European Journal of Operational Research 319 (2024) 332–344

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Innovative Applications of O.R.

Risk-aversion versus risk-loving preferences in nonparametric frontier-based fund ratings: A buy-and-hold backtesting strategy ☆

Tiantian Ren [a], Kristiaan Kerstens [b,*], Saurav Kumar [c]

[a] School of Business, Xiangtan University, Xiangtan 411105, China
[b] Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, Lille F-59000, France
[c] Indira Gandhi Institute of Development Research, Mumbai 400065, India

ARTICLE INFO

Keywords:
Shortage function
Frontier
Fund rating
Risk-loving preferences

ABSTRACT

The eventual risk-loving nature of preferences of investors has largely been ignored in the existing frontier-based fund rating literature. This contribution develops a series of nonparametric frontier-based methods to rate mutual funds accounting for both mixed risk-loving and mixed risk-aversion preferences. These new methods are proposed by defining the corresponding shortage functions that can allow for increases in all moments, or increases in odd moments and reductions in even moments. The empirical part designs a buy-and-hold backtesting to test the out-of-sample performance of the proposed rating methods corresponding to different risk preferences on the actual MF selection. The evidence indicates that the backtesting strategies based on the output frontier-based rating models with risk-loving preferences exhibit an overwhelming dominance compared to most existing frontier-based and traditional financial ratings.

1. Introduction

A mutual fund (MF) is one of the typical pooled investment vehicles that allows investors to aggregate smaller amounts of capital into a larger amount for investment. MFs have become a popular investment option for individuals and institutions, leading to a rapid growth in terms of both the amount and the diversity (see Bogle, 2005). In the actual investment process, investors rely heavily on performance measures to identify MFs worthwhile investing in among the numerous ones available. They are increasingly concerned with the ratings and/or rankings of MFs determined by explicit performance measures. Clearly, an effective MF performance appraisal provides strategic support for investors' MF screening, but also investment benchmarking for MF managers can improve the performance of their managed portfolios.

Since the foundational work of Markowitz (1952) on modern portfolio theory, it has been recognized that portfolio performance should be measured by the trade-off between portfolio return and risk, mainly based on the mean–variance (MV) portfolio optimization problem of simultaneously maximizing returns and minimizing risk. This stems from the theory of optimal investment choice, i.e., the ability to manage assets that maximize expected utility (EU) under risk. Within the EU framework, the utility function $U(x)$ describes the risk appropriation of decision makers, where the curvature of the utility function reflects whether one experiences risk-aversion (a concave utility function, i.e., $U''(x) < 0$) or risk-loving (a convex utility function, i.e., $U''(x) > 0$), where the inequalities hold for all x belonging to the domain of the function $U(x)$. However, it is fair to say that the overwhelming majority of theoretical and empirical work in portfolio analysis typically maintains the assumption that the utility function of wealth is concave (i.e., its marginal utility is diminishing) and hence investors are absolutely and globally risk-averse (see Haering, Heinrich, & Mayrhofer, 2020). The classic financial efficiency indicators (e.g., the Sharpe ratio, the Sortino ratio, among others) have basically been developed under this modern portfolio theory and serve for investment decisions of investors with risk-averse preferences solely.

More recently, the properties of the utility functions for risk-aversion and risk-loving preferences in a more general setting have been studied. Eeckhoudt and Schlesinger (2006) discuss the broad class of mixed risk-aversion (RA) utility functions that are characterized by a preference for odd moments and an aversion for even moments. In combination with the traditional risk preference like risk aversion, the mixed risk-aversion utility function is related to the prudence (i.e., $U'''(x) > 0$) and temperance (i.e., $U''''(x) < 0$) of investors corresponding to the third and fourth moments of the returns distribution. Crainich, Eeckhoudt, and Trannoy (2013) systematically investigate the properties of mixed risk-loving (RL) utility functions,

☆ We acknowledge the most constructive comments of three referees. The usual disclaimer applies.
* Corresponding author.
E-mail addresses: tiantianren@xtu.edu.cn (T. Ren), k.kerstens@ieseg.fr (K. Kerstens), saurav@igidr.ac.in (S. Kumar).

https://doi.org/10.1016/j.ejor.2024.06.013
Received 9 August 2023; Accepted 10 June 2024
Available online 22 June 2024

which agree with the RA ones for odd moments, but differ from them in that they also have a preference for even moments.

Frontier-based methods used for assessing the performance of MF have recently gained some popularity for portfolio performance evaluation in the context of RA preferences. With the help of efficiency measures transposed from production theory, these frontier methodologies measure the efficiency of a MF by estimating the distance between an observed portfolio and its reference portfolio on a portfolio frontier (or rather the nonparametric estimators of this frontier) along a given projection direction. In this framework, the portfolio frontier is regarded as a benchmark to measure the portfolio efficiency, and the projection direction serves to characterize investors' preferences for return, risk and higher-order moments.

Among the alternative efficiency measures, the shortage function (Luenberger, 1995) has proven to be an excellent tool for gauging MF performance compatible with general investor preferences by virtue of its ability to seek for improvements in multidimensional directions simultaneously.[1] Briec, Kerstens, and Lesourd (2004) are likely the first to develop the shortage function to measure portfolio efficiency in the context of RA preferences, whereby the investors can both expand expected return and contract variance simultaneously in the MV case. Using the shortage function, the efficiency of a given portfolio is measured based on the distance between it and the MV portfolio frontier (theoretical frontier). This evaluation method based on the shortage function has been extended to more generalized portfolio frameworks to be compatible with general RA investor preferences. Briec, Kerstens, and Jokung (2007) use a general shortage function to look for improvements in efficiency in MV-Skewness (MVS) space by looking for simultaneous expansions in mean return and positive skewness and reductions in risk. Even more general, for the class of RA utility functions, Briec and Kerstens (2010) assess portfolio performance for the general moments case by simultaneously looking for improvements in odd moments and reductions in even moments. Examples of studies based on the shortage function include in alphabetic order, e.g., Adam and Branda (2020), Boudt, Cornilly, and Verdonck (2020), Branda (2013, 2015), Jurczenko, Maillet, and Merlin (2006), Jurczenko and Yanou (2010), Khemchandani and Chandra (2014), Krüger (2021), Lin and Li (2020) and Massol and Banal-Estañol (2014), among others.

All of the above methods are based on diversified portfolio frontiers. These are also referred to as diversified portfolio models in the literature. These diversified models require nonlinear programming in most cases, and their potential computational burden make these models rather unsuitable for large-scale evaluations. Probably in view of the complexity of these diversified portfolio models, nonparametric production frontiers have been transposed into the financial literature in an effort to offer alternative MF ratings, which are called nonparametric (production) frontier-based rating methods. Intuitively, based on a sample of observed units, one estimates nonparametric frontiers of any multi-dimensional choice set and uses an efficiency measure to position the benchmark of each observation on the boundary of such choice set. For instance, Kerstens, Mounir, and Van de Woestyne (2011) launch a new proposal in favour of the use of shortage function in terms of convex/nonconvex nonparametric frontiers, and systematically test for the need of nonparametric frontier specifications (i.e., returns to scale, higher-order moments and convexity) in defining the efficiency measures. Liu, Zhou, Liu, and Xiao (2015) state that a convex variable returns to scale (VRS) nonparametric frontier estimator provides an inner approximation to the portfolio frontier derived from the traditional MV diversified model. These arguments are of great relevance in providing theoretical supports for the application of nonparametric frontiers in MF rating in the context of general RA investor preferences.

In particular, there has been a great development on the shortage function in combination with nonparametric frontier techniques for assessing portfolio performance with RA preferences, e.g., Brandouy, Kerstens, and Van de Woestyne (2015), Kerstens, Mazza, Ren, and Van de Woestyne (2022), Matallín-Sáez, Soler-Domínguez, and Tortosa-Ausina (2014), Nalpas, Simar, and Vanhems (2017), Xiao, Zhou, Ren, and Liu (2022) and Zhou, Gao, Xiao, Wang, and Liu (2021). In this literature, it is assumed that all investors behave similarly towards the uneven moments and even moments, i.e., favouring uneven moments and disliking even moments.

Summarizing the above discussion, it is prudent to conclude that almost all these studies on portfolio performance appraisal build upon the assumption of RA preferences of investors following the traditional paradigm in modern finance. However, a plethora of empirical studies shows that the limitations of modern finance to explain some market anomalies (such as the equity premium, small firm effect, and the diversification puzzle) is to some extent related to such a strong assumption on risk preferences (see, e.g., Statman, 2004). As behavioural portfolio theory emerged (see Shefrin & Statman, 2000 for a survey), and especially as prospect theory gradually became well understood (see the seminal work of Kahneman & Tversky, 1979), the RL (or risk-seeking) preferences of investors towards return and risk have been increasingly studied in the literature. One of the seminal articles that inspired behavioural finance theory is Friedman and Savage (1948) who note that both RA and RL preferences share roles in investment behaviour: investors who buy insurance policies often also buy lottery tickets. It implies that there exists both risk lovers and risk averters in actual investment.

Relatively recently, there has emerged experimental evidence that both RA and RL decision makers exist, and the characteristics of RL preferences have become better understood. For instance, Deck and Schlesinger (2014) show that a nonnegligible minority of individuals make consistently second-order RL choices in their experiments, and most of the risk lovers tend to exhibit prudent and intemperate behaviour. Based on the binary lotteries in the context of experimental studies, the behaviour of risk lovers is interpreted as a preference for taking a chance on the "good" outcomes or all the "bad" ones, rather than combining some of the "good" outcomes with some of relatively "bad" ones (i.e., the preferences of RA). Combining the classic EU theory, Crainich et al. (2013) formally introduces the concept of RL preferences, i.e., a preference for all odd and even moments. In fact, mixed risk lovers are distinct from mixed risk averters by the signs of even derivatives of their utility function, while they agree on the signs of odd derivatives. These innovative articles have given rise to an extensive discussion on RL preferences: e.g., Åstebro, Mata, and Santos-Pinto (2015), Bleichrodt and van Bruggen (2022), Hongwei and Wei (2019), Jokung and Mitra (2019), Nocetti (2016), among others. Overall, these observations inspire our fundamental thought of accounting for RL preferences of general investors in portfolio performance evaluation.

