



A robust approach for inference on style analysis coefficients

Domenico Vistocco¹ 

Accepted: 19 March 2024 / Published online: 17 April 2024
© The Author(s) 2024

Abstract

Style analysis is an asset class factor model aiming at obtaining information on the internal allocation of a financial portfolio and at comparing portfolios with similar investment strategies. The aim of the paper is to investigate the use of quantile regression to draw inferences on style coefficients. In particular, we compare an approximation widely used from practitioners, the Lobosco–Di Bartolomeo solution, with robust estimators based on constrained median regression. The inference process exploits a rolling window procedure based on subsampling theory. Different sets of outliers have been simulated in order to show differences in the efficiency, in the coverage error and in the length of the resulting confidence intervals. The proposed solution shows better performance in presence of outliers in Y , in X , or in X and Y , in terms both of empirical coverage and of interval lengths, i.e. whereas the performance of the classical solution deteriorates.

Keywords Style analysis · Inference for style coefficients · Quantile Regression · Subsampling

1 Introduction

Style analysis, as originally proposed by Sharpe (1992), is a return based analysis aimed at decomposing the performance of a financial portfolio with respect to the contribution of different constituents. The constituents can be related to industrial sectors, geographical markets, financial markets, among others. Each constituent is represented by an index whose returns are available. The model regresses portfolio returns on constituents returns, in order to decompose the portfolio performance with respect to the different constituents used to describe the portfolio. The classical model is based on a least squares regression model, but different constraints can

✉ Domenico Vistocco
domenico.vistocco@unina.it

¹ Dipartimento di Scienze Politiche, Università degli Studi di Napoli Federico II, Via Leopoldo Rodinò, 22, Napoli, Italy

be imposed on the coefficients. Following Horst et al. (2004), we distinguish three types of style models:

- *Weak style analysis*: the coefficients are estimated using an unconstrained regression model,
- *Semi-strong style analysis*: the coefficients are imposed to be positive,
- *Strong style analysis*: the coefficients are imposed to be positive and to sum up to one.

The three types of style models are typically estimated as regression through the origin. The use of the double constraint (strong style analysis), and the absence of the intercept, lead to interpret the regression coefficients in terms of composition quotas. Therefore, style analysis is used to estimate the internal composition of the portfolio Conversano and Vistocco (2010). In the framework of classical regression, the estimated coefficients measure the sensitivity of portfolio expected returns to constituents returns. However, classical inferential procedures should be interpreted with caution, due to the imposition of inequality linear constraints Judge and Takayama (1966). Some general results are available for the normal linear regression model Geweke (1986), and a different bayesian procedure is formulated in Davis (1978). A commonly applied solution is the approximation proposed by Lobosco and Bartolomeo (1997), that is a second order Taylor series - based approximation of the confidence intervals of style weights. The proposed solution works well except when the parameters are on the boundaries, i.e. when one or more parameters are near 0 and/or when a parameter falls near 1. To overcome this drawback, Kim et al. (2005) propose two approximate solutions: the former based on the method of Andrews (1999), and the latter based on the bayesian method proposed by Geweke (1986). A further bayesian approach has been proposed by Christodoukalis (2005). Among practitioners, the solutions of Lobosco and Di Bartolomeo is commonly adopted, for its conceptual and computational simplicity, and its very good results in regular settings. And not forgetting, the wide dissemination among financial practitioners of the journal where Lobosco and Di Bartolomeo published their proposal.

Since the Lobosco and Di Bartolomeo solution is essentially based on a least squares estimation procedure, it is sensitive to outliers. In this paper we investigate the use of quantile regression Davino et al. (2013); Koenker (2005) to draw robust inferences on style coefficients. In particular we compare the approximation proposed by Lobosco and Di Bartolomeo with robust estimators based on constrained median regression. The proposal exploits a rolling window procedure based on the subsampling theory Politis et al. (1999). Different sets of outliers have been simulated both in constituents returns and in portfolio returns. The procedures are then compared with respect to efficiency, coverage percentage and length of the resulting confidence intervals.

The paper is organized as follows: Sect. 2 shortly introduces the classical Sharpe style model, along with the Lobosco–Di Bartolomeo approximation commonly used for drawing inferences on style coefficients. Section 3 describes the quantile regression approach to style analysis, highlighting the subsampling procedure used at inference purposes. The simulation scheme and the main results are discussed in

Sect. 4. Conclusions and discussion of further research developments to be explored are in Sect. 5.

2 Sharpe–style regression model

Style analysis model regresses portfolio returns on the returns of a variety of investment classes. Namely, the method identifies the portfolio style in the time series of its returns and of constituents returns Horst et al. (2004); Sharpe (1998). The use of past returns is a Hobson’s choice since no other information is typically available to external investors.

Let us denote by $\mathbf{r}^p$ the vector of portfolio returns along time, and by $\mathbf{R}^c$ the matrix containing the returns along time, j -th portfolio constituent being on the j -th column ($j = 1, \dots, S$). Data are observed on T subsequent time periods. The style analysis model regresses portfolio returns on the returns of the S constituents:

$$\underset{T \times 1}{\mathbf{r}^p} = \underset{T \times S}{\mathbf{R}^c} \underset{S \times 1}{\mathbf{w}^c} + \underset{T \times 1}{\mathbf{e}} \quad \text{s.t.: } \mathbf{w}^c \geq 0, \quad \mathbf{1}^T \mathbf{w}^c = 1. \quad (1)$$

Within the framework of the general linear regression model with normal error terms (Kutner et al. 2005), the observations $\mathbf{r}^p$ are independent normal variables, with mean $\mathbf{R}^c \mathbf{w}^c$ and with constant variance $\sigma^2 \mathbf{I}$, since $\mathbf{e}$ is a vector of uncorrelated normal random variables with expectation $E(\mathbf{e} | \mathbf{R}^c) = \mathbf{0}$, and variance–covariance matrix $\text{var}(\mathbf{e} | \mathbf{R}^c) = \sigma^2 \mathbf{I}$. Under such assumptions, the model focuses on the conditional expectation of the response, i.e., in financial terms, on the conditional expectation of portfolio returns distribution, $E(\mathbf{r}^p | \mathbf{R}^c)$. Regression coefficients express the sensitivity of portfolio expected returns to constituents returns, the vector $\mathbf{e}$ can be interpreted as the tracking error of the portfolio, and the two constraints provide exhaustive and non negative coefficients, interpretable in terms of compositional data. Indeed, from the linearity assumption stating that the systematic component of the model, $\mathbf{R}^c \mathbf{w}^c$, is a linear combination of covariates, it follows that the estimated coefficients measure constituents quotas in composing the portfolio. That is, the $\mathbf{R}^c \mathbf{w}^c$ term of the equation can be interpreted as the return of a weighted portfolio: the portfolio with optimized weights is a portfolio with the same style as the observed portfolio. It differs from the former as estimates about the internal composition are available Conversano and Vistocco (2010). We refer the interested reader to the paper of Kim et al. (2005) for the assumptions on portfolio returns and on constituents returns adopted in style models.

