North American Journal of Economics and Finance 51 (2020) 101055

Contents lists available at ScienceDirect

North American Journal of Economics and Finance

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

Lasso-based index tracking and statistical arbitrage long-short strategies

Leonardo Riegel Sant'Anna[a,*], João Frois Caldeira[b], Tiago Pascoal Filomena[a]

[a] School of Business, Federal University of Rio Grande do Sul, 855 Washington Luis Street, Porto Alegre, RS 90010-460, Brazil
[b] Department of Economics, Federal University of Santa Catarin, Campus Universitario Reitor João David Ferreira Lima, Florianopolis, SC 88040-900, Brazil

ARTICLE INFO

JEL classification:
C52
C55
C58
G11

Keywords:
Lasso
Index tracking
Long-short
Portfolio selection
Statistical arbitrage

ABSTRACT

In this paper, we apply the lasso-type regression to solve the index tracking (IT) and the long-short investing strategies. In both cases, our objective is to exploit the mean-reverting properties of prices as reported in the literature. This method is an interesting technique for portfolio selection due to its capacity to perform variable selection in linear regression and to solve high-dimensional problems (which is the case if we consider broader indexes such as the S&P 500 or the Russell 1000). We use lasso to solve IT and long-short with three market benchmarks (S&P 100 and Russell 1000 – US stock market; and Ibovespa – Brazilian market), comprising data from 2010 to 2017. Also, we formed IT portfolios using cointegration (a method widely used for index tracking) to have a basis for comparison of the results using lasso. The findings for IT showed similar overall performance between portfolios using lasso and cointegration, with a slight advantage to cointegration in some cases. Nonetheless, lasso-based IT portfolios presented average monthly turnover at least 40% smaller, indicating that lasso generated portfolios that had not only a consistent tracking performance but also a considerable advantage in terms of transaction costs (represented by the average turnover).

1. Introduction

Stock index tracking (IT) is a passive investment strategy that consists in building a portfolio of stocks to replicate (or *track*) as close as possible the cumulative return of a stock market benchmark, such as the Standard & Poor's 500 Index. The most straightforward approach to track an index would be to make a full replication, i.e., to invest in all index constituents according to their weights in the index composition. However, such choice tends to lead to a more significant level of transaction and management costs especially in the case of broader indexes such as the S&P 500. For this reason, the option for a full replication is feasible only when the structure of the index is smaller, is kept constant for extensive periods, and the tracking investment fund is not affected by large inflows or outflows of money. Hence, in spite of their varying solving methods, IT models usually present a standard feature, which is the need to perform variable selection with the aim of forming portfolios with a reduced number of assets relative to the index.

Furthermore, a natural extension of IT is the market-neutral long-short strategy, which consists of a self-financing method that seeks to explore temporary market inefficiencies to generate alpha (excess return relative to the market benchmark). Its application involves buying undervalued stocks and short selling overvalued ones relative to the index (Alexander & Dimitriu, 2002). Even

* Corresponding author.
E-mail addresses: leonardo.santanna@ufrgs.br (L.R. Sant'Anna), emaildocaldeira@gmail.com (J.F. Caldeira), tpfilomena@ea.ufrgs.br (T.P. Filomena).

https://doi.org/10.1016/j.najef.2019.101055
Received 14 July 2018; Received in revised form 21 August 2019; Accepted 26 August 2019
Available online 29 August 2019
1062-9408/

though this strategy lacks some broader implementation among many hedge funds due to its short exposure (Badrinath & Gubellini, 2011), it is an attractive option as a result of its self-financing and market neutral features. Moreover, due to the similarities between IT and long-short regarding their implementation approach, forming portfolios long-short also requires the capacity of optimization models to perform variable selection to choose a reduced number of stocks to compose each portfolio. Thereby, in this paper we seek to present an implementation of the so-called lasso-type regression (least absolute shrinkage selection operator) to solve both the IT and long-short investing problems, due to this method's capacity to make variable selection in linear regression and provide good-quality solutions for high-dimensional datasets (which is the case of broader indexes such as the Russell 1000).

The choice for a passive investment strategy such as IT relies on both theoretical (market efficiency) and empirical (performance and costs) reasons. If the market is efficient, it is not possible to obtain superior risk-adjusted returns through active management portfolio strategies. In line with the Efficient Market Hypothesis (EMH) and studies such as Fama and French (2010), active investment funds, in general, do not beat the overall market performance consistently in the long run. Zenios (2008) reported that the average return of 769 all-equity actively managed funds was between 2% and 5% lower than the S&P 500 from 1983 to 1989. Additionally, Kwon and Wu (2017) mentioned that more than 82% of the actively managed large-cap funds in the US were outperformed by the S&P 500 from 2005 to 2014. For this reason, investors should be better off choosing a passive strategy, as this type of investment seeks to minimize costs while expecting that the market exhibits the mean-reverting properties and its constituents present positive returns over time.

Given its practical importance, the IT problem has been constantly attracting some attention particularly in the literature regarding finance and operations research. Many methods have already been proposed to solve this problem. Konno and Wijayanayake (2001) and Mezali and Beasley (2013) introduced models based on optimization and quantile regression. Guastaroba and Speranza (2012) and Scozzari, Tardella, Paterlini, and Krink (2013) described a heuristic based model that includes fixed and variable transaction costs, as well as a constraint on the total transaction costs incurred when each portfolio is updated. Alexander (1999) and Alexander and Dimitriu (2005) proposed the use of cointegration to build tracking portfolios. Wu, Yang, and Liu (2014) and Yang and Wu (2016) explored the lasso-type regression and its capacity of performing variable selection to solve the IT problem. More recently, Mutunge and Haugland (2018) focused on the use of a heuristic approach (mixed integer quadratic programming – MIQP) to solve the index tracking problem designed as a quadratic function based on a coefficient matrix (keeping their focus on the optimization for a short period of time with instances up to 2000 stocks). García, Guijarro, and Oliver (2018) also held their attention on the index tracking solving process through the use of a heuristic methodology. Similarly to Mutunge and Haugland (2018), first those authors showed that the optimization problem is NP-hard, thus requiring some heuristic method to solve it in a reduced time; then, two distinct methods were used to solve the problem in a static manner, without attention to the portfolio rebalancing process. In contrast, Strub and Baumann (2018) aimed to handle the IT problem with focus on the portfolio updating process, based on the use of a mixed-integer linear programing formulation that considered the trade-off between transaction costs and the index fund performance over time.[1]

Nevertheless, despite their conceptual distinctions, the definition of a cardinality constraint to limit the size of each tracking portfolio is a standard feature among most of the studies cited above. As already mentioned, the most practical way to select a bundle of stocks to track an index would be to do a full replication. However, as such choice would tend to produce more substantial costs particularly in the case of broader indexes such as the S&P 500 or the Russell 1000, some papers form tracking portfolios by choosing the stocks with the most significant weights in the index (for instance, Alexander & Dimitriu, 2005; Dunis & Ho, 2005). In this sense, picking the stocks to compose each tracking portfolio becomes a decision exogenous to the solving process, which might be efficient if we consider market benchmarks composed by a small number of stocks. For instance, some smaller indexes are the Dow Jones Industrial Average (DJIA), composed by the 30 largest firms listed either in the New York Stock Exchange (NYSE) or in the Nasdaq Stock Market; the DAX, composed by the 30 largest German firms listed in the Frankfurt Stock Exchange; and the Ibovespa, formed by the most liquid stocks in the Brazilian stock market. For example, regarding the Ibovespa, the index was composed by 59 stocks during September-December 2017, and the top 10 stocks with the largest weights were enough to account for about 57.7% of the index.[2]

However, as we move to larger indexes like the S&P 100, such choice for the most relevant stocks tends to become more difficult due to the lower index concentration. In the case of the S&P 100, it had a total of 102 stocks in December 2017, and the top 10 major stocks accounted for about 31% of the index.[3] In the case of a bigger index such as the Russell 1000, this contrast becomes even more evident, as the largest stock weight composing this index in December 2017 corresponded to only 2.23% of the index, and the top 10 stocks accounted for only 14.7%.[4]

As a result, we can understand the need of imposing a constraint on the size of the tracking portfolios, which leads to the usefulness of a method that has the capacity of performing variable selection, such as lasso. In the context of index tracking and portfolio optimization, the use of a statistical model that selects the most relevant coefficients is appropriate since it makes the portfolio selection process endogenous to the optimization problem. Moreover, the lasso regression has the potential to solve the

[1] For a complete literature review on index tracking, as well as another examples of recent studies related to this topic, we also refer the reader to: Andriosopoulos and Nomikos (2014), Chavez-Bedoya and Birge (2014), Filippi, Guastaroba, and Speranza (2016), Gnägi and Strub (2018), Mezali and Beasley (2013), Strub and Trautmann (2019).

[2] Source, access in March 04th, 2019:http://www.bmfbovespa.com.br.

[3] Source, access in March 04th, 2019:https://latam.spindices.com/indices/equity/sp-100.

[4] Source, access in March 04th, 2019:http://www.ftse.com/analytics/factsheets/Home/ConstituentsWeights.

index tracking problem effectively for small market indexes as well as for more extensive benchmarks such as the Russell 1000, due to its capacity to provide solutions with high-dimensional datasets.

Regarding past studies using lasso and index tracking, Wu et al. (2014) proposed the so-called Nonnegative Lasso, which consists of computing the lasso regression constrained by having all coefficients equal or larger than zero, thereby avoiding short positions in the portfolios. The authors used the Chinese index CSI 300 and its stock constituents (with data from August 2008 to November 2011), and their empirical analysis showed the capacity of lasso to generate good quality portfolios concerning annual tracking error in-sample (fitted error) as well as out-of-sample (predicted error). Later, Yang and Wu (2016) extended the former approach and proposed the Nonnegative Adaptive Lasso. Moreover, they also applied a two-stage approach combining nonnegative adaptive lasso and nonnegative least squares, and their results concerning tracking error also presented overall consistent performance regarding fitted and predicted errors.

Thus far, the studies mentioned above focus on the introduction of two statistical approaches, while their empirical analysis related to the index tracking problem is quite limited. So, our study differs from the previous literature as we focus on the financial context and apply the lasso-type regression to different markets using diversified sample sizes. Particularly, we use three datasets in our empirical tests: S&P 100 and Russell 1000 (US stock market, with databases composed by 102 and 907 stocks, respectively), and the Ibovespa Index (Brazilian stock market, dataset with 55 stocks) – all three indexes with data from January 2010 to September 2017. Additionally, we also estimate tracking portfolios for the Russell 2000 with a dataset comprising up to 1567 stocks; however, due to particularities regarding this index, in this case, we restrict our empirical analysis to the period from November 2014 to September 2018, and for this reason we present these results in Appendix A. Finally, to complement the empirical analysis with data from 2010 to 2017, we have also carried out similar tests for index tracking with data from 2002 to 2017 for the S&P 100, the Russell 1000 and the Ibovespa. The motivation for the analysis with this broader time frame results from the fact that this period contains the subprime mortgage crisis in 2007–2008 and therefore represents a situation with abnormal increasing market volatility. Thus, we estimated tracking portfolios for the three indexes mentioned above with data from 2002 to 2017 and included the results in an Electronic Appendix attached to this paper.[5]

As a result, we aim at exploring the lasso regression in different market contexts (US: a financial market with robust stability; and Brazil: a more volatile emerging market) and with varying sample sizes (datasets ranging from 55 to 907 stocks). Furthermore, as we use the Russell 1000, we also seek to analyze the capacity of lasso to solve a high-dimensional problem. According to the documentation on IT dealing with more substantial datasets, past studies usually employ optimization models that require some heuristic design to solve the problem more quickly. Thus, we carry out empirical tests based on the Russell 1000 as a tentative to assess the capacity of the lasso methodology to solve the index tracking problem with a more substantial dataset while avoiding the need to use some heuristic approach as done by other studies concerning IT. Lastly, since cointegration is one of the methods that has already been widely implemented in previous literature to solve the IT problem, we also used this approach in our empirical analysis so that we have a basis for comparison and validation of the results obtained using lasso.

Finally, we extrapolate index tracking and also use lasso to form portfolios based on the long-short strategy. This type of investment is known for its market-neutral characteristic, thus having low correlation with the market benchmark. Also, portfolios long-short are self-financed, which is done by short selling overvalued stocks and assuming long positions in undervalued stocks among the index constituents.[6] The most widely used approach to build long-short strategies is cointegration (Alexander, 1999; Alexander & Dimitriu, 2002, 2005; Li & Bao, 2014). Still, such method has a relevant drawback as it does not have the capacity of performing variable selection, thus requiring an ex-ante choice of the portfolio constituents. As a result, the use of lasso may be useful due to not only its more straightforward application (relative to cointegration, which requires a two-step approach) but also its capacity to select the stocks to compose each portfolio by choosing the most fitted coefficients in the linear regression.

For each of the three indexes selected for the empirical tests, we analyzed portfolios with two different sizes and three distinct updating frequencies. Overall, the lasso-based IT portfolios performed well concerning cumulative return and tracking error, particularly with the US benchmarks. Still, as we compare the results for index tracking obtained using lasso with those obtained using cointegration, we notice somewhat superior performance using cointegration in most of the cases. Nonetheless, despite having some disadvantage in performance relative to cointegrated portfolios, lasso-based portfolios have average monthly turnover at least 40% smaller than the average monthly turnover of portfolios using cointegration, which implies transaction costs at least about 40% lower. Such outcome is interesting since the reduced transaction costs fulfill an essential expectation regarding passive investments: to diminish costs while keeping a satisfactory overall performance (especially in the long run). Finally, the results for the long-short strategy also confirmed the quality of lasso, as we were able to obtain consistent cumulative returns noticeably with the S&P 100 and the Ibovespa.