Given the literature on portfolio performance appraisal reviewed previously, even though the nonparametric frontier-based models have been intensively used in the finance literature, we are unaware of any discussion applying these methods to MF performance evaluation in the context of RL preferences. Therefore, the aim of this contribution is threefold. First, we develop a series of nonparametric frontier-based MF rating methods to handle MFs rating in the contexts of RA and RL preferences. On the one hand, this contribution defines performance measures by generalizing the shortage function within the traditional nonparametric input–output frontier allowing for both RA and RL preferences of investors. While a large body of contributions use the traditional input–output frontier-based methods to assess MF performance, all of these use RA preferences by default. On the other hand, given the characteristics of RL preferences, i.e., the preference for odd and even moments, we further propose new shortage functions based on the output nonparametric frontiers (see Lovell & Pastor,

[1] Actually, the shortage function has also been named as the directional distance function.

1999 for output frontiers) as the benchmark compatible with general RL investor preferences. To the best of our knowledge, the idea of introducing a nonparametric output frontier to tackle MF evaluation in the framework of RL preferences is novel.

Second, applying these proposed models to a large database of actual MFs, we test the impacts of these risk preferences and benchmark settings (i.e., nonparametric frontiers) in MF appraisal. The ability of the shortage function to seek for improvements in several directions simultaneously makes it an excellent tool for gauging performance compatible with investor preferences. Therefore, we employ a Li-test statistic to empirically evaluate differences in inefficiency results computed relative to the above proposed nonparametric frontiers under different RA and RL risk preferences.

Third, our key research question is that we explore the potential benefit of the proposed nonparametric frontier-based methods on MF rating and selection by a buy-and-hold backtesting analysis. To the best of our knowledge, we offer a first detailed backtesting analysis comparing the out-of-sample performance of the proposed frontier-based MF ratings with RA and RL preferences, as well as some traditional financial performance measures. The superior performance of the output frontier-based MF ratings with RL preferences is verified in extensive backtesting exercises and their robustness checks. These backtesting results may make some investors adopt RL behaviour and may make MF management companies to consider introducing new MF managed using RL behaviour because of the high gains.

The structure of this contribution is as follows. Section 2 presents the discussion on the nonparametric frontier methodology dealing with MF rating in the context of both RA and RL preferences. Section 3 describes the details of the backtesting setup based on the proposed nonparametric frontier-based methods, as well as the traditional finance performance measures. Section 4 presents an empirical illustration using actual MF data. Conclusions and issues for future work are summarized in the final section.

2. Methodology

This contribution focuses on gauging MFs performance for both RA and RL preferences of investors by integrating the shortage function with nonparametric frontiers. Therefore, we start by introducing a mathematical formulation of the shortage function and some basic notation. As stated previously, the shortage function is a perfectly general efficiency measure that is compatible with general investor preferences and it can easily be employed to handle negative data commonly occurring in a financial context.

Basically, there are two basic issues that need to be distinguished: (i) the choice of reference (projection) direction vector for a MF to be evaluated and (ii) the identification of the nonparametric frontier linking the different dimensions as a benchmark. First, the reference direction depends on the investor's preferences, which can be either RA or RL preference structure: the former aims to increase odd moments and decrease even moments, while the latter seeks increase in both even and odd moments. Second, the nonparametric frontier benchmarking indicates the potential improvements in multidimensional performance for this observed MF along the given reference direction.

In the remainder, we develop different types of nonparametric frontier models based on an extended shortage function for MF rating in the frameworks of RA or RL preferences. In particular, in Section 2.1 we develop a traditional frontier (TF) with inputs and outputs and both the possibility to evaluate RA and RL preferences. In Section 2.2 we develop an output frontier (OF) without input dimensions that only serves to evaluate RL preferences.

2.1. Nonparametric frontier models with inputs and outputs (TF)

To introduce some basic notations and definitions, consider that there are n MFs under evaluation, where the jth MF ($j \in \{1,\dots,n\}$) can be characterized by m input-like values x_{ij} ($i \in \{1,\dots,m\}$) and s output-like values y_{rj} ($r \in \{1,\dots,s\}$). Input-like variables need to be minimized and output-like variables need to be maximized. Hereafter, we refer to the input-like and output-like variables as inputs and outputs for short and for consistency with the nonparametric frontier technology in a production context.

We employ one widely used nonparametric production frontier-based model with VRS and strong disposability. A unified algebraic representation of convex and nonconvex input–output possibility sets for this sample of n MFs is:

$$P_{\Lambda}^{IO} = \Big\{(x,y) \in \mathbb{R}^m \times \mathbb{R}^s \mid \forall i \in \{1,\dots,m\} : x_i \geq \sum_{j=1}^{n} \lambda_j x_{ij}, \quad \forall r \in \{1,\dots,s\} : y_r \leq \sum_{j=1}^{n} \lambda_j y_{rj}, \lambda \in \Lambda\Big\}, \tag{1}$$

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.

If there exists an input–output combination $(\sum_{j=1}^{n} \lambda_j x_{ij}, \sum_{j=1}^{n} \lambda_j y_{rj})$ in the convex or nonconvex set using less inputs and producing more outputs than the observed MF, then this MF is considered inefficient since it can improve on its inputs and/or outputs. MFs are efficient if no improved input–output combinations can be found. The input–output combinations of these efficient MFs are all located on the boundary of P_{Λ}^{IO} which is called the convex or nonconvex VRS nonparametric TF.

Using the nonparametric TF defined in (1), the shortage function of any observed MF is now defined as follows:

Definition 2.1. Let $g = (g_x, g_y) \in \mathbb{R}^m \times \mathbb{R}^s$ and $g \neq 0$. For any observation $z = (x,y) \in \mathbb{R}^m \times \mathbb{R}^s$, the shortage function S_{Λ}^{IO} in the direction of vector g is defined as:

$$S_{\Lambda}^{IO}(z;g) = \sup\{\beta \in \mathbb{R} \mid z + \beta g \in P_{\Lambda}^{IO}\}. \tag{2}$$

In Definition 2.1, if $g = (g_x, g_y) \in \mathbb{R}_-^m \times \mathbb{R}_+^s$, then the variables to be reduced (inputs) and variables to be expanded (outputs) can, for example, be a vector of even and odd moment characteristics for RA preferences. This shortage function simultaneously permits the enhancement of output-like variables and the reduction of input-like variables. In contrast, if $g = (g_x, g_y) \in \mathbb{R}_+^m \times \mathbb{R}_+^s$, then the corresponding shortage function seeks for the increases in all variables (e.g., both even and odd moments) for RL preferences.

Generally speaking, if the shortage function value $S_{\Lambda}^{IO}(z_o; g_o) = 0$, then z_o is efficient and located on the frontier. By contrast, if the shortage function value $S_{\Lambda}^{IO}(z_o; g_o) > 0$ for the input–output combination $z_o = (x_o, y_o)$ of a specific MF, then z_o is inefficient and not located on the frontier of P_{Λ}^{IO}. Hence, its inputs and/or outputs can be improved according to the specific risk preferences to catch up with the VRS nonparametric TF.[2]

Consider a MF with index $o \in \{1,\dots,n\}$ in need of assessment by means of the shortage function with direction vector $g_o = (g_{xo}, g_{yo}) \in \mathbb{R}^m \times \mathbb{R}^s$. Combining (1) and Definition 2.1, the inefficiency value of this MF can be determined by solving the following model:

$$\begin{aligned}
\max \quad & \beta_{IO} \\
\text{s.t.} \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io} + \beta_{IO} g_{io}, \quad i = 1,\dots,m, \\
& \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro} + \beta_{IO} g_{ro}, \quad r = 1,\dots,s, \\
& \sum_{j=1}^{n} \lambda_j = 1, \quad \beta_{IO} \geq 0, \\
& \forall j = 1,\dots,n : \begin{cases} \lambda_j \geq 0, & \text{under convexity}, \\ \lambda_j \in \{0,1\}, & \text{under nonconvexity}. \end{cases}
\end{aligned} \tag{3}$$

[2] The shortage function is actually a measure of inefficiency. However, it is sometimes, perhaps misleadingly, called a measure of efficiency for MFs or portfolios in several literature (see, e.g., Krüger, 2021 or Xiao et al., 2022).

Model (3) is not new (see, e.g., Kerstens et al., 2011 in TF MF rating), but it can be interpreted as a general model for both RA and RL preferences simultaneously: the latter aspect is new. More specifically, for the RA preferences that aim to increase odd moments and decrease even moments, model (3) projects the benchmarked MF with index o in the direction $g_o = (-|x_{1o}|, \dots, -|x_{mo}|, |y_{1o}|, \dots, |y_{so}|)$, whereby all input-like values x_{io}, $(i = 1, \dots, m)$ are decreased, and all output-like values y_{ro}, $(r = 1, \dots, s)$ are increased in proportion to their initial values. By contrast, with respect to the RL preference seeking to increase all moments, model (3) projects this MF along the direction $g = (|x_{1o}|, \dots, |x_{mo}|, |y_{1o}|, \dots, |y_{so}|)$ such that all input- and output-like values are increased in proportion to their initial values. Furthermore, the optimal value of β^*_{IO} indicates the amount of inefficiency for MF with index o, whereby an efficient MF obtains a zero-valued shortage function ($\beta^*_{IO} = 0$). Thus, the MF_o is more efficient if its inefficiency value β^*_{IO} is closer to zero.

We can now particularize the above formulation to characterize the efficient frontier based on the nonparametric TF with either the classic moments (e.g., mean and variance) or the higher-order moments (e.g., skewness, kurtosis, etc.) for both RA and RL preferences. Since the mean return and skewness can be negative, we take the directions allowing for negative values by the absolute value of the coordinates of the position vector of the initial points, as has been proposed in Kerstens and Van de Woestyne (2011). This guarantees the proportionality and transforms the shortage function into the proportional distance function, which satisfies the commensurability property in productivity measurement (see Briec, Dumas, Kerstens, & Stenger, 2022 for the proof).

