors, *Multi-moment Asset Allocation and Pricing Models*, pages 51–66. Wiley, New York.

Jurczenko, E. and Yanou, G. (2010). Fund of hedge funds portfolio selection: A robust non-parametric multi-moment approach. In Watanabe, Y., editor, *The Recent Trend of Hedge Fund Strategies*, pages 21–56. Nova Science, New York.

Kerstens, K., Mazza, P., Ren, T., and Van de Woestyne, I. (2022). Multi-time and multi-moment nonparametric frontier-based fund rating: Proposal and buy-and-hold backtesting strategy. *Omega*, 113:102718.

Kerstens, K., Mounir, A., and Van de Woestyne, I. (2011a). Geometric representation of the mean-variance-skewness portfolio frontier based upon the shortage function. *European Journal of Operational Research*, 210(1):81–94.

Kerstens, K., Mounir, A., and Van de Woestyne, I. (2011b). Non-parametric frontier estimates of mutual fund performance using C- and L-moments: Some specification tests. *Journal of Banking & Finance*, 35(5):1190–1201.

Krüger, J. (2021). Nonparametric portfolio efficiency measurement with higher moments. *Empirical Economics*, 61(3):1435–1459.

Kuosmanen, T. (2007). Performance measurement and best-practice benchmarking of mutual funds: Combining stochastic dominance criteria with data envelopment analysis. *Journal of Productivity Analysis*, 28(1-2):71–86.

Lai, T. Y. (1991). Portfolio selection with skewness: A multiple objective approach. *Review of Quantitative Finance and Accounting*, 1(3):293–305.

Lamb, J. and Tee, K.-H. (2012). Data Envelopment Analysis models of investment funds. *European Journal of Operational Research*, 216(3):687–696.

Lamb, J. D. and Tee, K.-H. (2024). Using stochastic frontier analysis instead of data envelopment analysis in modelling investment performance. *Annals of Operations Research*, 332(1):891–907.

Lamb, J. D. and Tee, K.-H. (2026). Data envelopment analysis with shrinkage estimators. *OR Spectrum*, 48(1):117–142.

Lin, R., Chen, Z., Hu, Q., and Li, Z. (2017). Dynamic network DEA approach with diversification to multi-period performance evaluation of funds. *OR Spectrum*, 39(3):821–860.

Lin, R. and Li, Z. (2024). Directional distance based cross-efficiency evaluation and decomposition for parallel two-stage systems: An application to equity funds. *Expert Systems with Applications*, 242:122760.

Liu, W., Zhou, Z., Liu, D., and Xiao, H. (2015). Estimation of portfolio efficiency via DEA. *Omega*, 52(1):107–118.

Lozano, S. and Guttiérez, E. (2008a). Data Envelopment Analysis of mutual funds based on second-order stochastic dominance. *European Journal of Operational Research*, 189(1):230–244.

Lozano, S. and Guttiérez, E. (2008b). TSD-consistent performance assessment of mutual funds. *Journal of the Operational Research Society*, 59(10):1352–1362.

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

Massol, O. and Banal-Estañol, A. (2014). Export diversification through resource-based industrialization: The case of natural gas. *European Journal of Operational Research*, 237(3):1067–1082.

Matallín-Sáez, J., Soler-Domínguez, A., and Tortosa-Ausina, E. (2014). On the informativeness of persistence for evaluating mutual fund performance using partial frontiers. *Omega*, 42(1):47–64.

Morey, M. and Morey, R. (1999). Mutual fund performance appraisals: A multi-horizon perspective with endogenous benchmarking. *Omega*, 27(2):241–258.

Qin, X. (2012). Can l-moments beat central moments in modelling risk? an empirical analysis. *Applied Economics Letters*, 19(15):1441–1447.

Ren, T., Kerstens, K., and Kumar, S. (2024a). Risk-aversion versus risk-loving preferences in non-parametric frontier-based fund ratings: A buy-and-hold backtesting strategy. *European Journal of Operational Research*, 319(1):332–344.

Ren, T., Xiao, H., and Zhou, Z. (2024b). A literature review for nonparametric frontier methods applied to portfolio analysis. In Edalatpanah, S., Hosseinzadeh Lotfi, F., Kerstens, K., and Wanke, P., editors, *Analytical Decision Making and Data Envelopment Analysis: Advances and Challenges*, pages 235–257. Springer, Singapore.

Ren, T., Zhou, Z., and Xiao, H. (2021). Estimation of portfolio efficiency considering social responsibility: Evidence from the multi-horizon diversification DEA. *RAIRO-Operations Research*, 55(2):611–637.

Sarkar, S. (2022). Performance evaluation of mutual funds using frontier methods. In Ray, S., Chambers, R., and Kumbhakar, S., editors, *Handbook of Production Economics*, volume 2, pages 1281–1307. Springer, Singapore.

Scott, R. and Horvath, P. (1980). On the direction of preference for moments of higher order than the variance. *Journal of Finance*, 35(4):915–919.

Sengupta, J. (1989). Nonparametric tests of efficiency of portfolio investment. *Journal of Economics*, 50(3):1–15.

Sillito, G. P. (1969). Derivation of approximants to the inverse distribution function of a continuous univariate population from the expectations of order statistics. *Biometrika*, 56(3):641–650.

Tolikas, K. and Gettinby, G. (2009). Modelling the distribution of the extreme share returns in Singapore. *Journal of Empirical Finance*, 16(2):254–263.

Tsolas, I. (2014). Precious metal mutual fund performance appraisal using DEA modeling. *Resources Policy*, 39(1):54–60.

Xiao, H., Zhou, Z., Ren, T., and Liu, W. (2022). Estimation of portfolio efficiency in nonconvex settings: A free disposal hull estimator with non-increasing returns to scale. *Omega*, 111:102672.