Hence, the contribution of this paper is twofold. First, we add to the index tracking literature by widely testing a statistical model (lasso) that has only been used a few times in past research (and with limited empirical analysis). To expand previous studies, we adopt market benchmarks of different sizes (from 55 to 907 stocks) as well as from two financial markets with distinct characteristics (US and Brazil). Also, we estimate IT portfolios using an alternative approach (cointegration), so that we have a basis for the analysis and validation of the results obtained using lasso. Second, the empirical testing also presents innovations as we employ lasso to

[5] We have stored the Electronic Appendix at: https://www.dropbox.com/s/4qzv700jm4t60hp/Electronic_Appendix_Lasso-based_IT_NAJEF.zip?dl=0.

[6] In addition to this approach, long-short also may be developed by pairs trading or trading strategies that involve a stock and an ETF (Exchange-Traded Funds) (Avellaneda & Lee, 2010).

simulate the market neutral long-short strategy. Consequently, we also contribute to the finance literature focused on portfolio selection by showing how a different statistical approach can be consistently used for index tracking and long-short, considering the more substantial simplicity in the use of lasso relative to cointegration (which is a two-step method that requires a more extended analysis, as referred to in Section 3.2).

This study is organized as follows. Initially, Section 2 describes the method associated with the lasso-type regression. Then, Section 3 presents the methodology of the study, including the guidelines for the index tracking and long-short investing strategies, as well as the description of the cointegration approach based on simulations. Finally, Section 4 describes the results, and Section 5 concludes the study.

2. Lasso – least absolute shrinkage and selection operator

In this Section, we discuss the lasso-type regression methodology. First, Section 2.1 describes the general concepts regarding lasso. Second, Section 2.2 focuses on the guidelines concerning the *K*-fold cross-validation algorithm as a viable method to solve the lasso regression.

2.1. Lasso: general concepts

Konzen and Ziegelmann (2016) point out that the central goal of a linear regression analysis consists of estimating the coefficients for the model $y_i = \beta_0 + X_i^{\intercal}\boldsymbol{\beta} + \varepsilon_i$, where $y_i \in \mathbb{R}$ is the dependent variable to be predicted, $X_i = (x_{1i},...,x_{ki})^{\intercal} \in \mathbb{R}^k$ is the vector of independent variables, the union of β_0 and $\boldsymbol{\beta}$ is the set of predictors $(\beta_0, \beta_1,...,\beta_k)^{\intercal}$, and ε_i is the error term, in the context of a model with variables $j \in 1...k$, and time frame $i \in 1...N$. To compute such model, some approaches are available; among them, one of the most popular is OLS (Ordinary Least Squares), which is based on the minimization of the sum of the squared residuals (SSR), such that its estimator is calculated as follows:

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\beta_0,\beta_1,...,\beta_k}{\operatorname{argmin}} \sum_{i\in N} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 \tag{1}$$

However, as highlighted by Tibshirani (1996), the OLS approach has some inconsistencies[7] (particularly if we increase the number of independent variables and move to high-dimensional models), which might be overcome with the use of two specific techniques: subset selection and ridge regression. Still, both techniques have downsides as well.

In the case of subset selection, the procedure consists basically in the use of discrete choice to drop or add variables to the model as one aims at locating the best combination of input information for the model. Thus, the ideal situation in this case would be to test all 2^k possible combinations of the variables (Konzen & Ziegelmann, 2016). Yet, such analysis has a strong drawback in terms of computing time necessary to test all combinations.[8]

Concerning the ridge regression, Tibshirani (1996) points out its stability in terms of coefficients, in comparison to subset selection, as ridge regression consists of a continuous process that shrinks the regression coefficients. To carry out such process, the model receives a penalty on the sum of the squared residuals:

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\beta_0,\beta_1,...,\beta_k}{\operatorname{argmin}} \sum_{i\in N} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 \tag{2}$$

Subject to:

$$\sum_{j=1}^{k} \beta_j^2 \leqslant t \tag{3}$$

$$t \geqslant 0 \tag{4}$$

which is equivalent to:

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\beta_0,\beta_1,...,\beta_k}{\operatorname{argmin}} \left[\sum_{i\in N} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right] \tag{5}$$

In Eqs. (2)–(4), parameter $t \geqslant 0$ works as a control for the penalty, which means t has the same role as λ in Eq. (5). Thus, increasing λ in Eq. (5) strengthens the shrinkage process, while setting $\lambda = 0$ equalizes $\hat{\beta}_{Ridge}$ and $\hat{\beta}_{OLS}$. Different from subset selection, however, the ridge regression approach does not involve variable selection. As Nasekin (2013) emphasizes, regression analyses

[7] According to Tibshirani (1996), the OLS estimates has basically two issues: (1) prediction accuracy, which results in parameters with large variance, and (2) interpretation, which is the case specially in large models since the method does not perform variable selection and thus make the interpretation of the results more difficult and inaccurate.

[8] It is possible to find some algorithms in the literature to solve the subset selection problem, such as forward and backward elimination (Hastie, Tibshirani, & Friedman, 2009), and the Dantzig Selector (Candes & Tao, 2007).

usually face a situation where many independent variables are irrelevant for the model and may actually decrease its prediction power.

As a result, Tibshirani (1996) proposes the so-called lasso approach, which consists of a shrinkage method that aims at combining features from both the subset selection and the ridge regression. In this sense, lasso imposes a penalty on the coefficients (similar to the ridge regression); meanwhile, its estimating procedure works similarly to calculating the subset selection process continuously. Thus, the method results in the shrinkage of some of the coefficients while setting others to zero, achieving the final goal of performing variable selection in the regression model.

Tibshirani (1996) defines the lasso estimates in the form of the following optimization problem[9]:

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \sum_{i\in N} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 \tag{6}$$

Subject to:

$$\sum_{j=1}^{k} |\beta_j| \leqslant t \tag{7}$$

$$t \geqslant 0 \tag{8}$$

where the variables and parameters have the same definitions from the models for $\hat{\beta}_{OLS}$ and $\hat{\beta}_{Ridge}$. Additionally, we have the assumption that x_{ki} are standardized, thus resulting in $\sum_{i\in N} x_{ki} = 0$ and $\frac{\sum_{i\in N} x_{ki}^2}{N} = 1$ for each k. However, even though Eqs. (2) and (6) are similar, their Constraints (3) and (7) (which are applied on the penalty parameter t) are slightly different. As a consequence of Constraint (7), the optimization in Eqs. (6)–(8) takes the following form using the Lagrangian:

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \left[\sum_{i\in N} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right] \tag{9}$$

According to Tibshirani (1996) and Hastie et al. (2009), the model in Eq. (6) might be re-parametrized by standardizing the predictors, so that the solution for β_0 equals $\bar{y}$; thereby, we can suppose $\bar{y} = 0$, thus omitting β_0. Furthermore, in a similar way to the ridge regression, parameter t in Constraint (7) works as the penalty imposed on the coefficients. Nevertheless, while the ridge regression imposes a penalty of L_2 norm with $\sum_{j=1}^{k} \beta_j^2$, the lasso regression is characterized by a penalty of L_1 norm with $\sum_{j=1}^{k} |\beta_j|$ (Hastie et al., 2009). As a consequence, while the ridge regression makes use of a squared magnitude, the lasso regression considers the magnitude in absolute value. Therefore, a key feature resultant from the L_1 norm is its characteristic to generate sparse coefficients, i.e. provide coefficients equal to zero, thereby allowing the method to perform variable selection by picking only the most relevant variables. In contrast, the use of the L_2 norm does not have the sparsity property, thus not producing coefficients equal to zero as it happens in the lasso regression.

In Eqs. (6)–(8), as $t \geqslant 0$ represents the penalty on the coefficients and works as a control of the amount of shrinkage applied on the estimates, Tibshirani (1996) defines $\hat{\beta}_j^0$ as the full least square estimates (OLS coefficients) and $t_0 = \sum_{j=1}^{k} |\hat{\beta}_j^0|$. Therefore, setting $t \leqslant t_0$ leads to a shrinkage of the solutions in convergence to zero, with some coefficients equal to zero. On the other hand, for $t \geqslant t_0$, the lasso regression estimates will be equal to the OLS estimates. For instance, letting $t = t_0/2$ has the effect of (roughly) shrinking the OLS coefficients by 50% on average (Hastie et al., 2009; Konzen & Ziegelmann, 2016). For this reason, parameter t should be selected in a dynamic process to minimize an estimate of the expected prediction error.

Finally, concerning Eq. (9), it is worth noting that $\lambda = 0$ (in the same way as $t \geqslant t_0$) results in lasso coefficients equal to the OLS ones. Moreover, increasing λ implies a larger penalty that forces the coefficients to converge towards zero. Hence, the choice for λ (or, equivalently, the choice for t) becomes an important step for the lasso-type regression to achieve good quality results (Nasekin, 2013), and is related to estimating the prediction error. As Tibshirani (1996) emphasizes, one option is to choose the value of the penalty parameter to minimize the prediction error, which is based on the construction of a cross-validation style statistic. In this study, we opted to employ the K-fold cross-validation method since it is traditionally used in the literature (Hastie et al., 2009), as described in the next section.

2.2. K-fold cross-validation

Hastie et al. (2009) describe the K-fold cross-validation as the simplest method to estimate the prediction error. As Efron and Tibshirani (1993) emphasize, in a regression model, the prediction error (PE) consists of the expected squared difference between a variable y_i and its prediction $\hat{y}_i$: $\mathrm{PE} = \mathrm{E}(y_i - \hat{y}_i)^2$. Then, the mean squared error (MSE) in-sample is defined as follows:

$$\mathrm{MSE} = \frac{1}{N} \sum_{i\in N} (y_i - \hat{y}_i)^2 \tag{10}$$

[9] To keep the description of the lasso-type regression short, we omit the explanation regarding the properties of $\hat{\beta}_{lasso}$. For instance, we refer the reader to Zhao and Yu (2006) and Konzen and Ziegelmann (2016) for a complete description of the lasso estimator consistency.

However, a more robust application would be to split the data into training and testing samples, thus using the fitted model from the training sample to estimate the MSE of the testing sample (Efron & Tibshirani, 1993; Tibshirani, 1996). Based on this idea, Efron and Tibshirani (1993) presented the following Algorithm 1 for cross-validation:

Algorithm 1 K-fold Cross-validation (Efron and Tibshirani, 1993)

Step 1: Split the data into K roughly equal-sized parts
Step 2: For the k-th part, fit the model to the other $K-1$ parts of the data, and calculate the prediction error of the fitted model when predicting the k-th part of the data
Step 3: Do the above for $k \in 1, \ldots, K$ parts, and combine the K estimates of prediction error

For instance, if we set $K = 5$, then for each $k \in 1...K$, the model will be fitted for the data of all $K-1$ parts, and the fitted model will be used to verify the MSE of the k-th part of the sample. As described by Efron and Tibshirani (1993), if we let $k(i)$ be the part containing the i-th observation of the data, and define $\hat{y}_i^{-k(i)}$ as the fitted value for the i-th observation (estimated with the fitted model and the $k(i)$-th part of the data removed), then the cross-validation estimate for the prediction error (or cross-validated MSE) will be as follows:

$$\text{CVMSE} = \frac{1}{N}\sum_{i \in N}(y_i - \hat{y}_i^{-k(i)}). \tag{11}$$

So, in the lasso-type estimation, the K-fold cross-validation is used to compute the CVMSE statistic in Eq. (11) employing different values for λ. Hence, the chosen value for λ will be the one that results in the least value for the cross-validation error. Fig. 1 illustrates the process, where the y-axis represents the CVMSE. As λ increases (x-axis), the results present an increasing number of coefficients equal to zero, which tends to lead to larger error, and the best value for λ, as already mentioned, is the one that minimizes the cross-validated error – identified by the vertical dotted line in the figure. In our basic simulation to generate this example, as well as in our empirical tests described in Section 4, we use $K = 10$, i.e. 10-fold cross-validation, based on Breiman and Spector (1992) and Kohavi (1995), who claim that $K = 5$ or $K = 10$ are satisfactory choices to solve the lasso-type regression in general cases.

3. Methodology of the study

First, this Section describes the basic methodology for the portfolio selection using both index tracking and long-short investing strategies (Sections 3.1.1 and 3.1.2). Later, Section 3.2 describes the essential guidelines to solve the index tracking portfolio selection using cointegration.

3.1. Index tracking and long-short investing strategies

3.1.1. Index tracking

As mentioned in the Introduction, the index tracking problem is commonly presented with a constraint on the size of each portfolio, due to the increasing costs that would result from a full index replication. Consequently, IT models are usually evaluated by their tracking error (TE), which is defined according to Eq. (12) (Beasley, Meade, & Chang, 2003; Guastaroba & Speranza, 2012):

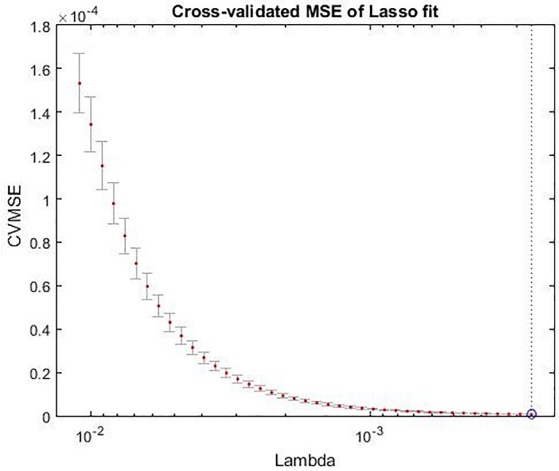


Fig. 1. Cross-validated MSE (CVMSE) of lasso fit.

$$TE = \frac{1}{T}\left[\sum_{t=1}^{T} (r_t^p - R_t)^2\right]^{1/2}, \tag{12}$$

where T is the time frame (for instance, one month), $t \in 1...T$ corresponds to each business day in our dataset, r_t^p is the portfolio daily return, and R_t is the index daily return.