In the mean–variance–skewness–kurtosis (MVSK) space, suppose that there are n MFs to be evaluated and each MF j can be identified by its random return R_j, where $j \in \{1, \dots, n\}$. Let the variance, the skewness and the kurtosis be defined as follows: $V(R_j) = E[(R_j - E(R_j))^2]$, $S(R_j) = E[(R_j - E(R_j))^3]$, and $K(R_j) = E[(R_j - E(R_j))^4]$.[3] Therefore, the direction vector in the MVSK framework can be specified as $g = (|E(R_o)|, V(R_o), |S(R_o)|, K(R_o))$ for RL preferences, whereby all four moments are simultaneously increased in proportion to their initial values respectively, and $g = (|E(R_o)|, -V(R_o), |S(R_o)|, -K(R_o))$ for RA preferences, whereby all output-like values (i.e., mean and skewness) and input-like values (i.e., variance and kurtosis) are simultaneously increased and decreased in proportion to the corresponding initial values, respectively.

The estimation of MVSK are expressed as follows: let $r_{j,t}$ denote the sample of historical raw returns over a given time period t $(t = 1, \dots, T)$ for the jth MF $(j \in \{1, \dots, n\})$. Then, the first four moments can be estimated by using the observed sample data as follows:

$$E(R_j) = \frac{1}{T}\sum_{t=1}^{T} r_{j,t}, \tag{4}$$

$$V(R_j) = \frac{1}{T}\sum_{t=1}^{T} (r_{j,t} - E(R_j))^2, \tag{5}$$

$$S(R_j) = \frac{1}{T}\sum_{t=1}^{T} (r_{j,t} - E(R_j))^3, \tag{6}$$

$$K(R_j) = \frac{1}{T}\sum_{t=1}^{T} (r_{j,t} - E(R_j))^4. \tag{7}$$

These estimators for the first four moments are simple and straightforward, and thus are commonly used in fund (or portfolio) evaluation when higher-order moments are considered. In the literature on portfolio optimization emphasizing the use of sample historical returns to estimate the future characteristics of asset (or portfolio) returns, alternative methods for estimating the higher-order moments of funds (or portfolios) can also be found: for instance, robust estimators based on quantiles (see Kim & White, 2004) and shrinkage estimation methods (see Krüger, 2021).

[3] Note that skewness and kurtosis sometimes refer to specific transformations of the third and fourth central moments presented here. These transformations serve a statistical purpose comparable to standardization. Since these transformations are not essential in developing a multi-moment portfolio framework, we stick to central moments to avoid unnecessary complexity.

2.2. Nonparametric output frontier models without inputs (OF)

Based on the notations and definitions in the above general case, the MFs with the RL preferences look for possible extensions in all moments: it assumes that both even moments and odd moments are considered as output-like variables (to be increased). Therefore, there is in our view a need to define a new performance measure that is capable to handle the nonparametric OF (without inputs) as a benchmark for the potential expansion of all output-like variables with the use of shortage function. In a production context, the nonparametric OF without explicit inputs have been rather widely discussed in the literature (see, e.g., Lovell & Pastor, 1999).

To the best of our knowledge, we are the first to discuss the adoption of nonparametric OF in a MF performance evaluation context. Motivated by the above existing studies, by analogy we first define a unified algebraic representation of convex and nonconvex OF under the VRS assumption for a sample of n MFs. Let jth MF $(j \in \{1, \dots, n\})$ be characterized by p output-like values l_{kj} $(k \in \{1, \dots, p\})$: these output-like variables need to be maximized. The OF for this sample of n MFs is:

$$P^O_\Lambda = \left\{ l \in \mathbb{R}^p \mid \forall k \in \{1, \dots, p\} : l_k \le \sum_{j=1}^{n} \lambda_j l_{kj}, \lambda \in \Lambda \right\}, \tag{8}$$

where: $\Lambda \equiv \Lambda^C = \{\lambda \in \mathbb{R}^n \mid \sum_{j=1}^n \lambda_j = 1 \text{ and } \forall j \in \{1, \dots, n\} : \lambda_j \ge 0\}$ if convexity is assumed, and $\Lambda \equiv \Lambda^{NC} = \{\lambda \in \mathbb{R}^n \mid \sum_{j=1}^n \lambda_j = 1 \text{ and } \forall j \in \{1, \dots, n\} : \lambda_j \in \{0, 1\}\}$ under nonconvexity. To compare it with (1), since we consider the same MFs for both RL and RA preferences in later sections, the dimensions for observations are the same as (1), i.e., $p = m + s$.

If there exists an output combination $\sum_{j=1}^n \lambda_j l_{kj}$ in this convex or nonconvex OF larger than the observed MF, then this MF is considered inefficient under the RL preferences (since its output-like variables can be increased). MFs are efficient if no expanded output combinations can be found in that OF. The output combinations of these efficient MFs are all located at the boundary of P^O_Λ which is called the convex or nonconvex VRS nonparametric OF.

Based on the OF proposed in (8), the shortage function of any observed MF accounting for RL preferences is defined as:

Definition 2.2. Let $g = g_l \in \mathbb{R}^p_+$. For any observation $z = l \in \mathbb{R}^p$, the shortage function S^O_Λ in the direction of vector g is defined as:

$$S^O_\Lambda(z; g) = \sup\{\beta \in \mathbb{R} \mid z + \beta g \in P^O_\Lambda\}. \tag{9}$$

This shortage function permits the enhancement of all output-like variables (since there are no input-like variables). If the shortage function value $S^O_\Lambda(z_o; g_o) > 0$ for the output $z_o = l_o$ of a specific MF, then z_o is inefficient and not located on the OF of P^O_Λ. Hence, its outputs can be improved to catch up with the VRS nonparametric OF. By contrast, if the shortage function value $S^O_\Lambda(z_o; g_o) = 0$, then z_o is efficient and located on this OF.

Turning now to consider a MF with index $o \in \{1, \dots, n\}$ in need of assessment by means of the shortage function with direction vector $g_o = g_{lo} \in \mathbb{R}^p_+$, combining (8) and Definition 2.2, one can compute the following mathematical programming problem to determine the

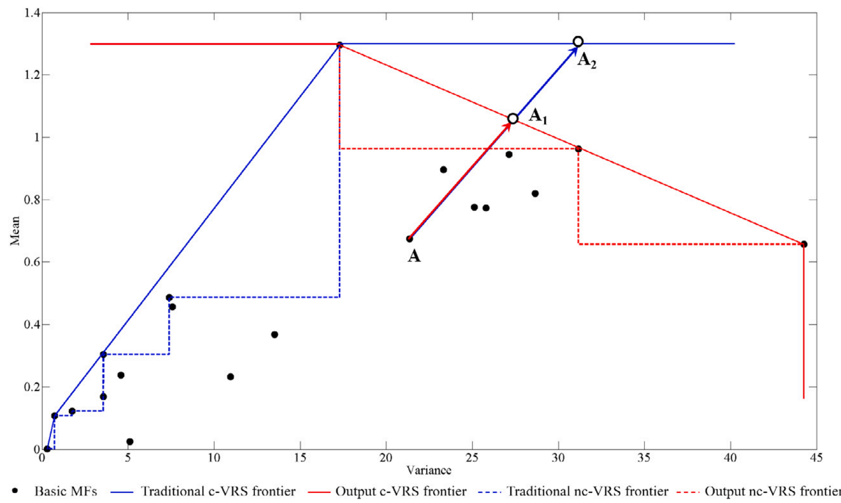

Fig. 1. Traditional and output c-VRS/nc-VRS nonparametric frontiers.

inefficiency value of this MF under evaluation:

$$\begin{aligned} \max \quad & \beta_O \\ \text{s.t.} \quad & \sum_{j=1}^{n} \lambda_j l_{kj} \geq l_{ko} + \beta_O g_{ko}, \quad k = 1, \dots, p, \\ & \sum_{j=1}^{n} \lambda_j = 1, \quad \beta_O \geq 0, \\ & \forall j = 1, \dots, n : \begin{cases} \lambda_j \geq 0, & \text{under convexity,} \\ \lambda_j \in \{0, 1\}, & \text{under nonconvexity.} \end{cases} \end{aligned} \tag{10}$$

Observe that there is no input constraint included in model (10).[4] Model (10) allows a scaling of outputs upwards to the nonparametric OF along the given direction vector g_{lo}. If the direction vector is set as $g_{lo} = (|l_{1o}|, \dots, |l_{po}|)$ (i.e., we have RL preferences), then the optimal value β_O^* of model (10) measures the maximum proportion for a MF with respect to extension in y_o along this direction based on a set of nonparametric OF. When $\beta_O^* = 0$, then this means the evaluated MF is on the VRS nonparametric OF. When this optimum β_O^* is non-zero, then this MF is inefficient and located away from the OF.

By analogy, we explain in detail how to adopt model (10) to compute the inefficiency of MF o with an example of the MVSK case. In a RL preference context, both even moments and odd moments are considered as output-like variables to be maximized. Therefore, the corresponding direction vector can be specified as $g = (|E(R_o)|, V(R_o), |S(R_o)|, K(R_o))$ such that the output-like variables (i.e., MVSK) are simultaneously increased in proportion to the initial position of o with respect to the VRS nonparametric OF. The applications of model (10) in MV and MVS cases can be specified in a similar vein.

To intuitively compare the two types of nonparametric frontiers related to model (3) and model (10) for assessing MF performance under RL preferences, we consider a simple case with an MV setting depicted in Fig. 1. In particular, we generate the nonparametric TF and the nonparametric OF based upon 20 French MFs in MV space, i.e., the convex (c-VRS)/nonconvex (nc-VRS) nonparametric efficient TF (marked as blue lines in solid and in dashed, respectively) and the c-VRS/nc-VRS nonparametric efficient OF (marked as red lines in solid and in dashed, respectively).

