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

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

[a] *Department of Economics, Federal University of Rio Grande do Sul*[1]
[b] *School of Business, Federal University of Rio Grande do Sul*[2]

Abstract

In this paper, we apply the lasso-type regression for the index tracking (IT) and long-short investing strategies. Due to its capacity of both (1) performing variable selection in linear regression, and (2) being adequate for high-dimensional problems, lasso becomes an interesting technique for portfolio selection. We consider three market benchmarks (S&P 100 and Russell 1000, from the US stock market, and the Ibovespa Index, from the Brazilian market), with data from 2010 to 2017. Also, to assess the quality of lasso-based tracking portfolios, we also solved the IT problem using cointegration to have a basis for comparison of the results obtained using lasso. The findings for IT showed similar performance between portfolios based on lasso and cointegration. Nevertheless, portfolios lasso presented average monthly turnover at least 40% smaller, which indicates a considerable advantage regarding transaction costs (represented by the turnovers) in comparison with cointegration.

Keywords: lasso; index tracking; long-short; portfolio selection; statistical arbitrage

Resumo

Nesse artigo, aplicamos o modelo de regressão lasso para as estratégias de investimento de index-tracking (IT) e long-short. Devido à sua capacidade de (1) realizar seleção de variáveis em regressão linear, e (2) ser adequado para problemas de alta dimensionalidade, o método lasso torna-se uma técnica interessante para seleção de carteiras. Em nossa análise, consideramos três índices de mercado (S&P 100 e Russell 1000, do mercado dos EUA, e índice Ibovespa, do mercado brasileiro), com dados de 2010 a 2017. Além disso, para analisar a qualidade do método lasso para carteiras de index tracking, também resolvemos o problema de IT usando cointegração, de forma a ter uma base para comparação dos resultados obtidos com lasso. Os resultados para IT demonstram desempenho similar entre carteiras baseadas em lasso e cointegração. Porém, carteiras utilizando lasso apresentaram turnover médio mensal pelo menos 40% menor, o que indica uma vantagem considerável em termos de custos de transação em comparação com cointegração.

Palavras-chave: lasso; index tracking; long-short; seleção de carteiras; arbitragem estatística

JEL Codes: C52; C55; C58; G11

Área de submissão do artigo: Econometria.

[1] Address: 52 João Pessoa Ave, Porto Alegre/RS, Zip Code 90010-281, Brazil.
[2] Address: 855 Washington Luiz Ave, Porto Alegre/RS, Zip Code 90010-460, Brazil.
[*] Corresponding author
Email addresses: leonardo.santanna@ufrgs.br (Leonardo Riegel Sant'Anna), joao.caldeira@ufrgs.br (João Frois Caldeira), tpfilomena@ea.ufrgs.br (Tiago Pascoal Filomena)

1. Introduction

Stock index tracking (IT) is a passive investment management that consists in building a portfolio of stocks to replicate (*or track*) as close as possible the performance of a market benchmark, such as the Standard & Poor's 100. Many methods have already been proposed in the literature to solve the IT problem, and most of them determine the tracking portfolio by minimizing its tracking error: the difference between the historical returns of the tracking portfolio and the index over time. Moreover, a natural extension of index tracking is the long-short strategy (also referred to as market neutral), which is a self-financing strategy that aims at exploring temporary market inefficiencies through buying undervalued stocks and short selling overvalued stocks (Alexander and Dimitriu, 2002). Even though such approach lacks a broader implementation among many hedge funds due to its short exposure (Badrinath and Gubellini, 2011), it is an attractive tool for investors as a result of its self-financing and market neutral characteristics. Thus, in this paper, we 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 its capacity of both performing variable selection in linear regressions and providing good-quality solutions for high-dimensional datasets.

Different formulations for the IT problem have been proposed in the literature, such as optimization (Konno and Wijayanayake, 2001; Mezali and Beasley, 2013), optimization combined with simulation (Consiglio and Zenios, 2001), heuristic methods (Beasley et al., 2003; Scozzari et al., 2013), cointegration (Alexander, 1999; Alexander and Dimitriu, 2005), and lasso-type regression (Wu et al., 2014; Yang and Wu, 2016). In spite of their methodological distinctions, however, past studies usually carry out their analysis considering a standard feature, which is the use of a cardinality constraint to limit the size of the tracking portfolios and diminish transaction costs. Therefore, such constraint requires a methodology that performs variable selection to determine the combination of stocks that minimizes the difference between the performance of the portfolio and index.