The procedure of Lobosco and Bartolomeo (1997) to estimate confidence intervals for style weights consists of the following steps:

- ▶ estimation of the style coefficients using the strong style model
- ▶ computation of the standard error of style analysis using the coefficients obtained at the previous step

for each constituent do

- estimation of the style coefficients through the semi-strong style model using the j -th constituent as response and the others as regressors

- computation of the unexplained Sharpe style volatility for index j , namely the variance of the vector of the differences between the j -th constituent returns and their estimates as obtained in the previous point
- computation of the standard error of the j -th constituent combining the standard error of the style analysis and the unexplained Sharpe volatility

end

A description of the procedure is provided in Algorithm 1, using the following notation:

$\leftarrow$	denotes the assignment operator,
$\ \cdot\ $	denotes the L_2 norm of a vector,
<code>strong_style</code>	is the function used to estimate the style coefficients through the strong model. The function takes the vector of portfolio and the matrix of constituents as input, returning the style coefficients as output,
<code>cardinality</code>	is the function computing the cardinality of a vector,
<code>[i]</code>	is the indexing operator, working in a different manner for vectors and for matrices. In particular, <code>variable[i]</code> extracts The i -th element if <code>variable</code> is a vector, and the i -th column if <code>variable</code> is a matrix.

Algorithm 1 Function for the Lobosco–Di Bartolomeo approximation of the standard errors of style coefficients

```

Data: r.p (portfolio returns), R.c (constituents returns)
Result: se_ldb_weights (standard errors of style coefficients)

Function se_lobosco_dibartolomeo(r.p, R.c):
  #estimate the style coefficients (strong model)
  style_weights ← strong_style(r.p, R.c)
  #compute the standard error of style analysis
  T ← cardinality(r.p)
  se_style ←  $\sqrt{\frac{\|R.c \times \text{style\_weights} - r.p\|}{T-2}}$ 
  #compute the Lobosco–Di Bartolomeo standard error
  k ← cardinality(style_weights ≠ 0)
  for i = 1 to cardinality(style_weights) do
    #estimate the style coefficient through the semi-strong style
    #model using constituent i as response and the others
    #constituents as regressors
    aux_w[i] ←  $\min_{\text{aux\_w}} \left( R.c[i] - \sum_{j \neq i} R.c[j] \times \text{aux\_w}[j] \right)^2 \text{ s.t.: } \mathbf{1}^T \text{aux\_w} = 1$ 
    #Lobosco–Di Bartolomeo approximation of standard errors
    se_weights[i] ←  $\|R.c[i] - \text{aux\_w}[i]\|$ 
    se_ldb_weights[i] ←  $\frac{\text{se\_style}}{\text{se\_weights}[i] \sqrt{T-k-1}}$ 
  end
  return se_ldb_weights

```

The procedure of Lobosco and Di Bartolomeo is a convenient method for estimating confidence intervals for style coefficients based on a Taylor expansion. The method works well except when the true parameters are on the boundary of the parameter space: one parameter is near 1, and/or one or more parameters are near 0. Indeed, in such a case it is not possible to exploit the classical Taylor expansion to obtain the asymptotic distribution of the style weights Kim et al. (2005). In the following, we restrict our attention to regular cases, i.e. when the parameters do not lie on the boundaries of parametric space. And because of that, the median regression approach presented in the next section is compared only with the Lobosco–Di Bartolomeo procedure. Even if this could seem a limitation, a portfolio where there is one asset with a large weight is not desirable in practice for the sake of diversification, with the aim of reducing risk.

3 An alternative approach to estimate the standard errors of style analysis coefficients

Quantile regression (QR) extends classical regression to a set of quantile functions of a response variable conditional on a set of covariates. Originally proposed by Koenker and Basset (1978), QR estimates separate models for different asymmetries $\tau \in [0, 1]$, where τ denotes the particular quantile of interest. Alternatives to approach the same problem have also been proposed in the literature Antoch and

Janssen (1989). The use of QR in the style analysis context has originally been proposed by Basset and Chen (2001), and retrieved by Attardi and Vistocco (2006, 2007, 2010). It offers an useful complement to the standard model, since it allows to discriminate portfolios that would be otherwise judged equivalent. Indeed, classical least squares regression estimates the influence of style exposure on expected returns, while QR focuses at places other than the expected value providing so useful insights, since the style exposure could affect returns in different ways at different locations of the portfolio returns distribution.

The QR style model for a given conditional quantile τ follows:

$$Q_\tau(\mathbf{r}^p \mid \mathbf{R}^c) = \mathbf{R}^c \mathbf{w}_\tau^c \quad \text{s.t.: } \mathbf{w}_\tau^c \geq 0, \mathbf{1}^T \mathbf{w}_\tau^c = 1,$$

where $Q_\tau(\cdot)$ denotes the conditional quantile function, and $\tau(0 < \tau < 1)$ the particular quantile of interest. The separate models associated to each quantile pose the only assumption $P(e_\tau \leq 0) = \tau$, that is the τ -th quantile of error term is 0. Put differently, instead of assuming that the average error is zero as in the standard linear regression, QR makes assumptions on the error term of the quantile τ of interest. Just as the assumption on the expectation of the error term in the standard linear regression implies that the regression coefficients impact the expectation of the response, the corresponding QR assumption on the quantile of the error term implies that QR coefficients impact on conditional quantiles of the response. Therefore, the $w_\tau^{c_j}$ coefficient can be interpreted as the rate of change of the τ -th conditional quantile of the portfolio returns distribution per unit change in the j -th constituent returns, holding the values of the other constituents constant.