Let us consider the measurement of inefficiency with regard to these different types of nonparametric frontiers in Fig. 1. Consider the MF operating at point A with a given MV level, and recall that this MF simultaneously looks for increases of both return and risk dimensions. As explained previously, here we set the direction vector as the absolute value of the coordinates of the position vector of observed point A. As presented in Fig. 1, taking the convex case as an example, the inefficiency amount of observed point A based on the nonparametric TF is the proportional distance from an observed point to a projection point A_2 on the TF, whereas the inefficiency value of this point calculated by the nonparametric OF is the proportional distance to A_1. Clearly, there is a significant distinction in MFs rating under RL preferences obtained by these two types of nonparametric frontiers as benchmarks. Keeping the difference of these two rating methods in mind, we further explore the potential benefits of the proposed rating methods based on TF and OF in the following agenda.

3. Backtesting strategy

One of the main aims of this contribution is to compare and rank the out-of-sample performance of different rating methods in the actual selection of MFs. In this regard, we adopt a comparative approach based on a backtesting analysis, which has been commonly used to evaluate the out-of-sample performance of investment strategies or portfolio models using historical data of assets. Recent examples include Brandouy et al. (2015), DeMiguel, Garlappi, and Uppal (2009), Kerstens et al. (2022), Tu and Zhou (2011) and Zhou, Xiao, Jin, and Liu (2018), among others.

As introduced in Section 2, one can evaluate a MF under the RL preference using either the TF or the new OF. Instead of extensively discussing each of the alternative input-like and output-like variables of these nonparametric models, we specify the rating scenarios as a convex or nonconvex frontier model with the first two (MV), three (MVS), or four (MVSK) moments, respectively, so as to narrow down the number of potential models worthwhile considering.

For a comprehensive comparison, we also consider several typical nonparametric frontier rating models corresponding to the RA preference, where MF ratings under the RA preference are based on the combination of shortage function with the TF, as in most of the existing literature. In addition, we collect three traditional financial indicators (in particular, Sharpe, Sortino and Omega ratios) into our comparison. In total, we end up with 21 rating methods considered in our backtesting analysis as summarized in Table 1. The details on how to implement these 18 frontier rating models are available in Section 2. The exact definitions for Sharpe and Sortino ratio can be found in Feibel (2003, p. 187 and p. 200), and the definitions for all three traditional

[4] In simple terms, one can set $x_{ij} = 0$ and $g_{io} = 0$ for all $i \in m$ and $j \in n$ in model (3). Consequently, the input-like constraints are always satisfied. Thus, these constraints can be removed from model (3) to yield model (10).

Table 1
List of various rating models considered.

#	Models	Abbreviation
Traditional financial measures.		
1	Sharpe ratio	Sharpe
2	Sortino ratio	Sortino
3	Omega ratio	Omega
Traditional frontier rating corresponding to the RA preferences.		
4	Model (3) under convexity in MV framework with convexity	TF-RA: MVc
5	Model (3) under convexity in MVS framework with convexity	TF-RA: MVSc
6	Model (3) under convexity in MVSK framework with convexity	TF-RA: MVSKc
7	Model (3) under nonconvexity in MV framework with nonconvexity	TF-RA: MVnc
8	Model (3) under nonconvexity in MVS framework with nonconvexity	TF-RA: MVSnc
9	Model (3) under nonconvexity in MVSK framework with nonconvexity	TF-RA: MVSKnc
Traditional frontier rating corresponding to the RL preferences.		
10	Model (3) under convexity in MV framework with convexity	TF-RL: MVc
11	Model (3) under convexity in MVS framework with convexity	TF-RL: MVSc
12	Model (3) under convexity in MVSK framework with convexity	TF-RL: MVSKc
13	Model (3) under nonconvexity in MV framework with nonconvexity	TF-RL: MVnc
14	Model (3) under nonconvexity in MVS framework with nonconvexity	TF-RL: MVSnc
15	Model (3) under nonconvexity in MVSK framework with nonconvexity	TF-RL: MVSKnc
Output frontier rating corresponding to the RL preferences.		
16	Model (10) under convexity in MV framework with convexity	OF-RL: MVc
17	Model (10) under convexity in MVS framework with convexity	OF-RL: MVSc
18	Model (10) under convexity in MVSK framework with convexity	OF-RL: MVSKc
19	Model (10) under nonconvexity in MV framework with nonconvexity	OF-RL: MVnc
20	Model (10) under nonconvexity in MVS framework with nonconvexity	OF-RL: MVSnc
21	Model (10) under nonconvexity in MVSK framework with nonconvexity	OF-RL: MVSKnc

financial indicators can be found in Eling and Schuhmacher (2007, p. 2634 and p. 2635).

Following the backtesting setup in Kerstens et al. (2022), we design a simple buy-and-hold backtesting strategy where investors select the 10, 20, or 30 best performing MFs depending on the ranking of these 21 rating methods and hold these selections to the end of sample period. This strategy is a useful tool to compare the out-of-sample performance of different rating methods for MF selection in the actual investment process, as it is constructed with a specific asset allocation method and this selection is maintained without changing the asset components over the given holding period.

Based on the fundamental setting of backtesting, we opt for two main indicators to evaluate the out-of-sample performance of the corresponding buy-and-hold strategy: (i) the realized terminal wealth, and (ii) the shortage function with identical risk preferences as the ones used in the MF rating. First, the realized terminal wealth mainly focuses on the potential gains that can actually be realized out of sample per buy-and-hold backtesting strategy. This is regarded as a universal performance indicator for all the different backtesting strategies: a greater terminal value is better for both RA and RL preferences of investors. Second, the shortage function as an assessment tool that is compatible with different investor preferences is adopted to gauge the multidimensional performance associated with the holding return that can be achieved out of sample for each strategy. It is capable to assess either RA or RL preferences by a variety of model specifications.[5]

To empirically examine the performance of these backtesting strategies, we first collect 750 active French MFs with monthly returns from February 2011 to August 2021. The detailed description for these sample MFs is presented in the following Section 4. Then, our backtesting analysis is performed multiple times relying on a rolling time window (similar to the "rolling sample" approach in DeMiguel et al., 2009). To be specific, we split the period from the beginning of the sample period to the end of January of 2019 with a time window of 5 years (60 months). The first backtesting starts from February 2016, where the frontier-based and financial ratio-based ratings are computed with the use of 5 years of historical return data from February 2016 onward used to construct the corresponding buy-and-hold strategy for the first backtesting exercise. This exercise is repeated 36 times by sliding the time window one month at the time.

From a statistical perspective, it is commonly assumed that a sample data size larger than 30 (recall that we have 60 months) is already statistically sufficient to estimate the first four moments of the return distribution. Of course, the larger the sample of observations to estimate the MVSK of MFs, the smaller the potential bias. Therefore, given the whole sample period length of our data set in Section 4, we select an appropriate 5-year estimation window length (i.e., 60 monthly historical returns) to estimate the MVSK per MF for each backtesting exercise. Furthermore, Appendix A discusses the empirical results pertaining to the performance of the same 21 buy-and-hold backtesting strategies when using only a 3-year estimation window length (i.e., 36 monthly historical returns) to calculate the MVSK per MF.

Depending on an updated set of ratings in each period thereafter, the 10, 20 or 30 best performing MFs are selected for the backtesting exercise, and then one holds these selected MFs till the end of the whole sample period. In each of the above three selecting scenarios, the out-of-sample holding return per buy-and-hold backtesting strategy is computed and stored.[6] For each of the buy-and-hold scenarios, the out-of-sample performance of all these 21 strategies are evaluated and compared in terms of the terminal wealth, including the average values and overall distribution of realized out-of-sample terminal wealth across the 36 backtesting time windows. Moreover, these strategies are divided into three groups based on their initial risk preferences associated with the rating models, which are separately assessed using three types of shortage functions: (i) the three financial ratio-based and six TF-RA frontier-based strategies are evaluated by the shortage functions computed by TF-RA models, (ii) the six TF-RL frontier-based strategies are evaluated by the shortage functions computed by TF-RL models, and (iii) the six OF-RL frontier-based strategies are assessed by the shortage functions computed by OF-RL models. The average inefficiency scores and the number of efficient units for the values of shortage function resulting from the 36 backtesting time windows are

[5] Remark that while the financial indicators (e.g., Sharpe, Sortino and Omega ratios) are also commonly used to evaluate the out-of-sample performance of backtesting strategies (see Brandouy et al., 2015), these are traditionally conceived as only suitable for the RA world.

[6] Note that each buy-and-hold strategy starts with the same initial capital (e.g., 1 Euro), and this initial capital is invested equally in these selected MFs.

Table 2
Descriptive statistics for all 750 MFs over the entire period.

	Mean	Variance	Skewness	Kurtosis	Jarque–Bera
Min.	−0.4596	0.0017	−7919.3325	0.0000	1.6930
Q1	0.2488	3.7136	−128.9169	139.5807	64.4734
Median	0.4699	14.3466	−44.7947	1630.7688	203.9937
Mean	0.4662	14.5948	−108.4061	5619.3998	1441.8449
Q3	0.6656	22.5781	−9.5462	4248.9220	860.4685
Max.	1.3121	107.0492	962.7422	785 104.9172	74 124.5492

adopted to the inter-group comparisons. Remark that the buy-and-hold strategies driven by Sharpe, Sortino and Omega ratios are regarded as relevant for RA preferences.

4. Empirical testing

4.1. Sample description

We illustrate the RA and RL preferences of investors by using a sample of French MFs with at least 10 years of historical prices available in the Datastream database (Thomson Reuters). The dataset contains monthly prices of 750 MFs from February 2011 to August 2021. Prices are converted into a common currency (i.e., Euro) from which the monthly returns are computed. It needs to be stated that we initially specify these nonparametric frontier-based rating methods following the idea of Kerstens et al. (2011) that higher order moments and cost components are included. However, since the cost data is unavailable in this database, our empirical analysis is limited to focus on the characteristics of the return distributions for these MFs (while ignoring cost factors).

