A hybrid approach to the discrepancy in financial performance's robustness

Sally G. Arcidiacono · Damiano Rossello

Received: 19 July 2021 / Revised: 3 February 2022 / Accepted: 29 March 2022 / Published online: 3 May 2022
© The Author(s) 2022, corrected publication 2022

Abstract
Performance measurement is a crucial ingredient in the industry of investment funds. Mainly grounded on indices of risk-adjusted returns, it requires historical data to estimate the relevant statistics such as the Sharpe ratio. Therefore the measurement process is sensitive to outliers in the time series underlying historical data. Since alternative measures are available for performance evaluation, we propose an iterative methodology for a set of eleven indices (including the Sharpe ratio) in order to: (a) quantify their intrinsic degree of statistical robustness; (b) find different sensitivity to alternative outliers configuration. This methodology is a combination of a reasonable definition of breakdown point and the definition of discrepancy of a finite point set. A suitable Monte Carlo simulation provides numerical evidence of changing sensitivity among all considered performance measures, instead the classical definition of breakdown point only shows lack of robustness among all indices without further specification. Our approach may be useful in choosing the most robust performance measure to be employed in investment management, especially when robust portfolio optimization has to be used.

Keywords Robustness · Portfolio optimization · Performance measures · Breakdown point · Outliers

1 Introduction
Competing financial institutions use performance analysis to judge the skillfulness of investment fund managers. High values of a performance criterion reveal the ability of a manager in processing information not necessarily reflected by market
prices, especially when he use active investment strategies.\footnote{Passive management tries to achieve returns similar to a specific benchmark.} Standard performance measures are needed to compare funds, such as those provided by the Association for Investment Management and Research (AIMR). Therefore, portfolio performance evaluation aims at verify if fund managers (either active or passive) has met return and risk requirements set by the clients, separately from the movements of capital which they do not control. For a standard survey see AIMR (1995) and AIMR (1997). A reasonable measure of portfolio performance should be increasing with respect to the expected return and decreasing with respect to the riskiness of the investment process. A prototypical of such measure is the Sharpe ratio, a reward-to-variability index related to the classical mean-variance model of portfolio selection (see Sect. 2.1). A decision maker who faces the problem of comparing and ranking alternative investment funds can estimate one Sharpe ratio for each fund, based on predicted return and risk characteristics of the involved portfolio. Although the Sharpe ratio is considered as the reference performance measure both researchers and practitioners has tailored other indices by replacing the expected return (used as a reward measure) and/or the standard deviation of returns (used as a risk measure) with different parameters, see for example Amenc and Le Sourd (2003) for a survey at a textbook level.

There are new trends in general portfolio management based on robust procedures concerning the estimation of portfolio risk and return, or portfolio optimization. In the field of investment science, this is a novel application of a well developed approach to statistical and modelling methods. To clarify the matter, let \( r \) be a random variable representing the terminal return of an investment fund with cumulative distribution function \( F_r(x) = \mathbb{P}(r \leq x) \) for every real number \( x \in \mathbb{R} \), given a final date \( T > 0 \). Here it is assumed that a manager holds a portfolio/fund over a fixed time horizon, with initial value \( V_0 \) taken as a function \( f(S^1_0, \ldots, S^n_0) \) of asset prices \( S^i_0 \) for \( i = 1, \ldots, n \), at the date \( t = 0 \) and with terminal value \( V_T = f(S^1_T, \ldots, S^n_T) \). The final asset prices \( S^i_T \) are all random variables giving \( r = \frac{V_T - V_0}{V_0} \). Regardless the problem of the stochastic dependence among \( S^i_T \) and how to model the univariate cumulative distribution function \( F_r(x) \) given some multivariate cumulative distribution function of the random vector \( (S^1_T, \ldots, S^n_T) \), our focus in the present article is on to what extent a performance measure such as the Sharpe ratio depends upon the random return \( r \) and in turn upon its cumulative distribution function. As we will see, ex-ante performance measures are statistical models based on some characteristic of \( F_r(x) \) such as the expected value and the standard deviation in the case of the Sharpe ratio, or other summary statistics like covariances and correlation coefficients, quantiles, lower and upper partial moments in the case of different performance measures (see Sect. 2.1). Now as is typical in data analysis, we are faced with the problem of

\footnote{In classical portfolio theory this function is linear, i.e. \( V_0 = f(S^1_0, \ldots, S^n_0) = \sum_{i=1}^n h_i S^i_0 \), where \( h_i \) are the holdings in each asset with individual return \( r_i = \frac{S^i_T - S^i_0}{S^i_0} \). The terminal portfolio value is \( V_T = f(S^1_T, \ldots, S^n_T) = \sum_{i=1}^n h_i S^i_T \) since the holdings are not rebalanced over the time horizon. The portfolio return is also a linear combination of asset returns, \( r = \sum_{i=1}^n w_i r_i \), since it is assumed that there is no leverage, \( \sum_{i=1}^n w_i = 1 \) and \( w_i = \frac{h_i S^i_T}{V_0} \).}
computing the ex-post value of a performance measure given historical data on asset prices $S_t^i$ and then on their returns $r_t$, each considered as a population. One may assume that the statistical model for each $S_t^i$ is Gaussian so that each $r_t$ is log-normal distributed, i.e. a given data generating process (DGP) is conjectured and the problem is to forecast the future values $r_t$ which are comprised in the definition of $r$. This requires parameter estimation of quantities such as the means $E(r_t)$, the standard deviations $SD(r_t)$, and so on. We admittedly limit ourselves to the univariate distribution of the portfolio return, but also in this case we must deal with estimation error.\(^3\) Skewed and fat-tailed models for the cumulative distribution $F_{r}(x)$ should be more adequate than the Gaussian one because some historical data called outliers cannot be adequately described as the bulk of the data, and sometimes they have no normal pattern at all. Even a single outlier may have a serious distorting influence on the good fit of a performance measure to the bulk of the historical returns. We are aware that thinking of outliers as bad data is misleading: extreme return observations could reveal future market opportunity. But our aim is to discover differences in robustness among point estimates of selected performance measures in such a way the fund manager’s ability is better reflected by the typical pattern of historical returns rather than outliers, what we call normal market conditions. To illustrate the point, consider the following elementary example.

Example 1 Let the daily rate of return\(^4\) $r$ be a Gaussian random variable having probability distribution $N(\mu, \sigma^2)$ with daily mean return $\mu = 0.0014$ and daily volatility $\sigma = 0.0142$. Consider the ‘contaminated’ model $\tilde{r} = (1 - \lambda)r + \lambda z$, where $\lambda$ is a Bernoulli random variable with parameter $\lambda = 0.002$ (success probability) and $z$ is a degenerate random variable with point mass distribution at 0.08. The latter is an excess daily return of 8% which should be considered an extremely rare value out of normal market conditions. It is assumed that $r$ and $z$ are independent of $\lambda$, thus the underlying cumulative distribution function of the new model\(^5\) of daily return comprising the outlier 8% is $F_{\tilde{r}}(x) = (1 - \lambda)F_r(x) + \lambda F_z(x)$. Now, to compute the performance of the fund’s return $r$ we use the Sharpe measure defined as the ratio of the expected return $\mu$ to the standard deviation $\sigma$:

$$\text{SR}(r) = \frac{\mu}{\sigma} = 0.1035,$$

where SR(·) stands for the Sharpe ratio of a given random return (see Sect. 2.1). If we compute this performance ratio for the contaminated model then:

\(^3\) We do not treat the related problem of accumulated estimation errors across the asset universe, which is another delicate issue.

\(^4\) Typically, it would account for the excess return of an investment with respect to a target return (e.g. a risk-free rate).

\(^5\) The cumulative distribution function of $z$ can equivalently be given by the unit step function $u(x - 0.08)$, which is equal to 1 if $x \geq 0.08$ and is zero otherwise.
We used the following relations:

- \( \bar{E}(\tilde{r}) = (1 - \lambda)E(r) + \lambda E(z) \);
- \( \bar{E}(\tilde{r}^2) = (1 - \lambda)[\mu^2 + \sigma^2] + \lambda[E^2(z) + V(z)] \);
- \( V(\tilde{r}) = E(\tilde{r}^2) - \bar{E}(r)^2 \).

Observe that \( E(z) = 0.08 \) and \( V(z) = 0 \). After the unexpected market shock we should have a rather high daily return of 8%. This contaminates the original model \( r \) with a small probability \( \lambda 100\% = 0.2\% \) of having one outlier, and yields a higher daily mean return but almost identical daily volatility, i.e. \( SD(\tilde{r}) \approx \sigma \). As a result, the Sharpe ratio in the presence of data contamination if very high.

This example shows what can happen when we try to estimate the Sharpe ratio using historical daily returns. If the DGP is Gaussian with no deviation from normal market conditions, then plausible observations \( r_1, \ldots, r_T \) sampled from \( r \sim N(\mu, \sigma^2) \) may produce an estimate \( SR \approx 0.1035 \). On the other hand, if estimation error is taken into account\(^6\) then ex-post computation of the Sharpe ratio is spoiled by a single outlier and should be no more reliable, i.e. it seems that the true managerial skill is hide by a very extreme market movement. Moreover, this anomaly may affect the ranking of alternative investment funds as explained in the next example.

**Example 2** Suppose a second fund is under investigation and the modeler believes that its random return \( r' \) has a Gaussian distribution \( r' \sim N(\mu', \sigma^2) \), where the expected daily return is \( \mu' = 0.0016 > \mu = 0.0014 \) but the volatility is the same as that of the random return \( r \) in the previous example. Ranking the two funds through the Sharpe ratio we have:

\[
SR(r) < SR(r') = \frac{0.0016}{0.0142} = 0.1152.
\]

The second fund has assigned a higher performance because of an additional expected daily return of only \( (\mu' - \mu)100\% = 0.017\% \), the standard deviation being the same. If the modeler tries to compare the contaminated random return \( \tilde{r} \) with the second random return \( r' \) the ranking is reversed, \( SR(\tilde{r}) > SR(\tilde{r'}) \).

When estimation of alternative Sharpe ratios is based on historical daily returns, once again one outlier spoiled the reliability of the whole ranking process: under normal market conditions one should prefer the second fund based on its Sharpe ratio, but after data contamination this preference is reversed. Both examples emphasize to what extent the computation of a performance measure, in its own or

---

\(^6\) The theoretical contaminated model \( \tilde{r} \) should be fitted to a sample of historical daily returns with just one aberrant high value of 8%. This is what we called model risk.
in ranking investment funds, is sensitive to outliers. The estimation error affects the Sharpe ratio twice: the expected return parameter and the volatility/riskiness parameter are both not robust, see Lo (2002) for a financial explanation and Huber and Ronchetti (2009) for statistical reasoning. The same is true when other performance measures than the Sharpe ratio are used, see for example Rossello (2015) for an analysis restricted to four performance indices carried out with the influence function approach.

From the statistical point of view, different estimation procedures (historical, ordinary least squares, maximum likelihood, etc.) of the cumulative distribution $F_r(x)$’s parameters used in the definition the Sharpe ratio or other performance measures may deliver estimators exhibiting different sensitivities to the given dataset of historical observations. While it is possible to get robust version of them, we instead seek a procedure for quantify their degree of robustness and not merely restrict our investigation to the trade-off between robust and not robust estimators of financial performance. For the case of the risk measure’s estimator alone, the latter approach has been developed in Cont et al. (2010). On the other hand, the alternative approach studied in Krätschmer et al. (2014) inspired us in applying the concept of comparative robustness to the current context of performance measures’ sensitivity to outliers and the consequential impact on performance measurement reliability. Our novel contribution is build around two pillars: (a) the comprehensive definition of breakdown point provided by Genton and Lucas (2003; b) the definition of discrepancy of a finite set of real numbers, see Niederreiter (1992). The advantage of using (a) for robustness analysis is the ability to handle serial correlation (dependent data) and the lack of equality in distribution usually found in financial time series of returns.