In formal terms, QR assumes that the cumulative distribution function F_{e_τ} of the error term e_τ fulfills $F_{e_\tau}(0) = \tau$, namely that the conditional quantile of the error term is zero. This implies that the conditional quantile τ of the response is given by $\mathbf{R}^c \mathbf{w}_\tau^c$:

$$\begin{aligned} \tau = F_{e_\tau}(0) &= P(e_\tau \leq 0) = \\ &= P(\mathbf{R}^c \mathbf{w}_\tau^c + e_\tau \leq \mathbf{R}^c \mathbf{w}_\tau^c) = \\ &= P(\mathbf{r}^p \leq \mathbf{R}^c \mathbf{w}_\tau^c) = F_{\mathbf{r}^p}(\mathbf{R}^c \mathbf{w}_\tau^c) \end{aligned}$$

An important feature of QR to stress is its generality with respect to the error distribution: QR only assumes the independence between individual observations and the quantile restriction, and therefore it is a distribution-free regression approach.

The conditional quantiles are estimated through an optimization function minimizing a sum of weighted absolute deviation. The general conditional quantile estimator is:

$$\hat{\mathbf{w}}^c(\tau) = \sum_{i=1}^T \underset{\mathbf{w}^c(\tau)}{\operatorname{argmin}} \rho_\tau(\mathbf{r}^p - \mathbf{R}^c \mathbf{w}_\tau^c), \quad (2)$$

where $\rho_\tau(\cdot)$ is a function which asymmetrically weights positive and negative residuals:

$$\rho_{\tau}(u) = \begin{cases} \tau u, & \text{if } u > 0 \\ (\tau - 1)u, & \text{if } u \leq 0. \end{cases}$$

The machinery for solving the quantile regression problem originally exploited linear programming Davino et al. (2013), and in particular the simplex algorithm. In fact, the original algorithm for solving the quantile regression problem extended the Barrodale and Roberts algorithm Barrodale and Roberts (1973) to conditional quantiles. While Barrodale and Roberts algorithm was introduced for solving the median regression problem, Koenker and Bassett (1978) slightly modified the original L_1 problem placing asymmetric weights on positive and negative residuals, introducing quantile regression. Different solutions exploiting barrier methods have been also proposed. See Furno and Vistocco (2018) for a detailed treatment of QR solution methods. Whatever the particular solving method, the linear programming formulation of the QR problem makes relatively easy to impose inequality constraints. In particular, we refer to the algorithm proposed in Koenker and Ng (2005) for computing inequality constrained quantile regression. In the following we use a semiparametric approach since we assume a linear model in order to compare QR estimates with the classical style model. Moreover, we restrict our attention to the median regression by setting $\tau = 0.5$. The use of absolute deviations ensures that conditional quantile estimates are robust. It is worthwhile to mention that the use of different values of τ would make it possible to obtain a set conditional quantile estimators that could be easily linearly combined in order to construct L-estimator, gaining so in efficiency Koenker and Bassett (1987). To draw inferences in the described framework, we refer to subsampling. This approach was first introduced by Politis and Romano (1994) and it can be considered as the most general theory for the construction of first order asymptotically valid confidence intervals or regions. The basic idea is to approximate the sampling distribution of the statistic of interest through the values of the statistic (suitably normalized) computed over smaller subsets of the data. This approach has been shown to be valid under very weak assumptions and, when compared to other resampling schemes such as the bootstrap, it does not require that the distribution of the statistic is somehow locally smooth as a function of the unknown model. Indeed, the subsampling can also be applied in situations which represent counterexamples to the bootstrap. To highlight why the subsampling approach is more general than other resampling schemes it is worthwhile to observe that each subseries of size b is a sample of size b from the true model. Hence, under the weak convergence hypothesis, the sampling distributions based on subseries of size b and T should be close. The bootstrap, on the other hand, is based on recomputing a statistic over a pseudo series of size T from some estimated model and the asymptotic properties of the method are related to how close is the resampling model to the true model.

The subsampling can be implemented in our framework as follows:

1. Fix b , the subseries length.
2. Generate subseries of length b , $\mathbf{r}_{i,b}^p$ and $\mathbf{R}_{i,b}^c$ for $i = 1, \dots, T - b + 1$.
3. For each j ($j = 1, \dots, S$), compute the constrained quantile estimator $\hat{\mathbf{w}}_{j,b}^c(\tau)$.

4. The subsampling approximation of the unknown sampling distribution, for the j th style coefficient, is given by:

$$L_T(x, \tau) = \frac{1}{T-b+1} \sum_{i=1}^{T-b+1} \mathbf{I} \left\{ b^{1/2} \left(\hat{w}_{i,b,j}^c(\tau) - \hat{w}_j^c(\tau) \leq x \right) \right\}$$

with $\mathbf{I}$ denoting the indicator function.

In the following, we compare the Lobosco–Di Bartolomeo solution with three solutions based on subsampling, namely an approximate two-sided equal-tailed confidence interval, a two-sided symmetric confidence interval, and an interval based on asymptotic Gaussian distribution, exploiting subsampling only to estimate the standard error of the style coefficient estimators.

In particular, a *two-sided equal-tailed confidence interval* (EQ) for the j -th style coefficient, of nominal level $1 - \alpha$, can be obtained as:

$$\left[\hat{w}_j^c(\tau) - T^{-1} \int_{T,b}(1 - \alpha/2); \quad \hat{w}_j^c(\tau) - T^{-1} \int_{T,b}(\alpha/2) \right]$$

where $\int_{T,b}(\alpha)$ is the quantile of order α of the subsampling approximation. This interval has approximately the same area ($\alpha/2$) in each tail.

As an alternative, *two-sided symmetric confidence intervals* (SYM) can be constructed as follows:

$$\left[\hat{w}_j^c(\tau) - T^{-1} \tilde{\int}_{T,b}(1 - \alpha); \quad \hat{w}_j^c(\tau) + T^{-1} \tilde{\int}_{T,b}(1 - \alpha) \right]$$

where $\tilde{\int}_{T,b}(\alpha)$ is the quantile of order α of the distribution:

$$L_T(x, \tau) = \frac{1}{T-b+1} \sum_{i=1}^{T-b+1} \mathbb{I} \left\{ b^{1/2} \left| \hat{w}_{i,b,j}^c(\tau) - \hat{w}_j^c(\tau) \right| \leq x \right\}$$

Hall (1988) showed that symmetric confidence intervals may present enhanced coverage and they can be shorter than equal-tailed confidence intervals even in asymmetric circumstances.