The lasso approach has been introduced by Tibshirani (1996) and is a method that makes variable selection automatically in linear regression modeling through the generation of sparse estimates of the coefficients (Zeng et al., 2012). Additionally, it is a method successfully used in statistical modeling especially with high-dimensional datasets. Such features make this technique interesting for the index tracking problem, mainly as a result of the need to impose a cardinality constraint on the size of the tracking portfolios as well as the possibility of forming efficient portfolios to track larger indexes that are composed by hundreds or even thousands of stocks.

Regarding the studies that employed lasso for the IT problem, 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. Later, Yang and Wu (2016) extended this approach and introduced the Nonnegative Adaptive Lasso. Nevertheless, the studies mentioned above focus on the introduction of two statistical approaches, while their empirical analysis to the index tracking problem is quite limited.

Thus, our study differs from the previous literature as we focus on the financial environment and apply the lasso-type regression to different markets using diversified sample sizes. More specifically, we use three datasets in our empirical tests: S&P 100 and Russell 1000 (US stock market – respectively, databases with 102 and 907 stocks), and the Ibovespa Index (Brazilian stock market – dataset with 55 stocks). Also, because the most widely used approach to solve the index tracking and long-short strategies is cointegration (for instance, Alexander and Dimitriu, 2005; Li and Bao, 2014), we also solve the IT problem using this approach, as we seek to have a basis for comparison and validation of the results obtained using lasso.

Finally, we extrapolate index tracking and also use the lasso regression to construct long-short market neutral strategies, since the process for obtaining portfolios long-short is very similar to the index tracking one. Such strategy aims at building market-neutral portfolios, thus having low correlation with the market benchmark. Moreover, 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[3].

[3]In addition to this approach, long-short also may be developed by pairs trading or trading strategies that involve a stock and

Overall, the lasso-based index tracking performed well regarding returns and tracking error in all analyzed cases, especially for the US market. Furthermore, as we compare the results for index tracking obtained using lasso with those obtained using cointegration, we notice very similar performance in all cases. However, despite having comparable cumulative returns and volatility, portfolios lasso have average monthly turnover at least 40% smaller than the average monthly turnovers of portfolios using cointegration, which implies transaction costs at least about 40% lower for portfolios using lasso. Such outcome is interesting since the reduced turnover implies a substantial difference in transaction costs, thereby fulfilling the expectations regarding a passive investment: to diminish costs while keeping a satisfactory performance.

As a result, 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 with different sizes (from 55 to 907 stocks) as well as from distinct financial environments (US and Brazil). Also, we compute index tracking portfolios using an alternative approach (cointegration), so that we can compare and validate the results obtained using lasso. Second, the empirical testing also presents innovations as we use lasso to explore a market neutral long-short strategy. Consequently, we also contribute to the finance studies by showing how a different statistical approach can be consistently used for 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 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 empirical tests and our results, and Section 5 concludes the study.

2. Lasso – Least Absolute Shrinkage and Selection Operator

2.1. Lasso: General Concepts

As Konzen and Ziegelmann (2016) point out, the central goal of a linear regression analysis consists of estimating the coefficients for the model $y_i = \beta_0 + X_i^\intercal \beta + \varepsilon_i$, where $y_i \in \mathbb{R}$ is the dependent variable to be predicted, $X_i = (x_{1i}, \ldots, x_{ki})^\intercal \in \mathbb{R}^k$ is the vector of independent variables, the union of β_0 and β is the set of predictors $(\beta_0, \beta_1, \ldots, \beta_k)^\intercal$, and ε_i is the error term – considering 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) as follows:

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

However, as pointed out by Tibshirani (1996), the OLS approach presents some inconsistencies, specially as we increase the number of independent variables and move to high-dimensional models[4]. For this reason, Tibshirani (1996) cites two specific techniques that attempt to overcome the OLS inconsistencies: subset selection and ridge regression.

Nonetheless, 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

an ETF (Exchange-Traded Funds) (Avellaneda and Lee, 2010).

[4]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 especially in large models since the method does not perform variable selection and thus make the interpretation of the results more difficult and inaccurate.

test all 2^k possible combinations of the variables (Konzen and Ziegelmann, 2016). Yet, such analysis has a strong drawback in terms of computing time necessary to test all combinations[5].