Regarding the lasso regression, the index tracking problem is implemented as follows. The dataset contains a time series of daily log returns for the market index and N stocks, where r_{jt}^l represents the daily log return of the j-th stock on the t-th day, and R_t^l represents the index daily log return. Then, we implement Eq. (9) in the following equivalent form:

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta_0,\beta_1,\ldots,\beta_N}{\operatorname{argmin}} \left[\sum_{t\in T}\left(R_t^l - \beta_0 - \sum_{j=1}^{N} \beta_j r_{jt}^l\right)^2 + \lambda \sum_{j=1}^{N} |\beta_j|\right], \tag{13}$$

where $R_t^l = \log(P_t/P_{t-1})$, P_t is the index price on the t-th day, $r_{jt}^l = \log(p_{j,t}/p_{j,t-1})$, and p_{jt} is the stock price of the j-th stock, $j \in 1...N$.

The value for λ is computed using K-fold cross-validation in line with Algorithm 1; as already mentioned, we choose $K = 10$, i.e. 10-fold cross-validation. After computing Eq. (13), the IT portfolio is defined by normalizing the coefficients β_j, $j \in 1...N$, to sum up to one; as a result, the stock weight of the j-th asset in the portfolio equals the normalized value of the j-th coefficient.

Finally, concerning the lasso predictors, we set up two definitions. First, we impose a constraint on the number of lasso coefficients that may take value different from zero, which means to restrict the size of each portfolio. Second, in contrast with prior literature (Wu et al., 2014; Yang & Wu, 2016), we do not impose a nonnegative constraint on the coefficients β_j, which means we allow the tracking portfolios to have short positions. In this case, since short positions are often associated with cost issues because it depends on the amount of stocks available for rent, our results already account for transaction costs as explained in Section 4.1.

3.1.2. Long-short market neutral portfolios

Based on the indexing approach introduced in the previous section, we now explore the long-short strategy by trading on two tracking portfolios simultaneously, instead of building only one portfolio as it is done for index tracking. An equity statistical arbitrage long-short strategy is an investing strategy, used primarily by hedge funds, that involves buying one equity portfolio and shorting another. Specifically, a long-short strategy takes long positions in stocks that are expected to increase in value and short positions in stocks that are expected to decrease in value. They are often sold as a way to add a premium with special diversification benefits that arise because the premium is not highly correlated with the rest of an investor's equity portfolio.

The long-short strategy may be viewed as a natural extension of the index tracking strategy. First, two new indexes are constructed to mimic the evolution of the original index; however, one of those new indexes has a higher yield at the end of the period ("plus benchmark") and the other has a lower yield ("minus benchmark"). Next, for each of the new indexes, two portfolios are built: the "plus portfolios" and "minus portfolios", which are then combined in a long-short strategy that has a low correlation with the market returns. Consequently, these portfolios are expected to generates "double alpha" according to the plus/minus spread and (ideally) exhibit low volatility. Moreover, as the spread is considered to be uncorrelated with the market returns, this statistical arbitrage strategy will generate a market-neutral portfolio.

Once the indexes plus and minus are constructed, by adding to and subtracting from the benchmark returns an annual excess return that will be uniformly distributed to daily returns (α%), the portfolios that track the "plus" and the "minus" benchmarks are estimated using the lasso approach, as in the index tracking strategy. So, we use Eq. (13) to estimate two portfolios, the first of them using the index plus y_i^+ instead of the original index time series, and the second one using the index minus y_i^-, as follows:

$$\hat{\boldsymbol{\beta}}^+ = \underset{\beta_0^+,\beta_j^+}{\operatorname{argmin}} \left[\sum_{i=1}^{N}\left(y_i^+ - \beta_0^+ - \sum_{j=1}^{k} \beta_j^+ x_{ji}\right)^2 + \lambda \sum_{j=1}^{k} |\beta_j^+|\right], \tag{14}$$

$$\hat{\boldsymbol{\beta}}^- = \underset{\beta_0^-,\beta_j^-}{\operatorname{argmin}} \left[\sum_{i=1}^{N}\left(y_i^- - \beta_0^- - \sum_{j=1}^{k} \beta_j^- x_{ji}\right)^2 + \lambda \sum_{j=1}^{k} |\beta_j^-|\right], \tag{15}$$

where y_i^+ and y_i^- represent the two artificial plus and minus benchmarks that are tracked.

Finally, as in the case of index tracking, the stock weights are obtained after normalizing the coefficients to sum up to one. Then, the outcomes will be two portfolios (plus and minus), and the final weight of the i-th stock in the long-short portfolio will be the difference between its weights in the portfolios plus and minus. Naturally, the larger the excess return α, the more difficult it will become to construct the portfolios, which might result in increasing portfolio volatility over time.

According to Alexander and Dimitriu (2005), the conceptual background that supports the choice for long-short strategy is its self-financing characteristic, since investing in the long-short portfolio is the equivalent to selling the short portfolio to obtain the resources necessary to buy the long portfolio, hence forming a zero-sum portfolio. Then, the implementation of the long-short equity strategy in practice requires, most importantly, the ability of the investor involved to identify correctly a set of mispriced securities, at which point the lasso-type regression can play an important role due to its capacity concerning variable selection.

3.2. Cointegration approach based on simulations for index tracking

The concept of cointegration was introduced by Granger (1981) in time-series analysis and formalized by Engle and Granger (1987). Since then, empirical studies (for instance, Alexander, Giblin, & Weddington, 2002; Alexander & Dimitriu, 2005) have shown that financial assets can be found to be cointegrated quite often, and this has actually motivated an alternative approach to equity trading and portfolio construction. By using all information embedded in prices, it may be possible to detect a long run equilibrium between a portfolio and a benchmark which can be used to indicate the optimal strategic asset allocation.

Cointegration is a statistical feature defined formally by Hamilton (1994) as follows: if there is an $(n \times 1)$ vector time series $\boldsymbol{y_t}$, then this time series is defined as cointegrated if (i) each of the time series taken individually is non-stationary with unit root, i.e. $I(1)$ and (ii) there is a non-zero $(n \times 1)$ vector a such that some linear combination of the series $\boldsymbol{a'y_t}$ is stationary, i.e. $I(0)$.

When applied to prices in a stock market index, cointegration occurs when there is at least one portfolio of stocks that has a stationary tracking error, e.g., when there is mean reversion in the price spread between the portfolio and the index. This property does not provide any information for forecasting the individual prices in the system, or the position of the system at some point in the future, but it provides the valuable information that, irrespectively to its position, the prices of the portfolio and the index are expected to stay together in the long run.

The design for the use of cointegration in asset allocation is based on a two-step approach as follows. The first step for the selection of a tracking portfolio requires the analysis to confirm that each price series is $I(1)$ (i.e. the price series of the index and each stock in the sample) in a predefined time frame of in-sample data. Then, we estimate a linear regression for the same time frame of in-sample data to infer the portfolio weights. The most used approach to solve the linear regression is OLS (ordinary least squares), and the linear regression is estimated according to Eq. (16):

$$\log(P_t) = \beta_{0,t} + \sum_{i=1}^{n} \beta_{i,t} \log(p_{i,t}) + \varepsilon_t \tag{16}$$

where P_t denotes the index price on the t-th day, $p_{i,t}$ denotes the stock price of the i-th stock, $i \in 1...N$, and ε_t is a zero-mean "tracking error". By normalizing the cointegration coefficients β_i, $i \in 1...N$, to sum up to one, we determine the proportional weights of the i-th stock in the portfolio.

The second step is to apply the unit root test on the series of residuals $\hat{\varepsilon}_t$ resulting from Eq. (16) to confirm that the linear combination of the price series of N stocks $I(1)$ is a stationary time series with order $I(0)$. To confirm if such linear combination is stationary, we apply the Augmented Dickey-Fuller (ADF) test on $\hat{\varepsilon}_t$ to test the null hypothesis of no cointegration. If we let q be the order of the autoregressive (AR) process, $\hat{\varepsilon}_t$ be the estimated error term from Eq. (16), and $\Delta\hat{\varepsilon}_t$ be the change between two error terms, then the ADF regression takes the following form:

$$\Delta\hat{\varepsilon}_t = \gamma\hat{\varepsilon}_{t-1} + \sum_{i=1}^{q} \phi_i \Delta\hat{\varepsilon}_{t-i} + u_t. \tag{17}$$

By rejecting the null hypothesis, we confirm the time series of estimated residuals is stationary, thereby attesting that the variables used on the regression are cointegrated. We consider the critical values suggested by MacKinnon (2010) at 1% level of significance for the ADF test. Then, as the null hypothesis is rejected, the portfolio obtained from Eq. (16) consists in a valid portfolio to track the market benchmark.

Finally, as described by Alexander and Dimitriu (2005), cointegration fits in the context of portfolio selection and IT strategy due to its features as an appropriate method to detect long run asset price dynamics. However, a drawback of past studies lies in the issues relative to asset selection to compose each portfolio, which is usually exogenous to the portfolio optimization process, since the OLS method does not make variable selection. For instance, some studies make the selection of the stocks to compose the tracking portfolio based on the weights of the stocks in the index composition, in which case the portfolios have the stocks with the largest weights (for instance, Alexander & Dimitriu, 2005; Dunis & Ho, 2005). In this case, if we would like to form a 10-stock cointegrated tracking portfolio for the DJIA index (composed by 30 stocks), such portfolio would be built with the 10 stocks with the largest weights in the index composition. Nonetheless, selecting the top 10 stocks with the largest weights might become a tricky choice in the case of broader indexes that have lower concentration, such as the S&P 100 or the Russell 1000, as already discussed.

Under those circumstances, considering we would like to form a tracking portfolio with s stocks out of a sample with N available stocks (which are the index constituents), the ideal option would be to test all possible combinations of s stocks to choose the best portfolio. For example, to build a 10-stock portfolio using a sample with 100 stocks, it is necessary first to select the 10 stocks to estimate the cointegration analysis. Hence, the best-case scenario would be to test each possible combination of 10 stocks and choose the best one according to some criteria. Yet, such analysis would face a challenge specially in terms of computing time, since the number of all possible combinations of 10 stocks in this case would be $1.73E^{13}$.

Then, since making the selection of the stocks to compose each tracking portfolio can be a difficult task, we attempt to mitigate this problem by using a series of random simulations to select the components of each cointegrated portfolio, thus making the stock selection endogenous to the solving process (not an *ex-ante* choice). In this random process, first we form a sequence of M different portfolio candidates for each in-sample subset, where each portfolio candidate is composed by s stocks randomly selected (thus, s corresponds to the size of each portfolio). Second, after constructing M different portfolio candidates and discarding the ones that do not meet the cointegration requirements previously described, we select the portfolio whose estimation of Eq. (16) resulted in the

smallest mean squared error (Eq. (10)).[10]

4. Empirical analysis

First, Section 4.1 presents the details regarding the databases and the background definitions to compute the tests. Then, Section 4.2 discusses the results for index tracking with S&P 100 and Ibovespa, and Section 4.3 discusses the results for a high-dimensional dataset: index tracking with the Russell 1000. Later, Section 4.4 describes the comparison between the results for index tracking using lasso and cointegration, and Section 4.5 sums up the findings regarding index tracking with a discussion of the results. Finally, Section 4.6 shows the results for long-short strategy using lasso.

4.1. Database and testing setup

For the empirical analysis, we choose three datasets. The first dataset refers to the S&P 100 (one of the main benchmarks in the US market) and is composed by 101 stock price series plus the index price series; the second dataset refers to the index Ibovespa (the main index in the Brazilian stock market) and is composed by 55 stock price series; finally, the third dataset refers to the Russell 1000 (which is formed by approximately the 1000 largest firms listed in the US stock market) and is composed by 907 stock price series. The datasets referring to the S&P 100 and the Ibovespa were extracted from software Economatica, a financial database widely used in Brazil by both market participants and academicians; in the meantime, the dataset referring to the Russell 1000 was obtained from Compustat. Our database includes daily stock prices from January 2010 to September 2017, with a total number of 1921 data points (each data point being one business day). Prices are adjusted for (i) splits, mergers, and other corporate actions and (ii) dividend payments. Additionally to the three indexes mentioned above, we also carried out an analysis using the Russell 2000, which is an index composed mostly by small- and mid-cap firms listed in the US stock market, and presents larger volatility in comparison with the S&P 100 and the Russell 1000. However, since we could not reconstruct the index due to the lack of information regarding its historical composition, in addition to the fact that there is a significant turnover each time this index is updated, our choice was to select a dataset starting in Nov 2014, resulting in a shorter time window in comparison to the other three benchmarks. The description of the analysis for the Russell 2000 is presented in Appendix A.

Additionally, as a complement to the analysis done in this paper with data from 2010 to 2017, we also present an Electronic Appendix in which we describe empirical tests for the S&P 100, the Ibovespa and the Russell 1000 with data from 2002 to 2017, thus extending our time frame to roughly 16 years. Given that we could not reconstruct any of these indexes due to the lack of data regarding their historical composition (which is the same situation described above for the Russell 2000), all tracking analysis carried out in this paper was based on the structure of each index in 2018. Nonetheless, considering that the use of an extended time frame starting in 2002 is very interesting as this period involves considerably higher market volatility (including the subprime mortgage crisis in 2007–2008), we have included the description of the results with data from 2002 to 2017 in an Electronic Appendix attached to this paper.[11]

For each dataset, we select two sizes for the tracking portfolios. To track the S&P 100, we form portfolios limited to 15 and 25 stocks; regarding the Ibovespa index, we estimate portfolios up to 8 and 12 stocks; finally, regarding the Russell 1000, we form portfolios limited to 20 and 30 stocks. Additionally, we select in-sample intervals equal to 480 data points (similar to Alexander & Dimitriu, 2002), whereas out-of-sample intervals equal 60, 120, and 240 data points (which means to perform portfolio updates roughly every three months, six months, and one year – i.e. quarterly, semiannual, and annual updates).