In fact, as revealed by computing the breakdown point of eleven performance measures considered in this article (see Sect. 3.1 and Appendix B), it is useless to make a trade-off between robust and not robust indices. Instead, we propose a hybrid methodology which enable us to classify the performance measures according to their intrinsic robustness, looking for a degree of sensitivity with respect to dataset contaminated by outliers. As the ultimate application, this should help a decision maker (e.g. fund manager) in choosing that performance measure which is more resistant to estimation error. To the best of our knowledge, our iterative methodology is new and makes advantage of integrating the two notions of breakdown point and discrepancy. Moreover, our methodology is able to provide the most robust performance measure to be used in robust portfolio optimization leading to a reinforced robustness in forming optimal asset allocation, but without modify the economic definition of the chosen index.

---

7 For a short review of the robustness concept see Appendix A.
8 For example, we can replace the numerator of the Sharpe ratio (mean return) with the median return or the trimmed mean. Also, we can replace the denominator with a proper quantile of $F_r(x)$ as a robust risk measure.
9 The typical hypothesis in sampling from a population is that observations are independent and identically distributed, written i.i.d.
The paper is organized as follows. Section 2 is on performance measures and its connection to portfolio optimization. Section 2.1 lists eleven performance measures chosen for our investigation, as indices intimately linked to optimal asset allocation. Section 2.2 reviews the relevant works on robust portfolio optimization and the related literature on robust return-to-risk measurement. Section 3.1 reviews the Genton and Lucas (2003) definition of breakdown point to be applied for the estimation of our selected performance measures. Since this definition only detects zero robustness, to discriminate their sensitivity for different outliers configurations Sect. 3.2 proposes an ad hoc finite breakdown methodology for finding a different degree of robustness among the eleven performance measures, mixing the reasonable definition à la Genton and Lucas (2003) and the notion of discrepancy for sequences of biases. Section 4 presents a numerical analysis of our methodology through a Monte Carlo simulation, based on different outliers configurations that contaminate ‘clean’ AR(1) and GARCH(1,1) models for asset returns. Section 5 contains some concluding remarks.

2 Financial performance: a review

2.1 Selected performance indices

In this article we focus on eleven measures of financial performance defined as functions of the random variable $r$ modelling a periodic return and the related cumulative distribution $F_r(x)$’s parameters:10

1. **Sharpe ratio**, $\text{SR}(r) = \frac{\mu_r}{\sigma_r}$ where from now on $\mu_r$ denotes the expected return $E(r)$, and $\sigma_r$ is the corresponding standard deviation;
2. **Treynor ratio**, $\text{TR}(r) = \frac{\mu_r}{\beta_{r,M}}$ where as usual $\beta_{r,M} = \frac{\text{cov}(r,M)}{\sigma_M^2}$ is the systematic risk sensitivity of $r$ with respect to the market mean portfolio’s return $\mu_M = E(M)$ and variance $\sigma_M^2$;
3. **Yitzhaki-Gini ratio**, $\text{YGR}(r) = \frac{\mu_r}{\text{Gi}_r}$ where $\text{Gi}_r = \frac{1}{2}E(|r - r'|)$ is the Gini inequality measure, defined as a dispersion parameter with $r'$ being an independent copy of $r$; this definition is equivalent to the difference between $\mu_r$ and twice the mean multiplied by $\int_0^1 L_{E(r)}(c)dc$;
4. **Calmar ratio**, $\text{CR}(r) = \frac{\mu_r}{E(\text{D}(r))}$ where the numerator corresponds to the mean $\mu_r$ of the final return at the horizon $t_v = T$ over the expected maximum drawdown (see below);
5. **Mean-Absolute Deviation ratio**, $\text{MADR}(r) = \frac{\mu_r}{\text{MAD}_r}$ where the denominator is just the mean absolute deviation $E(|r - \mu_r|)$;

---

10 We assume a zero risk-free rate and do not consider excess returns.
6. **Gain-Loss ratio**, GLR(r) = \( \frac{\mu_r}{\mu_{r-}} \) where \( \mu_{r-} := \mathbb{E}(-\min\{r, 0\}) \) is the first lower partial moment; considering a zero threshold between gains and losses;

7. **Omega ratio**, OR(r) = \( \frac{\mu_r}{\mu_{r+}} \) where \( \mu_{r+} := \mathbb{E}(\max\{0, r\}) \) is the first upper partial moment, considering a zero threshold between gains and losses;

8. **Value-at-Risk ratio**, \( V\@RR(r) = \frac{\mu_r}{\sqrt{\mathbb{V}ar_{\alpha}(r)}} \) where the denominator is the Value-at-Risk of \( r \) given by the negative of the \( \alpha \)-quantile \( q_\alpha(r) \) of its cumulative distribution function (see below);

9. **Average Value-at-Risk ratio**, \( AV\@RR(r) = \frac{\mu_r}{\mathbb{V}ar_{\alpha}(r)} \) where the denominator is the coherent version of \( V\@R_{\alpha}(r) \) given by \(-\frac{1}{\alpha} \int_0^r q_\alpha(r)dc\);

10. **Jensen Alpha**, \( JA(r) = \mu_r - \mu_M\beta_{r,M} \);

11. **Morningstar Risk-Adjusted Return**, \( MRAR(r) = E\left((1 + r)^{-A}\right)^{\frac{12}{A}} - 1 \) with nonzero \( A \geq -1 \) representing a risk tolerance parameter (Morningstar’s analysts use \( A = 2 \), resulting in fund ranking consistent with risk tolerance dealt by typical retail investors).

As seen above, some performance measures require additional distributional parameters.

- The \( \alpha \)-quantile of the cumulative distribution, \( q_\alpha(r) = \inf \{x \in \mathbb{R} : F_r(x) \geq \alpha\} \), with \( \alpha \in (0, 1) \).
- The Lorenz curve \( L_{F_r}(\alpha) \) of \( F_r \) with finite expectation, defined to be the ratio of the integrated quantile function \( \int_0^r q_\alpha(r)dc \) to the mean \( \mu_r \), and interpreted as the fraction of total random return \( r \) attributed to the \( \alpha \)100\% percentage of worst scenarios.
- Assuming \( r = (r_t)_{t \in \mathbb{N}} \) is the discrete-time process for the return (in contrast to the univariate random variable \( r \) in the one-period setting), the associated drawdown process is \( D_k(r) = \max_{0 \leq i \leq k} \{r_i\} - r_{t_k} \) as a functional of the trajectory \( r \) representing the drop of a trading position’s value at time \( t_k \) with respect to its maximum preceding that time.
- The maximum drawdown of \( r = (r_t)_{t \in \mathbb{N}} \) up to time \( t_v = T \) is \( \overline{D}(r) = \max_{1 \leq k \leq n} \{D_k(r)\} \), another functional of the path \( r \) dealing with the maximum drop as defined above and giving that a risk manager handles an investment period \([0, T]\) divided into subintervals \([t_{k-1}, t_k]\).

Portfolio/fund management historically traced back to the mean-variance model of Markowitz (1952), where investors care of both the return and risk of their investments defined as a fixed level of expected return \( \mu_r = c \) such that a minimal value of the standard deviation \( \sigma_r \) is achieved. Typically, the random portfolio return is a linear combination \( r = \sum_{i=1}^n w_i r_i \) of individual random returns \( r_i \) from asset \( i = 1 \) to \( i = n \), with weight \( w_i \). Hence, this is an optimization model with control variables \( w_i \) and equality constrains \( \sum_{i=1}^n w_i \mu_i = c \) and \( \sum_{i=1}^n w_i = 1 \), where \( \mu_i \) is the expected return of the \( i \)th asset. Also, the objective function is strictly convex in the portfolio weights \((w_1, \ldots, w_n)\) and is given by \( \sigma_r = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \rho_{ij} \sigma_i \sigma_j w_i w_j} \), where \( \sigma_i \) is the standard deviation of the \( i \)th asset, \( \sigma_j \) is the standard deviation of the \( j \)th asset and \( \rho_{ij} \).
is the corresponding Pearson linear correlation coefficient. Recall that the quadratic optimization problem can be formulated as

\[ \min_{(w_1, \ldots, w_n) \in \mathcal{W}} \sqrt{ \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \sigma_i \sigma_j w_i w_j }, \]

(1)

where \( \mathcal{W} := \{ (w_1, \ldots, w_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} w_i \mu_i = c, \sum_{i=1}^{n} w_i = 1 \} \). The latter set can be modified to account for short-sale restrictions or bounds on asset/sector allocations.\(^\text{11}\) Once the parameters \( \mu_i, \sigma_i, \rho_{ij} \) are estimated using historical observations of asset returns, there is a unique solution to the quadratic programming problem above called efficient portfolio. Since this solution depends on the exogenous parameter \( c \in \mathbb{R} \), one usually sets \( c_{\text{min}} \leq c \leq c_{\text{max}} \) where \( c_{\text{min}} = \sum_{i=1}^{n} w_i^* \mu_i \) is the smallest expected return achieved with optimal weights \( (w_1^*, \ldots, w_n^*) \in \mathcal{W} \), while \( c_{\text{max}} \) is the corresponding maximum expected return for a feasible portfolio. The Sharpe ratio comes into play because problem (1) is equivalent to\(^\text{12}\)

\[ \max_{(w_1, \ldots, w_n)} \frac{\sum_{i=1}^{n} w_i \mu_i}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \sigma_i \sigma_j w_i w_j }}, \]

(2)

Taking \( \mu_r = \sum_{i=1}^{n} w_i \mu_i \) and \( \sigma_r = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \sigma_i \sigma_j w_i w_j } \) we recover the definition of the Sharpe ratio given at item 1 above. The Treynor ratio, the Jensen Alpha and the Sharpe ratio are related as follows:

- \( \text{SR}(r) = \frac{jA(r)}{\sigma_r} + \frac{\mu_r}{\sigma_r} \)
- \( \text{TR}(r) = \frac{jA(r)}{\beta_{r,M}} + \mu_{M} \).

These relations are originally due to the Capital Asset Pricing Model deduced from the Markowitz model.\(^\text{13}\) The Gain-Loss ratio has been developed to account for return values below and/or above the mean. Defining the \( q \)th lower partial moment \( \mathbb{E}((-\min(0, r))^q) \) as a measure of portfolio downside risk, for \( q = 1 \) we get the corresponding Gain-Loss ratio. For \( q = 2 \) we recover the classical semi-variance. The lower partial moment is useful whenever the cumulative distribution \( F_r(x) \) is not symmetrical. To model a risk seeking investor one can set \( 0 < q \leq 1 \), while for a risk-averse investor one can choose \( q > 1 \). The Omega ratio is a modification of the Gain-Loss ratio, and can be related to the Sharpe ratio provided that we assume a Gaussian portfolio return \( r \sim N(\mu_r, \sigma_r^2) \):

\(^{11}\) For example, the expected return level can be given as a lower bound, \( \sum_{i=1}^{n} w_i \mu_i \geq c. \)

\(^{12}\) This is no longer a quadratic optimization problem since the objective function is not concave.

\(^{13}\) Actually, one needs to reintroduce a riskless asset so that the Two-Fund Separation Theorem holds: any efficient portfolio is a linear combination of the riskless asset and of the market portfolio with random return \( M \), see Amenc and Le Sourd (2003), (Ch 4).
A hybrid approach to the discrepancy in financial performance's…

where $F'_r(\cdot)$ is the Gaussian density function of the random return $r$ evaluated at the Sharpe ratio and at its negative. The Ytzhaki-Gini ratio is studied in Ytzhaki (1983) and is based on the well known Gini inequality index. A more sophisticated models is studied by Ji et al. (2017) where the Gini’s mean differences are minimized and a new Mean-Gini ratio is developed and optimized. The Calmar ratio is a modification of the Sharpe index that uses maximum drawdown rather than standard deviation, originally introduced in the context of hedge fund performance where investors might prefer the maximum possible loss from a peak as a more adequate risk measure. The Mean-Absolute Deviation ratio is based on a linear programming formulation of the classical quadratic mean-variance portfolio optimization as studied in Konno and Yamazaki (1991), where again the risk measure in the denominator of the Sharpe ratio is properly replaced. The Value-at-Risk ratio is a further example of a Sharpe-type index, see Rachev et al. (2008), (Ch 10) and Dowd (2000). The Average Value-at-Risk ratio is a modification where the risk measure $\text{AV@R}(r)$ accounts for all the potential losses given by negative values of the random return $r$, with a specified confidence. The Morningstar Risk-Adjusted Return is actively used for performance analysis related to ranking of funds belonging to some peer group, and is related to the corresponding rating system using ‘stars’. In summary, given the Sharpe ratio as the benchmark index of performance, the remaining ten indices selected in this article are more or less its modifications, see Zakamouline and Koekebakker (2009) for a review. These modifications are motivated by not symmetrically distributed returns that exhibit skewness and fat-tails, see Donnelly and Embrechts (2010). For further references on alternative performance measures see also Christopherson et al. (2009), Rachev et al. (2008), Cogneau and Hübner (2015).