Regarding 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:

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

Subject to:

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

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

which is equivalent to:

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

In Equations (2)-(4), the parameter $t \geq 0$ works as a control for the penalty, which means t has the same role as λ in Equation (5). In the case of Equation (5), increasing λ strengthens the shrinkage process, while setting $\lambda = 0$ results in $\hat{\beta}_{Ridge} = \hat{\beta}_{OLS}$.

Different from subset selection, however, the ridge regression approach does not involve variable selection. As Nasekin (2013) highlights, the regression analyses 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, the lasso-type regression 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[6]:

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

Subject to:

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

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

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 $(1/N) \sum_{i \in N} x_{ki}^2 = 1$ for each k. However, even though Equations (2) and (6) are similar, their Constraints (3) and (7) (which are

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

[6]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's consistency.

applied on penalty parameter t) are slightly different. As a consequence of Constraint (7), the optimization in Equations (6)-(8) takes the following form using the Lagrangian:

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

As Tibshirani (1996) and Hastie et al. (2009) point out, the model in Equation (6) might be re-parametrized by standardizing the predictors, so that the solution for β_0 is $\beta_0 = \bar{y}$; 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).

In Equations (6)-(8), as $t \geq 0$ 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 $t \leq t_0$ leads to a shrinkage of the solutions in convergence to zero, with some coefficients equal to zero. On the other hand, for $t \geq t_0$, 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). For this reason, the parameter t should be selected in a dynamic process to minimize an estimate of the expected prediction error.

Finally, concerning Equation (9), it is worth noting that $\lambda = 0$ (in the same way as $t \geq t_0$) 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 the calculation of 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).

2.2. *K-fold Cross-validation*

Hastie et al. (2009) describe the K-fold cross-validation as the simplest and most used method to estimate the prediction error. According to Efron and Tibshirani (1993), starting from a simple regression model, the prediction error (PE) consists of the expected squared difference between a future response and its prediction from the model: $\text{PE} = \text{E}(y_i - \hat{y}_i)^2$. Then, the in-sample mean squared error is $\text{MSE} = (1/N) \sum_{i \in N} (y_i - \hat{y}_i)^2$.

However, a more realistic 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 and 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 with 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:

$$\text{CVMSE} = \frac{1}{N}\sum_{i\in N}\left(y_i - \hat{y}_i^{-k(i)}\right) \tag{10}$$

In the lasso-type estimation, the K-fold cross-validation is used to compute the CVMSE statistic in Equation (10) 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. As λ increases, the results should present an increasing number of coefficients equal to zero, which tends to lead to larger error; then, the best value for λ, as already mentioned, is the one that minimizes the cross-validated error. 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

In this Section, first we present 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, we describe the essential guidelines to solve the index tracking portfolio selection using cointegration (Section 3.2).

3.1. Index Tracking and Long-Short Investing Strategies

3.1.1. Index Tracking

According to most of the past literature, index tracking portfolios are commonly evaluated by their tracking error (TE), which is defined as the standard deviation of the difference between portfolio and index returns in a specific time interval (Beasley et al., 2003; Guastaroba and Speranza, 2012):

$$TE = \frac{1}{T}\left[\sum_{t=1}^{T}(r_t^p - R_t)^2\right]^{1/2} \tag{11}$$

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.

Concerning the lasso regression, the IT 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 Equation (9) in the following equivalent form:

$$\hat{\beta}_{lasso} = \underset{\beta_0,\beta_1,\ldots,\beta_N}{\operatorname{argmin}}\left[\sum_{t\in T}\left(R_t^l - \beta_0 - \sum_{j=1}^{N}\beta_j r_{jt}^l\right)^2 + \lambda\sum_{j=1}^{N}|\beta_j|\right] \tag{12}$$

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 according to Algorithm 1, with $K = 10$, i.e. 10-fold cross-validation. After computing Equation (12), 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 and Wu, 2016), we do not impose

a nonnegative constraint on the parameters. Hence, we allow the IT portfolios to have short positions. Usually, IT models avoid short positions due to liquidity and cost issues because shorting stocks might be difficult as a result of the potential lack of stocks available for rent, thereby leading to larger transaction costs. However, because the indexes selected for the empirical tests in our study are composed by the most liquid stocks in the markets, we opt to allow portfolios to have short positions. Furthermore, our results already account for the larger costs associated with short selling, as explained in Section 4.