As a result, the first portfolio will be obtained by estimating a regression with data from $t = 1$ to $t = 480$, and its results will be observed over the data for the next 60, 120, or 240 data points in a rolling horizon framework. Then, if we consider the case with quarterly updates for instance, the second portfolio will be formed with data form $t = 61$ to $t = 540$, and so on. Consequently, we obtain 24 portfolios to cover all the dataset interval in the case of quarterly updates, 12 portfolios in the case of semiannual updates, and 6 portfolios for annual updates. Moreover, we also consider a buy-and-hold case in which we do not update the portfolios over time, i.e. the first portfolio is optimized with data from $t = 1$ to $t = 480$ and held until the last day in the sample.

Concerning the lasso-type regression, the empirical analysis consists in evaluating Eq. (13) with index and stocks daily returns (in natural logarithm). In contrast, the tests based on cointegration are estimated with index and stocks daily prices (also in natural logarithm), as described in Section 3.2.[12]

Finally, we highlight that the results presented in the next Sections already account for transaction costs as we use Eq. (18) to compute the daily returns in the rolling window projections (Han, 2005; Do & Faff, 2012):

[10] In this study, we select parameter $M = 50{,}000$, so that we form 50 thousand distinct portfolios to select the best one based on the sum of the squared residuals. The choice for 50,000 is due to the fact that this was the maximum number of different combinations that we were able to form. As M increases, there is a larger use of physical memory (RAM) by the CPU, thereby imposing a limit on this parameter.

[11] We have stored the Electronic Appendix at: https://www.dropbox.com/s/4qzv700jm4t60hp/Electronic_Appendix_Lasso-based_IT_NAJEF.zip?dl=0.

[12] Concerning the use of log returns to estimate the lasso-based tracking portfolios, we refer the reader to Appendix B, where we present the results for an empirical analysis in which we substitute log returns for simple returns.

$$r_{i,t} = \log\left(\frac{p_{i,t}}{p_{i,t-1}}\right) + \log\left(\frac{1-C}{1+C}\right) - d \tag{18}$$

where C represents the transaction costs, and d refers to the costs related to short positions. In our empirical tests, we set $c = 0.5\%$ (which refers mainly to brokerage fees), and $d = 2\%$ per year (which refers basically to rental costs). Both costs are discounted from the return of stock i every day the portfolio is updated.

4.2. Index tracking using lasso – indices S&P 100 and Ibovespa

We start the empirical analysis using lasso regression to solve the index tracking problem for S&P 100 and Ibovespa. The tracking error (TE) was evaluated as in Eq. (12), and the portfolios were compared using the following performance measures: (i) Annual average returns; (ii) Cumulative returns; (iii) Annual volatility; (iv) Daily TE average; (v) Daily TE volatility; and (vi) Monthly average turnover, which is defined as follows:

$$\frac{\sum_{p=2}^{np}\left(\frac{\sum_{i=1}^{N}|x_i^p - x_i^{p-1}|}{2}\right)}{np-1} \times \frac{1}{f} \tag{19}$$

where np is the number of portfolios estimated per portfolio size and updating frequency (for instance, considering quarterly updates, we form a total of 24 portfolios), p and $p-1$ are time instants where sequential rebalancing were carried out, and f equals 3 for quarterly rebalancing, 6 for semiannual rebalancing, and 12 for annual rebalancing.

The results are in Table 1, Figs. 2 and 3. Regarding the S&P 100, we can initially notice in Table 1 the good quality of the results in terms of tracking performance specially in the case of portfolios up to 15 stocks with quarterly update, and up to 25 stocks with semiannual update, as they presented cumulative returns very close to the index. Also, we can observe the outstanding results of portfolios buy-and-hold, considering that these portfolios are held constant throughout the entire out-of-sample interval (roughly 5.5 years); in both cases (portfolios up to 15 and 25 stocks), the choice for buy-and-hold results in annual average returns (respectively 12.43% and 12.30%) very close to the index average annual return (11.13%).

Additionally, increasing the size of the portfolios from 15 to 25 stocks results in smaller portfolios' average tracking error for all updating frequencies (except quarterly), as it would be naturally expected (intuitively, larger portfolios should track the index more accurately). Moreover, increasing the size of the portfolios also results in larger correlation with the benchmark index and smaller average monthly turnover.

Lastly, as an example, Fig. 2a shows the cumulative value of tracking portfolios up to 15 and 25 stocks with semiannual updates. In this case, the figure exhibits the good quality in terms of tracking performance in both cases, as they remain very close to the index over time. Moreover, we can also reach out this conclusion by observing the monthly return of those portfolios in Fig. 3a; here, we notice small detachments from the index in a few months, specially Dec 2012, Jul 2016, Aug 2016 and Oct 2016.

Regarding the Ibovespa, first we highlight the considerably larger volatility of the Brazilian index in comparison with the S&P 100. In fact, Table 1 shows that the Ibovespa has annual volatility equal to 23.04%, almost twice as large as the annual volatility of the S&P 100 (12.47%). The consequence of such volatility is noticed in the portfolios' average tracking error, where the values for the Ibovespa tracking portfolios are in general twice as large as the values of the portfolios tracking the S&P 100. Nonetheless, we may also see the good quality results for Ibovespa tracking portfolios in terms of cumulative returns, specially in the case of portfolios up to 8 and 12 stocks with semiannual and annual updating frequency. In those cases, the difference between the portfolio's cumulative return and the index's cumulative return remains below 10 percentage points.

Furthermore, we point out to the fact that increasing the number of stocks in the portfolio results in smaller values for portfolios' average tracking error, annual volatility and average monthly turnover, as well as larger correlation with the index. Such results are in line with the conclusions drawn from the tracking portfolios for the S&P 100.

Finally, Figs. 2b and 3b allow one to observe the quality of the results for the Ibovespa in the case of cumulative returns as well as monthly returns, considering semiannual updating frequency. Concerning cumulative return, it is noticeable the increasing detachment of the portfolio from the index specially between June 2015 and July 2016, with a peak in the difference between monthly returns in February 2016, as shown in Fig. 3b. However, these findings are mitigated by the fact that, from June 1st 2015 to Jan 26th 2016, the index had a cumulative return of -29.29%, then bouncing back to accumulate 52.83% from Jan 26th 2016 to July 29th 2016. Such strong volatility shows how difficult it was for tracking portfolios to follow the index during this period and justifies the separation between monthly returns of the index and portfolios from June 2015 and August 2016.

4.3. Index tracking in a high-dimensional dataset – index Russell 1000

In the previous Section 4.2, we described the results for the tracking portfolios using lasso regression and two market benchmarks: S&P 100 and Ibovespa. However, neither of those indexes is composed by a large number of stocks (101 stocks concerning the dataset for the S&P 100, and 55 stocks for the Ibovespa). In contrast, according to the literature on the lasso methodology (for example, Tibshirani, 1996; Nasekin, 2013; Konzen & Ziegelmann, 2016), a relevant characteristic of this statistical approach is its capability to

Table 1
Overall results for index tracking using lasso – **S&P 100** and **Ibovespa**.[1]

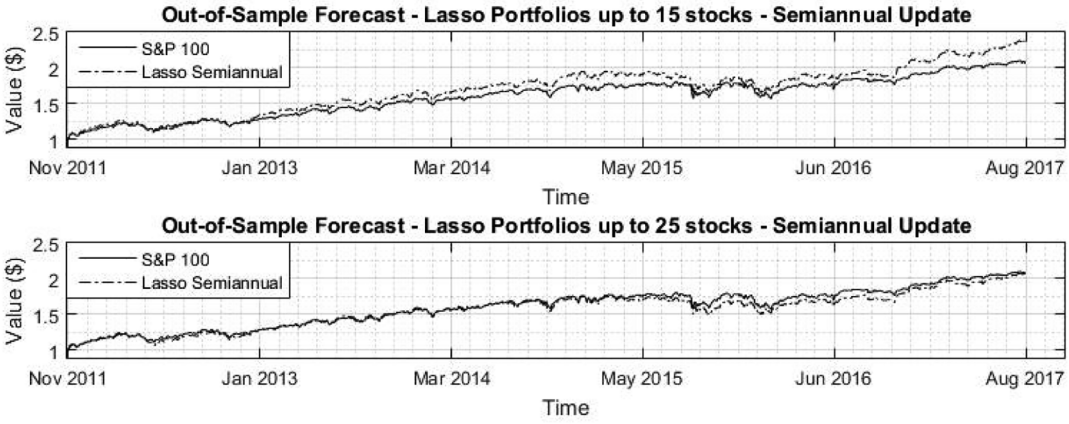
		Portfolios up to 15 stocks			
	S&P 100	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.13%	11.12%	13.73%	13.95%	12.43%
Cumulative Return	106.04%	103.08%	138.48%	140.38%	119.62%
Portfolios' Average Tracking Error	–	0.043%	0.030%	0.021%	0.051%
Annual Volatility	12.47%	14.54%	14.45%	14.48%	14.51%
Correlation	–	0.931	0.936	0.938	0.939
Average Monthly Turnover	–	6.05%	4.31%	3.29%	0.00%
		Portfolios up to 25 stocks			
	S&P 100	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.13%	8.02%	11.61%	12.95%	12.30%
Cumulative Return	106.04%	66.84%	109.78%	127.74%	118.35%
Portfolios' Average Tracking Error	–	0.043%	0.026%	0.017%	0.043%
Annual Volatility	12.47%	14.48%	14.16%	14.14%	14.17%
Correlation	–	0.932	0.949	0.956	0.956
Average Monthly Turnover	–	5.70%	4.27%	3.19%	0.00%
		IBOVESPA			
		Portfolios up to 8 stocks			
	Ibovespa	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	5.42%	5.71%	8.32%	8.33%	9.55%
Cumulative Return	29.98%	14.43%	36.80%	35.44%	37.51%
Portfolios' Average Tracking Error	–	0.084%	0.060%	0.044%	0.115%
Annual Volatility	23.04%	29.22%	29.18%	29.60%	29.25%
Correlation	–	0.943	0.942	0.941	0.930
Average Monthly Turnover	–	4.88%	3.63%	2.94%	0.00%
		Portfolios up to 12 stocks			
	Ibovespa	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	5.42%	3.92%	6.23%	7.58%	8.95%
Cumulative Return	29.98%	4.32%	20.75%	30.12%	40.76%
Portfolios' Average Tracking Error	–	0.072%	0.050%	0.035%	0.094%
Annual Volatility	23.04%	28.03%	27.83%	27.88%	28.22%
Correlation	–	0.954	0.957	0.958	0.953
Average Monthly Turnover	–	4.46%	3.24%	2.49%	0.00%

[1] Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Portfolios' Average Tracking Error refers to the average of the tracking error calculated for each portfolio according to Eq. (12). Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Correlation refers to the correlation between daily returns of each strategy and daily returns of the index during the entire out-of-sample period. Average Monthly Turnover is calculated according to Eq. (19).

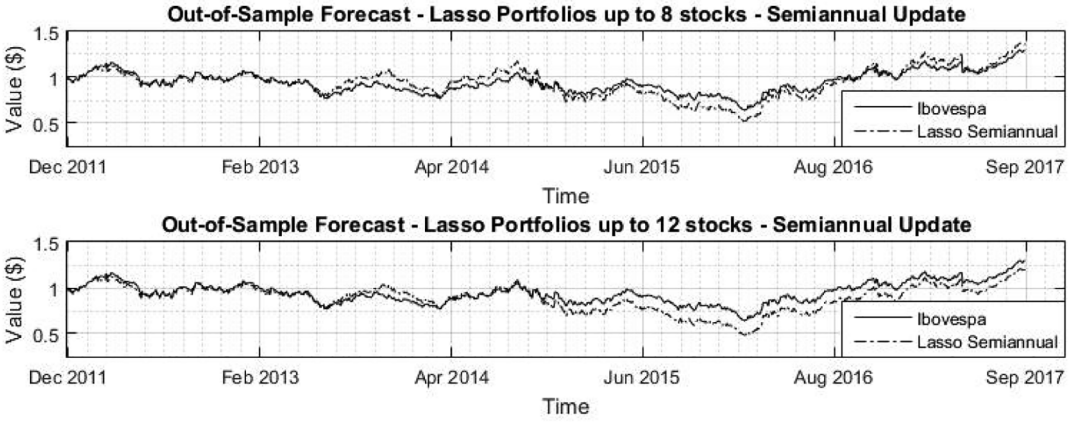
solve high-dimensional problems. Such feature is a result of the capacity of lasso to perform variable selection through its penalty function that is imposed on the coefficients, which leads the model towards a shrinkage process that selects only the most relevant coefficients in the regression.

For this reason, we also opted to carry out an empirical analysis of index tracking using two larger market benchmarks: the Russell 1000 and the Russell 2000. The Russell 1000 is theoretically composed approximately by the 1000 largest firms listed in the US equity market (the top 1000 firms in the Russell 3000, a capitalization-weighted index); the Russell 2000 is formed by 2000 small- to mid-cap firms listed in the US stock market (the bottom 2000 firms in the Russell 3000). In our specific analysis, the dataset for the Russell 1000 has a total of 907 stocks, thereby imposing a challenge for the index tracking problem since the Russell 1000 constituents have minimal concentration in the index portfolio, as mentioned in the Introduction. In the meantime, the dataset for the Russell 2000 has 1567 stocks and a larger volatility in comparison to the Russell 1000.

In this Section, we focus on the results for the Russell 1000, meanwhile our findings for the Russell 2000 are described in Appendix A (since we had to adopt a dataset with a shorter time window for this index as already mentioned). We describe the results for the tracking portfolios for the Russell 1000 in Table 2 and Fig. 4. Initially, we can infer from Table 2 once again the good quality of the tracking solutions in terms of both the average annual returns and the cumulative returns. In the case of portfolios up to 30 stocks using quarterly updates, the cumulative returns are very low and the tracking performance is poorer relatively to the other updating



(a) S&P 100



(b) Ibovespa

Fig. 2. Out-of-sample forecast per index and portfolio updating frequency.