A third approach, *confidence interval based on asymptotic Gaussian distribution* (ASY), can be obtained using the subsampling just to estimate the standard error of the style coefficient estimator and keep on using the asymptotic Gaussian distribution. The standard error for the j -th style coefficient is:

$$\text{Var} \left(\hat{w}_j^c(\tau) \right) = \frac{b}{T} \sum_{i=1}^{T-b+1} \left(\hat{w}_{i,b,j}^c(\tau) - \bar{w}_{\cdot,b,i}^c(\tau) \right)^2$$

where $\bar{w}_{\cdot,b,i}^c(\tau)$ is the sample mean of the subsampling replicates. Clearly, a key issue is the selection of a proper subseries length. Suggestions and algorithms for optimal selection of this parameter can be found in Politis et al. (1999). Next section discusses such issue along with the presentation of the simulation results.

Table 1 The 7 scenarios of the the simulation design, obtained combining different levels of presence of outliers (columns) with different percentages of contamination (rows)

		Outliers				
		None	in Y	in X	in Y and X	
Percentage of contamination	None	Case A	–	–	–	
	1%	–	Case B1	Case C1	Case D1	
	5%	–	Case B2	Case C2	Case D2	

4 Simulation results

In this section the finite sample properties of the proposed procedure is investigated via a Monte Carlo study. Artificial fund returns are generated as follows:

$$\mathbf{r}_t^p = \mathbf{R}_t^{c'} \mathbf{w}^c + \sigma \mathbf{e}_t, \quad t = 1, 2, \dots, T.$$

We considered a portfolio with 5 constituents generated using GARCH(1,1) processes to simulate the behavior of true time series returns. The true style weights have been set to $w_j^c = 0.2$, $j = 1, 2, \dots, 5$. The choice of using here the same value for the constituents weights aims to simplify the presentation of the results, since we can focus only on one constituent. We went through several tests using different values and different ratios for the weights confirming the same results reported below. The scaling factor σ has been fixed in order to have R^2 close to 0.90 while $e_t \sim N(0, 1)$. The nominal level has been set to 90%. The choice of such a value for the R^2 responds to typical situations for such type of models. Indeed, in practical applications, the constituents used to explain portfolio's composition represent sectorial quotas of the investing market. Such constituents correspond to a sectorial stratification or a geographical stratification of the portfolio able to reproduce a benchmark index and to explain a large part of the observed variation.

Additive outliers at randomly chosen positions were added both in the constituents series and in the portfolio returns. The percentage of outliers contamination has been set to 1% and 5%. A synthesis schema of the simulation design is offered in Table 1, showing the 7 resulting scenarios obtained from the combination of the presence of outliers (none, in Y , in X , in Y and X) with the percentage of outliers contamination (none, 1%, and 5%). For each considered scenario, we set $T \in \{250, 500\}$ with 1000 Monte Carlo runs for each simulation of the experimental set up. Finally, we used different values for the block length b ($b = 20, \dots, 80$) to test its effect on the three solutions based on subsampling.

Performance evaluation has been based on two closely connected aspects:

- the efficiency of the robust estimator with respect to the classical one;
- the length of the confidence intervals and their coverage rate, which measures how often the constructed confidence intervals contains the true style weights.

A comparison between classical style analysis estimator based on least squares (LS) and on median regression (QR) is offered in Fig. 1 through violin plots Hintze

and Nelson (1998). Such representation allows to depict simultaneously the full distribution of a set of data, and the number of the data considered. Indeed, the height of each violin informs about the range of the values, while the position of the peak is provided by the width of the violin. The point inside each violin depicts the average value, and the segments are located at ± 1 standard deviation. In particular, the violins in Fig. 1 depict the distributions (vertical axis) of the two estimators along the 1000 considered replications, for $T = 250$ and $T = 500$ (horizontal axis), the panels being associated to the 7 considered scenarios. As expected, the impact of outlying observations can be very serious on least squares estimates of style coefficients, especially when considering outliers in the constituents series. Even if both the estimators are unbiased (all the violin plots are centered on the true parameter value 0.2), obviously LS estimator is more efficient when no outliers are considered (first panel). LS estimator degrades in efficiency adding outliers in Y , and even more when outliers are inserted in X . The joint effect of outliers in Y and in X does not differ much from the case of outliers only in X . In general, the LS estimates variability increases very much and this can have serious drawbacks since the style coefficients are supposed to vary in the unit interval. QR estimator is instead robust for the different cases of outliers considered, showing only a slight increase in the variability of the estimates moving from scenario A to scenario D2. This result is partly unexpected, since quantile regression is robust only for outliers in Y . We guess the robustness also when outliers in X are considered may be due to the presence of the double constraints which restrict the parameter space. In any case, efficiency increases as sample size increases, as evident comparing each class of estimator between the two cases of $T = 250$ and $T = 500$.

The different behavior of the two estimators has an obvious impact on the width of confidence intervals (Fig. 2) and on the observed coverage (Fig. 3). Both the figures show the effect of the block length used for subsampling (horizontal axis). The two charts go hand in hand, since a proper calibration might allow the selection of a

Table 2 Average empirical coverage of confidence intervals in the 1000 considered replications, using the classical Lobosco–Di Bartolomeo solution and the three subsampling options (rows) along the 7 considered scenarios (columns), when the block length b is set equal to 34

	Case A no outliers	Case B1 outliers in $Y - 1\%$	Case B2 outliers in $Y - 5\%$	Case C1 outliers in $X - 1\%$	Case C2 outliers in $X - 5\%$	Case D1 outliers in $X, Y - 1\%$	Case D2 outliers in $X, Y - 5\%$
LDB	0.90 [0.024]	0.90 [0.059]	0.88 [0.121]	0.88 [0.109]	0.91 [0.240]	0.88 [0.121]	0.87 [0.263]
SYM	0.89 [0.031]	0.90 [0.032]	0.91 [0.033]	0.90 [0.032]	0.90 [0.033]	0.91 [0.032]	0.89 [0.035]
EQ	0.84 [0.030]	0.85 [0.031]	0.85 [0.032]	0.83 [0.031]	0.83 [0.032]	0.83 [0.031]	0.84 [0.034]
ASY	0.90 [0.031]	0.90 [0.031]	0.91 [0.033]	0.90 [0.031]	0.88 [0.032]	0.90 [0.031]	0.88 [0.034]