3.1.2. Long-Short

Alexander and Dimitriu (2005) describe the long-short strategy as a natural extension of the IT optimization using cointegration. However, in the case of long-short, we take the original index returns and use it to build enhanced indexes by adding (index plus) and subtracting (index minus) an annual excess return equal to $\alpha\%$. For instance, if we set $\alpha = 5\%$, then the construction of the index plus consists in adding an annual excess return of 5% (uniformly distributed over daily returns) to the original index daily returns. Likewise, the index minus is constructed by subtracting 5% from the original index returns. Once the indexes plus/minus are built, we estimate the long-short portfolio with lasso by using Equation (12) to calculate two models, the first of them using the index plus instead of the original index time series, and the second one using the index minus. For each regression, the coefficients should be used to form a portfolio normalized to sum up to one (similar to the index tracking methodology). As a result, the outcomes will be two portfolios (plus and minus), and the final wight of the i^{th} stock in the long-short portfolio will be the difference between its weights in the portfolios plus and minus.

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 (constructed using the index minus) to obtain the resources necessary to buy the long portfolio (constructed using the index plus). Then, portfolios long-short are expected to produce positive, low-volatility returns that are uncorrelated with market returns. As a result, such strategy follows Roll (1992) and Stucchi (2015), who argue that indexes may be inefficient, thus giving the investor the possibility of forming portfolios to outperform the market.

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 (Alexander et al., 2002; Alexander and Dimitriu, 2005) have shown that financial assets can be found to be cointegrated quite often, and this has 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 then can be used to indicate the optimal strategic asset allocation.

Cointegration is a statistical feature which defines that a set of time series that are integrated of order 1, i.e. $I(1)$, can be linearly combined to produce one time series which is stationary, $I(0)$. Formally, if we set $S_{1,t}, S_{2,t}, \ldots, S_{n,t}$ to be a sequence of $I(1)$ time series, and if there are nonzero real numbers $\beta_1, \beta_2, \ldots, \beta_n$ such that $\beta_1 S_{1,t}, \beta_2 S_{2,t}, \ldots, \beta_n S_{n,t}$ becomes $I(0)$, then $S_{1,t}, S_{2,t}, \ldots, S_{n,t}$ are said to be cointegrated (Hamilton, 1994).

When applied to prices in a stock market index, cointegration occurs when there is at least one portfolio has a stationary tracking error, i.e., when there is a mean reversion tendency 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 will stay together on a long-run basis.

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 $I(1)$ in a predefined time frame of in-sample data. Then, we estimate the linear regression in Equation (13) (given a predefined in-sample calibration period) to infer the portfolio weights. The estimation may

be carried out using OLS or an alternative approach such as non-negative least squares (NNLS), hence ensuring non-negativity on the regression coefficients. The linear regression consists of:

$$\log(P_t) = \beta_{0,t} + \sum_{i=1}^{n} \beta_{i,t} \log(p_{i,t}) + \varepsilon_t \tag{13}$$

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 (for $i \in 1..N$) 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 $\widehat{\varepsilon}_t$ resulting from Equation (13) to confirm that the linear combination of the price series of N stocks $I(1)$ is a stationary combination with order $I(0)$. To confirm if such stationary combination occurs, we apply the Augmented Dickey-Fuller (ADF) test on $\widehat{\varepsilon}_t$ to test the null hypothesis of no cointegration, where γ is the coefficient of the lagged fitted error term $\hat{\varepsilon}_{t-1}$ in Equation (14). If we let q be the order of the autoregressive (AR) process, $\widehat{\varepsilon}_t$ be the estimated error term from Equation (13), and $\Delta\widehat{\varepsilon}_t$ be the change between two error terms, then the ADF regression takes the following form:

$$\Delta\hat{\varepsilon}_t = \gamma\hat{\varepsilon}_{t-1} + \sum_{i=1}^{q} \phi_i \Delta\hat{\varepsilon}_{t-i} + u_t. \tag{14}$$

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 Equation (13) 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 for 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. Thus, we seek to mitigate the difficult concerning portfolio selection through the use of a series of simulations to form each cointegrated portfolio, as we aim at making the portfolio selection endogenous to the solving process. In this process, to obtain the portfolio for each in-sample subset, first we form a sequence of M different portfolio candidates, where each portfolio is composed by s stocks randomly selected, i.e. s corresponds to the limit 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 Equation (13) resulted in the smallest fitted sum of the squared residuals[7].