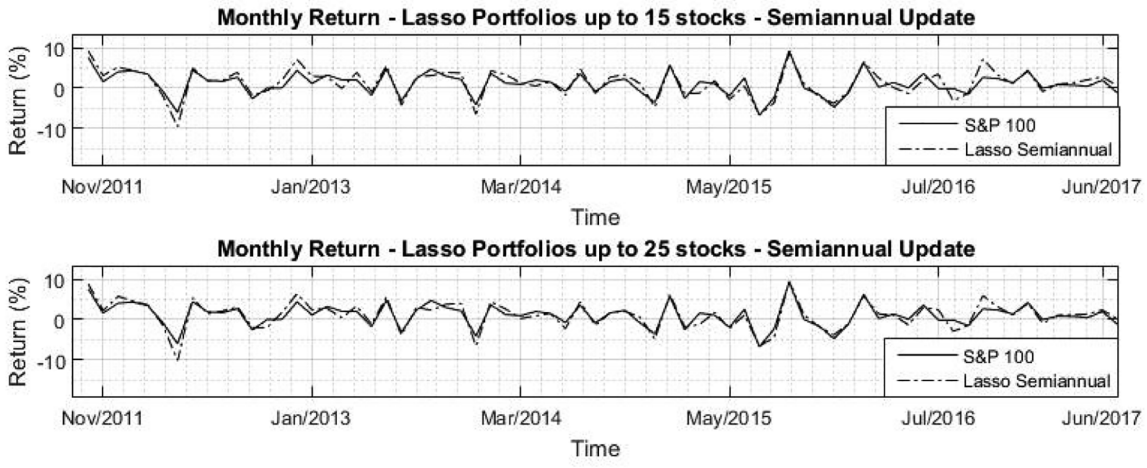
frequencies, since the more frequent portfolio updates combined with the larger size of the portfolios resulted in increasing transaction costs that penalized the portfolio's performance. However, as the update interval increases, the results become consistent for all remaining portfolios (semiannual and annual updates, as well as the buy-and-hold strategy).

Regarding quarterly updating portfolios up to 20 stocks, the results show average annual returns of 11.91% and cumulative return equal to 111.16%, which is a difference smaller than 1 percentage points from the index returns. Moreover, we notice better performance as we increase the limit size of the tracking portfolios, as it would be expected (except for quarterly updating frequency); in the case of semiannual and annual intervals, we see a significant increase in performance in the cumulative returns, specially for semiannual updates. Also, increasing the size of each portfolio (per updating frequency) resulted in lower portfolios' average tracking error and annual volatility, as well as larger correlation with the index (again, except for quarterly updates). Such findings are in accordance with the results for the S&P 100 and the Ibovespa, where we also obtained slightly better performance with larger tracking portfolios.

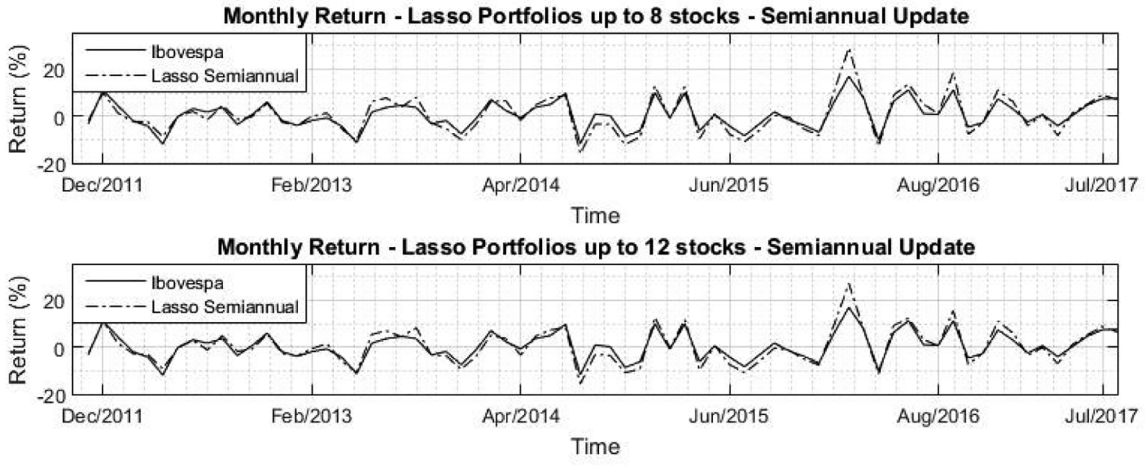
Finally, the results relative to the Russell 1000 are also introduced in Fig. 4a and b, that show respectively the cumulative performance and the monthly returns of tracking portfolios using semiannual updates (Fig. 4a). Regarding the cumulative performance, we notice good tracking results specially for the portfolios limited to 30 stocks; in this case, we see in Fig. 4b more significant differences between monthly returns of the portfolios and index mainly in August 2016 and October 2016. Nonetheless, the performance remains adequate during the remaining sample interval, which highlights the good quality of the tracking results obtained using lasso for a high-dimensional dataset.

4.4. Validation of the lasso-type regression: comparison with cointegration based portfolios

As discussed in the previous sections, the application of lasso to solve the index tracking problem resulted in promising conclusions regarding the capacity of this method to perform portfolio selection. Still, a comparison with another statistical method



(a) S&P 100



(b) Ibovespa

Fig. 3. Monthly return per index and portfolio updating frequency.

might be useful as an attempt to shed some new light in the discussion related to the previous findings. So, due to the extensive use of cointegration in the previous literature on index tracking, we also opted to estimate the tracking portfolios using this method, as we sought to have a basis for comparison and validation of the results obtained using lasso.

To carry out the cointegration tests, we followed the methodology described in Sections 3.2 and 4.1. In addition, we highlight that the use of OLS to estimate the regression in Eq. (16) and obtain each cointegrated portfolio would most likely result in negative and positive OLS estimates, i.e. long and short positions in each portfolio. Nevertheless, none of the portfolios obtained using lasso presented short positions. For this reason, we chose to estimate cointegration using non-negative least squares (OLS-NN) instead of OLS, thereby avoiding short positions (negative regression coefficients) in each cointegrated portfolio, as the use of short positions is associated with larger transaction costs that can potentially difficult the portfolios' tracking performance.

The results for cointegration (hereafter, referred to as OLS-NN) and lasso are described in Table 3, Figs. 5 and 6. Initially, Table 3 has a summary of the results using both methods for each of the three indexes. In the case of the S&P 100, we observe mixed results if we compare portfolios based on lasso and OLS-NN with the same size and rebalancing frequencies. Regarding average annual returns and cumulative returns, all lasso-based portfolios presented results closer to the index than portfolios OLS-NN (except in the case of portfolios up to 15 stocks and semiannual rebalancing interval). In contrast, all portfolios OLS-NN presented smaller average tracking error and somewhat lower volatility, which implies OLS-NN was able to produce portfolios less risky consistently over time.

In the case of tracking portfolios for the Ibovespa index, the results present a pattern quite similar to the findings regarding the S&P 100. In terms of performance represented by the average annual returns and cumulative returns, lasso-based portfolios up to 12 stocks produced superior results in comparison to portfolios OLS-NN. However, in the case of smaller portfolios (up to 8 stocks), we obtained better results using the cointegration approach (except for the buy-and-hold strategy). Concerning the portfolios' average

Table 2
Overall results for index tracking using lasso – **Russell 1000**.[1]

	Russell 1000				
		Portfolios up to 20 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.53%	11.91%	15.22%	15.18%	13.49%
Cumulative Return	110.24%	111.16%	158.59%	159.55%	135.42%
Portfolios' Average Tracking Error	–	0.049%	0.033%	0.023%	0.059%
Annual Volatility	12.70%	15.76%	15.56%	15.41%	15.26%
Correlation	–	0.927	0.934	0.936	0.926
Average Monthly Turnover	–	8.21%	6.27%	4.74%	0.00%
		Portfolios up to 30 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.53%	7.98%	13.58%	13.91%	13.12%
Cumulative Return	110.24%	64.84%	135.05%	140.63%	129.08%
Portfolios' Average Tracking Error	–	0.052%	0.031%	0.020%	0.049%
Annual Volatility	12.70%	15.69%	15.12%	14.90%	14.69%
Correlation	–	0.915	0.938	0.947	0.947
Average Monthly Turnover	–	8.40%	6.29%	4.50%	0.00%

[1] Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Portfolios' Average Tracking Error refers to the average of the tracking error calculated for each portfolio according to Eq. (12). Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Correlation refers to the correlation between daily returns of each strategy and daily returns of the index during the entire out-of-sample period. Average Monthly Turnover is calculated according to Eq. (19).

tracking error and annual volatility, again we observe portfolios OLS-NN with more accurate results in all cases, in line with the findings for the S&P 100.

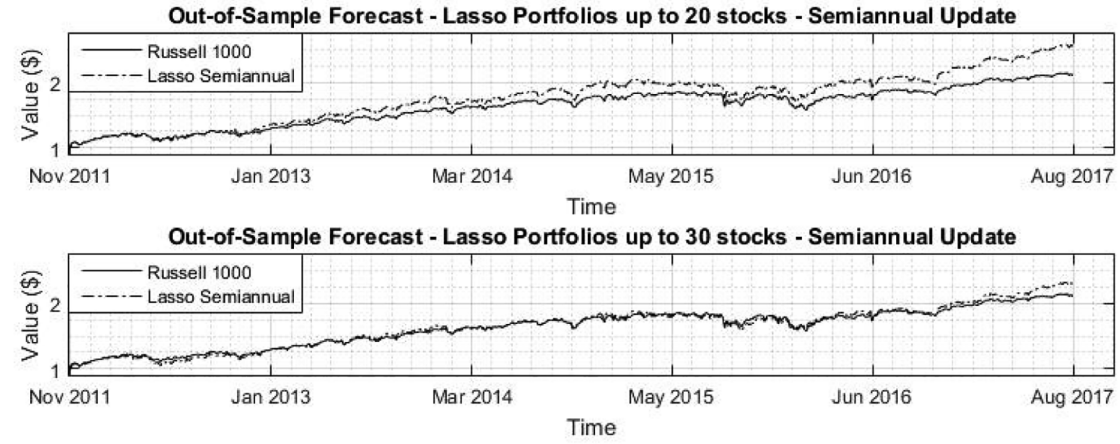
Finally, we can draw similar conclusions from the results regarding the Russell 1000. In terms of portfolios' average tracking error and annual volatility, once again portfolios OLS-NN resulted in superior results in comparison with lasso-based portfolios. Concerning average annual returns and cumulative returns, we see moderately better performance for portfolios OLS-NN in the case of larger portfolios (up to 30 stocks); in the meantime, smaller portfolios (up to 20 stocks) favor the lasso approach. However, we can also notice that such differences in performance are very small specially considering portfolios up to 30 stocks, in which case the difference between portfolios' average tracking error for portfolios OLS-NN and lasso with the same updating intervals remain below 0.005 percentage points for quarterly, semiannual, and annual updating frequencies.

As the findings for portfolios OLS-NN and lasso are hardly distinguishable in terms of overall performance, we turn our attention to the portfolio concentration and average monthly turnover, since both measures might be translated into portfolio risk and costs. Fig. 5 compares the concentration of the stock weights in the portfolios for each index. In this analysis, we consider all 24 portfolios obtained per index and size (as already mentioned, in the case of quarterly updates, we formed a total number of 24 portfolios), so that we are able to verify the concentration of the stock weights.

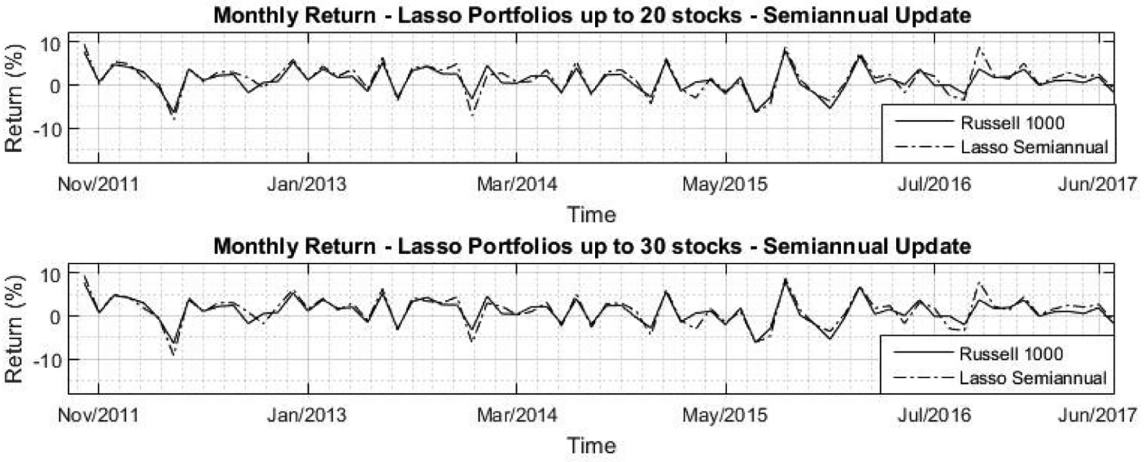
In Fig. 5a, we see the tracking portfolios for the S&P 100 have similar stock weights between both methods if we compare portfolios with the same size. However, lasso-based portfolios present more extreme (outliers) weights, which justifies the larger annual volatility values in Table 3. Also, similar conclusions can be noticed in the results for the Ibovespa (Fig. 5b) and the Russell 1000 (Fig. 5c). Overall, portfolios lasso have a larger number of stocks with weights recognized as outliers specially in the case of the two US indexes, supporting the fact that the results obtained using lasso presented larger volatility in comparison to cointegration.

Nonetheless, despite the slightly better results of portfolios OLS-NN regarding concentration of stock weights, we see a consistent advantage of portfolios using lasso by observing Fig. 6. Here, we compare the average monthly turnover and the portfolios' average tracking error (organized by index and size of the portfolios). As a result, the figure shows that the average tracking error per portfolio is slightly smaller for portfolios using OLS-NN. For instance, portfolios OLS-NN using the S&P 100 and limited to 15 stocks have average tracking error equal to 0.036%, 0.024%, and 0.017% respectively in the cases of quarterly, semiannual, and annual updating frequencies; in the meantime, portfolios lasso have average tracking error equal to 0.043%, 0.030%, and 0.021%. However, as we observe the average monthly turnover, the values for lasso-based portfolios are at least 50% inferior: 6.05%, 4.31%, and 3.29%, against 25.65%, 12.37%, and 6.59% for portfolios OLS-NN.