Remark 1 Non-Gaussian models of random returns are not necessarily intended to represent a DGP that is contaminated by outliers. They in general aim at modelling returns under normal market conditions for which Gaussian models are too unrealistic, since asset prices do not necessarily follow a random walk pattern and then yield historical returns for which the i.i.d. and the symmetry in distribution hypotheses do not hold. These are features of alternative DGP’s, capable to produce extreme returns with greater probability than a Gaussian model. Therefore, a non-Gaussian cumulative distribution $F_r(x)$ can be different from a contaminated model where the probability of very extreme values can be negligible. But among different non-Gaussian models, some are in fact designed to represent outliers.

Remark 2 The Sharpe ratio’s reliability represents a typical issue in fund management. It is advised that this index is fully reliable only in bull markets than in bear or mixed markets. This is also true for other performance indices (e.g. Treynor ratio, 14 Some authors use the same term for the ratio of the quantile-based risk measure $V@R(r)$ to the total size of the portfolio. We do not follow this usage.
Jensen Alpha). In fact there is a truly economic reason regardless the statistical point of view: they are designed assuming never ending bull markets. The reliability problem can lead to serious fiduciary and legal issues.

2.2 Related works

As explained in Sect. 2.1, the Sharpe ratio and the other selected performance measures are given by estimators based on relevant parameters of the underlying return distribution such as the mean, the standard deviation, the correlation coefficient. Deviations from the ideal return’s DGP affect the corresponding inference procedures that use historical observations of asset prices and returns. A strand of literature is devoted to robust portfolio selection, which in its basic form can be given as

\[
\max_{(w_1, \ldots, w_n) \in W} \left\{ \min_{(\mu_i, \sigma_i, \rho_{ij}) \in \mathcal{U}} \sum_{i=1}^{n} w_i \mu_i - \gamma \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \sigma_i \sigma_j w_i w_j} \right\},
\]

where \( \mathcal{U} \) is the so called uncertainty set containing those values of the parameters \( \mu_i, \sigma_i, \rho_{ij} \), for all \( i, j = 1, \ldots, n \), varying over some specified intervals to account for worst-case realizations of their values. The dependence of the optimal solution on the estimation error affecting \( \mu_i, \sigma_i, \rho_{ij} \) is handled in such a way inaccuracies in the input parameters do not ruin the efficiency in portfolio allocation. The uncertain parameters are embedded in the objective function and can affect the optimality of the solution, thus this methodology aims to ensure the closest possible proximity of the feasible solutions to the optimum. The size of \( \mathcal{U} \) is fixed by the modeler in order to determine the desired level of robustness and can be generated using different methodologies. For example, Tütüncü and Koenig (2004) find the ranges for the mean returns, the standard deviations and the correlation coefficients using percentiles of bootstrapped samples of historical data as well as the percentiles of moving averages. As in Goldfarb and Iyengar (2003), this approach to robustness in portfolio optimization does not rely on using robust estimates of these parameters. In fact, there is a well established research field known as robust optimization mainly dedicated to find optimal solutions that are the best for all feasible parameters’ realizations in given uncertainty sets, taking into account deviations from their nominal values. For a classical approach to robust formulation of convex and linear programming problems see Ben-Tal and Nemirovski (1998), Ben-Tal and Nemirovski (2000) and at a textbook level Ben-Tal et al. (2009). See also Bertsimas and Brown (2009) and the corresponding survey Bertsimas et al. (2011). For applications to multi-objective optimization see Fliege and Werner (2014). Complete surveys of recent advances in robust optimization are given in Gabrel et al. (2014) and Roy (2010).

15 The objective function in this robust portfolio selection problem is equivalent to that of the original quadratic programming in (1), with \( \gamma \) being a risk-aversion parameter. In general, the latter is exogenously fixed.
Using robust optimization methodologies does not require the cumulative distribution function underlying the random return’s DGP, then no probabilistic structure is assumed and optimization problems are generally computationally tractable. Since issues can arise when the number of uncertain parameters is large, additional techniques and mixed statistical procedures have been proposed in the literature for modelling the uncertainty set \( \mathcal{U} \), see for example Gregory et al. (2011) and Fabozzi et al. (2007). Moreover, the traditional approach for defining uncertainty sets can fails in delivering well diversified and not too much conservative portfolio allocations, then some authors proposed to build \( \mathcal{U} \) based both on non-stochastic confidence intervals and statistical procedures applied to the estimation of the uncertain parameters given as first, second and joint moments of the DGP’s distribution, see for example Bertsimas and Sim (2004), Lu (2006) and the expanded version Lu (2011). On one hand, tools from robust statistics (see Huber and Ronchetti (2009) for a standard reference) are available for replacing the parameters \( \mu_i, \sigma_i, \rho_{ij} \) in problem (3) with their robust counterparts. The main idea is to develop statistical procedures which are still reliable under small deviations from the assumed DGP given by outliers. Reliable procedure best fit the majority of the historical return observations in general. Another possibility is to use hybrid methodologies such as those proposed by Pinar and Paç (2014) or more recently by Momen et al. (2020). The former authors replace the risk measure in portfolio optimization problems similar to (3) with lower partial moments indices, and then introduce ambiguity (uncertainty) at the level of the DGP’s distribution and of the mean. The latter authors propose the spectral risk measures instead of the standard deviation of portfolio returns, thus suggest the Black and Litterman (1991) model to get a robust version of the mean return that is, in addition, more compatible with behavioral portfolio selection and allows for biases in investor’s modelling. For a very recent and comprehensive review on alternative methodologies of robust portfolio optimization see Ghahtarani et al. (2022).

We propose a methodology, detailed studied in Sect. 3, for selecting the more robust performance index, which we denote \( \text{PERF} \) at once, from the proposed set of eleven candidates in order to reduce the cost of robustness in asset allocation problem, but without any attempt to suggest modified versions of the parameters’ estimators such those in problem (3). We aim at providing an additional methodology that further supports robust portfolio optimization, in the sense that possibly one \( \text{PERF} \) will be chosen according to its degree of robustness and then used in robust portfolio optimization. Indeed each \( \text{PERF} \) posseses an intrinsic value of robustness depending on its own economic definition, especially considering the different conception of the involved portfolio risk measures such as those volatility-based or quantile-based.

16 Their approach rely on the idea of mean return ambiguity given by Delage and Ye (2010).

17 They are a whole class of coherent risk indices having the expected shortfall (what we called Average Value-at-Risk in Sect. 2.1) as their main representative. It is showed by Cont et al. (2010) that they are qualitatively less robust than Value-at-Risk measures, although the latter is not coherent and then may deliver sub-optimal portfolio allocations in the sense of diversification.

18 Technically, \( \text{PERF} \) could be any of the eleven performance measures listed in Sect. 2.1, e.g. \( \text{PERF}(\cdot) = \text{SR}(\cdot) \). With little abuse of notation, we refer to it as the estimator of a performance index, i.e. a function of the periodic return’s sample, see Sect. 3.1.
It is well known how the standard deviation $\sigma_i$ of returns considers positive as well as negative deviation from the mean return as risk, what should be an undesirable property when a Gaussian distributed DGP is not assumed and an almost linear formulation of the objective function in (3) is quested to make the solution algorithm more tractable. Other risk measures such as the mean absolute deviation, the lower partial moment, the Value-at-Risk, the Average Value-at-Risk or the maximum drawdown have been introduced in the literature on robust portfolio optimization. For example, Chang et al. (2009) consider also variance with skewness. See also El Gahoui et al. (2003) for a mean-Value-at-Risk robust portfolio problem of the type (3), or Deng et al. (2013) for a version with modified Sharpe ratio using Value-at-Risk, of the type (4) below. Even a more refined version of the latter work is Zhu and Fukushima (2009) where the conditional Value-at-Risk (what we called Average Value-at-Risk) is employed. A work that in addition considers both robust estimators of several risk measures and robust portfolio optimization is Scutellà and Recchia (2013). Omega ratio used in combination with worst-case scenarios for uncertainty parameters is studied in Kaspos et al. (2014). A modification that considers Average Value-at-Risk is Sharma et al. (2017). Other related studies are concerned with yet alternative performance measures not listed in our set, as Ji et al. (2022) that analyze a robust linearized stable tail adjusted return ratio used in a portfolio maximization problem under a worst-case scenario. Tong and Wu (2014) develop a reward–risk ratio models under partially known message of random variables’ distribution. Goel et al. (2017) refined this model handling the problem of multivariate dependence among portfolio assets and its interaction with portfolio weights. For a more detailed list of recent contributions to robust reward-to-risk optimization see again Ghahtaran et al. (2022).

From the mathematical statistics point of view Krätschmer et al. (2014) showed different degrees of robustness among selected risk measures that inspired us to apply the same idea to our set of performance measures in order to classify them accordingly. After this classification is made we are able to provide the decision maker with a performance index which is more resistant to data contamination, that as byproduct suggests a renewed maximization problem,

$$
\max_{(w_1, \ldots, w_n) \in W} \left\{ \min_{\theta \in U} \text{PERF}(w_1, \ldots, w_n, \theta) \right\},
$$

handling a second level of robustness given by the proper robust optimization procedure he would like to adopt for handling the uncertainty set $\mathcal{U}$. In problem (4) above the objective function can be any of the eleven performance indices (actually the more robust) introduced in Sect. 2.1 and the vector of uncertainty parameters $\theta$ can comprise $\mu_i, \sigma_i, \rho_{ij}$ and further parameters such as the lower partial moment, the mean absolute deviation, the $\alpha$-quantile or the maximum drawdown of the portfolio random return $r = \sum_{i=1}^n w_i r_i$. For $\text{PERF}(\cdot)$ equals to $\text{SR}(\cdot), \text{TR}(\cdot), \text{JA}(\cdot)$ all the individual asset return’s parameters $\mu_i, \sigma_i, \rho_{ij}$ are in $\theta$. When $\text{PERF}(\cdot)$ corresponds to $\text{YGR}(\cdot), \text{CR}(\cdot), \text{MADR}(\cdot), \text{GLR}(\cdot), \text{OR}(\cdot), \text{V@RR}(\cdot), \text{AV@RR}(\cdot)$ then the mean returns $\mu_i$ in their numerator are in $\theta$. The remaining parameters correspond to the denominators (risk measures) of the performance indices from 3 to 9 that are
nonlinear function of the portfolio weights $w_i$. In the case of $\text{MRAR}(\cdot)$ we have only one parameter corresponding to the performance index itself.

In selecting the more robust $\text{PERF}$ to be used in problem (4) we by no mean tackle the annexed portfolio allocation issue, i.e. we assume the random return $r$ has a univariate probability distribution and rule out the role of portfolio weights $w_i$. Future research agenda will handle the problem of the multivariate distribution of asset returns $(r_1, \ldots, r_n)$ and their interaction with weights $w_i$. A similar concern is afforded\(^{19}\) by Lauprete et al. (2002), that study the estimation problem underlying the classical risk-minimizing portfolio problem as influenced by marginal heavy tails modelled by a univariate Student-$t$ distribution, and multivariate tail-dependence modeled by a multivariate Student-$t$ copula. With such departures from normality, the authors propose robust alternatives to the variance portfolio estimator having lower risk. Eventually, our methodology depends on sampling contaminated univariate returns to be compared with non-contaminated ones sampled from the ideal DGP of $r$. We stress the importance of discriminating the degree of robustness among alternative reward and risk measures derived by their changing sensitivity to estimation errors, and on the other hand we emphasize the importance of unifying the two perspectives of portfolio selection (optimal allocation) and statistical modelling (robustness of estimators). To appreciate the financial perspective see for example Best and Grauer (1991), Chan et al. (1999), Chopra and Ziemba (1993), Jagannathan and Ma (2003), Israelsen (2005) to cite a few. A related strand of literature treats comparison of alternative performance measures and/or their statistical robustness. Bradrania and Pirayesh Neghab (2021) analyze changes in market conditions and their impact on performance that deviates from the ranges predicted by long-term averages of means and covariances. Since non-normal market conditions affect the truly economic definition of performance indices, other authors carry on a more in-depth analysis of alternative measures in term of their implicit robustness to outliers. León et al. (2019) study the effects of non-normality on rank correlations between orderings induced by alternative performance measures. Indeed, Caporin et al. (2014) provide a unified framework to classify performance indices. De Capitani and Pasquazzi (2015) analyze the precision of point estimators of some performance measures. Similarly, Mamatzakis and Tsionas (2021) tackle the estimation of errors in parameters of a Bayesian panel model for persistence in US funds’ performance. The methodology we develop in the current article is inspired by these works in that it tries to spotlight the more robust (in the sense explained in Sects. 3.1 and 3.2) performance measure taken from a fixed set and eventually to deliver it to a robust optimization procedure. Our approach can be applied to a larger set of performance measures without further modifications.