4. Empirical Tests

4.1. Database and Testing Setup

We select three databases: the S&P 100 (one of the main benchmarks in the US market) and 101 stocks; the Ibovespa index (reference benchmark in the Brazilian market) and 55 stocks; and the Russell 1000 index (composed approximately by the 1,000 largest firms in the US equity market) and 907 stocks. All three datasets were extracted from software Economatica, a financial database widely used in Brazil by both market participants and academicians. Our database includes daily stock prices from January 2010 to September 2017, and contain 1,921 trading days. Prices are adjusted for (1) splits, mergers, and other corporate actions and (2) the payment of dividends.

[7] In this study, we select M=50,000, so that we form 50 thousand distinct portfolios to select the best one based on the sum of the squared residuals. We use 50,000 because 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, thus imposing a limit on the number of M.

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 30 and 40 stocks. Additionally, to compute the tests, we choose in-sample intervals equal to 480 data points (similar to Alexander and Dimitriu, 2002), each data point being one business day, 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). Consequently, we obtain a total number of 24 portfolios in the case of quarterly updates, 12 portfolios with 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.

Concerning the lasso-type regression, the empirical analysis consists in evaluating Equation (12) with index and stocks daily returns. In contrast, the tests based on cointegration are estimated with index and stocks daily prices as described in Section 3.2. Finally, we highlight that the results presented in the next Sections already account for transaction costs as we compute the daily returns in the rolling window projections (Han, 2005; Do and Faff, 2012) as $r_{i,t} = \log(\frac{p_{i,t}}{p_{i,t-1}}) + \log(\frac{1-C}{1+C}) - d$, 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 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 defined as follows:

$$\left[\sum_{p=2}^{np}\left(\frac{\sum_{i=1}^{N}|x_i^p - x_i^{p-1}|}{2}\right)\right] \times \frac{1}{f} \tag{15}$$

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 and Figure 1. Concerning 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 and quarterly update, and up to 25 stocks and semiannual update, as they present 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 12.41% and 12.48%) very close to the index average annual return (11.43%).

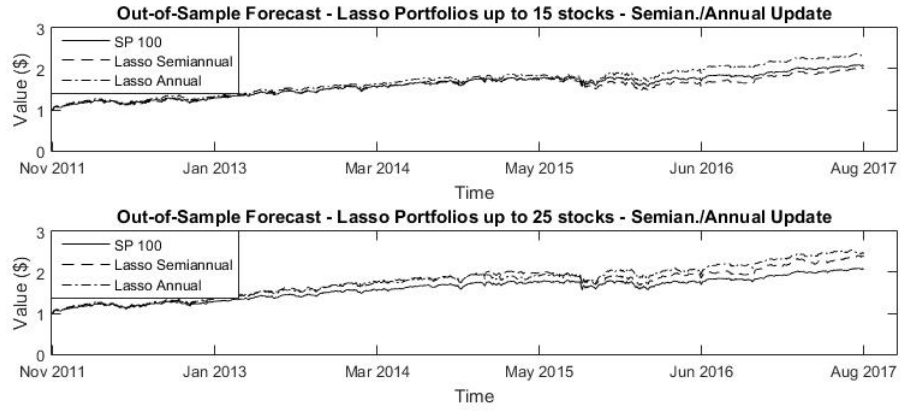
TABLE 1 HERE

Additionally, increasing the size of the portfolios from 15 to 25 stocks results in smaller portfolios' average tracking error for all updating frequencies (comparing portfolios with the same updating frequency), 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.

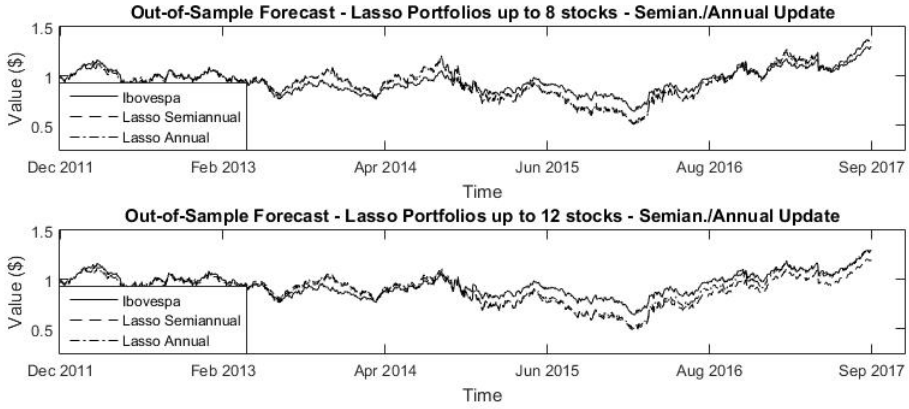
Regarding the Ibovespa index, first we highlight the considerably larger volatility of the Brazilian index in comparison with the S&P 100. In fact, Table 1 shows that the Ibovespa has annual volatility equal to 23.05%, almost twice as large as the annual volatility of the S&P 100 (12.47%). The consequence 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 S&P 100 tracking portfolios.