The complete list of results for average monthly turnover is presented in Table 3, and the same pattern mentioned earlier for the S&P 100 can be noticed in the results relative to the Ibovespa and the Russell 1000. For instance, lasso-based tracking portfolios for the Ibovespa and limited to 8 stocks resulted in average monthly turnover equal to 4.88%, 3.63%, and 2.94% (respectively, quarterly, semiannual and annual updates); in contrast, portfolios OLS-NN have turnover equal to 19.39%, 9.45%, and 5.16%. In this case, portfolios lasso have turnover at least 43% smaller (in the case of annual update). Concerning the Russell 1000, the least average monthly turnover for portfolios OLS-NN is 7.61% in the case of portfolios up to 30 stocks with annual update; however, the turnover



(a) Out-of-sample forecast per index and portfolio updating frequency



(b) Monthly return per index and portfolio updating frequency

Fig. 4. Russell 1000.

for portfolios lasso with 30 stocks and annual updates equals 4.5%, hence signaling transaction costs about 40.9% smaller. On average, the turnover using lasso is about 63.5% smaller for portfolios using lasso in comparison to OLS-NN in the case of the S&P 100, 62.9% in the case of the Ibovespa, and 43.9% in the case of the Russell 1000.

As a result, Fig. 6 shows that, on the one hand, portfolios formed using lasso and OLS-NN are very similar concerning overall performance (represented by average tracking errors). On the other hand, the substantial difference regarding average monthly turnover implies that portfolios using lasso have considerably lower costs. Thus, we can infer from these results the good quality of the lasso regression solutions for index tracking; although portfolios lasso have slightly inferior performance in some cases, this approach resulted in portfolios with overall costs at least 40% lower than portfolios OLS-NN.

4.5. Discussion and overall results regarding index tracking

For the experimental exercises, four indexes were adopted: the S&P 100, the Ibovespa, the Russell 1000, and the Russell 2000. Thus, these indexes could be classified in terms of size/volatility as low/low, low/high, high/low, and high/high.[13] As a result, in the empirical analysis we were able to test the two methods in distinct market environments, with indexes that present either more stability or a relevant degree of volatility over time, as well as with different sample sizes.

First, when comparing the results across methods in Section 4.4, we notice a consistent pattern regarding tracking performance and portfolio volatility: as we observe the results for portfolios' average tracking error and annual volatility, all portfolios OLS-NN

[13] We thank the Referees for suggesting the use of those market benchmarks based on this classification.

Table 3
Overall results for index tracking per market benchmark (**S&P 100**, **Ibovespa**, and **Russell 1000**) and statistical model (lasso and OLS Non-Negative).[1]

LASSO – S&P 100

		Portfolios up to 15 stocks				Portfolios up to 25 stocks			
	S&P 100	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.13%	11.12%	13.73%	13.95%	12.43%	8.02%	11.61%	12.95%	12.30%
Cumulative Return	106.04%	103.08%	138.48%	140.38%	119.62%	66.84%	109.78%	127.74%	118.35%
Annual Volatility	12.47%	14.54%	14.45%	14.48%	14.51%	14.48%	14.16%	14.14%	14.17%
Portfolios' Average Tracking Error	–	0.043%	0.030%	0.021%	0.051%	0.043%	0.026%	0.017%	0.043%
Average Monthly Turnover	–	6.05%	4.31%	3.29%	0.00%	5.70%	4.27%	3.19%	0.00%

OLS NON-NEGATIVE – S&P 100

		Portfolios up to 15 stocks				Portfolios up to 25 stocks			
	S&P 100	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.13%	10.63%	13.23%	14.79%	14.18%	7.64%	11.74%	13.64%	15.92%
Cumulative Return	106.04%	100.24%	135.29%	157.65%	147.98%	64.69%	113.34%	142.27%	175.13%
Annual Volatility	12.47%	13.37%	13.35%	13.68%	13.92%	13.23%	12.83%	12.79%	13.25%
Portfolios' Average Tracking Error	–	0.036%	0.024%	0.017%	0.038%	0.037%	0.021%	0.013%	0.031%
Average Monthly Turnover	–	25.65%	12.37%	6.59%	0.00%	23.39%	11.66%	6.44%	0.00%

LASSO – IBOVESPA

		Portfolios up to 8 stocks				Portfolios up to 12 stocks			
	Ibovespa	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	5.42%	5.71%	8.32%	8.33%	9.55%	3.92%	6.23%	7.58%	8.95%
Cumulative Return	29.98%	14.43%	36.80%	35.44%	37.51%	4.32%	20.75%	30.12%	40.76%
Annual Volatility	23.04%	29.22%	29.18%	29.60%	29.25%	28.03%	27.83%	27.88%	28.22%
Portfolios' Average Tracking Error	–	0.084%	0.060%	0.044%	0.115%	0.072%	0.050%	0.035%	0.094%
Average Monthly Turnover	–	4.88%	3.63%	2.94%	0.00%	4.46%	3.24%	2.49%	0.00%

OLS NON-NEGATIVE – IBOVESPA

		Portfolios up to 8 stocks				Portfolios up to 12 stocks			
	Ibovespa	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	5.42%	5.97%	6.08%	6.31%	9.95%	9.45%	9.61%	9.06%	10.82%
Cumulative Return	29.98%	18.66%	31.88%	31.77%	55.83%	56.59%	66.33%	64.45%	84.70%
Annual Volatility	23.04%	26.21%	25.94%	26.22%	27.84%	25.18%	24.95%	25.33%	23.42%
Portfolios' Average Tracking Error	–	0.071%	0.048%	0.035%	0.094%	0.062%	0.041%	0.031%	0.063%
Average Monthly Turnover	–	19.39%	9.45%	5.16%	0.00%	18.87%	9.51%	5.59%	0.00%

LASSO – RUSSELL 1000

		Portfolios up to 20 stocks				Portfolios up to 30 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.53%	11.91%	15.22%	15.18%	13.49%	7.98%	13.58%	13.91%	13.12%
Cumulative Return	110.24%	111.16%	158.59%	159.55%	135.42%	64.84%	135.05%	140.63%	129.08%
Annual Volatility	12.70%	15.76%	15.56%	15.41%	15.26%	15.69%	15.12%	14.90%	14.69%
Portfolios' Average Tracking Error	–	0.049%	0.033%	0.023%	0.059%	0.052%	0.031%	0.020%	0.049%
Average Monthly Turnover	–	8.21%	6.27%	4.74%	0.00%	8.40%	6.29%	4.50%	0.00%

OLS NON-NEGATIVE – RUSSELL 1000

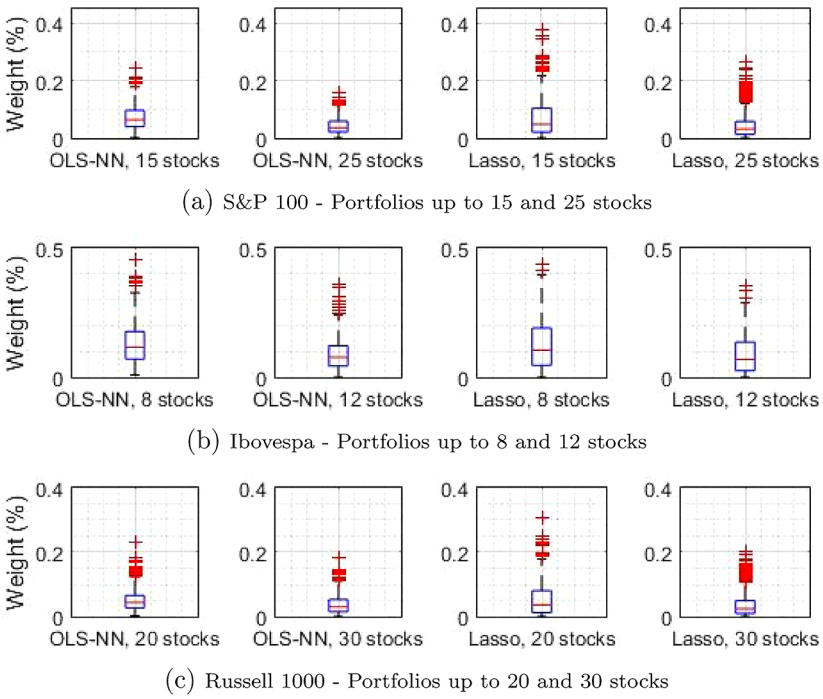
		Portfolios up to 20 stocks				Portfolios up to 30 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.53%	9.91%	12.32%	16.33%	14.81%	7.85%	11.34%	12.27%	11.83%
Cumulative Return	110.24%	89.88%	120.60%	180.00%	157.17%	65.18%	106.20%	119.38%	109.98%
Annual Volatility	12.70%	14.13%	13.57%	13.58%	13.01%	14.45%	14.02%	13.96%	13.93%

(*continued on next page*)

Table 3 (*continued*)

Portfolios' Average Tracking Error	–	0.038%	0.024%	0.016%	0.034%	0.047%	0.027%	0.018%	0.041%
Average Monthly Turnover	–	30.75%	15.97%	7.65%	0.00%	31.74%	15.54%	7.61%	0.00%

[1] Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Portfolios' Average Tracking Error refers to the average of the tracking error calculated for each portfolio according to Eq. (12). Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Average Monthly Turnover is calculated according to Eq. (19).



(a) S&P 100 - Portfolios up to 15 and 25 stocks

(b) Ibovespa - Portfolios up to 8 and 12 stocks

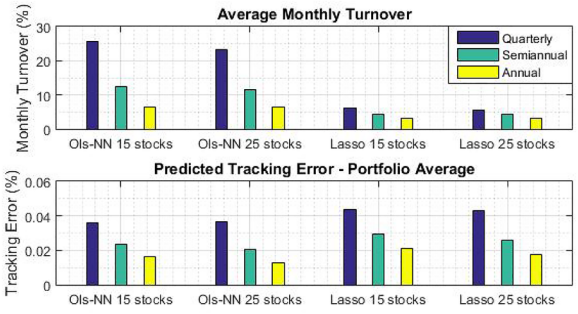
(c) Russell 1000 - Portfolios up to 20 and 30 stocks

Fig. 5. Distribution of the stock weights in the portfolios per index, size of portfolio and statistical model.

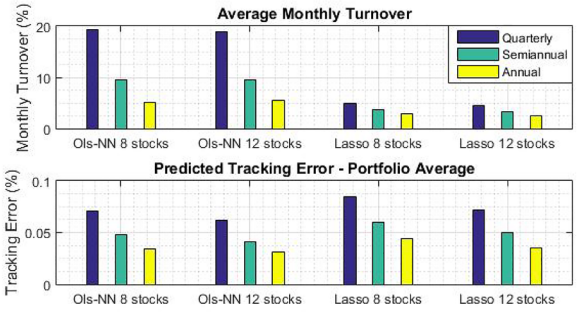
presented superior results in comparison to lasso-based portfolios. A similar pattern can also be noticed in the results for the Russell 2000 in Appendix A with only two exceptions (lasso-based portfolios delivered better portfolios' average TE in the case of portfolios up to 20 stocks with annual updates, and portfolios up to 30 stocks using the buy-and-hold strategy). Likewise, the results described in the Electronic Appendix with data from 2002 to 2017 for the S&P 100, Ibovespa and Russell 1000 also showed the quality of cointegrated portfolios to generate better tracking error and lower volatility. Hence, we can infer the quality of cointegration, when applied with a random process to compose the tracking portfolios, to produce portfolios less risky, which would be expected since the method attempts to identify a long run equilibrium to exploit the mean reversion characteristic in the price spread between portfolio and benchmark (as discussed in Section 3.2).

In contrast with the findings described above for portfolios' average turnover and annual volatility, the results for average annual returns and cumulative returns are mixed when comparing both methods, with some advantage to cointegration. In Section 4.4, we see that lasso-based portfolios had overall superior performance when tracking the S&P 100. In parallel, portfolios OLS-NN delivered more consistent results for the Ibovespa with portfolios up to 8 stocks, and for the Russell 1000 with portfolios up to 30 stocks, while lasso-based portfolios had better results for the Ibovespa with portfolios up to 12 stocks, and for the Russell 1000 with portfolios up to 20 stocks. Additionally, the Electronic Appendix confirms the quality of cointegration in terms of average and cumulative returns. Considering the S&P 100, distinct from the findings described in Section 4.4, lasso-based portfolios were superior only with portfolios up to 15 stocks, while portfolios up to 25 stocks had mixed results between the methods. The same conclusion can be reached in the case of the Russell 1000: with data from 2002 to 2017, portfolios up to 20 stocks had mixed results, while portfolios up to 30 stocks were similar to the results with data from 2010 to 2017 (favoring the cointegration approach).