Values in brackets refer to the average interval lengths

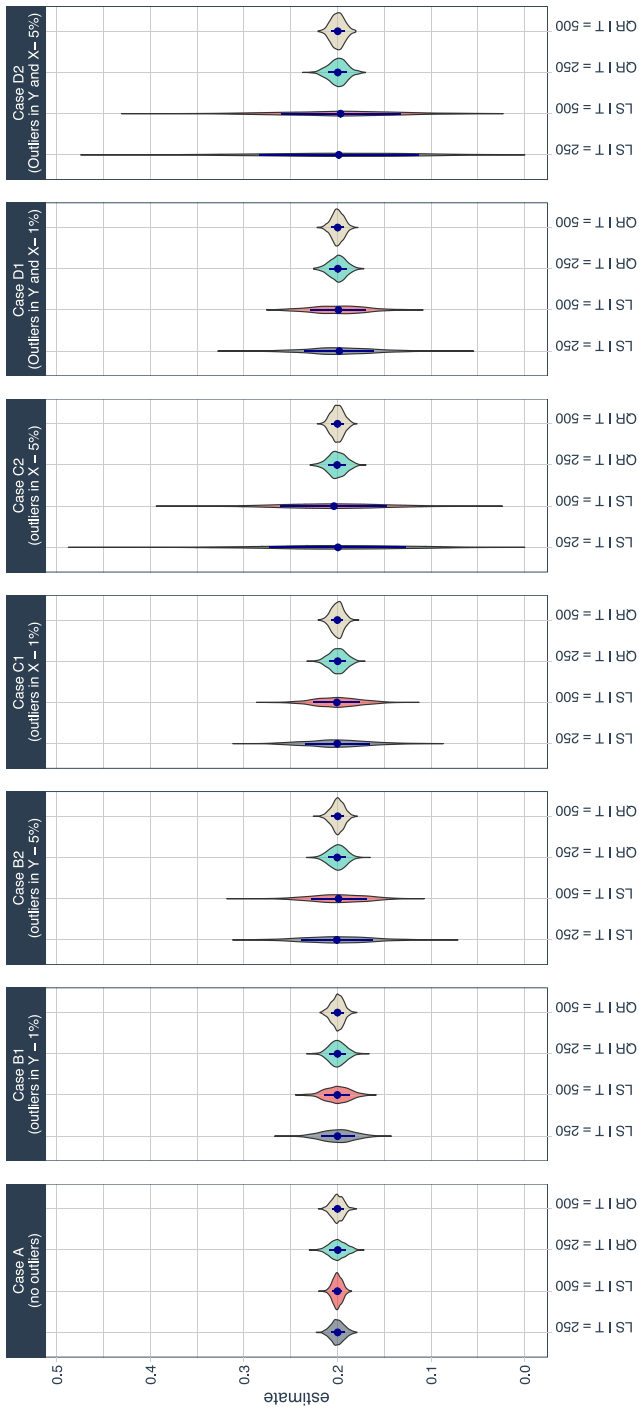


Fig. 1 Comparison of classical style analysis estimator based on least squares (LS) and estimator based on median regression (QR) for the two considered sample sizes ($T = 250$, and $T = 500$). Violin plots show the distribution on the 1000 considered replications, the dot in each violin being located at the average value and the segment at ± 1 standard deviation. The true coefficient is equal to 0.2