(a) S&P 100



(b) Ibovespa

Figure 1: Out-of-sample forecast per index and portfolio updating frequency

4.3. Index Tracking in a High-dimensional Dataset – Index Russell 1000

According to the literature on lasso (for example, Tibshirani, 1996; Nasekin, 2013; Konzen and Ziegelmann, 2016), a common characteristic of this statistical approach is its capability to solve especially high-dimensional problems. Such feature is a result of the capacity of the lasso regression to perform variable selection through its penalty function 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 a larger market benchmark: the Russell 1000, which is theoretically composed approximately by the 1,000 largest firms listed in the US equity market. 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.

We describe the results for the tracking portfolios in Table 2 and Figure 2. 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 using quarterly updates, the cumulative returns are very low and the tracking performance is poorer relatively to the other updating frequencies, since the more frequent portfolio updates resulted in larger transaction costs that penalized the portfolio's cumulative 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). 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. 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.

TABLE 2 HERE

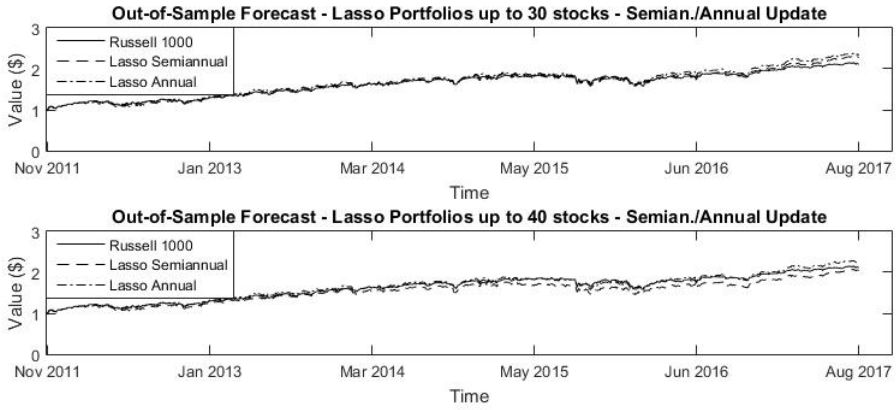


Figure 2: Out-of-sample forecast per index and portfolio updating frequency

4.4. Validation of the Lasso-type Regression: Comparison with Cointegration Based on Simulations

As discussed in the previous subsections, the application of lasso regression 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 might be useful as an attempt to shed some 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 Section 3.2. Also, we highlight that the use of the OLS regression 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, thereby avoiding short positions in the cointegrated portfolios.

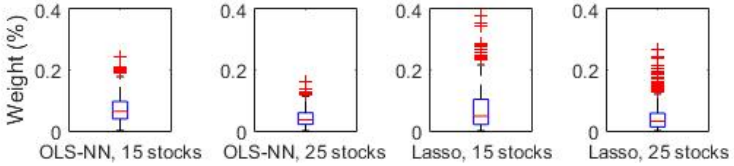
The results for cointegration (hereafter, referred to as OLS-NN) and lasso are described in Table 3 and Figures 3 and 4. Initially, Table 3 has a summary of the results using lasso and OLS-NN for each of the three indexes. As a result, for all three indexes, we can notice very similar performance among portfolios lasso and portfolios OLS-NN, observing either cumulative returns, tracking error or the volatility results.

TABLE 3 HERE

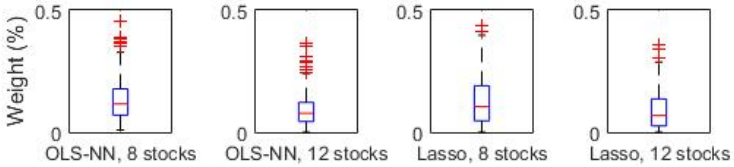
As the findings for portfolios OLS-NN and lasso are hardly distinguishable in terms of overall performance, we turn our attention to the portfolio concentration and average monthly turnover, because both

measures might be translated into portfolio risk and costs. Figure 3 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 of portfolio, so that we are able to verify the concentration of the stock weights.