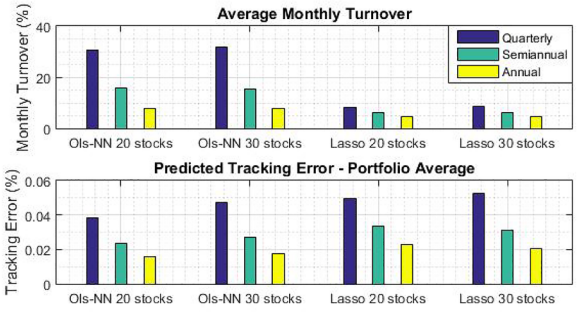
In summary, the results for the S&P 100, Ibovespa and Russell 1000 with data from either 2010 to 2017 or 2002 to 2017, in addition to the results for the Russell 2000 with data from 2014 to 2018, show better performance with the use of cointegration to



(a) S&P 100



(b) Ibovespa



(c) Russell 1000

Fig. 6. Comparison between Average Monthly Turnover and Portfolios' Average Predicted Tracking Error.

form the tracking portfolios. Nonetheless, as already discussed, we see a relevant contrast between the methods when looking at the results for average monthly turnover, which represent transaction and management costs over time. In this case, with data from 2010 to 2017, lasso-based portfolios had average turnover about 63.5% lower than portfolios OLS-NN in the case of the S&P 100, 62.8% for the Ibovespa, and 57.6% for the Russell 1000. Furthermore, with data from 2002 to 2017 in the Electronic Appendix, lasso-based portfolios had average turnover about 57.4% lower for the S&P 100 and 62.5% for the Ibovespa; however, in the case of the Russell 1000, this percentage fell to only 38.4%. When combined, these results suggest that lasso-based portfolios had a loss not only in their performance when moving from the smaller indexes (S&P 100 and Ibovespa) to the large ones (Russell 1000 and 2000), but also in their advantage in terms of average turnover, which diminished considerably. Still, the analysis demonstrates the viability of lasso as an approach to solve the index tracking problem due to (i) its ability to perform variable selection, which avoids the use exogenous methods to select each portfolio; and (ii) its advantage regarding transaction costs, which is a relevant issue for passive investment funds as they seek to diminish costs over time.

4.6. Results for lasso-based long-short strategies

As described in Section 3.1.2, the goal of long-short strategy is to explore temporary market failures by assuming long positions in undervalued stocks and short positions in overvalued stocks among the index constituents. The selection of those stocks is made through the use of benchmarks plus and minus obtained by adding/subtracting an annual percentage $\alpha\%$ to the index (uniformly

Table 4
Overall results for Long-Short using lasso – **S&P 100**.[1]

		Minus 0%	Minus 2.5%	Minus 5%	Minus 10%
Plus 0%	Average Annual Return	–	3.93%	2.24%	0.82%
	Cumulative Return	–	23.52%	12.89%	4.44%
	Annual Volatility	–	6.91%	6.54%	6.87%
	Correlation	–	0.071	0.001	−0.058
	Skewness	–	−0.005	−0.128	0.211
	Kurtosis	–	7.946	6.563	2.336
Plus 2.5%	Average Annual Return	1.99%	3.25%	2.35%	1.08%
	Cumulative Return	8.64%	18.31%	13.05%	5.85%
	Annual Volatility	7.08%	6.67%	6.50%	6.92%
	Correlation	0.050	0.066	0.033	−0.032
	Skewness	0.604	0.431	−0.267	0.044
	Kurtosis	7.369	5.912	5.456	2.387
Plus 5%	Average Annual Return	1.91%	2.83%	2.23%	1.39%
	Cumulative Return	9.65%	16.02%	12.58%	7.95%
	Annual Volatility	7.08%	6.75%	6.62%	7.16%
	Correlation	0.088	0.085	0.063	0.000
	Skewness	0.873	0.524	−0.112	−0.020
	Kurtosis	8.960	5.636	4.606	2.573
Plus 10%	Average Annual Return	1.37%	2.49%	2.36%	1.67%
	Cumulative Return	7.43%	14.16%	13.23%	9.37%
	Annual Volatility	6.74%	6.85%	7.07%	7.84%
	Correlation	0.129	0.120	0.094	0.034
	Skewness	0.648	0.555	−0.022	0.068
	Kurtosis	4.761	5.443	5.435	3.184

[1] The empirical analysis for portfolios long-short is done according to the methodology described in Section 3.1.2 and the data described in Section 4.1. Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Correlation refers to the correlation between daily returns of each strategy and daily returns of the index during the entire out-of-sample period. Skewness (Kurtosis) refers to the skewness (kurtosis) between daily returns of each strategy and daily returns of the index during the entire out-of-sample period from 2011 to 2017.

distributed over daily returns). After creating those synthetic benchmarks, the next step consists in forming portfolios long and short separately (tracking the indexes plus/minus); finally, it is necessary to subtract the long portfolio from the short portfolio to obtain the net positions for each stock. To estimate the long-short portfolios per benchmark, we selected α equal to 2.5%, 5%, and 10%, so that we formed portfolios plus 0% combined with minus 2.5%, 5%, and 10%, plus 2.5% combined with minus 0%, 2.5%, 5%, and 10%, and so on.

Moreover, we define that the portfolios long-short are limited to 40 stocks based on the S&P 100 and Ibovespa (maximum of 20 stocks for each of the portfolios long and short separately), and limited to 50 stocks based on the Russell 1000. Lastly, different from the definitions for the empirical analysis with index tracking, we now consider monthly updates for the portfolios, i.e., each portfolio long-short is re-estimated every 20 business days. Then, we set the out-of-sample interval from November 2012 to September 2017, during which period we form a total of 60 portfolios for the tests based on each of the three indexes, as now we consider the time frame in-sample equal to 720 data points (similar to Dunis & Ho, 2005).

The results are presented in Tables 4 (S&P 100), 5 (Russell 1000), and 6 (Ibovespa), in which we describe the results out-of-sample. Initially, regarding the US market indexes, we can notice that the results obtained using long-short produced overall interesting performance specially for the S&P 100, in which case the average annual return remained above 2.2% in all cases using the index minus equal to 2.5% and 5%. In fact, such results are particularly compelling as we have that portfolios long-short are theoretically a zero-cost strategy. Also, portfolios long-short presented low correlation with the market benchmark, as shown in Table 4 for example (results for the S&P 100). In this case, the correlation oscillates from −0.058 (plus 0%, minus 10%) to 0.129 (plus 10%, minus 0%), which supports the independence of this strategy from the overall market results.

A similar pattern concerning correlation is also found in Table 5 (results for the Russell 1000). In this case, the findings are not as consistent as the results for the S&P 100. Still, we obtained sound results if we analyze for instance the portfolios using minus 10%, where the average annual return remained equal to at least 1.71%. As a complement, portfolios long-short based on the Russell 1000 also have low annual volatility, which equals at most 7.08% (plus/minus 10%) and confirms the reduced risk associated with long-short relative to the market benchmark.

Finally, portfolios long-short related to the Ibovespa index present solid results particularly concerning portfolios with plus equal to 2.5%, as it can be seen in Table 6. In this case, the average annual return equals at least 2.93%, with cumulative return during the entire time frame out-of-sample as large as 26.05%. However, in contrast with the results for both the US indexes, all portfolios long-short had a more significant correlation with the index, which equals approximately −0.27 on average.

Table 5
Overall results for Long-Short using lasso – **Russell 1000**.[1]

		Minus 0%	Minus 2.5%	Minus 5%	Minus 10%
Plus 0%	Average Annual Return	–	2.04%	0.82%	2.81%
	Cumulative Return	–	12.64%	4.96%	17.64%
	Annual Volatility	–	5.79%	5.58%	5.72%
	Correlation	–	−0.027	−0.008	0.013
	Skewness	–	−0.165	0.174	−0.265
	Kurtosis	–	4.160	3.230	3.713
Plus 2.5%	Average Annual Return	0.24%	1.46%	0.65%	2.72%
	Cumulative Return	0.09%	8.44%	3.71%	17.05%
	Annual Volatility	5.68%	5.34%	5.29%	5.81%
	Correlation	−0.004	−0.012	−0.005	0.012
	Skewness	0.299	−0.053	0.130	−0.031
	Kurtosis	2.407	3.775	2.773	4.177
Plus 5%	Average Annual Return	−0.41%	0.77%	0.38%	2.31%
	Cumulative Return	−3.25%	4.30%	2.15%	14.23%
	Annual Volatility	5.91%	5.51%	5.38%	6.22%
	Correlation	0.068	0.050	0.055	0.063
	Skewness	0.281	−0.024	0.079	−0.017
	Kurtosis	3.102	4.060	3.003	6.277
Plus 10%	Average Annual Return	−0.75%	0.32%	−0.03%	1.71%
	Cumulative Return	−5.40%	1.18%	−0.65%	9.95%
	Annual Volatility	6.22%	6.01%	6.07%	7.08%
	Correlation	0.093	0.086	0.095	0.097
	Skewness	0.199	0.201	0.167	−0.055
	Kurtosis	2.914	3.373	2.957	6.840

[1] The empirical analysis for portfolios long-short is done according to the methodology described in Section 3.1.2 and the data described in Section 4.1. Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Correlation refers to the correlation between daily returns of each strategy and daily returns of the index during the entire out-of-sample period. Skewness (Kurtosis) refers to the skewness (kurtosis) between daily returns of each strategy and daily returns of the index during the entire out-of-sample period from 2011 to 2017.

All in all, the results for long-short using the lasso approach are consistent in most cases in terms of average annual return and cumulative return. In addition, it is also noticeable the low correlation between each portfolio and its market benchmark, as well as the reduced annual volatility. Therefore, we obtained portfolios that are market-neutral, and were capable of generating excess return in most cases while retaining low risk and small dependence from the market benchmarks.

5. Conclusions

In this study, our goal was to further develop previous papers (Wu et al., 2014; Yang & Wu, 2016) and extend the use of lasso to solve the index tracking optimization problem. The previous literature introduced the lasso-type regression as an innovative method to solve linear regression by making use of a penalty function to perform variable selection, thereby expanding the traditional OLS approach. Moreover, due to its capacity to select the most fitted coefficients in a regression, lasso became a statistical model suited specially for high-dimensional problems (Konzen & Ziegelmann, 2016; Tibshirani, 1996), based on the generation of sparse estimates of the coefficients (Zeng, He, & Zhu, 2012).

Hence, we selected a wide variety of datasets from different market environments (United States and Brazil) as well as with distinct sizes (ranging from 55 to 907 stocks, plus an analysis for a dataset composed by 1567 stocks with a short time frame). Thus, we sought to assess the performance of lasso to solve the index tracking problem in different financial environments (a financial market with robust stability – United States – as well as a more volatile emerging market – Brazil). Also, as we selected a particular case with larger dataset (index Russell 1000, with a database composed by 907 stocks), we aimed at exploring the capacity of lasso to deal with high-dimensional data. Finally, we also carried out an empirical analysis with the use of lasso to construct a market neutral long-short strategy, which is an investment choice that attempts to explore temporary maker inefficiencies to form zero-cost market-neutral portfolios in a methodology similar to the one used for index tracking.

The results described in Section 4 showed overall good quality solutions in all the empirical tests. In the case of index tracking, we noticed the capacity of lasso to form portfolios that tracked consistently both US indexes (S&P 100 and Russell 1000) regardless of the size of the portfolios or their updating frequencies. Further, we also estimated tracking portfolios using the cointegration methodology due to its extensive use in previous literature regarding index tracking. As a result, we were able to verify that, in most of the empirical tests, portfolios estimated using cointegration presented superior performance than lasso-based portfolios in general. Nonetheless, we obtained promising results for tracking portfolios using lasso in terms of transaction costs (in comparison with

Table 6
Overall results for Long-Short using lasso – **Ibovespa Index**.[1]

		Minus 0%	Minus 2.5%	Minus 5%	Minus 10%
Plus 0%	Average Annual Return	–	2.08%	2.42%	2.67%
	Cumulative Return	–	9.33%	10.62%	12.60%
	Annual Volatility	–	13.98%	13.48%	13.06%
	Correlation	–	−0.185	−0.262	−0.254
	Skewness	–	0.138	−0.050	−0.191
	Kurtosis	–	3.210	1.755	1.193
Plus 2.5%	Average Annual Return	4.87%	4.26%	3.66%	2.93%
	Cumulative Return	26.05%	21.35%	17.15%	13.14%
	Annual Volatility	13.30%	13.09%	13.04%	12.84%
	Correlation	−0.219	−0.236	−0.281	−0.274
	Skewness	0.013	0.055	0.031	−0.059
	Kurtosis	2.100	1.894	1.525	0.969
Plus 5%	Average Annual Return	2.11%	1.96%	2.16%	2.51%
	Cumulative Return	8.03%	6.69%	7.23%	9.87%
	Annual Volatility	13.14%	12.73%	12.77%	12.80%
	Correlation	−0.239	−0.259	−0.294	−0.292
	Skewness	−0.081	−0.018	0.010	0.004
	Kurtosis	1.951	1.899	1.531	0.982
Plus 10%	Average Annual Return	1.47%	1.81%	1.98%	2.27%
	Cumulative Return	3.87%	5.30%	5.92%	7.78%
	Annual Volatility	12.52%	12.28%	12.37%	12.64%
	Correlation	−0.311	−0.329	−0.349	−0.339
	Skewness	−0.054	−0.020	0.017	0.039
	Kurtosis	2.065	2.102	1.894	1.449

[1] The empirical analysis for portfolios long-short is done according to the methodology described in Section 3.1.2 and the data described in Section 4.1. Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Correlation refers to the correlation between daily returns of each strategy and daily returns of the index during the entire out-of-sample period. Skewness (Kurtosis) refers to the skewness (kurtosis) between daily returns of each strategy and daily returns of the index during the entire out-of-sample period from 2011 to 2017.

cointegrated portfolios); by observing average monthly turnover values for each portfolio, the outcomes showed that lasso-based portfolios had turnover at least 40% smaller than the turnover of cointegrated portfolios. Such results pointed us towards the conclusion that portfolios lasso had, in general, transaction costs at least 40% lower than portfolios using cointegration, in spite of the slightly superior tracking performance in favor of cointegration.

Finally, we extrapolated the index tracking problem and employed lasso to solve the long-short investing strategy. The results obtained during the empirical analysis presented consistent performance specially for the S&P 100 and the Ibovespa, whereas the results for the Russell 1000 exhibited a slightly more inconsistent patter. Thus, such findings are interest particularly for the Ibovespa, in which case it is relevant to observe the strong bullish and bearish market conditions faced in the Brazilian market during 2014 and 2015, which immediately imposes a challenge in terms of portfolio optimization for both the index tracking and the long-short strategies.