In the following, we first make a basic analysis of the monthly return characteristics of the 750 MFs in the sample over the whole sample period. The return distributions of the sample have been tested for normality using Jarque–Bera tests. Normality is rejected for 99.47% of MFs in the data set. Analysing the characteristics of the return distribution for this sample of MFs over the whole period (February 2011 to August 2021), Table 2 reports descriptive statistics on the first four moments and the Jarque–Bera statistics of the monthly returns series over the entire sample period which are used for the subsequent computations.

From Table 2, one can observe positive mean monthly returns for the majority of MFs (at least 75%) in the sample. The dispersion is quite high, as evidenced by the large positive variance, by the negative skewness, and by the positive kurtosis of these MFs throughout. It reveals that the monthly return distributions of most MFs in our sample are more negatively skewed and more fat tailed compared to the normal distribution, where both skewness and kurtosis lead to strong rejections of normality for most MFs by the Jarque–Bera tests. When negative skewness is present in the data, this implies that the payoffs of MFs are subject to the downside risk more than normally distributed MFs, while the large positive kurtosis signifies a higher probability of big gains or losses than traditional normal distributions.

4.2. Evaluation results of frontier-based methods

In this subsection, we empirically compare the differences among inefficiencies computed by various nonparametric frontier-based rating methods under the RA and RL preferences (i.e., the 18 frontier-based models listed in Table 1 of Section 2), as well as the differences among the corresponding rankings. We extract the monthly returns of these 750 MFs for the first 5 years and calculate the central moments from these return data and then we project each MF in turn against each of the 3 frontier-based ratings (i.e., TF-RA, TF-RL, and OF-RL) resulting in the inefficiencies (i.e., the values of the shortage function) and rankings of each MF relative to this sample. The resulting descriptive statistics, Li-test statistics and Kendall rank correlations of the computed frontier-ratings are shown in Table 3.

Table 3 is structured in the following way. First, the basic descriptive statistics for the inefficiencies computed by frontier-based rating methods are reported in the columns 3–8, including all models corresponding to RA and RL preferences specified in the rating frameworks of MV, MVS and MVSK with convexity and nonconvexity. Second, explaining the rows in Table 3, the first two blocks of numbers contain summary statistics for the inefficiencies of these MFs obtained by the TF corresponding to both RA and RL preferences, respectively. The third block of numbers contains summary statistics for the inefficiencies of all 750 French MFs in the sample computed by the OF corresponding to the RL preferences. The first row in each of these three horizontal blocks reports the number of efficient observations (i.e., the number of times the relevant measure of performance is estimated to be 0). The next three rows in each horizontal block report the averages, standard deviations, and median of the relevant measures. The last two horizontal blocks contain test results with respect to the first three horizontal blocks and are explained in detail below.

Before analysing the empirical results, two remarks need to be made. First, the descriptive statistics in Table 3 are specific to the shortage function calculated by the various frontier-based models: i.e., TF-RA, TF-RL, and OF-RL. Larger inefficiency values correspond with larger distances to the corresponding efficient frontier. In general, the choice of direction depends on the investor's preferences. Let us look at the three types of frontier-based rating methods TF-RA, TF-RL and OF-RL in turn. First, TF-RA inefficiency results indicate the possible reductions in its input-like variables (i.e., variance and kurtosis) and increases in its output-like variables (i.e., mean and skewness) relative to its projection on the nonparametric TF using RA preferences. Second, TF-RL inefficiency results show the potential increase in all of its input-like and output-like variables with respect to the TF using RL preferences. Third, OF-RL inefficiency results equally indicate the expansion in all input-like variables and output-like variables for this MF, but now relative to the nonparametric OF using RL preferences. Thus, the TF-RA rating seems to provide useful information for investors with RA preferences, whereas the TF-RL and OF-RL ratings are more likely to serve investors with RL preferences.

Second, the calculation of the shortage function value depends on both nonparametric frontier specifications and the choice of direction vector. For the three types of frontier-based rating methods (i.e., TF-RA, TF-RL and OF-RL), the TF-RA and TF-RL rating methods join the same nonparametric TF, but have different directions of projection. By contrast, the TF-RL and OF-RL ratings concern two different nonparametric frontiers, but with the same direction of projection respecting RL preferences. Overall, these three rating methods can be probably best compared two by two: (i) TF-RA and TF-RL share the same frontier with different projections and allow to assess the role of these different projections; and (ii) TF-RL and OF-RL have a different frontier, but the same projection, and allow to assess the role of these different frontiers.

Several conclusions can be drawn from the results reported in Table 3. The descriptive statistics for the inefficiencies computed by various frontier-based rating methods (i.e., TF-RA, TF-RL and OF-RL) all show certain differences in each rating framework (i.e., the MV, MVS and MVSK frameworks with convexity and nonconvexity). First, we focus on the comparison between the inefficiencies calculated by the TF-RA and TF-RL models. The inefficiencies of most MFs calculated by the TF-RL model is substantially higher than those calculated by the TF-RA model in all rating frameworks. This is reflected in both a higher average inefficiency score and a smaller number of efficient observations under the TF-RL model compared to those under the TF-RA model. Thus, it is clear that the difference between holding the RA and RL preferences has a great implication for the current sample of MFs in terms of ratings. This result can be seen as one indication of a

Table 3
Descriptive statistics, Li-test and Kendall rank correlations of various frontier-based ratings.

	Sample	MVc	MVSc	MVSKc	MVnc	MVSnc	MVSKnc
TF-RA	# Eff. Obs	12	19	32	37	65	66
	Average	0.4622	0.4544	0.4458	0.3961	0.3720	0.3714
	Stand. Dev	0.2502	0.2516	0.2538	0.2585	0.2639	0.2644
	Med.	0.4313	0.4261	0.4201	0.3675	0.3523	0.3518
TF-RL	# Eff. Obs	11	18	32	21	58	59
	Average	30.9006	16.6501	1.1084	30.1994	4.8654	0.6626
	Stand. Dev	268.8115	114.9035	0.8368	267.5325	58.6348	0.4496
	Med.	1.5817	1.0872	0.9676	1.5382	0.7159	0.7129
OF-RL	# Eff. Obs	2	6	6	8	32	67
	Average	20.0169	16.9509	16.9425	18.3108	7.8775	7.8056
	Stand. Dev	108.5379	93.9800	93.9815	101.7275	43.0974	43.1099
	Med.	1.6037	1.2836	1.2640	1.3202	0.5705	0.4889
TF-RA vs. TF-RL	Li-test	86.6477	52.8880	41.5621	82.5265	28.6757	28.4953
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	Rank cor	0.6160	0.6525	0.7057	0.5843	0.7191	0.7151
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
TF-RL vs. OF-RL	Li-test	4.1523	4.0902	9.3959	22.9907	24.2094	32.7104
	p-value	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
	Rank cor	0.6603	0.6597	0.4337	0.4936	0.3773	0.2814
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

greater improvement on the inefficiency for a MF along the direction of RL preference versus along the RA preference relative to the TF.

Second, we turn to an analysis between the inefficiencies obtained by the TF-RL and OF-RL models, whereby the MFs under evaluation are assumed to be consistent with RL preference, and are appraised through the nonparametric TF and OF, respectively. Although the inefficiencies obtained by the OF-RL model can in theory be either higher or lower than those obtained by the TF-RL model, Table 3 reveals that in our application the former model yields in most cases higher inefficiency scores compared to the latter model. This is reflected in the higher average inefficiency score and the smaller number of efficient observations under the OF-RL model in most rating frameworks. For instance, in the MVSK case with convexity, the number of efficient observations and the average inefficiency score for the OF-RL frontier-based rating amounts to 6 and 16.9509, while the TF-RL frontier-based rating yields the number of efficient observations and the average inefficiency score of 32 and 1.1084, respectively. To some extent, this result signifies that a greater improvement in inefficiency can be obtained for MFs under the RL preference based on the nonparametric OF compared to the TF.

To formally assess the reported differences among these various frontier-based ratings in these rating frameworks (TF-RA vs. TF-RL, and TF-RL vs. OF-RL in the MV/MVS/MVSK framework with and without convexity, respectively), the distributions of inefficiency values computed for all these models are compared by means of Li-tests. We employ the modified version of the Li-test by Li, Maasoumi, and Racine (2009): it is a nonparametric test to compare the inefficiency distributions based on the shortage function of the 18 frontier-based rating methods. It tests for the eventual statistical significance of differences between two kernel-based estimates of density functions f and g of a random variable x. The null hypothesis proposes the equality of both the density functions almost everywhere: $H_0 : \ f(x) = g(x) \ \forall \ x$. The alternative hypothesis negates this equality: $H_1 : \ f(x) \neq g(x)$ for some x. Since the Li-test statistic measures the deviation between two inefficiency distributions obtained by different frontier-based rating methods, the higher the value of the Li-test statistic, the bigger the differences between both inefficiency distributions. In addition, the Kendall rank correlations is applied to test the degree of concordance in rankings determined by these rating methods. The results of comparisons mainly aim to illustrate the difference among these frontier-based ratings for our data. The corresponding results on both Li-test statistic and Kendall rank correlations are reported in the last two horizontal blocks in Table 3.

We start by conducting Li-tests to test the null hypothesis that the inefficiency distributions computed by the TF-RA versus TF-RL models, and the TF-RL versus OF-RL models are equal in each rating framework. From Table 3, one can draw two findings. First, the inefficiencies obtained under the TF-RA model and those obtained under the TF-RL model are significantly different at the 1% significance level for all rating frameworks. A similar result can also be observed for the comparisons of the other computed frontier-ratings (i.e., TF-RL vs. OF-RL).

Second, there are variations in the Li-test statistics under different rating frameworks. Regarding the cases of the TF-RA versus TF-RL models, one can notice that the difference of the inefficiency distributions calculated by the TF-RA and TF-RL models is more pronounced under the MV framework than under the MVS/MVSK frameworks. Furthermore, the obtained inefficiency distributions between both models in the convex case show a more considerable difference compared to the corresponding nonconvex model results. In contrast, for the case of the TF-RL versus OF-RL models, the difference between their computed inefficiency distributions under the MV framework is less significant than those under the MVS/MVSK frameworks. It to some extent reveals that the identification of the nonparametric benchmarking frontier has a greater effect on the inefficiency estimates under the RL preferences than under the RL preferences towards return and risk solely. In addition, the difference between the TF-RL and OF-RL results is more significant in the case of nonconvexity than in the case of convexity.