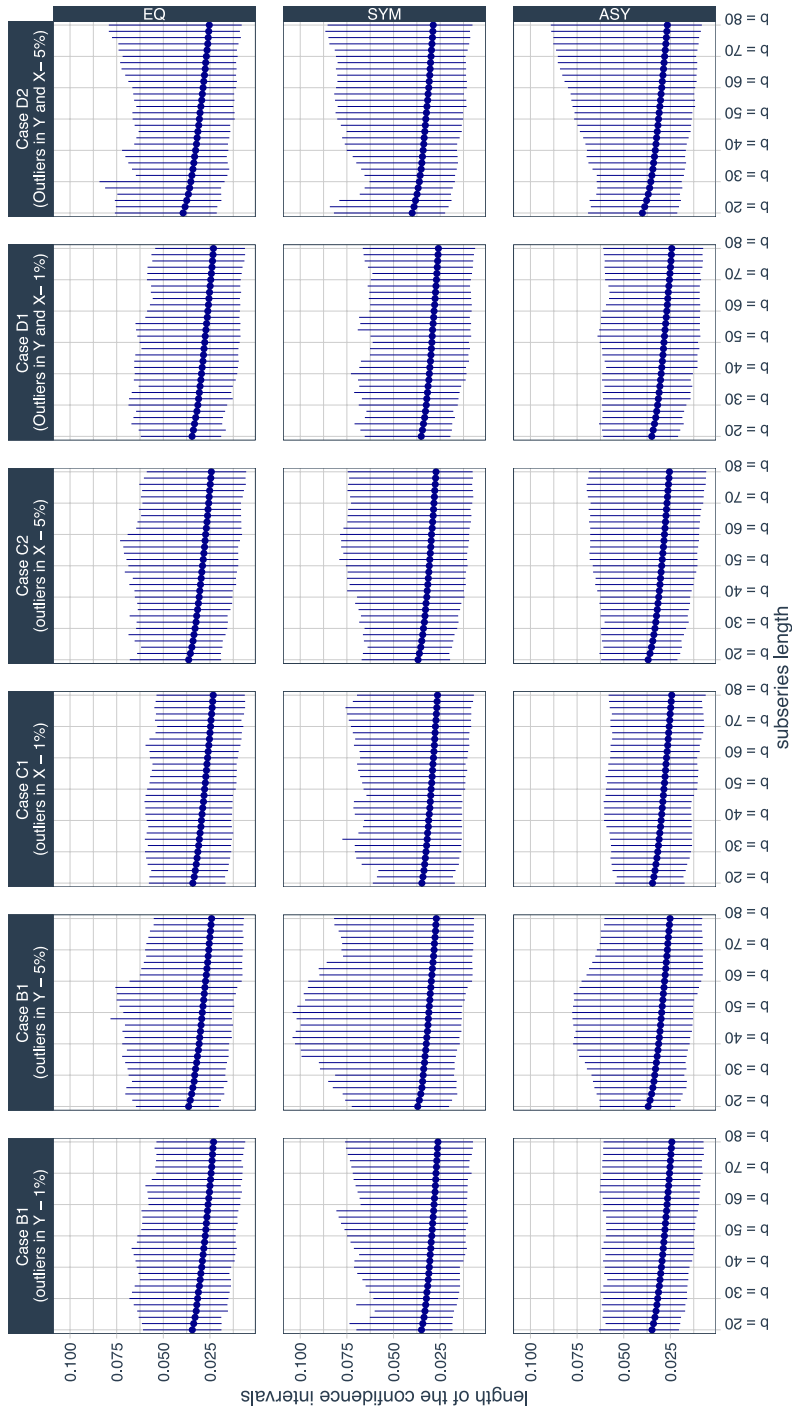


Fig. 2 Effect of the block length used for subsampling (horizontal axis) on the length of the confidence intervals (vertical axis). Each panel of the plot refers to a different scenario (columns) and a different solution used for subsampling (rows). The dots are located at the average of the 1000 considered replications, the segments at the minimum and maximum values. The plot depicts the case $T = 250$, results for $T = 500$ do not differ in terms of patterns here highlighted

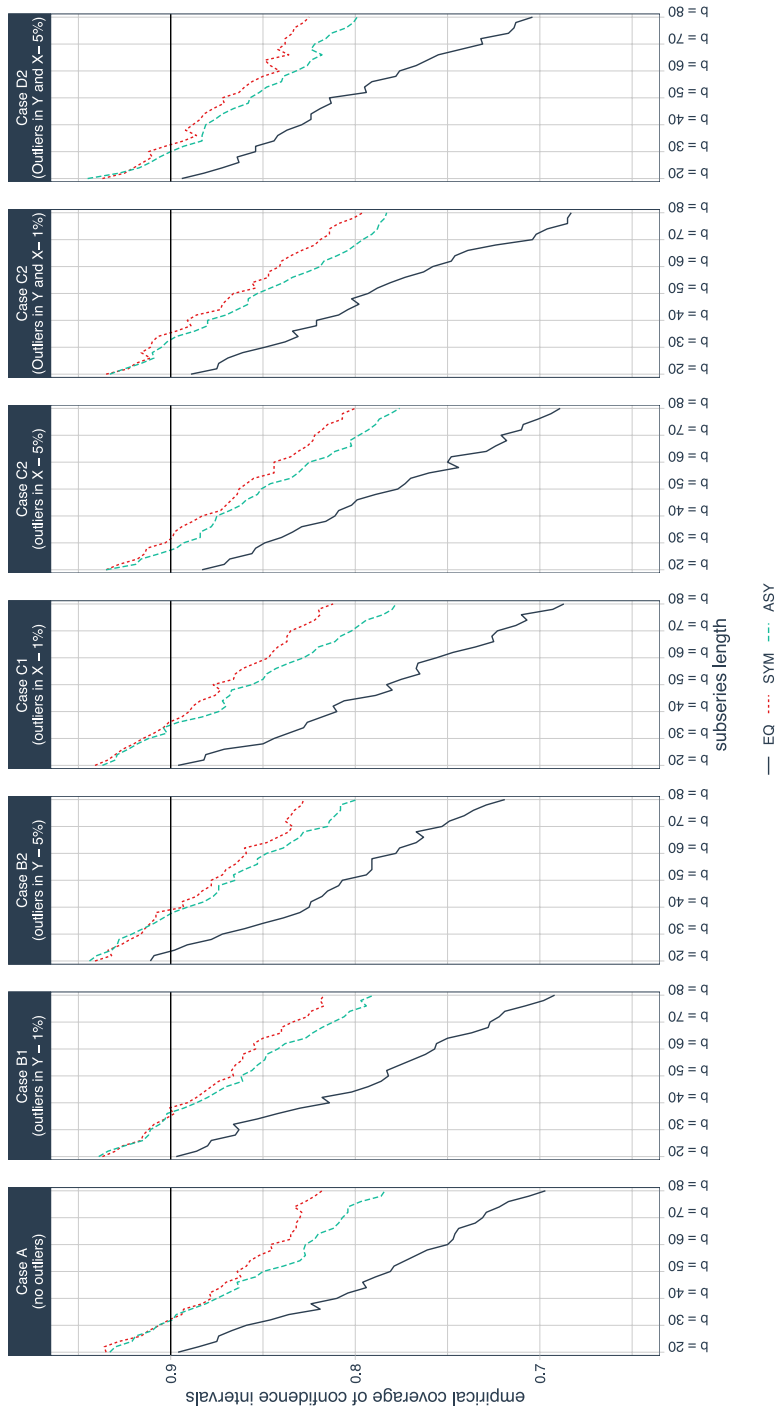


Fig. 3 Effect of the block length used for subsampling (horizontal axis) on the empirical coverage of the confidence intervals (vertical axis). The nominal level has been set to 90% (horizontal line). Each panel of the plot refers to a different scenario. The three lines refer to the three solutions considered for subsampling. The plot depicts the case $T = 250$, results for $T = 500$ do not differ in terms of patterns here highlighted