Appendix A. Complementary dataset – index Russell 2000

As described in Section 1, this paper adopts datasets for three indexes: the S&P 100 and the Russell 1000 (US market), and the Ibovespa (Brazilian market). Then, as the literature described in Sections 1 and 2 introduces the lasso methodology as a relevant statistical method for high-dimensional data, it justifies our choice for the use of a large dataset such as the Russell 1000. Still, as an attempt to go further, in addition to the Russell 1000, this Appendix presents the results for a complementary empirical analysis with a broader index: the Russell 2000, which is theoretically composed by small- and mid-cap listed in the US market (the bottom 2000 stocks that constitute the index Russell 3000).

Nonetheless, while the analysis for the three indexes previously mentioned are computed with datasets that cover the time window between 2010 and 2017, the data concerning the Russell 2000 is composed by daily closing prices for the index and 1,567 underlying stocks from Nov 24th 2014 to Sep 18th 2018 (totaling 961 data points). Such choice for a reduced time interval is justified by the larger turnover this index presents relative to the other three main indexes in the paper.[14] It is worth mentioning that, for this study, we were able to obtain the Russell 2000 underlying components in 2012, therefore also being able to verify that only 44% of

[14] Source, access in March 04th, 2019:https://www.ftserussell.com/impacts-reconstitution.

Table A.1
Overall results for index tracking using lasso and OLS Non-negative – **Russell 2000**.[1]

Lasso – Russell 2000					
		Portfolios up to 20 stocks			
	Russell 2000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.70%	6.64%	8.46%	6.06%	4.38%
Cumulative Return	39.33%	20.42%	27.04%	18.49%	12.20%
Portfolios' Average Tracking Error	–	0.069%	0.046%	0.029%	0.040%
Annual Volatility	13.39%	16.61%	16.39%	16.17%	16.09%
Average Monthly Turnover	–	11.96%	9.77%	5.92%	0.00%
		Portfolios up to 30 stocks			
	Russell 2000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.70%	5.32%	7.94%	6.76%	4.98%
Cumulative Return	39.33%	16.03%	25.37%	21.12%	14.26%
Portfolios' Average Tracking Error	–	0.068%	0.044%	0.027%	0.036%
Annual Volatility	13.39%	16.54%	16.06%	15.80%	15.66%
Average Monthly Turnover	–	12.27%	9.21%	5.59%	0.00%
OLS Non-negative – Russell 2000					
		Portfolios up to 20 stocks			
	Russell 2000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.70%	9.09%	7.98%	18.38%	11.63%
Cumulative Return	39.33%	29.55%	25.48%	65.33%	38.27%
Portfolios' Average Tracking Error	–	0.067%	0.042%	0.031%	0.040%
Annual Volatility	13.39%	16.35%	15.56%	15.95%	15.16%
Average Monthly Turnover	–	32.90%	16.67%	8.29%	0.00%
		Portfolios up to 30 stocks			
	Russell 2000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.70%	8.24%	13.38%	11.14%	13.46%
Cumulative Return	39.33%	26.25%	45.59%	36.98%	46.05%
Portfolios' Average Tracking Error	–	0.058%	0.037%	0.025%	0.040%
Annual Volatility	13.39%	15.72%	15.09%	14.49%	14.81%
Average Monthly Turnover	–	32.05%	14.99%	7.95%	0.00%

[1] Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Portfolios' Average Tracking Error refers to the average of the tracking error calculated for each portfolio according to Eq. (12). Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period. Correlation refers to the correlation between daily returns of each strategy and daily returns of the index during the entire out-of-sample period. Average Monthly Turnover is calculated according to Eq. (19).

the Russell 2000 composition in 2012 remained in the index in 2018 (as a reference, this percentage equals 78% for the S&P 100, and 60% for the Russell 1000).

Therefore, as we could not reconstruct backwards the index due to the lack of information between 2012 and 2018, and due to its high turnover each time the index is updated, our choice was to select a dataset starting in Nov 2014, resulting in a shorter time window in comparison to the other indexes mentioned earlier. The methodology for the analysis based on the Russell 2000 follows the same approach previously employed for the three other indexes, as we also estimated tracking portfolios using both lasso and OLS methodologies with two sizes for the tracking portfolios (up to 20 and 30 stocks) and the same updating frequencies.

The results are introduced in Table A.1 and present an unexpected contrast relative to the findings described in Section 4.4 in terms of tracking performance. As described, the analysis of the S&P 100 produced lasso-based portfolios with superior performance (average annual returns and cumulative returns) if compared with portfolios OLS-NN; in parallel, we obtained mixed results for the Ibovespa and Russell 1000, in which case portfolios lasso and OLS-NN generated somewhat similar performance. However, as we observe the portfolios' average tracking error and annual volatility for all three indexes, cointegrated portfolios dominated lasso-based portfolios with superior results in all cases. Nonetheless, lasso-based portfolios compensated the more substantial volatility and loss of performance with a considerably lower monthly average turnover, points towards the conclusion that portfolios using lasso can be associated with lower costs over time.

In contrast, the results for the Russell 2000 presented superior performance using cointegration in comparison to lasso. As we observe average returns and cumulative returns across methods, we notice that portfolios OLS-NN up to 20 stocks have better results

using quarterly updates and the buy-and-hold strategy, whereas lasso-based portfolios are superior in the case of semiannual and annual updates. However, if we observe portfolios up to 30 stocks, we see a clear advantage using cointegration, thus demonstrating that lasso was not able to form tracking portfolios with a large dataset (such as the Russell 2000) as accurately as it was in the case of the other three indexes. Furthermore, if we observe the results for portfolios' average tracking error and annual volatility, the conclusions remain the same: portfolios OLS-NN have superior results in all cases in comparison to lasso-based portfolios.

Still, in line with the findings in Section 4.4, lasso-based portfolios resulted in considerably lower monthly average turnover. In the case of portfolios up to 20 stocks, the turnover obtained using lasso is on average 44.5% smaller than the turnover obtained using OLS-NN; likewise, in the case of portfolios up to 30 stocks, the average turnover using lasso is 43.3% smaller on average. Such findings for the Russell 2000 confirm lasso as a viable method for the generation of tracking portfolios as it favors lower transaction costs over time. Still, the lower costs can be associated with a significant loss in performance in comparison to the other three indexes adopted in the research, which was not initially expected since the lasso methodology is, in theory, a consistent approach especially in the case problems using more massive data sets, as discussed in the Introduction.

Appendix B. Empirical analysis considering the use of simple returns to estimate the lasso-based tracking portfolios

As explained in Section 4.1, all tracking portfolios estimated using lasso make use of daily log returns for each stock. Nonetheless, it is relevant to emphasize that the choice for log returns instead of simple returns does not have a considerable impact on the results previously obtained with the datasets composed by log returns.

To demonstrate such conclusion, we estimated the tracking portfolios up to 30 stocks for the Russell 1000 for the entire period from 2010 to 2017, with the exact same methodology described in Section 4. The only change we made was to use a dataset composed by simple returns, instead of log returns.

The results shown in Table B.2 demonstrate that the choice for simple returns or log returns does not affect considerably our findings, as we can notice that the portfolios' average tracking error remains virtually equal in both cases.

Table B.2
Comparison of the results obtained using either log returns or simple returns.[1]

Lasso – Results with log returns					
	Index	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.53%	7.98%	13.58%	13.91%	13.12%
Cumulative Return	110.24%	64.84%	135.05%	140.63%	129.08%
Annual Volatility	12.70%	15.69%	15.12%	14.90%	14.69%
Portfolios' Average Tracking Error	–	0.052%	0.031%	0.020%	0.049%
Lasso – Results with simple returns					
	Index	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.53%	8.23%	13.60%	14.05%	13.17%
Cumulative Return	110.24%	67.63%	135.49%	142.70%	129.77%
Annual Volatility	12.70%	15.65%	15.10%	14.91%	14.66%
Portfolios' Average Tracking Error	–	0.052%	0.031%	0.020%	0.048%

[1] Average Annual Return refers to the average of the cumulative returns for each year from 2011 to 2017. Cumulative Return refers to the return calculated cumulatively during the entire out-of-sample period. Portfolios' Average Tracking Error refers to the average of the tracking error calculated for each portfolio according to Eq. (12). Annual Volatility refers to $\sigma \times \sqrt{252}$, where σ is the standard deviation of daily returns verified during the entire out-of-sample period.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.najef.2019.101055, https://www.dropbox.com/s/4qzv700jm4t60hp/Electronic_Appendix_Lasso-based_IT_NAJEF.zip?dl=0.

References

Alexander, C. (1999). Optimal hedging using cointegration. *Philosophical Transactions of the Royal Society Series A, 357*(1758), 2039–2058.
Alexander, C., & Dimitriu, A. (2002). The cointegration alpha: Enhanced index tracking and long-short equity market neutral strategies. *ISMA Finance Discussion Paper No. 2002-08 8*.
Alexander, C., & Dimitriu, A. (2005). Indexing and statistical arbitrage: Tracking error or cointegration? *The Journal of Portfolio Management, 31*(2), 50–63.
Alexander, C., Giblin, I., & Weddington, W. (2002). Cointegration and asset allocation: A new active hedge fund strategy. *Research in International Business and Finance,*

16, 65–90 URL:http://sro.sussex.ac.uk/id/eprint/40609.
Andriosopoulos, K., & Nomikos, N. (2014). Performance replication of the spot energy index with optimal equity portfolio selection: Evidence from the UK, US and Brazilian markets. *European Journal of Operational Research, 234*(2), 571–582.
Avellaneda, M., & Lee, J.-H. (2010). Statistical arbitrage in the us equities market. *Quantitative Finance, 10*(7), 761–782.
Badrinath, S., & Gubellini, S. (2011). On the characteristics and performance of long-short, market-neutral and bear mutual funds. *Journal of Banking & Finance, 35*(7), 1762–1776.
Beasley, J. E., Meade, N., & Chang, T.-J. (2003). An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research, 148*(3), 621–643.
Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The x-random case. *International Statistical Review, 60*(3), 291–319.
Candes, E., & Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics,* 2313–2351.
Chavez-Bedoya, L., & Birge, J. R. (2014). Index tracking and enhanced indexation using a parametric approach. *Journal of Economics Finance and Administrative Science, 19*(36), 19–44.
Do, B., & Faff, R. (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research, 35*(2), 261–287.
Dunis, C. L., & Ho, R. (2005). Cointegration portfolios of european equities for index tracking and market neutral strategies. *Journal of Asset Management, 6*(1), 33–52.
Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Boca Raton, FL: Chapman & Hall/CRC.
Engle, R. F., & Granger, C. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica, 55*, 251–276.
Fama, E., & French, K. (2010). Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance, 65*(5), 1915–1947.
Filippi, C., Guastaroba, G., & Speranza, M. (2016). A heuristic framework for the bi-objective enhanced index tracking problem. *Omega, 65*(C), 122–137.
García, F., Guijarro, F., & Oliver, J. (2018). Index tracking optimization with cardinality constraint: A performance comparison of genetic algorithms and tabu search heuristics. *Neural Computing and Applications, 30*(8), 2625–2641.
Gnägi, M., & Strub, O. (2018). Tracking and outperforming large stock-market indices. *Omega*.
Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics, 16*(1), 121–130.
Guastaroba, G., & Speranza, M. G. (2012). Kernel search: An application to the index tracking problem. *European Journal of Operational Research, 217*, 54–68.
Hamilton, J. D. (1994). *Time series analysis.* Princeton, New Jersey, USA: Princeton University Press.
Han, Y. (2005). Asset allocation with a high dimensional latent factor stochastic volatility model. *The Review of Financial Studies, 19*(1), 237–271.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *Springer series in statistics*New York, NY: Springer.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence (IJCAI)* Stanford, CA.
Konno, H., & Wijayanayake, A. (2001). Minimal cost index tracking under nonlinear transactions costs and minimal transactions unit constraints. *International Journal of Theoretical and Applied Finance, 4*, 939–958.
Konzen, E., & Ziegelmann, F. A. (2016). Lasso-type penalties for covariate selection and forecasting in time series. *Journal of Forecasting, 35*(7), 592–612.
Kwon, R. H., & Wu, D. (2017). Factor-based robust index tracking. *Optimization and Engineering, 18*(2), 443–466.
Li, Q., & Bao, L. (2014). Enhanced index tracking with multiple time-scale analysis. *Economic Modelling, 39*, 282–292.
MacKinnon, J. G. (2010). *Critical values for cointegration tests*. Queen's Economics Department Working Paper 1227, Kingston, Ont. URL:http://hdl.handle.net/10419/67744.
Mezali, H., & Beasley, J. E. (2013). Quantile regression for index tracking and enhanced indexation. *Journal of the Operational Research Society, 64*(11), 1676–1692.
Mutunge, P., & Haugland, D. (2018). Minimizing the tracking error of cardinality constrained portfolios. *Computers & Operations Research, 90*, 33–41.
Nasekin, S. (2013). *High-dimensional lasso quantile regression applied to hedge funds' portfolio* (Master's thesis)Humboldt-Universität zu Berlin.
Scozzari, A., Tardella, F., Paterlini, S., & Krink, T. (2013). Exact and heuristic approaches for the index tracking problem with UCITS constraints. *Annals of Operations Research, 205*, 235–250.
Strub, O., & Baumann, P. (2018). Optimal construction and rebalancing of index-tracking portfolios. *European Journal of Operational Research, 264*(1), 370–387.
Strub, O., & Trautmann, N. (2019). A two-stage approach to the ucits-constrained index-tracking problem. *Computers & Operations Research, 103*, 167–183.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 267–288.
Wu, L., Yang, Y., & Liu, H. (2014). Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis, 70*, 116–126.
Yang, Y., & Wu, L. (2016). Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling. *Journal of Statistical Planning and Inference, 174*, 52–67.
Zeng, P., He, T., & Zhu, Y. (2012). A lasso-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics, 21*(1), 92–109.
Zenios, S. A. (2008). *Practical financial optimization. Decision making for financial engineers.* Malden, MA: Wiley-Blackwell.
Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research, 7*, 2541–2563 URL:http://www.jmlr.org/papers/v7/zhao06a.html.