Looking at the results of the Kendall ranking correlations presented in Table 3, to a large extent the findings obtained from the Kendall ranking correlations are compatible with those obtained from the Li-test statistic. In each rating framework, all the different frontier-based ratings yield a low correlation in terms of rankings for our MF samples. For the case of TF-RA versus TF-RL models in different rating frameworks, the correlation between the computed rankings is lower in the MV setting than that in the MVS/MVSK settings. For the case of TF-RL versus OF-RL models, the correlation between the rankings calculated by these two models is lower when higher order moments are included. Furthermore, the correlation between the rankings computed by TF-RL and OF-RL models in a nonconvex framework is lower than that in a convex framework. Similar observations also can be found for the case of TF-RA versus TF-RL. This indicates that nonconvexity has a stronger discrimination in the impacts of investor's preferences and the frontier benchmarking on the MF ratings compared to convexity.

Overall, we find that the TF-RA rating method has the lowest inefficiencies on average and thus has the best fit with the data. When comparing the TF-RL and OF-RL rating methods, we find on average that OF-RL improves the fit in MV compared to TF-RL, but this is reversed in higher moments (i.e., MVS and MVSK). Thus, the traditional

Table 4
Performance results for 21 backtesting strategies: Descriptive statistics of the values of terminal wealth.

	MF(10)		MF(20)		MF(30)	
	Average	Rank	Average	Rank	Average	Rank
Sharpe	103.6187	19	106.3515	18	108.0535	17
Sortino	102.3691	20	105.5446	20	107.3213	19
Omega	102.0042	21	105.1201	21	106.7430	21
TF-RA: MVc	113.3696	9	116.7350	8	119.6208	8
TF-RA: MVSc	106.3280	18	107.5696	16	110.9056	16
TF-RA: MVSKc	106.3564	17	105.8614	19	106.7540	20
TF-RA: MVnc	115.2394	7	113.3482	12	115.2600	10
TF-RA: MVSnc	113.2660	10	114.0830	9	113.6405	13
TF-RA: MVSKnc	112.7304	11	113.6401	11	114.2276	11
TF-RL: MVc	114.7282	8	123.2749	7	127.2493	7
TF-RL: MVSc	108.0001	16	108.6046	15	113.9245	12
TF-RL: MVSKc	108.3181	15	107.2575	17	107.5854	18
TF-RL: MVnc	112.6617	12	113.9459	10	118.5963	9
TF-RL: MVSnc	110.5223	14	110.9234	14	111.9707	14
TF-RL: MVSKnc	110.5960	13	111.6203	13	111.2350	15
OF-RL: MVc	131.0051	3	138.0682	1	140.0498	1
OF-RL: MVSc	124.6991	5	136.4635	2	138.6793	2
OF-RL: MVSKc	124.6991	5	136.3746	3	138.6793	2
OF-RL: MVnc	127.1522	4	136.1215	5	138.2332	4
OF-RL: MVSnc	131.3708	2	135.2009	6	137.4538	6
OF-RL: MVSKnc	133.8089	1	136.1764	4	137.9792	5

TF-RA rating seems to offer by far the best fit with the data. It is an open question to which extent these excellent results of the TF-RA rating method carry over to the backtesting results.

4.3. Backtesting results

As stated in Section 3, we have on purpose designed a simple buy-and-hold backtesting strategy to empirically compare the out-of-sample performance of 21 rating methods listed in Table 1, i.e., 18 frontier-based rating methods discussed above, as well as 3 traditional financial performance gauges. To assess the magnitude of the potential benefits that can actually be realized by the investors, it is necessary to analyse the out-of-sample performance of the strategies from the rating methods. In our backtesting exercises, the performance of all these buy-and-hold backtesting strategies is gauged by evaluating and ranking two types of indicators: (i) the realized terminal value starting with a capital of unity, and (ii) the shortage function with identical risk preferences as the ones applied in the MF rating. To some extent, the former indicates the level of the gain or loss realized by each backtesting strategy, and the latter reveals the impacts of the model specifications (i.e., higher-order moments and convexity) on various frontier-based ratings.

4.3.1. Terminal wealth

We first calculate the average value of the terminal wealth for each buy-and-hold strategy across 36 time windows of backtesting as displayed in Table 4. It is to be noted that the rankings listed in this table are based on the average values of terminal wealth over all strategies for each of the three alternative selection scenarios (i.e., a selection of the 10, 20 or 30 best rated MFs).

Analysing Table 4 yields the following key conclusions. First, focusing on the comparison between the two families of ratings (frontier-based vs. financial ratings), the buy-and-hold strategies based on the frontier-based ratings outperform the strategies based on the financial ratings in the majority of cases. This reveals that the frontier-based ratings under the RA and RL preferences systematically guarantee a better gain than the traditional financial ratings. In particular, the OF-RL strategies largely perform better than the financial ratio-based strategies in all six rating frameworks (i.e., the MV/MVS/MVSK frameworks with convexity and nonconvexity). Taking the selection of 10 best ranked MFs, for example, the percentage gap in the average terminal wealth between the strategy driven by the OF-RL frontier-based rating and that based on the Sharpe-driven rating is 27.39% (i.e., (131.0051 - 103.6187)/100) and 23.53% (i.e., (127.1522 - 103.6187)/100) in the MV models with convexity and nonconvexity, respectively. With respect to the TF-RL frontier-based ratings, the strategies driven by the TF-RL frontier-based ratings present superior results compared to those driven by all three financial ratings, except for the case of MVSK with convexity. For instance, in the case of MV with convexity and nonconvexity, the terminal wealth of the TF-RL strategy is on average 11.11% (i.e., (114.7282 - 103.6187)/100) and 9.04% (i.e., (112.6617 - 103.6187)/100) higher than that of Sharpe-driven strategy when the 10 best MFs are selected, for instance. A similar pattern emerges for the comparisons between the TF-RA and financial ratings. This conclusion is confirmed when buying the 20 and 30 best MFs.

Second, when one moves to the comparisons of different frontier-based ratings (i.e., TF-RA vs. TF-RL, TF-RA vs. OF-RL, and TF-RL vs. OF-RL) in these rating frameworks, it rather clearly appears that the strategies constructed by the OF-RL frontier-based rating exhibit a significant dominance over those constructed by both TF-RA and TF-RL frontier-based ratings in all rating frameworks under consideration. In each rating framework, the out-of-sample terminal wealth of the OF-RL strategy is on average significantly higher than those of the strategies according to both TF-RA and TF-RL frontier-based ratings. This indisputable result can be seen as a strong indication that, at the aggregate level, the proposed OF-RL frontier-based rating allows for a greater potential gain to actually be realized for the investors in MF selection compared to both TF-RA and TF-RL frontier-based ratings. As for the comparison of TF-RA and TF-RL frontier-based ratings, only a slight difference can be observed between both of these in terms of average terminal wealth. All of this is valid when selecting the 10, 20 and 30 best MFs.

Third, we turn to the comparisons of the terminal wealth obtained under different rating frameworks for each frontier-based rating. Regarding the cases of TF-RA and TF-RL frontier-based ratings, the terminal wealth of the strategies constructed in the MV framework performs better than those constructed in the MVS/MVSK frameworks regardless of the convexity or nonconvexity. Furthermore, the non-convex TF-RA and TF-RL frontier-based ratings yield higher terminal wealths than their convex counterparts on average in the MVS/MVSK frameworks, while a contrary result emerges in the comparison of convex and nonconvex models in the MV framework. For the case of OF-RL frontier-based ratings, there is not a clear difference among these average terminal wealths obtained based on these OF-RL strategies across the six rating frameworks. Again, these results are confirmed when buying the 10, 20 and 30 best MFs for our data.

To compare the 21 buy-and-hold backtesting strategies intuitively, Fig. 2 provides box-plots to describe the entire distributions of the terminal wealth values per strategy across 36 time windows of backtesting exercises. In each figure, the sub-figures (a) to (c) correspond to the performance results of the buying scenarios with 10, 20 and 30 best MFs selected. The box of these box-plots indicates the interquartile range whereas the small vertical line reports the location of the median. Straightforwardly, the location of the box closer to the right indicates that the entire distribution of terminal wealth values for one strategy is somewhat at a higher level.

It is clear from Fig. 2 that the buy-and-hold backtesting strategies depending on the OF-RL frontier-based ratings establish an overwhelming dominance over the strategies depending on the financial ratings as well as the other two frontier-based ratings (i.e., the TF-RA and TF-RL frontier-based ratings) in all rating frameworks. From the above backtesting results, one can conclude that the proposed nonparametric OF models corresponding to RL preferences provide a useful measurement tool for the actual MF rating and selection, allowing investors to obtain higher potential earnings.

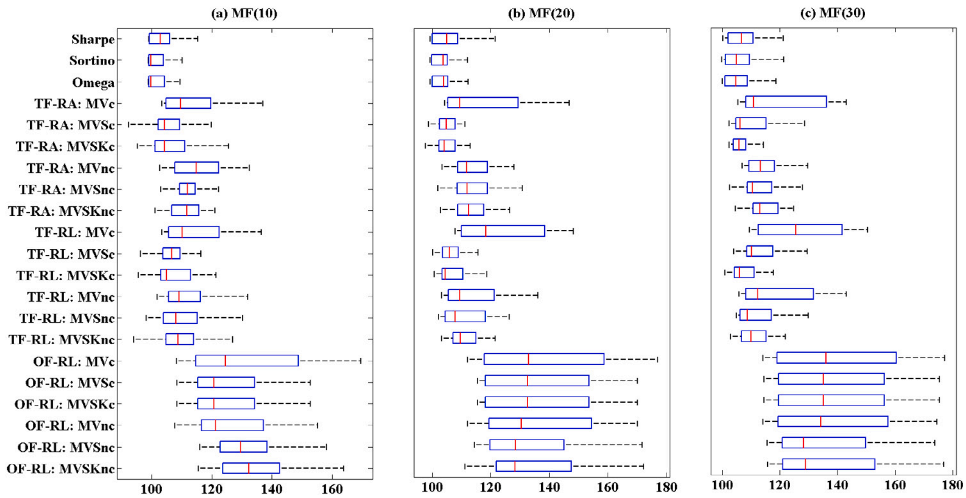

Fig. 2. Distributions of terminal wealth for the 21 buy-and-hold backtesting strategies.