\(^{19}\) However, these authors provide robust estimators of the variance portfolio parameters, in contrast to our approach to search for their ‘natural’ robustness.


3 Lack of robustness and sensitivity to outliers

3.1 Finite breakdown point of selected indices

In the current article we use the breakdown point of an estimator as a measure of statistical robustness, see Huber and Ronchetti (2009), (Ch 1) and Maronna et al. (2006), Sect. 3.2) for the classical theoretical definition. Besides this there is a finite-sample definition, see Maronna et al. (2006), (Sect. 3.2.5). Actually, there are variants of the classical definition provided by the literature. To develop our methodology we propose to apply (Genton and Lucas 2003, Definition 1). In particular, let \( r = (r_1, \ldots, r_T) \) be a sample of returns \( r_t \) for a given investment fund with dates \( t = 1, \ldots, T \), i.e. \( t \) is integer valued. We do not assume they form an i.i.d. random sample, and designate it as the uncontaminated set of observations that is free of outliers.\(^{20}\) Although not universally accepted, the definition of breakdown point (BP) of an estimator as given by Genton and Lucas (2003) is suitable for implementations regardless the i.i.d. hypothesis on \( r \) where other definitions of BP could fail, see also Davies and Gather (2005) and the subsequent discussion articles. We recast the definition of BP in Genton and Lucas (2003) for the current financial context. First, for a sample of \( T \) independent or dependent returns \( r \), let \( \hat{\theta} := \text{PERF}(r) \) be the estimator of a performance measure \( \text{PERF} \) selected from the eleven indices introduced in Sect. 2.1. Instead of using the concept of individual outliers, following Genton and Lucas (2003) we use a contaminating sample \( z \) with \( 1 \leq m < T \) nonzero components. For dependent observations taken from a time series of returns, \( z \) may not necessarily have \( T - m \) zeros and \( m \) nonzeros, instead it can have less than \( m \) individual outliers. The contaminating sample allows for additive outliers, replacement outliers and innovation outliers. In time series analysis, the first and the second kinds are usually attributed to error measurement and are responsible for the inaccuracy of the estimate \( \hat{\theta} \). The latter are caused by extreme events in financial markets and enter the feed-through mechanism of the return process, from which the sample \( r \) is picked. All these kinds of outliers are taken into account by letting \( r + z \) for every \( T \)-dimensional vector \( z \in \mathcal{Z} \subset \mathbb{R}^T \), where \( \mathcal{Z} \) is the class of allowable outlier contaminations. For example, a single additive outlier is obtained by considering \( z \) with all zero components but one equals to \( \zeta \in \mathbb{R} := \mathbb{R} \cup \{\pm \infty\} \), while a patch of \( m \) additive outliers of the same magnitude \( \zeta \) are the \( m \) nonzeros of the contaminating sample (they can have alternating signs); \( c_1 \zeta, \ldots, c_m \zeta \) is a patch of non identical outliers for finite constants \( c_i \) (though they depend on a single value \( \zeta \)). In case of a single replacement outlier, we can take \( z \) with all zero components but one equals to \( -I_{\{t\}} r_t + \zeta \), where \( I_{\{t\}} = 1 \) at the date \( t \) and zero otherwise, corresponding to the component \( r_t \) in the uncontaminated sample \( r \) for \( t = 1, \ldots, n \). Innovation outliers can be characterized by assuming an autoregressive model (AR) for the time series of returns, for example an AR(1) where just one outlier \( \zeta \) affects all

\(^{20}\) For example, with historical data on returns we should have used some procedure to detect and eliminate outliers.

\(^{21}\) We assume an at least weak stationary return process.
A hybrid approach to the discrepancy in financial performance's…

subsequent observations starting from a date \( t \) in such a way the components of \( z \) are zero up to \( t - 1 \) while the following are \( \zeta, \phi \zeta, \phi^2 \zeta, \ldots \) (the autoregressive parameter is \( \phi \)). We let the bias \( b = |\text{PERF}(\hat{r} + z) - \hat{r}| \) be the measure of badness and denote \( B(\hat{r}, Z) \) the badness set containing all possible values of \( b \in \mathbb{R}^+ := \mathbb{R}^+ \cup \{+\infty\} \) for alternative realizations \( \hat{r} \) of the uncontaminated sample. Therefore, the intuition behind (Genton and Lucas 2003, Definition 1) is that the estimator \( \hat{\theta} \) takes different values for alternative sample realizations, thus for \( r \) varying over \( \mathbb{R}^T \) and the estimator being continuous in \( r \) we expect \( \hat{\theta} \) lying in some subset of \( \mathbb{R} \), for example an interval. Different values of \( \hat{\theta} \) for a continuum of possible uncontaminated samples \( r \) needed to compute the badness \( b \) (together with different \( \hat{r} \)) can be corrupted by some outlier configuration \( z \in Z \) in such a way the badness set \( B(\hat{\theta}, Z) \) collapses to a finite subset of \( \mathbb{R} \), and the estimator breaks down because of its inability to distinguish between alternative uncontaminated sample (it is no longer informative but is driven only by outliers). Note that in our notation we leave out the functional dependence of the badness set on the measure of badness and on the neighborhood \( \mathbb{R}^T \) of the uncontaminated sample \( r \) (both are fixed in the current paper). Eventually, we arrive at the following definition of breakdown point (compare with Genton and Lucas (2003), Definition 1):

\[
\varepsilon_T(\hat{\theta}, Z) = \frac{1}{T} \min \{ m - 1 \mid \text{there is at least one } z \in Z \text{ with finite } B(\hat{\theta}, Z), \text{ for every } r \}.
\]

(5)

Let the stock return’s sample mean be \( \bar{r} = \frac{1}{T} \sum_{t=1}^{T} r_t \). For \( \text{PERF} = \text{SR} \) and using the above definition we have a zero BP. Given the estimator \( \hat{\text{SR}} = \sqrt{\frac{1}{2} \left( \sum_{t=1}^{T} r_t^{-1} \right)^{-1}} \) and a contaminating sample \( z \in Z \) with one isolated additive outlier \( \zeta \), evaluation at the corrupted sample yields \( \text{SR}(\hat{r} + z) \rightarrow \pm \frac{1}{\sqrt{n-1}} \) as \( \zeta \rightarrow \pm \infty \). The badness set is finite and contains only two biases, namely \( |\frac{1}{\sqrt{n-1}} - \hat{\text{SR}}| \) and \( |\frac{1}{\sqrt{n-1}} + \hat{\text{SR}}| \), so \( \varepsilon_T(\hat{\text{SR}}, Z) = 0 \) since it suffices only \( m = 1 \) additive outlier \( \zeta = \pm \infty \) to make the SR’s estimator uninformative. See Appendix B for the calculation of BP of the remaining 10 performance indices in the case of one additive outlier.\(^{22}\)

### 3.2 Performance indices’ breakdown: a new proposal

According to the analysis developed in Sect. 3.1, all the performance measures introduced so far have BP equal to zero. We try to see what happen in a more concrete situation when outliers in any configuration (viz. different choices of \( z \in Z \) with additive, patchy and innovation components) have a finite high magnitude. Since Genton and Lucas (2003), (Definition 1) mainly depends on taking \( \zeta \rightarrow \pm \infty \), we turn things around and introduce a new integer value \( L > 1 \), independent from \( m \) and \( T \), of possible \( z \) selected from the allowable outlier contaminations \( Z \). To this end we let them be indexed, i.e. \( z_j \) for \( j = 1, \ldots, L \). Now, we want to exhibit a finite

\(^{22}\) In a similar fashion and repeating elementary algebra, we can show how all estimators break down at zero also in the case of a patch of additive outliers and innovation outliers.
badness set $B(\hat{\theta}, Z)$ for each performance measure $\text{PERF}$, and the corresponding estimator $\hat{\theta}$, as containing exactly $L$ badness measures given by biases $b_j$ computed by re-sampling different uncontaminated returns $r_j, \tilde{r}_j$ together with $z_j$; the sample size $T$ and the number $m$ of possible outliers in each $z_j$ is fixed. Essentially, we handle an iterative procedure to evaluate

$$\mathbb{B}(\hat{\theta}) := \{ b_j \mid b_j = |\text{PERF}(\tilde{r}_j + z_j) - \hat{\theta}_j|, j = 1, \ldots, L \},$$

where to emphasize the ‘discreteness’ of the badness set we use a different notation; here $\hat{\theta}_j := \text{PERF}(r_j)$. We expect a performance measure is as robust as the $b_j$ are not very tied each others, and the corresponding estimator is influenced to some extent by the uncontaminated observations, and not so much driven by outliers. Being the biases spread as uniformly as possible, reveals an ‘acceptable’ number of clusters among the $b_j$ which in turns testifies an acceptable degree of robustness of $\hat{\theta}$; it is not ‘forced’ to assume the ‘same’ values. This reasoning is in line with the intuition behind the definition of Genton and Lucas (2003). On the other hand, the badness set is now discrete by construction and we have not yet a continuum of values for the estimator and for the badness measure. Therefore, we cannot use their definition of finite BP. This is where our contribution comes into play: to analyze the sameness of possible outliers in each $z_j$ contained in $\mathbb{B}(\hat{\theta})$, we propose the concept of discrepancy of a finite subset of real numbers, see for example Theorem 2.7, pag. 16 in Niederreiter (1992). For a finite sequence $0 \leq c_1 \leq \ldots \leq c_L \leq 1$ the discrepancy is defined as

$$D_L(c_1, \ldots, c_L) := \frac{1}{L} + \max_{1 \leq j \leq L} \left( \frac{j}{L} - c_j \right) - \min_{1 \leq j \leq L} \left( \frac{j}{L} - c_j \right).$$

Whence, $D_L(c_1, \ldots, c_L)$ quantifies the deviation of the fraction of points $c_j$ within any interval with respect to the length of that interval. A finite sequence is uniformly distributed whenever its discrepancy is bounded above by $K \cdot \frac{\ln L}{L}$, for some positive constant $K$. Thus, the lower is the discrepancy the higher is the spread of $c_j$. Our proposal is to interpret these points as the ordered biases contained in the badness set. But since the discrepancy refers to points lying into a subset of the unit interval we need to map $\mathbb{B}(\hat{\theta})$ onto $[0, 1]$. Let assume the $b_j$ have been ordered. Our choice for the one-to-one correspondence is $c_j = \frac{1}{\pi} \arctan b_j + \frac{1}{2}$, for $j = 1, \ldots, L$;\footnote{Other 1-to-1 correspondence can be use, but they should correspond to equivalent metrics.} we denote the transformed set $l(\hat{\theta})$. Introducing an ad hoc set of allowable outlier contaminations which we denote $Z^* = \{ z_1, \ldots, z_L \}$, we let

$$\epsilon_T(\hat{\theta}, Z^*) := D_L(l(\hat{\theta}))$$

be the ‘discrepancy breakdown point’ of a performance estimator (dis-BP for short), thus we propose the following iterative procedure:

1. We sample couples of uncontaminated returns $(r_j, \tilde{r}_j) \subset \mathbb{R}^T \times \mathbb{R}^T$ for $j = 1, \ldots, L$;
2. We generate contaminating samples $z_j$ for $j = 1, \ldots, L$;
3. We compute ordered biases $b_j$ for $j = 1, \ldots, L$ yielding the badness set $\mathbb{B}(\hat{\theta})$;
4. We get normalized biases $c_j$ from each $b_j$, using the chosen 1-to-1 mapping, and deduce the transformed badness set $\mathbb{I}(\hat{\theta})$ as a subset of the unit interval;
5. We compute dis-BP according to the above definition, and consider a performance estimator $\hat{\theta}$ as more robust as lower is this.