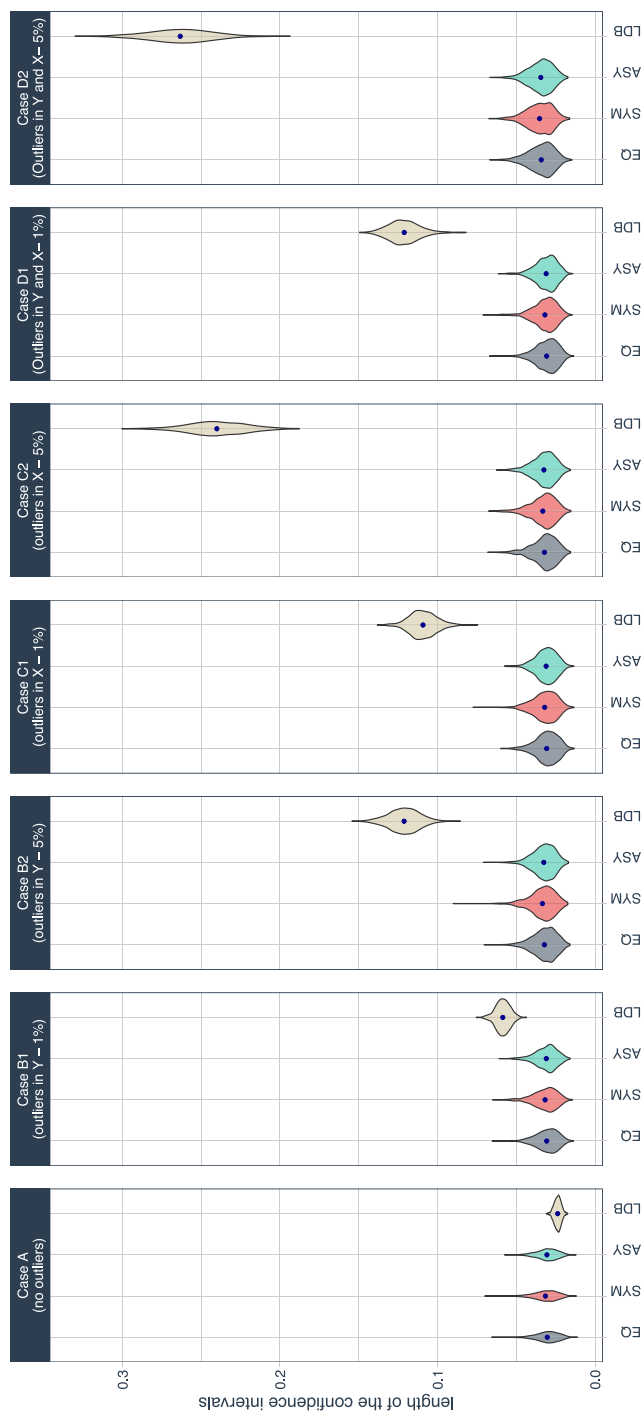


Fig. 4 Comparison of the three solutions considered for subsampling and the classical Lobosco-Di Bartolomeo procedure (horizontal axis) in terms of length of the confidence intervals (vertical axis), for the value $b = 34$ as block length for subsampling. Each panel of the plot refers to a different scenario. Violin plots show the distribution on the 1000 considered replications, the dots in each violin are located at the average of the 1000 considered replications. The plot depicts the case $T = 250$, results for $T = 500$ do not differ in terms of patterns here highlighted

proper subseries length to get low coverage errors. In particular, looking at the trend of points in Fig. 2, it is evident how the average block length decreases when the block length increases (horizontal axis) in all the considered scenarios (columns) and for the three alternative subsampling methods (rows). The three methods show instead different patterns if we look at the segments, depicting the minimum and maximum values of the intervals on the 1000 considered replications. Figure 3 compares the empirical coverage of the three subsampling methods (three lines) when block length increases (horizontal axis). Each panel refers to a different scenario, the horizontal line being set at the nominal level 90%. Two-sided symmetric confidence intervals (SYM) and asymptotic Gaussian intervals (ASY) have similar trends in all the scenarios, showing an empirical coverage closer to the nominal coverage when b is between 30 and 40, the precise value depending on the particular case. Instead, equal-tailed confidence intervals (EQ) show better coverage using smaller values of b . Clearly, when using subsampling, observed coverage and, as a consequence, confidence interval width, is related to the subseries length. However, this drawback can be even turned into an advantage. One could use subseries length as a calibration tool in order to get observed coverages close to nominal ones. Take for instance results in Table 2 and Fig. 4, reporting the case when b is set to 34. When there are no outliers in the data, the classical Lobosco–Di Bartolomeo approximation shows good performance both in terms of coverage and of interval length, but its performance deteriorates when outliers are considered. The three subsampling options show good coverages, with comparable interval lengths in almost all the cases, but the equal-tailed confidence intervals (EQ). This is clearly imputable to the choice of using $b = 34$, value of the block length most suitable for the methods SYM and ASY. Using this value for b , confidence intervals based on the asymptotic Gaussian approximation with standard error estimated via subsampling perform similarly to subsampling symmetric intervals. Finally, even if the observed coverage of the Lobosco procedure is still accurate when there are outliers in the series, the width of the confidence intervals increases very much. As expected, a much stronger increase in the length of confidence intervals is observed when increasing the percentage of contamination from 1% to 5%. This is more evident from the analysis of Fig. 4, which shows the distribution of the interval lengths (vertical axis) along the 1000 replications, through the use of violin plots.

5 Concluding remarks

Style analysis is widely used in financial practice in order to decompose portfolio performance with respect to a set of indexes representing the market in which the portfolio invests. The classical Sharpe method is commonly used for estimating purposes but requires corrections for drawing inferences because the model includes inequality constraints. Among the different proposals appeared in literature, the Lobosco–Di Bartolomeo approximation for computing corrected standard errors is widespread and it performs well for regular cases, i.e. when parameters are not on the boundaries of the parameter space. However, the Lobosco–Di Bartolomeo solution suffers from presence of outliers, being essentially based on a least squares

estimation procedure. This can have serious drawbacks since the style coefficients are supposed to vary in the unit interval.

In this paper we propose a robust procedure based on constrained median regression. For inference purposes we attain to subsampling theory. Median regression can be seen as a particular case of quantile regression. Such a framework opens up new possibilities for further improving the estimator efficiency, since weighted linear combinations of estimates related to different quantiles can be easily adopted for obtaining L -estimators. A simulation study gave us the possibility to explore the behavior of median regression coupled with three options to estimate confidence intervals for style coefficients. The design of the study combined different levels of presence of outliers with different percentages of contamination. Results showed that in case of no outliers, the classical Lobosco–Di Bartolomeo solution ensures good performance in terms both of empirical coverage and of interval length. Performance deteriorates when outliers (in Y , in X , or in Y and X) are inserted in the series.