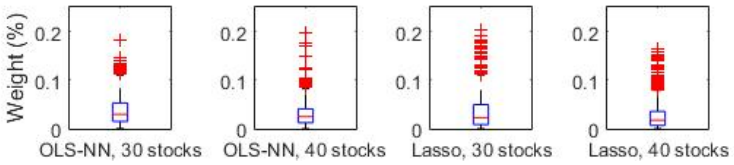
In Figure 3a, we can see that the tracking portfolios for the S&P 100 have slightly lower average weights using lasso, if we compare portfolios with the same size. Nonetheless, portfolios lasso also present more extreme (outliers) weights, which justifies the larger annual volatility values for lasso portfolios in Table 3. Moreover, similar conclusions can be drawn from the results for the Ibovespa (Figure 3b) and the Russell 1000 (Figure 3c). Overall, portfolios lasso have a larger number of stocks with weights recognized as outliers, supporting the fact that those portfolios resulted in larger volatility for all three indexes.



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



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

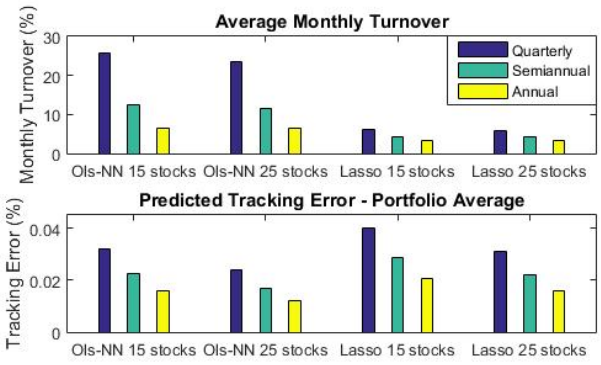


(c) Russell 1000 - Portfolios up to 30 and 40 stocks

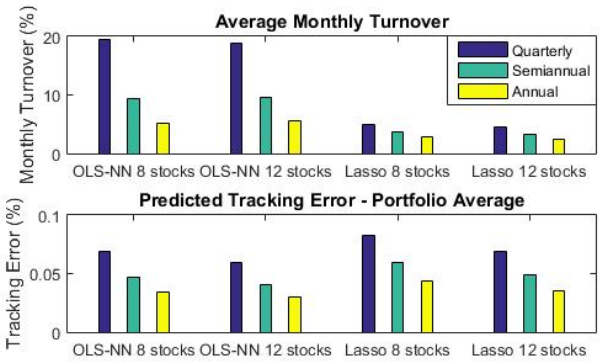
Figure 3: Distribution of the stock weights in the portfolios per index, size of portfolio and statistical model

Nonetheless, despite the slightly better results of OLS-NN portfolios regarding the concentration of stock weights in the portfolios, we see a remarkable advantage of portfolios using lasso by observing Figure 4. Here, we compare the average monthly turnover and the portfolios' average tracking error (organized by index and size of portfolios). Thus, 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 0.032%, 0.023%, and 0.016% respectively in the cases of quarterly, semiannual, and annual updating frequencies; in the meantime, portfolios lasso have average tracking error equal to 0.040%, 0.029%, and 0.020%. However, as we observe the average monthly turnover, the values for portfolios lasso are at least 50% inferior: 6.0%, 4.3%, and 3.3%, against 25.7%, 12.4%, and 6.6% for portfolios OLS-NN.

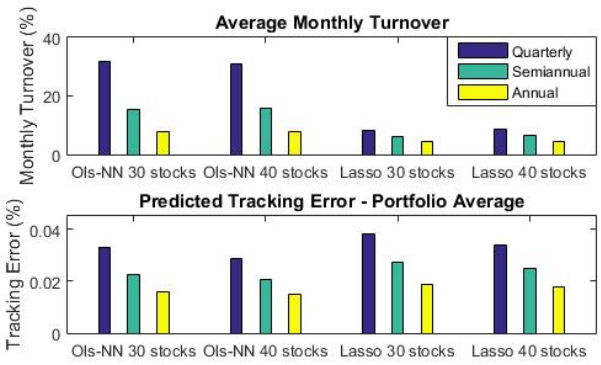
The complete list of results for average monthly turnover is presented in Table 3, and the same pattern mentioned above for the S&P 100 can be noticed in the results relative to the Ibovespa and the Russell 1000. As a result, Figure 4 shows that, on the one hand, portfolios formed using lasso and OLS-NN are very similar concerning overall performance (represented by average tracking errors). On the other hand, the substantial difference regarding average monthly turnovers 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 lasso portfolios have slightly inferior performance in some cases, this approach resulted in portfolios with overall costs at least 40% lower than portfolios OLS-NN.