Implementing the above procedure requires the model specification for the uncontaminated return series $r = (r_t)_{t=1,2,\ldots}$ and/or the associated contaminating samples $z_j$, given a number $m$ of outliers and a sample size $T$. More specifically, it is necessary to complement this procedure by considering the following probability models for outliers:

- **(Additive)** $y_t = r_t + z_t$, where the return process $r_t$ and the contaminating process $z_t$ are independent.
- **(Replacement)** $y_t = (1 - I_{[t]} )r_t + I_{[t]} z_t$, where $I_{[t]}$ is a process equal to 1 at $t$ and zero otherwise with $P( I_{[t]} = 0 ) = 1 - p$ and $0 < p < 1$; the replacement process $z_t$ is not necessarily independent of $r_t$.
- **(Innovation)** $r_t = \phi r_{t-1} + e_t$ and the i.i.d. innovation process $e_t$, with zero mean function and finite constant variance function, is assumed to have a fat-tailed univariate distribution for each $t$.

**Remark 3** In Sect. 4 we handle a substitute of replacement outlier configuration, simply by considering a patch of additive outliers of equal magnitude. □

We recall our previous assumption that $r_t$ is at least weak stationary with zero mean and finite variance. It seems that the only highly specialized form of outlier occur in the third case with the AR(1) model of uncontaminated returns. But see Sect. 4 for further distributional assumptions needed to simulate the input data in our procedure for all the possible outlier configurations. With these models in mind we are able to get finite samples $r_j, \hat{r}_j$ and $z_j$ of length $T$, accounting for all possible configuration of the contaminating sample $z_j$ (having $m$ outliers) whenever $j$ ranges over $\{1, \ldots, L\}$. A summary of the entire procedure described above is displayed in Fig. 1.

For example, the aforementioned procedure yields the following values of discrepancy in the case of SR's estimator.

These results are obtained by taking the daily return series of DAX30 Index from March 13 2018 to February 25 2019 (250 data points), and then considering five sub-samples of $T = 50$ returns in order to compute four biases without contamination as well as four biases with just $m = 1$ additive outlier of magnitude $\zeta = 0.50$ per sub-sample. Specifically, we estimate once SR with a fixed uncontaminated sub-sample (benchmark sample) and we recompute the estimates for the remaining four sub-samples, thus collecting $L = 4$ biases. We repeat this with the same benchmark uncontaminated sample, when the four sub-samples are contaminated. We apply the procedure twice because a comparison between the uncontaminated estimation and
the contaminated one enforces the understanding of the estimator’s robustness to outliers. Thus, we eventually get two badness sets each containing four normalized biases, but only the latter is given by \( \| \tilde{\theta} \| = \{ c_1, c_2, c_3, c_4 \} \) as defined above.

**PERF = SR.** To gain much more intuition about the sensitivity of SR’s estimator Fig. 2 displays the biases’ values \( c_j \) mapped onto the unit interval; the four biases on Fig. 2-top are computed without outlier. It is quite clear how the contamination with just one additive outlier (always the same \( z \) where all the components are zero but one which is \( \zeta = 0.50 \)) affects the behavior of SR, since it is forced to assume nearly the same values (viz. the four normalized biases \( c_j \) are tied each other within the unit interval see Fig. 2-bottom). In fact SR is very sensitive to outliers’ contamination (its original BP is zero). In the next section we apply this procedure to evaluate dis-BP for all the performance measures considered in this paper using Monte Carlo simulation.

### 4 Numerical comparison of relative robustness

The iterative procedure described in Sect. 3.2 and based on our concept of dis-BP (see Definition (8)) can be implemented via historical simulation, as we did in the numerical example at the end of Sect. 4, or via Monte Carlo simulation. In this section our aim is to run a controlled experiment and get the values of dis-BP \( \varepsilon_T(\tilde{\theta}, Z^*) \) given by \( D_L(\tilde{\theta}) \) for the eleven performance measures listed in Sect. 2.1, when additive and innovation outlier configurations are taken into account. We do not consider replacement outliers, instead we handle patchy additive outliers whose magnitude can accommodate replacement contamination. Definition (5) only provides a trade-off between robust and not robust indices and indeed none of the eleven performance indices is robust according to it, see Appendix B for detailed computations. Nonetheless, we claim that some differences among the performance measures’ robustness can be found by analyzing their sensitivity to outliers under alternative configurations. Henceforth, we employ Monte Carlo simulation to sample different return series, both uncontaminated and contaminated, assuming the AR(1) and the GARCH(1,1) models.

**Remark 4** Fixed values of the autocorrelation parameter \( \phi \) in the zero-mean AR(1) process are intended to provide an uncontaminated model for the underlying return series. It is well known that AR(1) models are not robust, in the sense that least squares estimates of \( \phi \) are sensitive to additive and replacement outliers. This lack of robustness holds true also in the case of AR(1) model with non-zero conditional mean, \( r_t = \mu + \phi r_{t-1} + e_t \). Maximum likelihood estimates are affected by the same deficiency. The situation is not different when GARCH(1,1) processes are used to model conditional variance (volatility). In all such cases, outliers typically result in larger AR(1) and smaller GARCH(1,1) estimates and deliver more variable volatility forecast. See Zhao et al. (2021) for a recent analysis of robust GARCH models.

\[ \blacksquare \]
4.1 Monte Carlo simulation with AR(1)

In this section we consider the AR(1) model

\[ r_t = \phi r_{t-1} + e_t, \quad \text{with i.i.d.} \quad e_t \sim N(0, \sigma^2) \quad (9) \]

for fixed values \( \phi \) of the autocorrelation coefficient. This model setting represents absence of outliers contamination, since estimation of the \( \phi \) parameter makes AR(1) not robust. Note that without loss of generality we assume a zero unconditional mean. To simulate returns \textit{without} outliers, the innovation component \( e_t \) remains
| $\zeta$ | SR   | TR   | YGR  | CR   | MADR | GLR  | OR   | V@RR | AV@RR | JA     | MRAR  |
|--------|------|------|------|------|------|------|------|------|-------|--------|-------|
| -0.8   | 0.9833 | 0.9990 | 0.9618 | 0.9894 | 0.9727 | 0.9480 | 0.9480 | 0.9861 | 0.9908 | 0.9993 | 0.9769 |
| -0.625 | 0.9825 | 0.9989 | 0.9652 | 0.9899 | 0.9740 | 0.9538 | 0.9538 | 0.9870 | 0.9908 | 0.9994 | 0.9944 |
| -0.45  | 0.9802 | 0.9984 | 0.9630 | 0.9896 | 0.9732 | 0.9503 | 0.9503 | 0.9872 | 0.9902 | 0.9991 | 0.9945 |
| -0.275 | 0.9810 | 0.9987 | 0.9661 | 0.9907 | 0.9755 | 0.9537 | 0.9537 | 0.9883 | 0.9908 | 0.9992 | 0.9957 |
| -0.1   | 0.9823 | 0.9986 | 0.9686 | 0.9914 | 0.9778 | 0.9566 | 0.9566 | 0.9895 | 0.9916 | 0.9991 | 0.9962 |
| 0.1    | 0.9829 | 0.9985 | 0.9697 | 0.9916 | 0.9786 | 0.9581 | 0.9581 | 0.9897 | 0.9916 | 0.9992 | 0.9963 |
| 0.275  | 0.9843 | 0.9987 | 0.9718 | 0.9922 | 0.9800 | 0.9606 | 0.9606 | 0.9905 | 0.9923 | 0.9991 | 0.9963 |
| 0.45   | 0.9855 | 0.9990 | 0.9745 | 0.9924 | 0.9813 | 0.9648 | 0.9648 | 0.9908 | 0.9926 | 0.9992 | 0.9962 |
| 0.625  | 0.9877 | 0.9993 | 0.9774 | 0.9934 | 0.9834 | 0.9688 | 0.9688 | 0.9919 | 0.9935 | 0.9993 | 0.9966 |
| 0.8    | 0.9889 | 0.9993 | 0.9789 | 0.9938 | 0.9845 | 0.9708 | 0.9708 | 0.9923 | 0.9937 | 0.9993 | 0.9966 |
A hybrid approach to the discrepancy in financial performance's

Gaussian as above. To get a contaminated model, we first corrupt the original simulated values $r_t$ with additive outliers (we set $j = 500$ from now on) with magnitude $\zeta = \{-0.8, -0.625, -0.45, -0.275, -0.1, 0.1, 0.275, 0.45, 0.625, 0.8\}$.

Table 1 contains the results, in terms of discrepancy, for the case of a single additive outliers and different outlier magnitudes. It is worth noting that the GLR and the OR always achieve the lower discrepancy, while the CAPM performance measures achieve the higher. This is confirmed by a graphical inspection of Fig. 3, which displays the pattern of dis-BP given by simulated values of $D_L(\hat{\theta})$ against the chosen additive outlier $\zeta$’s values. Recall that the lower is the discrepancy, the less sensitive to data contamination is the corresponding performance measure. Such behavior is also emphasized by Fig. 4, displaying the corresponding empirical densities of biases $b_j$ whose simulated values are mapped into the open unit interval $(0, 1)$: the densities of the GLR and OR biases are the most sparse, i.e., they are more uniformly distributed (lower discrepancy vs. higher robustness).

We repeat the previous analysis for patchy outliers and innovation outliers. Concerning the former type, for a patch of 3 consecutive outliers with the same magnitude $\zeta$ the results are listed in Table 2. Each row of Table 2 is computed by using the same outlier magnitude $\zeta$ repeated 3 times. A graphical inspection of Fig. 5 confirms the findings for additive outliers: the GLR and the OR achieve the lower discrepancy.
For innovation outliers, we assume a non-Gaussian specification for $e_t$ in the AR(1) model of returns’ DGP: a more skewed and fat-tailed Student-$t$ random variable $e_t \sim T(0, \sigma, d)$ with different degrees of freedom $d$. Notice that the ranking in terms of discrepancy does not change among the different degrees of freedom. Again, the GLR and OR yield the lower discrepancy, while the CAPM-based indices TR and JA have higher discrepancy (Table 3). The ranking of all PERF based on their dis-BP is not affected by alternative DGP specifications (Figs. 6 and 7).