The three subsampling methods used to obtain confidence intervals show comparable performance in terms of interval lengths, with empirical coverage being related to the choice of the block length used in the subsampling procedure. Indeed, when using subsampling, confidence interval lengths and empirical coverage are related to subseries length. However, this drawback can be even turned into an advantage. One could use subseries length as a calibration tool in order to get observed coverages close to nominal ones. Two alternative construction methods studied in this paper, i.e. subsampling symmetric intervals and intervals based on the asymptotic Gaussian approximation with standard error estimated via subsampling, perform similarly. Instead, a peculiar behavior is shown by equal-tailed intervals. Again, calibration might allow the selection of a proper subseries length to get low coverage errors also for this method. In general, as expected, a much stronger increase in the length of confidence intervals is observed when increasing the percentage of contamination from 1% to 5%. Finally, efficiency increases as sample size increases, confirming the consistency of the considered estimators.

It remains interesting to explore the unexpected robust behavior of the median regression estimator when outliers in X are considered, partly explainable with the presence of the double constraints, which restrict the parameteric space. Further investigations may concern the extension of the procedure to not regular cases, i.e. when the parameters lie on the boundaries of parametric space, the comparison with different robust estimators, and the estimation of the style weights at different conditional quantiles, providing so estimates that can be easily linearly combined in order to construct L -estimator and gain efficiency.

Finally, an issue worth investigating concerns the use of the proposed approach in combination with compositional data methods Aitchison (1982, 1986); Filzmoser et al. (2018). For example, the proposed approach can be applied on mutual funds or hedge funds, and then exploiting methods for classifying compositional data to obtain groups of portfolios with similar internal compositions. Indeed, for these types of financial assets, both the series of returns and of sectorial and/or geographical constituents are publicly available.

Acknowledgements The author wish to thank Professor Michele La Rocca for helpful comments on a previous draft of the paper, and the two anonymous referees for their valuable suggestions that helped to improve the final version of this paper. All computation and graphics were done in the R language R Core Team R: (2021) using the basic packages and the additional quantreg Koenker (2021), and ggplot2 Wickham (2016) packages.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

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

- Aitchison J (1982) The statistical analysis of compositional data. *J Roy Stat Soc Ser B (Methodol)* 44(2):139–177
- Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman and Hall, London
- Andrews DWK (1999) Estimation when the Parameter are on the Boundary. *Econometrica* 67:1341–1383
- Antoch J, Janssen P (1989) Nonparametric regression quantiles. *Statist Probab Lett* 8(4):355–362
- Attardi L, Vistocco D (2006) Comparing financial portfolio style through quantile regression. *Statistica Applicata* 18(2):6–16
- Attardi L, Vistocco D (2007) On estimating portfolio conditional returns distribution through style analysis models. In: Perna C, Sibillo M (eds) *Quantitative methods for finance and insurance*. Springer, Cham
- Attardi L, Vistocco D (2010) An index for ranking financial portfolios according to internal turnover. In: Lauro C, Palumbo F, Greenacre M (eds) *Data analysis and classification*. Springer, Cham, pp 445–452
- Barrodale I, Roberts F (1973) An improved algorithm for discrete l_1 linear approximation. *SIAM J Numer Anal* 10:839–848
- Basset GW, Chen HL (2001) Portfolio style: return-based attribution using quantile regression. In: Fitznerberger B, Koenker R, Machado JAF (eds) *Economic applications of quantile regression (Studies in Empirical Economics)*. Physica-Verlag, Wurzburg (Wien), pp 293–305
- Christodoukalis GA (2005) Sharpe style analysis in the MSCI sector portfolios: a monte carlo integration approach. In: Knight J, Satchell S (eds) *Linear factor models in finance*. Elsevier, Amsterdam, pp 83–94
- Conversano C, Vistocco D (2010) Analysis of mutual funds' management styles: a modeling, ranking and visualizing approach. *J Appl Stat* 37:1825–1845
- Davino C, Furno M, Vistocco D (2013) *Quantile regression: theory and applications*. Wiley, UK
- Davis WW (1978) Bayesian analysis of the linear model subject to linear inequality constraints. *J Am Stat Assoc* 78:573–579
- Filzmoser P, Hron K, Templ M (2018) *Applied compositional data analysis, with worked examples in R*. Springer, Springer Series in Statistics
- Furno M, Vistocco D (2018) *Quantile regression: estimation and simulation*. Wiley, Amsterdam
- Geweke J (1986) Exact inference in the inequality constrained normal linear regression model. *J Appl Economet* 1(2):127–141
- Hall P (1988) On symmetric bootstrap confidence intervals. *J Roy Statist Soc Ser B* 50:35–45

- Hintze JL, Nelson RD (1998) Violin plots: a box plot-density trace synergism. *Am Stat* 52:181–184
- Horst J.K. Ter, Nijman TH, De Roon FA (2004) Evaluating style analysis. *J Empir Financ* 11:29–51
- Judge GG, Takayama T (1966) Inequality restrictions in regression analysis. *J Am Stat Assoc* 61:166–181
- Kim TH, White H, Stone D (2005) Asymptotic and bayesian confidence intervals for sharpe-style weights. *J Financial Econom* 3:315–343
- Koenker R, Basset GW (1978) Regression quantiles. *Econometrica* 46:33–50
- Koenker R, Basset GW (1987) L-estimation for linear models. *J Am Stat Assoc* 82(399):851–857
- Koenker R (2005) Quantile regression, econometric society monographs
- Koenker R, Ng P (2005) Inequality constrained quantile regression, *Sankhya: The Indian. J Stat* 67(2):418–440
- Koenker R (2021) quantreg: quantile regression, R package version 5.86, <https://CRAN.R-project.org/package=quantreg>
- Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Applied linear statistical models. McGraw-Hill, New York
- Lobosco A, Di Bartolomeo D (1997) Approximating the confidence intervals for sharpe style weights. *Financial Anal J* 53:80–85
- Politis DN, Romano JP, Wolf M (1999) Subsampling. Springer-Verlag, New York
- Politis DN, Romano JP (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Ann Stat* 22:2031–2050
- R Core Team R: (2021) A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>
- Sharpe W (1992) Asset allocation: management styles and performance measurement. *J Portf Manage* 18(2):7–19
- Sharpe W (1998) Determining a fund's effective asset mix. *Invest Manage Rev* 2(6):59–69
- Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York

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