To verify the robustness of these findings, we test the performance of the same 21 buy-and-hold backtesting strategies when the estimation window length is set to 3 rather than 5 years in each backtesting exercise. Table A.1 and Figure A.1 in Appendix A present the performance results of the 21 backtesting strategies with the use of this 3-year estimation window. This sensitivity analysis yields in general backtesting results that are consistent with those under a 5-year estimation window length presented above. This implies that the effect of a smaller sample and its eventual resulting bias in estimation does not seem to impact our backtesting results. A more detailed discussion on this sensitivity analysis is provided in Appendix A.

In addition, we also report the performance of the same 21 strategies held for 1 year only to check the robustness of backtesting results with respect to the holding period. The performance of these backtesting strategies can be regarded as their short-term holding performance, whereas the above strategies held to the end of the whole sample period are concerned with their long-term holding performance. The performance results of the 21 strategies held for 1 year are presented in Table B.1 and Figure B.1 of Appendix B. Overall, one finds that the backtesting strategies driven by the OF-RL frontier-based ratings still show remarkably superior performance compared to the other strategies, and the other main findings also remain consistent with those in the buy-and-hold (held-to-end) long-term holding scenario.

4.3.2. Shortage function assessments

Having evaluated the performance of all the 21 buy-and-hold backtesting strategies based on their realized terminal wealth, in this subsection we focus specifically on the inter-group comparisons with identical risk preferences based on the shortage function by a variety of model specifications. Furthermore, we do some testing on the impacts of high-order moments and convexity on different frontier-based ratings.

Following to the detailed backtesting setup in Section 3, we adapt the shortage function to evaluate the MVSK performance of the holding period return series for each group of buy-and-hold strategies separately over 36 backtesting time windows. (i) The buy-and-hold strategies in Group 1 (i.e., three financial ratio-based and six TF-RA frontier-based strategies) are evaluated by the shortage functions computed by TF-RA models involving a total of 324 MVSK observations (9×36 observations). (ii) The strategies in Group 2 (six TF-RL frontier-based strategies) are evaluated by the shortage functions computed by TF-RL models involving a total of 216 MVSK observations (6×36 observations). (iii) Finally, the strategies in Group 3 (six OF-RL frontier-based strategies) are assessed by the shortage functions computed by OF-RL models also involving a total of 216 MVSK observations (6×36 observations). Note that the shortage functions used in these three inter-group comparisons are obtained in the MVSK rating frameworks (with convexity or nonconvexity).

Tables 5 to 7 present an overall analysis with respect to the resulting shortage functions for the strategies in each of the Groups 1 to 3 ((i)-(iii)) across 36 backtesting time windows, respectively. Each of the tables is structured as follows. The first series of four columns list the results with regard to the 10 MFs selected for the backtesting exercise, and the second and third series of four columns present the results for selecting the 20 and the 30 MFs, respectively. Within each selecting (buying) scenario, the first two columns report the average inefficiency scores and the number of efficient units for each strategy when evaluated using the corresponding shortage function in the MVSK framework with convexity (MVSKc), while the last two columns report these results in the MVSK framework with nonconvexity (MVSKnc).

Looking at the results of Table 5 corresponding to the comparisons among the buy-and-hold backtesting strategies in Group 1, one can draw the following observations in the selection of 10, 20 and 30 MFs for our data. First, most of the TF-RA frontier-based strategies have lower average inefficiency scores and a larger number of efficient units than the financial ratio-based strategies. As the number of MFs selected increases, the TF-RA frontier-based ratings have a clearer superiority over financial ratio-based ratings. This somewhat confirms earlier comparative results between frontier-based rating models and traditional financial ratios in Brandouy et al. (2015) and Kerstens et al. (2022). Second, combining the average inefficiency scores and efficient units in Table 5, the TF-RA frontier-based strategies in the MVS and MVSK rating frameworks perform better than those in the basic MV rating framework, and the strategies under the convex settings are superior to those under the nonconvex settings for our MF samples.

From the results in Table 6 corresponding to the comparisons among the strategies in Group 2, it is found that the consideration of higher-order moments makes a significant contribution to the MVSK performance of the holding period return series for TF-RL frontier-based strategies. This is evidenced in the observation that the TF-RL frontier-based strategies in the MVS and MVSK settings generally yield lower average inefficiency scores and more efficient units compared to the TF-RL frontier-based strategies in the MV setting. When comparing convex and nonconvex TF-RL frontier-based strategies, the buy-and-hold strategies determined by the nonconvex TF-RL frontier-based ratings outperform those determined by those convex TF-RL frontier-based ratings in the majority of cases, except the strategies constructed by the MVSK rating framework. All of these conclusions are found when selecting the 10, 20 and 30 best MFs.

Table 7 reports the descriptive statistics of the shortage functions corresponding to the comparisons among the six OF-RL frontier-based

Table 5
Descriptive statistics of the values of shortage function computed by TF-RA models in MVSK frameworks.

Methods	MF(10)				MF(20)				MF(30)			
	TF-RA: MVSKc		TF-RA: MVSKnc		TF-RA: MVSKc		TF-RA: MVSKnc		TF-RA: MVSKc		TF-RA: MVSKnc	
	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.
Sharpe	0.0481	1	0.0333	5	0.0455	3	0.0328	8	0.0556	3	0.0417	9
Sortino	0.0368	0	0.0292	3	0.0444	2	0.0350	8	0.0537	0	0.0443	4
Omega	0.0369	1	0.0298	2	0.0421	2	0.0335	7	0.0492	1	0.0387	4
TF-RA: MVc	0.0507	3	0.0376	12	0.0535	7	0.0451	12	0.0486	7	0.0380	13
TF-RA: MVSc	0.0436	4	0.0287	10	0.0436	4	0.0344	10	0.0482	4	0.0362	11
TF-RA: MVSKc	0.0436	1	0.0298	10	0.0271	5	0.0188	12	0.0270	5	0.0162	13
TF-RA: MVnc	0.0508	4	0.0365	15	0.0473	2	0.0350	14	0.0458	4	0.0354	13
TF-RA: MVSnc	0.0513	5	0.0375	13	0.0467	7	0.0322	12	0.0483	5	0.0385	12
TF-RA: MVSKnc	0.0481	5	0.0322	16	0.0497	3	0.0324	13	0.0455	4	0.0333	12

Table 6
Descriptive statistics of the values of shortage function computed by TF-RL models in MVSK frameworks.

Methods	MF(10)				MF(20)				MF(30)			
	TF-RL: MVSKc		TF-RL: MVSKnc		TF-RL: MVSKc		TF-RL: MVSKnc		TF-RL: MVSKc		TF-RL: MVSKnc	
	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.
TF-RL: MVc	0.0682	1	0.0445	13	0.0914	5	0.0738	12	0.0949	6	0.0796	14
TF-RL: MVSc	0.0415	5	0.0299	14	0.0726	4	0.0503	10	0.0889	4	0.0741	9
TF-RL: MVSKc	0.0437	4	0.0342	11	0.0320	3	0.0251	9	0.0323	5	0.0250	11
TF-RL: MVnc	0.0418	4	0.0328	11	0.0521	3	0.0422	10	0.0597	3	0.0496	12
TF-RL: MVSnc	0.0443	7	0.0363	14	0.0537	7	0.0462	12	0.0479	5	0.0425	10
TF-RL: MVSKnc	0.0568	6	0.0456	10	0.0514	3	0.0397	10	0.0436	3	0.0366	11

Table 7
Descriptive statistics of the values of shortage function computed by OF-RL models in MVSK frameworks.

Methods	MF(10)				MF(20)				MF(30)			
	OF-RL: MVSKc		OF-RL: MVSKnc		OF-RL: MVSKc		OF-RL: MVSKnc		OF-RL: MVSKc		OF-RL: MVSKnc	
	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.	Average	#Ef. Obs.
OF-RL: MVc	0.2304	2	0.2098	3	0.2183	1	0.2165	2	0.2116	1	0.2095	3
OF-RL: MVSc	0.2317	1	0.2120	1	0.2131	0	0.2131	0	0.2111	0	0.2097	0
OF-RL: MVSKc	0.2317	1	0.2120	1	0.2135	0	0.2135	0	0.2111	0	0.2097	0
OF-RL: MVnc	0.2324	0	0.2106	0	0.2175	0	0.2172	0	0.2142	1	0.2137	1
OF-RL: MVSnc	0.2071	0	0.1850	0	0.2087	0	0.2073	0	0.2062	0	0.2047	0
OF-RL: MVSKnc	0.1944	0	0.1710	0	0.2101	1	0.2093	1	0.2059	1	0.2046	1

strategies in Group 3. Combining average inefficiency scores and the number of efficient units given in Table 7, we notice that in most cases the buy-and-hold strategies based on the OF-RL frontier-based ratings in the MVS and MVSK frameworks seem to do better than those in the MV framework. This finding reveals the necessity for the addition of higher-order moments in OF-RL frontier-based rating and in the resulting selection of MFs. Again, for the comparison of the convex and nonconvex OF-RL frontier-based ratings, the strategies based on a selection of the best MFs using the nonconvex ratings outperform those using the convex ratings on average, except for the case of MV ratings. These results are again confirmed when buying the 10, 20 and 30 best MFs.