Repeating the iterative procedure for other values of the autocorrelation coefficient, $\phi = -0.75, -0.50, -0.25, 0.25, 0.75$, entails results consistent with those obtained for $\phi = 0.5$.\(^\text{24}\)

\(^{24}\) We also run the simulation in the case of a patch of outliers with alternating sign.
Table 2  Simulated values of dis-BP for patchy outliers $z_j$ with different magnitude $\zeta$. The vector $z_j$ contains a patch of 3 consecutive additive outliers of the same magnitude per row. All the values are calculated by fixing the autocorrelation coefficient $\phi = 0.5$ in the AR(1)

| $\zeta$ | SR    | TR   | YGR  | CR   | MADR | GLR  | OR   | V@RR | AV@RR | JA   | MRAR |
|--------|-------|------|------|------|------|------|------|------|-------|------|------|
| −0.800 | 0.9793| 0.9991| 0.9364| 0.9697| 0.9517| 0.9245| 0.9245| 0.9754| 0.9876| 0.9992| 0.9553|
| −0.625 | 0.9800| 0.9987| 0.9445| 0.9752| 0.9590| 0.9335| 0.9335| 0.9803| 0.9893| 0.9989| 0.9832|
| −0.450 | 0.9744| 0.9981| 0.9449| 0.9753| 0.9598| 0.9295| 0.9295| 0.9788| 0.9858| 0.9987| 0.9873|
| −0.275 | 0.9700| 0.9975| 0.9431| 0.9774| 0.9592| 0.9148| 0.9148| 0.9793| 0.9853| 0.9983| 0.9926|
| −0.100 | 0.9590| 0.9966| 0.9330| 0.9697| 0.9501| 0.9048| 0.9048| 0.9723| 0.9790| 0.9981| 0.9908|
| 0.100  | 0.9572| 0.9958| 0.9292| 0.9680| 0.9474| 0.8993| 0.8993| 0.9704| 0.9759| 0.9976| 0.9895|
| 0.275  | 0.9589| 0.9967| 0.9261| 0.9686| 0.9456| 0.8905| 0.8905| 0.9711| 0.9767| 0.9983| 0.9907|
| 0.450  | 0.9591| 0.9969| 0.9157| 0.9645| 0.9386| 0.8778| 0.8778| 0.9661| 0.9733| 0.9984| 0.9896|
| 0.625  | 0.9615| 0.9979| 0.9160| 0.9609| 0.9374| 0.8743| 0.8743| 0.9636| 0.9709| 0.9986| 0.9893|
| 0.800  | 0.9686| 0.9984| 0.9166| 0.9624| 0.9383| 0.8694| 0.8694| 0.9647| 0.9708| 0.9988| 0.9884|
Fig. 5 Simulated values of $D_L(l(\hat{\theta}))$ for indices PERF 1–11 and different patch outliers magnitudes $\zeta$ contaminating the AR(1) model with $\phi = 0.5$

Table 3 Simulated values of dis-BP for innovation outliers with a t-student distribution given by i.i.d. $e_t \sim T(0, \sigma, d)$. All the values are calculated by fixing $\phi = 0.5$ in the AR(1)

| $d$ | SR  | TR  | YGR | CR  | MADR | GLR | OR  | V@RR | AV@RR | JA  | MRAR |
|-----|-----|-----|-----|-----|------|-----|-----|------|-------|-----|------|
| 3   | 0.9621 | 0.9965 | 0.9282 | 0.9012 | 0.9458 | 0.8959 | 0.8959 | 0.9721 | 0.9807 | 0.9977 | 0.9897 |
| 5   | 0.9625 | 0.9966 | 0.9362 | 0.9123 | 0.9525 | 0.9040 | 0.9040 | 0.9743 | 0.9793 | 0.9976 | 0.9907 |
| 7   | 0.9611 | 0.9960 | 0.9351 | 0.9095 | 0.9515 | 0.9071 | 0.9071 | 0.9740 | 0.9792 | 0.9977 | 0.9902 |
| 9   | 0.9639 | 0.9967 | 0.9378 | 0.9162 | 0.9546 | 0.9080 | 0.9080 | 0.9766 | 0.9813 | 0.9974 | 0.9907 |

Fig. 6 Simulated values of $D_L(l(\hat{\theta}))$ for indices PERF 1–11 and innovation outliers with different degrees of freedom $d$ when $e_t \sim T(0, \sigma, d)$ in the AR(1) model with $\phi = 0.5$
Table 4  Simulated values of dis-BP for additive outliers $z_j$ with different magnitude $\zeta$. All the values are calculated by fixing the autocorrelation coefficient $\phi = 0.5$ in the AR(1) model.

| $\zeta$ | SR   | TR   | YGR  | CR   | MADR | GLR  | OR   | V@RR | AV@RR | JA   | MRAR |
|---------|------|------|------|------|------|------|------|------|-------|------|------|
| -0.800  | 0.9656 | 0.9544 | 0.9367 | 0.9726 | 0.9526 | 0.9085 | 0.9085 | 0.9746 | 0.9824 | 0.9368 | 0.8854 |
| -0.625  | 0.9672 | 0.9523 | 0.9453 | 0.9746 | 0.9588 | 0.9205 | 0.9205 | 0.9762 | 0.9830 | 0.9361 | 0.9669 |
| -0.450  | 0.9667 | 0.9443 | 0.9393 | 0.9722 | 0.9545 | 0.9157 | 0.9157 | 0.9791 | 0.9852 | 0.9351 | 0.9653 |
| -0.275  | 0.9619 | 0.9505 | 0.9343 | 0.9707 | 0.9513 | 0.9091 | 0.9091 | 0.9723 | 0.9806 | 0.9387 | 0.9706 |
| -0.100  | 0.9681 | 0.9499 | 0.9447 | 0.9769 | 0.9594 | 0.9174 | 0.9174 | 0.9783 | 0.9827 | 0.9430 | 0.9731 |
| 0.100   | 0.9659 | 0.9478 | 0.9401 | 0.9750 | 0.9560 | 0.9165 | 0.9165 | 0.9769 | 0.9829 | 0.9399 | 0.9743 |
| 0.275   | 0.9594 | 0.9458 | 0.9328 | 0.9680 | 0.9492 | 0.9031 | 0.9031 | 0.9721 | 0.9792 | 0.9456 | 0.9669 |
| 0.450   | 0.9689 | 0.9547 | 0.9439 | 0.9765 | 0.9591 | 0.9132 | 0.9132 | 0.9793 | 0.9842 | 0.9453 | 0.9745 |
| 0.625   | 0.9649 | 0.9551 | 0.9385 | 0.9740 | 0.9540 | 0.9084 | 0.9084 | 0.9757 | 0.9818 | 0.9418 | 0.9709 |
| 0.800   | 0.9654 | 0.9576 | 0.9376 | 0.9724 | 0.9542 | 0.9016 | 0.9016 | 0.9747 | 0.9816 | 0.9491 | 0.9696 |
4.2 Monte Carlo simulation with GARCH(1,1)

In this section we consider the GARCH(1,1) model with conditional mean given by Eq. (9) and conditional variance (volatility)

\[ \sigma_t^2 = 0.8220\sigma_{t-1}^2 + 0.1352e_t^2, \]  

where \( e_t = \sigma_t a_t \) and i.i.d. \( a_t \sim N(0, 1) \); the parameters’ values are based on (Tsay 2008, Example 3.3, p. 95). As in the previous section we let the autocorrelation coefficient in Eq. (9) for the conditional mean be equal to \( \phi = 0.5 \). Table 4 shows the simulated dis-BP values results for a single additive outlier and different outlier magnitudes. The numerical results are again in favor of the \( \text{AV}_@\text{OR} \) and \( \text{VAR}_@\text{OR} \), that achieve the lowest discrepancy, see also Fig. 8. Differently from the simple AR(1) model we note the \( \text{AV}_@\text{AR} \) achieves the highest level of discrepancy in this setting. Figure 8 compares the empirical densities \(^{25}\) of the dis-BP’s values computed for all performance indices, confirming the good behavior of the GLR and OR in terms of robustness (most scattered densities).

The results for the patchy outliers simulations (3 consecutive outliers with the same magnitude \( \zeta \)) are detailed in Table 5, see also Fig. 9. Likewise in the case of the simple AR(1) model the GLR and OR have the lowest discrepancy (more robustness) while the AV@RR achieves the worst.

For innovation outliers we assumed a skewed and fat-tailed Student-\( t \) random variable \( a_t \sim T(0, 1, d) \), with different degrees of freedom \( d \). The results are reported in Table 6 and in Fig. 10. By looking at the former we can see that the GLR and OR have the lowest discrepancy (higher robustness) for all \( d \)’s values but for \( d = 3 \), when the MRAR achieves the lowest discrepancy.

\(^{25}\) The densities are mapped into the open unit interval, as we did in the previous section.
A hybrid approach to the discrepancy in financial performance's…

5 Conclusion

Robust optimal solutions to portfolio optimization problems are mostly based on non-stochastic sets of uncertainty parameters. Some recent literature is devoted to reinforce this approach to asset allocation, by using robust version of the objective’s parameters or mixed procedures. Specializing to optimal return-to-risk management, indices of financial performance require statistical estimation and then a source of sensitivity to outliers in the dataset used comes into play. This can affect to some extent the ranking of investment funds. Measures of statistical robustness might be a sensible addition to the theory and practice of fund management, especially when the link to robust portfolio optimization is taken into account. We provide a

![Empirical densities of biases $b_j$ mapped into (0, 1) for indices PERF 1–11 and the GARCH(1,1) model with $\phi = 0.5$](image-url)
Table 5  Simulated values of dis-BP for patchy outliers $z_j$ with different magnitude $\zeta$. The vector $z_j$ contains a patch of 3 consecutive additive outliers of the same magnitude per row. All the values are calculated by fixing the autocorrelation coefficient $\phi = 0.5$ in the conditional mean of the GARCH(1,1) model.

| $\zeta$ | SR  | TR  | YGR | CR  | MADR | GLR | OR  | V@RR | AV@RR | JA   | MRAR |
|--------|-----|-----|-----|-----|------|-----|-----|------|-------|------|------|
| −0.800 | 0.9659 | 0.9624 | 0.9407 | 0.9743 | 0.9548 | 0.9157 | 0.9157 | 0.9765 | 0.9840 | 0.9488 | 0.8086 |
| −0.625 | 0.9688 | 0.9561 | 0.9417 | 0.9744 | 0.9574 | 0.9138 | 0.9138 | 0.9793 | 0.9856 | 0.9366 | 0.9535 |
| −0.450 | 0.9598 | 0.9599 | 0.9332 | 0.9699 | 0.9492 | 0.9033 | 0.9033 | 0.9722 | 0.9800 | 0.9423 | 0.9695 |
| −0.275 | 0.9636 | 0.9571 | 0.9358 | 0.9716 | 0.9528 | 0.9036 | 0.9036 | 0.9758 | 0.9817 | 0.9291 | 0.9717 |
| −0.100 | 0.9542 | 0.9519 | 0.9223 | 0.9660 | 0.9415 | 0.8893 | 0.8893 | 0.9694 | 0.9769 | 0.9344 | 0.9663 |
| 0.100  | 0.9611 | 0.9511 | 0.9316 | 0.9692 | 0.9498 | 0.8968 | 0.8968 | 0.9737 | 0.9799 | 0.9333 | 0.9704 |
| 0.275  | 0.9613 | 0.9514 | 0.9320 | 0.9687 | 0.9501 | 0.8976 | 0.8976 | 0.9737 | 0.9800 | 0.9384 | 0.9696 |
| 0.450  | 0.9569 | 0.9447 | 0.9248 | 0.9667 | 0.9434 | 0.8898 | 0.8898 | 0.9699 | 0.9778 | 0.9332 | 0.9655 |
| 0.625  | 0.9545 | 0.9506 | 0.9210 | 0.9639 | 0.9407 | 0.8821 | 0.8821 | 0.9679 | 0.9760 | 0.9412 | 0.9645 |
| 0.800  | 0.9578 | 0.9470 | 0.9237 | 0.9681 | 0.9442 | 0.8839 | 0.8839 | 0.9711 | 0.9772 | 0.9473 | 0.9687 |
Fig. 9 Simulated values of $D_l(l(\hat{\theta}))$ for indices PERF 1–11, for different patch outliers magnitudes $\zeta$ contaminating the GARCH(1,1) model with $\phi = 0.5$

Table 6 Simulated values of dis-BP for innovation outliers with a student-t distribution given by i.i.d. $e_t \sim T(0, \sigma, d)$ and $\phi = 0.5$ in the GARCH(1,1)

| $d$ | SR    | TR    | YGR   | CR    | MADR  | GLR   | OR    | V@RR  | AV@RR | JA    | MRAR  |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 3   | 0.9643| 0.9803| 0.9301| 0.9007| 0.9446| 0.9020| 0.9693| 0.9798| 0.9826| 0.8706|
| 5   | 0.9607| 0.9795| 0.9332| 0.9064| 0.9492| 0.9043| 0.9722| 0.9796| 0.9832| 0.9199|
| 7   | 0.9563| 0.9799| 0.9224| 0.8955| 0.9405| 0.8922| 0.9698| 0.9792| 0.9827| 0.9361|
| 9   | 0.9628| 0.9782| 0.9360| 0.9108| 0.9524| 0.9090| 0.9750| 0.9814| 0.9821| 0.9540|

Fig. 10 Simulated values of $D_l(l(\hat{\theta}))$ for indices PERF 1–11 and innovation outliers with different degrees of freedom $d$ when $e_t \sim T(0, \sigma, d)$ in the GARCH(1,1) model with $\phi = 0.5$
methodology aimed at finding the intrinsic degree of statistical robustness for a set of eleven performance indices. The quantitative characterization of robustness we are interested in is based on the concept of breakdown point of an estimator, because it enables us to deal with not necessarily i.i.d. observations of historical returns. Our methodology relies on the reasonable definition of breakdown point à la Genton and Lucas (2003). First, we note that a simple application of this device does not provide any information about the different degree of robustness among the chosen performance measures to alternative outliers configurations. Secondly, we instead suggest a new finite-sample definition of breakdown point to be mixed with the well-known notion of discrepancy for finite sets of biases measuring estimation errors in the involved parameters. This results in an iterative procedure for the numerical evaluation of the performance indices’ robustness. We implement it through Monte Carlo simulation using AR(1) and GARCH(1,1) models for the ‘clean’ return series, confirming that one can find differences in the indices’ sensitivity to outliers contaminating the return series. Future research agenda will include larger sets of performance indices and an in-depth analysis of the interaction between our methodology and portfolio weights’ optimality.

Appendix

This Appendix reviews some elementary concepts from robust statistics as applied to the current financial context. As we have seen in Examples 1, 2 and in Sect. 2, the eleven performance indices considered in this article are defined in term of the main parameters characterizing the cumulative distribution \( F_r(x) \) of the fund return. In a wide sense, robustness of these parameters can be defined with respect to changes in \( F_r(x) \) whenever one identifies the parameters’ estimators as functions of the sampled historical returns. Denoting \( \mathcal{Y} \) the set of all relevant \( F_r(x) \)’s parameters and writing \( \hat{\theta} \) for any estimate of them, we have a function \( \hat{\theta} := \theta(r_1, \ldots, r_T) \) of \( T \in \mathbb{N} \) historical returns sampled from a DGP with cumulative distribution \( F_r(x) \). Thus the estimate is an evaluation of the function \( \theta \) at the chosen sample and this can be equivalently be written as \( \theta(F_{r,T}(x)) \), where \( F_{r,T}(x) := T^{-1} \sum_{t=1}^{T} I_{\{r_t \leq x\}} \) is the empirical distribution function representing the sample counterpart of the population’s \( F_r(x) \). The indicator function \( I_{\{r_t \leq x\}} \) is equal to 1 if \( r_t \leq x \) under the appropriate scenario, and zero otherwise. Here with a little abuse of notation we identify \( (r_1, \ldots, r_T) \) with a set of \( T \) random variables, and in this case \( \theta(r_1, \ldots, r_T) \) is a random variable too (depending on the sample). Whence, there is a correspondence between a cumulative distribution and sample values. Asymptotically, as \( T \) gets large enough we have \( F_{r,T}(x) \to F_r(x) \) and accordingly the estimate \( \hat{\theta} \) converges to the population counterpart which is an element of \( \mathcal{Y} \). The first convergence is almost surely (see Huber and Ronchetti (2009), (Ch 2)) provided that the random returns \( r_t \) are i.i.d. (Strong Law of Large Number), and it is also uniform in \( x \in \mathbb{R} \) (Glivenko-Cantelli Theorem). The second convergence is in probability (and then almost surely). Now, the Hampel qualitative definition of robustness of \( \hat{\theta} \), for the given cumulative distribution \( F_r(x) \), considers how a small deviations from the ideal model \( F_r(x) \) result in small
deviations from the probability distribution of the estimator, or equivalently its cumulative distribution $F_{\hat{\theta}}(x)$. These deviations rely on defining appropriate metrics in the space of the corresponding cumulative/probability distributions. This also implies that outliers acted on the $r_i$'s realizations.

**Remark 5** All the parameters used in the current article to define a given performance measure are statistical functionals, i.e., functions of the population cumulative distribution, $\theta(F_r)$. For an i.i.d. sample from $F_r$, the law of large number implies $\text{plim}_{T \to \infty} \theta(F_{r,T}) = \theta(F_r)$, where plim stands for convergence in probability whenever $T \to \infty$. Intuitively, $\theta(F_{r,T})$ is robust if modifying a small proportion of the historical returns in the finite sample causes only a small change in the estimate, which is a sort of continuity of $\theta(\cdot)$. Thus, given a set of cumulative distribution functions (alternative DGP models, comprising degenerate distributions) one can endow it with a metric measuring the $d(F_r, G)$ distance between any two elements, one being the uncontaminated model of asset return $F_r$ and the other $G$ its contaminated version. If a modeler changes a small proportion of historical observations of returns and/or slightly modifies all of observations, then $d(F_{r,T}, F_{r,T})$ is negligible. Here $F_{r,T}$ is the empirical distribution function evaluated at the corrupted sample, that drawn from the cumulative distribution $F_r$ contaminated with outliers. Now $\theta$ is qualitative robust at $F_r$ if the statistical functional $\theta$ is continuous in the given metric, i.e., a distance $d(F_r, F_{r})$ small enough implies $|\theta(F_r) - \theta(F_{r})|$ is small too. In fact, robust estimates are also consistent since

$$\theta(F_r) = \theta \left( \text{plim}_{T \to \infty} \mathbb{F}_{r,T} \right) = \text{plim}_{T \to \infty} \theta(F_{r,T}) = \text{plim}_{T \to \infty} \hat{\theta},$$

meaning that $\theta(F_r)$ is the population parameter lying in $\mathfrak{F}$, for example the ratio of the mean $\mu_r$ to the standard deviation $\sigma_r$ (viz. Sharpe ratio).

The qualitative notion of robustness can be made quantitative using the influence function approach (see Huber and Ronchetti (2009), (Ch 1 and Ch 2) and Rossello (2015)) based on the notion of gross error given by the convex combination $(1 - \lambda)F_r(x) + \lambda F_z(x)$ of the population’s cumulative distribution $F_r(x)$ and the cumulative distribution $F_z(x)$ of the degenerate random variable $z$ which takes a single value with probability 1; here $0 \leq \lambda \leq 1$. The related influence function is defined to measure the quantitative effect of one single outlier $z$ on the estimator $\hat{\theta}$ for the given distribution. Another quantitative measure of robustness is given by the breakdown point $\varepsilon_{\hat{\theta}}(\hat{\theta}, \mathcal{Z})$ of an estimator $\hat{\theta}$, see Sect. 3.1 for the finite-sample definition of Genton and Lucas (2003) we use in our article. The set $\mathcal{Z}$ of admitted outliers configurations can drive $\hat{\theta}$ to be uninformative.

**Remark 6** A theoretical definition of breakdown point requires the notion of a neighborhood $\mathcal{N}_\delta(F_r) := \{ G \mid d(F_r, G) < \delta \}$ of the cumulative distribution $F_r$. First, one defines the maximum bias $b_\delta := \sup \{|\theta(F_r) - \theta(G)| \mid G \in \mathcal{N}_\delta(F_r)\}$ of the parameter $\theta$ at $F_r$. Since all the metrics considered are bounded, the breakdown point can
be defined for all admissible DGP as \( \varepsilon := \inf \{ \delta \mid b_{\delta} < b \} \), where all possible biases are taken into account, \( b = \sup \{ \| \theta(F_r) - \theta(G) \| \mid \text{all } G \} \). Our definition \( \varepsilon_T(\hat{\theta}, Z) \) is the finite-sample counterpart for a sample size \( T \) and the explicit appearance of outliers as elements of \( Z \), replacing the theoretical corrupted models \( G \). The influence function can be considered as the Fréchet derivative of an estimate \( \hat{\theta} \). Rewriting the gross error as \( F_r - \lambda(F_z - F_r) \), the influence function can be defined as

\[
\mathfrak{i}F_\theta(\varepsilon, F_r) := \lim_{\lambda \to 0} \frac{\theta(F_r - \lambda(F_z - F_r)) - \theta(F_r)}{\lambda},
\]

which is nothing but the derivative of the statistical functional \( \theta(\cdot) \) in the direction of the ‘amount of contamination’ \( F_z - F_r \).

\[ \square \]

Appendix

In this Appendix we list the elementary algebraic manipulations yielding BP of performance indices 2–11 from Sect. 2 (except SR), using Definition (5) in Sect. 3.1. We only consider one additive outlier in the contaminating sample \( z \). All the biases \( |\text{PERF}(\hat{r} + z) - \hat{\theta}| \) are computed for two uncontaminated samples \( r, \bar{r} \), where \( \hat{\theta} = \text{PERF}(r) \).

\text{PERF} = \text{TR} \quad \text{We consider a cross sectional sample based on a } T \times N \text{ matrix, where } r = (r_{i1}, \ldots, r_{iN}) \text{ is the sub-sample of the } i\text{th security’s return in the market portfolio } M, \text{ for } i = 1, \ldots, N \text{ and a total of } N \in \mathbb{N} \text{ securities. The estimator is } \hat{\text{TR}} = \frac{1}{T} \sum_{i=1}^{T} \bar{r}_i M_i - \bar{M}, \text{ where } \bar{r}_i \text{ is the } i\text{th’s security mean return and } M_i = \sum_{i=1}^{N} w_{ij} r_{ij}, \text{ with } w_i \text{ being the weight of the asset return } r_{ij}; \text{ here } \bar{M} = \frac{1}{T} \sum_{i=1}^{T} M_i \text{ denotes the market portfolio’s mean return. When the sample returns of security } i\text{th is corrupted, then } \text{TR}(\hat{r} + z) \to \pm \infty \text{ for } \zeta \to \pm \infty, \text{ respectively. Whence the badness set is } \{ +\infty \} \text{ and the BP is zero.}

\text{PERF} = \text{YGR} \quad \text{The estimator is } \hat{\text{YGR}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{t=1}^{T} |r_t - r_{\bar{t}}| \text{ where the denominator is } \hat{G}_i, \text{ the estimator of the Gini measure. Thus, } \zeta \to \pm \infty \text{ does imply } \text{YGR}(\hat{r} + z) \to \pm (T - 1) \text{ and the badness set is } \{ |\text{YGR} + 1 - T| | \text{YGR} + T - 1 \} \text{ so that the BP is zero.}

\text{PERF} = \text{CR} \quad \text{We set } t_k - t_{k-1} = 1 \text{ and for } \ell = n \cdot v \text{ we pick an overall sample } (u_1, u_2, \ldots, u_{\ell}). \text{ Thus, with the first } v \text{ observations we compute the first sample maximum drawdown } \hat{D}_1, \text{ then we repeat this step } T \text{ times to get the estimator } \frac{1}{T} \sum_{t=1}^{T} \hat{D}_t \text{ of the population maximum drawdown. The corresponding estimator of the Calmar ratio is } \hat{\text{CR}} = \frac{1}{T} \sum_{t=1}^{T} \hat{D}_t \text{, where the average return in the numerator is computed using the sample } r = (r_1, \ldots, r_T) \text{ representing}
random returns at the horizon, such that 
\[ r_1 = u_1, r_2 = u_2, \ldots, r_T = u_T. \]
For \( t = i \cdot n \), when \( \zeta \to +\infty \) and the corrupted observation \( u_t + \zeta \) is such that \( u_t \neq r_t \), then either \( \text{CR}(\hat{r} + z) \to 0 \) or \( \text{CR}(\hat{r} + z) \to \frac{T}{u_t} \) whenever \( u_t + \zeta \) is placed at the \( t \)th date of the corresponding sample drawdown. When \( u_t = r_t \), either \( \text{CR}(\hat{r} + z) \to 1 \) or \( \text{CR}(\hat{r} + z) \to +\infty \). In the case \( \zeta \to -\infty \) we have other four possible limits: \( \text{CR}(\hat{r} + z) \) converges to \( \frac{T}{u_\beta u_\alpha u_t} \), with \( \beta^* \) and \( t^* \) not necessarily equal to \( i \) and \( t \) respectively; \( \text{CR}(\hat{r} + z) \to 0 \), or \( \text{CR}(\hat{r} + z) \to -\infty \), or \( \text{CR}(\hat{r} + z) \to -1 \).

Thus, the badness set is \( \{ |\text{CR} - 1|, |\text{CR} + 1|, |\text{CR}|, +\infty \} \) and the BP is zero.

**PERF = MADR**

The estimator is \( \hat{MADR} = \frac{T}{\sum_{t=1}^{T} |r_t - r|} \) and \( \zeta \to +\infty \) does imply \( \text{MADR}(\hat{r} + z) \to \pm \frac{T}{2(T - 1)} \), then the badness set is \( \{ |\hat{MADR} - \frac{T}{2(T - 1)}|, |\hat{MADR} + \frac{T}{2(T - 1)}| \} \) and the BP is zero.