From Tables 5 to 7, it is rather unclear how the buy-and-hold backtesting strategies driven by the OF-RL frontier-based ratings lead to the highest terminal wealth. We provide another sensitivity analysis in Appendix C where the MVSK performance of the holding return series for all 21 buy-and-hold strategies are compared jointly by using the shortage functions computed by TF-RA, TF-RL and OF-RL models combined (or, 756 (= 324 + 216 + 216) is the sum of the samples above combined). The basic question remains. This issue requires some further investigations, since no earlier evidence about the success of OF-RL strategies is known to us.

5. Conclusions

The existing methods on MF ratings make the assumption that investors are all -without any exception- RA, while the eventual RL preferences (i.e., a preference for both increases in odd moments and even moments) of investors are completely ignored. In this contribution, a first attempt is made to bridge the connection between the portfolio evaluation and RL preferences. We now summarize the main contributions in terms of both methodologies and empirical investigations.

Theoretically, this contribution introduces a series of nonparametric frontier-based methods for measuring MF performance under RA and RL preferences by combining the shortage function with different types of nonparametric frontier technologies. To the best of our knowledge, we are the first to systematically discuss the shortage functions that can account for the RL preferences in the multidimensional portfolio appraisal. This makes the shortage function a general tool for gauging MF performance in line with general investor preferences, allowing for either RA preferences or RL preferences.

In particular, by extending the shortage function, we first propose a general performance measure based on the nonparametric TF that allows for evaluating the MF performance under RA and RL preferences along a multitude of dimensions. Furthermore, we develop a new performance measure that can be used with the nonparametric OF as a benchmark for assessing MF performance under RL preferences. A two dimensional MV plane has served to clarify the geometric intuition behind these new performance measures with respect to different types of benchmarking frontiers for handling the RL preference structure of investors.

Empirically, we illustrate how the proposed nonparametric frontier-based methods work in the ratings and selections of actual MFs in the context of RA and RL preferences. Our empirical investigation is devoted to making the following two issues clear. First, we identify and discuss whether the risk attitude of investors and the benchmarking frontiers have an impact on MF ratings. The results regarding

MFs performance evaluation confirm that there exist clear differences not only between TF MF ratings under RA and RL preferences, but also between the RL MF ratings by applying the TF and the OF as benchmarks. Moreover, analysing the inefficiency values calculated by different frontier-based models, we find that the improvements in efficiency of MFs are more substantial based on the OF-RL models than those based on both TF-RL and TF-RA models on average. Second, we test the potential benefits of the proposed frontier-based methods in selecting promising investment opportunities by a buy-and-hold backtesting analysis. In terms of the realized terminal wealth out of sample used for evaluating all different ratings, it rather clearly turns out that the frontier-based ratings under both RA and RL preferences generally outperform most of the finance-based ratings. Moreover, the proposed OF-RL frontier-based ratings always establish an overwhelming dominance compared to both TF-RL and TF-RA frontier-based ratings to select the best MFs. For the shortage function out of sample used for the inter-group comparisons, we find that the considerations of higher-order moments and nonconvexity makes a contribution to the application of the proposed OF-RL frontier-based ratings in actual MF selection.

To conclude, the proposed nonparametric frontier-based methodologies provide a new insight to evaluate MF performance accounting for the general risk preferences of investors. An interesting perspective for future research is to extend the RL preference to a dynamic evaluation context by integrating the index theory of production (see, e.g., Brandouy, Briec, Kerstens, & Van de Woestyne, 2010). This would enable analysing the performance evolutions of MFs over time and identifying the attributions of performance changes. Instead of passive backtesting strategy in the current work, it could also be intriguing to test the RL preference of investors in an active backtesting setting. Furthermore, it is worthwhile exploring the application of the methodologies regarding the RL preference to the performance appraisal of different asset classes, such as equities, bonds and ETFs.

It would also be intriguing to develop a general rating methodology that can handle the presence of both RA and RL preferences in MFs under evaluation. Future research could potentially deal with this issue by attempting to cross the current framework with the proposal in Jin, Basso, Funari, Kerstens, and Van de Woestyne (2024) that explores an application of a nonparametric metatechnology in a context of ethical and non-ethical funds performance assessment: one could attempt to develop a metafrontier-based rating procedure for making performance comparisons of MFs across heterogeneous risk preferences. But, this requires a plausible classification method to separate the risk preferences of MFs depending on the RA or RL of fund managers or individual investors. Another promising line of further research is the robust estimation of financial time series to come up with estimates describing the moment distribution of MF (or asset) returns that are more robust than the ordinary central moments adopted here, such as L-moments, truncated L-moments and the like (see, e.g., Kerstens et al., 2011; Yanou, 2013).

CRediT authorship contribution statement

Tiantian Ren: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Kristiaan Kerstens:** Conceptualization, Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. **Saurav Kumar:** Conceptualization, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.06.013.

References

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

Åstebro, T., Mata, J., & Santos-Pinto, L. (2015). Skewness seeking: Risk loving, optimism or overweighting of small probabilities? *Theory and Decision*, *78*(2), 189–208.

Bleichrodt, H., & van Bruggen, P. (2022). The reflection effect for higher-order risk preferences. *The Review of Economics and Statistics*, *104*(4), 705–717.

Bogle, J. C. (2005). The mutual fund industry 60 years later: For better or worse? *Financial Analysts Journal*, *61*(1), 15–24.

Boudt, K., Cornilly, D., & Verdonck, T. (2020). 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.

Brandouy, O., Briec, W., Kerstens, K., & Van de Woestyne, I. (2010). Portfolio performance gauging in discrete time using a Luenberger productivity indicator. *Journal of Banking & Finance*, *34*(8), 1899–1910.

Brandouy, O., Kerstens, K., & Van de Woestyne, I. (2015). Frontier-based vs. Traditional mutual fund ratings: A first backtesting analysis. *European Journal of Operational Research*, *242*(1), 332–342.

Briec, W., Dumas, A., Kerstens, K., & Stenger, A. (2022). Generalised commensurability properties of eficiency measures: Implications for productivity indicators. *European Journal of Operational Research*, *303*(3), 1481–1492.

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

Briec, W., Kerstens, K., & 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., & 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.

Crainich, D., Eeckhoudt, L., & Trannoy, A. (2013). Even (mixed) risk Lovers are prudent. *American Economic Review*, *103*(4), 1529–1535.

Deck, C., & Schlesinger, H. (2014). Consistency of higher order risk preferences. *Econometrica*, *82*(5), 1913–1943.

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.

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

Eling, M., & Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking & Finance*, *31*(9), 2632–2647.

Feibel, B. (2003). *Investment performance measurement*. New York: Wiley.

Friedman, M., & Savage, L. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*(4), 279–304.

Haering, A., Heinrich, T., & Mayrhofer, T. (2020). Exploring the consistency of higher order risk preferences. *International Economic Review*, *61*(1), 283–320.

Hongwei, B., & Wei, Z. (2019). The non-integer higher-order stochastic dominance. *Operations Research Letters*, *47*(2), 77–82.

Jin, Q., Basso, A., Funari, S., Kerstens, K., & 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.

Jokung, O., & Mitra, S. (2019). Risk Lovers, mixed risk loving and the preference to combine good with good. *International Journal of Applied Management Science*, *11*(4), 295–313.

Jurczenko, E., Maillet, B., & Merlin, P. (2006). Hedge funds portfolio selection with higher-order moments: A nonparametric mean-variance-skewness-kurtosis efficient frontier. In E. Jurczenko, & B. Maillet (Eds.), *Multi-moment asset allocation and pricing models* (pp. 51–66). New York: Wiley.

Jurczenko, E., & Yanou, G. (2010). Fund of hedge funds portfolio selection: A robust non-parametric multi-moment approach. In Y. Watanabe (Ed.), *The recent trend of hedge fund strategies* (pp. 21–56). New York: Nova Science.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 264–291.

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

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

Kerstens, K., & Van de Woestyne, I. (2011). Negative data in DEA: A simple proportional distance function approach. *Journal of the Operational Research Society*, *62*(7), 1413–1419.

Khemchandani, R., & Chandra, S. (2014). Efficient trading frontier: A shortage function approach. *Optimization*, *63*(10), 1533–1548.

Kim, T.-H., & White, H. (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters*, *1*(1), 56–73.

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

Li, Q., Maasoumi, E., & Racine, J. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics*, *148*(2), 186–200.

Lin, R., & Li, Z. (2020). Directional distance based diversification super-efficiency DEA models for mutual funds. *Omega*, *97*, Article 102096.

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

Lovell, C. A. K., & Pastor, J. T. (1999). Radial DEA models without inputs or without outputs. *European Journal of Operational Research*, *118*(1), 46–51.

Luenberger, D. G. (1995). *Microeconomic theory*. Mcgraw-Hill College.

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

Massol, O., & 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., & Tortosa-Ausina, E. (2014). On the informativeness of persistence for evaluating mutual fund performance using partial frontiers. *Omega*, *42*(1), 47–64.

Nalpas, N., Simar, L., & Vanhems, A. (2017). Portfolio selection in a multi-moment setting: A simple Monte-Carlo-FDH algorithm. *European Journal of Operational Research*, *263*(1), 308–320.

Nocetti, D. C. (2016). Robust comparative statics of risk changes. *Management Science*, *62*(5), 1381–1392.

Shefrin, H., & Statman, M. (2000). Behavioral portfolio theory. *Journal of Financial and Quantitative Analysis*, *35*(2), 127–151.

Statman, M. (2004). The diversification puzzle. *Financial Analysts Journal*, *60*(4), 44–53.

Tu, J., & Zhou, G. (2011). Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, *99*, 204–215.

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

Yanou, G. (2013). Extension of the random matrix theory to the L-moments for robust portfolio selection. *Quantitative Finance*, *13*(10), 1653–1673.

Zhou, Z., Gao, M., Xiao, H., Wang, R., & Liu, W. (2021). Big data and portfolio optimization: A novel approach integrating DEA with multiple data sources. *Omega*, *104*, Article 102479.

Zhou, Z., Xiao, H., Jin, Q., & Liu, W. (2018). DEA frontier improvement and portfolio rebalancing: An application of China mutual funds on considering sustainability information disclosure. *European Journal of Operational Research*, *269*(1), 111–131.