**PERF = GLR**

The estimator is \( \hat{GLR} = \frac{T}{\sum_{t=1}^{T} \min(0, r_t)} \).

If \( \zeta \to +\infty \) then \( \text{GLR}(\hat{r} + z) \to +\infty \), while if \( \zeta \to -\infty \) then \( \text{GLR}(\hat{r} + z) \to -1 \).

The badness set is \( \{ |\hat{GLR} + 1|, +\infty \} \) implying a zero BP.

**PERF = OR**

The estimator is \( \hat{OR} = \frac{T}{\sum_{t=1}^{T} \max(0, r_t)} \).

If \( \zeta \to +\infty \) then \( \text{OR}(\hat{r} + z) \to +\infty \), and for \( \zeta \to -\infty \) we have \( \text{OR}(\hat{r} + z) \to 0 \). The badness set is \( \{ +\infty, +\infty \} \) and the BP is zero.

**PERF = V@RR**

The estimator is \( \hat{V@RR} = -\frac{T}{\sum_{t=1}^{T} r_t(T_t + 1)} \), where \( (r_1, \ldots, r_{(T_t)}) \) denotes the ordered statistics and \( T_{\alpha} \) is \( T \alpha \) rounded to the nearest integer.

Given \( \alpha \in (0, 1) \), if \( \zeta \to +\infty \) then \( \text{V@RR}(\hat{r} + z) \to +\infty \), the badness set is \( \{ +\infty \} \) with a zero BP.

**PERF = AV@RR**

The estimator is \( \hat{AV@RR} = -\frac{T}{\sum_{t=1}^{T} r_t(\sum_{i=1}^{T} r_i + (T_{\alpha} + [T_{\alpha}]))} \).

We have \( \text{AV@RR}(\hat{r} + z) \to -\alpha \) for \( \zeta \to -\infty \) while for \( \zeta \to +\infty \) we have \( \text{AV@RR}(\hat{r} + z) \to +\infty \). The badness set is \( \{ |\hat{AV@RR} + \alpha|, +\infty \} \) and the BP is zero.

**PERF = JA**

The estimator is \( \hat{JA} = \hat{r} - \frac{1}{M} \sum_{t=1}^{T} r_t(M_t - TM) \), with the same notation as in the case of the TR. If we corrupt the sample returns of security \( i \)th in the market portfolio, then \( \text{JA}(\hat{r} + z) \to \pm\infty \) provided that \( \zeta \to \pm\infty \). The badness set is \( \{ +\infty \} \) and the BP is zero.

**PERF = MRAR**

The estimator is \( \hat{MRAR} = \left[ \frac{1}{T} \sum_{t=1}^{T} (1 + r_t)^{-2} \right]^{-1} - 1 \), where as usual we fix \( A = 2 \). If \( \zeta \to \pm\infty \) then \( \text{MRAR}(\hat{r} + z) \to +\infty \) and we get the badness set \( \{ +\infty \} \) implying a zero BP.
Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

Data availability The historical data used in the manuscript should be available on request.

Declarations

Conflict of interest The corresponding author states that there is no conflict of interest.

Code availability Custom pseudo-code (MATLAB-oriented) is used in the manuscript and should be available on request.

Informed consent The manuscript is not submitted to other journals.

Consent for publication The manuscript has not been published elsewhere.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Acerbi C (2002) Spectral measures of risk: a coherent representation of subjective risk aversion. J Bank Finance 26:1505–1518
AIMR: Performance Evaluation, Benchmarks, and Attribution Analysis. AIMR (CFA Institute), ICFA Continuing Education, First Edition (1995)
AIMR: Performance Presentation Standards Handbook (The AIMR-PPS Standards with Commentary and Interpretation). AIMR (CFA Institute), Second Edition (1997)
Amenc N, Le Sourd V (2003) Portfolio theory and performance analysis. Wiley, Hoboken
Ben-Tal A, Nemirovski A (1998) Robust convex optimization. Math Oper Res 23:769–805
Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertainty data. Math Program 88:411–424
Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization. Princeton University Press, Princeton
Bertsimas D, Brown DB (2009) Constructing uncertainty sets for robust linear optimization. Oper Res 57:1483–1495
Bertsimas D, Sim M (2004) The price of robustness. Oper Res 52:35–53
Bertsimas D, Sim M (2011) Theory and applications of robust optimization. SIAM Rev 53:464–501. https://doi.org/10.1080/1351847X.2021.1960404
Best MJ, Grauer RR (1991) On the sensitivity of mean-variance efficient portfolios to changes in asset weights: some analytical and computational results. Rev Financ Stud 4(2):315–342
Black F, Litterman R (1991) Asset allocation: combining investor views with market equilibrium. J Fixed Income 1(2):7–18
Bradrania R, Pirayesh Neghab D (2021) State-dependent asset allocation using neural networks. Eur J Finance. https://doi.org/10.1080/1351847X.2021.1960404
Caporin M, Jannin GM, Lisi F (2014) A survey on the four families of performance measures. J Econ Surv 28(5):917–942
A hybrid approach to the discrepancy in financial performance's…

Chan LKC, Karceski J, Lakonishok J (1999) On portfolio optimization: forecasting covariances and choosing the risk model. Rev Financ Stud 12(5):937–974
Chang T-J, Yang S-C, Chang K-J (2009) Portfolio optimization problems in different risk measures using genetic algorithm. Expert Syst Appl 36(7):10529–10537
Chopra VK, Ziemba WT (1993) The effect of errors in means, variances, and covariances on optimal portfolio choice. J Portfolio Manag Winter 19(2):6–11. https://doi.org/10.3905/jpm.1993.409440
Christopher JA, Cariño DR, Ferson WE (2009) Portfolio performance measurement and benchmarking. McGraw-Hill, New york
Cogneau P, Hüblner G (2015) The prediction of fund failure through performance diagnostics. J Bank Finance 50:224–241
Cont R, Deguest R, Scandolo G (2010) Robustness and sensitivity analysis of risk measurement procedures. Quant Finance 10(6):593–606
Davies PL, Gather U (2005) Breakdown and groups. Ann Stat 33(3):977–988
De Capitani L, Pasquazzi L (2015) Inference for performance measures for financial assets. Metron 73(1):73–98. https://doi.org/10.1007/s40300-014-0055-y
Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper Res 58(3):596–612
Deng G, Dulaney T, McCann C, Wang O (2013) Robust portfolio optimization with value-at-risk-adjusted sharpe ratios. J Asset Manag 14(5):293–305
Donnelly C, Embrechts P (2010) The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. ASTIN Bull J Int Actuar Assoc 40(01):1–33
Dowd K (2000) Adjusting for risk: an improved Sharpe ratio. Int Rev Econ Financ 9(3):560–577
El Gahoui L, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: a conic programming approach. Oper Res 51(4):543–556
Fabozzi FJ, Kolm PN, Pachamanova D, Focardi SM (2007) Robust portfolio optimization and management. Wiley, Hoboken
Fliege J, Werner R (2014) Robust multiobjective optimization and applications in portfolio optimization. Eur J Oper Res 234:422–433
Gabrel V, Murat C, Thiele A (2014) Recent advances in robust optimization. Eur J Oper Res 235:471–483
Genton MG, Lucas A (2003) Comprehensive definitions of breakdown points for independent and dependent observations. J R Stat Soc B 65(1):81–94
Ghatahrani A, Saif A, Ghasemi A (2022) Robust portfolio selection problems: a comprehensive review. Operational. Res Int J. https://doi.org/10.1007/s12351-022-00690-5
Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. Math Oper Res 28:1–38
Goel A, Sharma A, Mehr A (2017) Robust optimization of mixed CVaR starr ratio using copulas. J Comput Appl Math 347:62–83
Gregory C, Darby-Dowman K, Mitra G (2011) Robust optimization and portfolio selection: the cost of robustness. Eur J Oper Res 212:417–428
Huber PJ, Ronchetti EM (2009) Robust statistics. Wiley, Hoboken
Israelsen CL (2005) A refinement to the sharpe ratio and information ratio. J Asset Manag 5:423–427
Jagannathan R, Ma T (2003) Risk reduction in large portfolios: why imposing the wrong constraints helps. J Finance 58(4):1651–1683
Ji R, Lejeune MA, Prasad SY (2017) Properties, formulations, and algorithms for portfolio optimization using mean-Gini criteria. Ann Oper Res 248(1–2):305–343
Ji R, Lejeune MA, Fan Z (2022) Distributionally robust portfolio optimization with linearized STARR performance measure. Quant Finance. https://doi.org/10.1080/14697688.2021.1993623
Kaspos M, Christofides N, Rustem B (2014) Worst-case robust omega ratio. Eur J Oper Res 234(2):499–507
Krätschmer V, Schied A, Zhäle H (2014) Comparative and Qualitative Robustness for Law-Invariant Risk Measures. Finance Stochast 18:271–295. https://doi.org/10.1007/s00780-013-0225-4
Konno H, Yamazaki H (1991) Mean-absolute deviation portfolio optimization model and its application to Tokyo stock market. Manage Sci 37:519–531
Lauprete GJ, Samarow AM, Welsch RE (2002) Robust portfolio optimization. Metrika 55:139–149
León A, Navarro L, Nieto B (2019) Screening rules and portfolio performance. North Am J Econ Finance 48:642–662
Lo AW (2002) The statistics of sharpe ratios. Financ Anal J 58:36–52
Lu Z (2006) A new cone programming approach for robust portfolio selection. Department of Mathematics, Simon Fraser University, Burnaby, BC, Tech. Rep
Lu Z (2011) Robust portfolio selection based on a joint ellipsoidal uncertainty set. Opt Methods Softw 26(1):89–104
Mamatzakis E, Tsionas MG (2021) Testing for persistence in US mutual funds’ performance: a Bayesian dynamic panel model. Ann Oper Res 299(1):1203–1233
Maronna RA, Martin RD, Yohai VJ (2006) Robust Statistics. Wiley, Hoboken
Markowitz H (1952) Portfolio selection. J Financ 7:77–91
Momen O, Esfahanipour A, Seifi A (2020) A robust behavioral portfolio selection. Operational. Res Int J 20:427–446
Niederreiter H (1992) Random number generation and quasi-monte carlo methods. CBMS-NSF regional conference series in applied mathematics, SIAM
Pinar MÇ, Paç AB (2014) Mean semi-deviation from a target and robust portfolio choice under distribution and mean return ambiguity. J Comput Appl Math 259:394–405
Rachev ST, Stoyanov SV, Fabozzi FJ (2008) Advanced stochastic models, risk assessment, and portfolio optimization. Wiley, Hoboken
Rossello D (2015) Ranking of investment funds: acceptability versus robustness. Eur J Oper Res 245(3):828–836
Roy B (2010) Robustness in operational research and decision aiding: a multi-faceted issue. Eur J Oper Res 200:629–638
Scutellà M, Recchia R (2013) Robust portfolio asset allocation and risk measures. Ann Oper Res 204(1):145–169
Sharma A, Utz S, Mehra A (2017) Oemga-CVaR portfolio optimization and its worst case analysis. OR Spect 39(2):505–539
Tsay RT (2010) Analysis of financial time series. Wiley, Hoboken
Tong X, Wu F (2014) Robust reward-risk ratio optimization with application in allocation of generation asset. Optimization. https://doi.org/10.1080/02331934.2012.672419
Tütüncü RH, Koenig M (2004) Robust asset allocation. Ann Oper Res 132:157–187
Yitzhaki S (1983) On an extension of the Gini inequality index. Int Econ Rev 24(3):617–628
Zakamouline V, Koekebakker S (2009) Portfolio performance evaluation with generalized sharpe ratios: beyond the mean and variance. J Bank Finance 33(7):1242–1254
Zhao Q, Chen L, Wu J (2021) Robust and efficient estimation of GARCH models based on Hellinger distance. J Appl Stat. https://doi.org/10.1080/02664763.2021.1970120
Zhu S, Fukushima M (2009) Worst-case conditional value-at-risk with application to robust portfolio management. Oper Res 57(5):1155–1168

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