(a) S&P 100



(b) Ibovespa



(c) Russell 1000

Figure 4: Comparison between Average Monthly Turnover and Portfolios' Average Predicted Tracking Error

4.5. Results for Long-Short Using Lasso

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 distributed over daily returns). To estimate the long-short portfolios, we selected $\alpha = 2.5\%$ for each of the three indexes. Moreover, we also calculated long-short portfolios limited to 50 stocks based on the S&P 100 (maximum of 25 stocks for each of the portfolios long and short separately), portfolios limited to 30 stocks based on the Ibovespa, and limited to 80 stocks based on the Russell 1000.

The results are presented in Tables 4 and 5. Since all portfolios naturally have stocks with short positions and increasing volatility (in comparison with tracking portfolios), we diminished the updating frequency and adopted monthly, bimonthly and quarterly updating intervals. Additionally, even though short positions increase transaction costs, we emphasize all results already account for transaction and management costs.

Initially, we observe consistent results using both indexes concerning the American market. In Table 4, we notice positive average annual returns for the portfolios based on the S&P 100 and Russell 1000 regardless of the updating frequency, with best results for monthly and bimonthly updates. Concerning the S&P 100, some portfolios had negative cumulative returns in 2013, 2015, and 2017; and regarding the Russell 1000, the negative returns are restricted to 2015 and 2017. Nonetheless, the positive average annual returns are a consistent result, especially if we consider portfolios long-short are theoretically zero-cost portfolios[8]. Furthermore, the correlation between each portfolio long-short and the index is very close to zero in all cases (except for a few portfolios with the Ibovespa), thus stressing the market-neutral characteristic produced by this investing strategy.

TABLE 4 HERE

TABLE 5 HERE

5. Conclusions

In this study, our goal was to extend the use of lasso regression to solve the index tracking optimization problem and the long-short investing strategy. 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). Thus, we aimed at assessing the performance of the lasso regression to solve the index tracking problem in different financial environments (US and Brazil), as well as in the case of using a high-dimensional dataset (index Russell 1000, with a database composed by 907 stocks).

The results described in Section 4 showed overall good quality solutions in all the tests carried out. In the case of index tracking, we noticed the capacity of lasso to form portfolios that tracked consistently both the American and the Brazilian indexes. Then, regarding the comparison between lasso and cointegration, the outcomes showed that portfolios lasso had monthly turnovers at least 40% smaller than the turnovers 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 similar performance of both methods.

References

C. Alexander. Optimal hedging using cointegration. *Philosophical Transactions of the Royal Society Series A*, 357(1758): 2039–2058, 1999. doi: https://doi.org/10.1098/rsta.1999.0416.

C. Alexander and A. Dimitriu. The cointegration alpha: enhanced index tracking and long-short equity market neutral strategies. *ISMA Finance Discussion Paper No. 2002-08*, 8, 2002. doi: https://doi.org/10.2139/ssrn.315619.

C. Alexander and A. Dimitriu. Indexing and statistical arbitrage: tracking error or cointegration? *The Journal of Portfolio Management*, 31(2):50–63, 2005. doi: https://doi.org/10.3905/jpm.2005.470578.

C. Alexander, I. Giblin, and W. Weddington. Cointegration and asset allocation: A new active hedge fund strategy. *Research in International Business and Finance*, 16:65–90, 2002.

M. Avellaneda and J.-H. Lee. Statistical arbitrage in the us equities market. *Quantitative Finance*, 10(7):761–782, 2010. doi: https://doi.org/10.1080/14697680903124632.

S. Badrinath and S. Gubellini. On the characteristics and performance of long-short, market-neutral and bear mutual funds. *Journal of Banking & Finance*, 35(7):1762–1776, 2011. URL https://EconPapers.repec.org/RePEc:eee:jbfina:v:35:y:2011:i:7:p:1762-1776.

J. E. Beasley, N. Meade, and T.-J. Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3):621–643, 2003. doi: https://doi.org/10.1016/S0377-2217(02)00425-3.

L. Breiman and P. Spector. Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60(3):291–319, 1992. doi: https://doi.org/10.2307/1403680.

[8] In reality, portfolios long-short have a positive cost due to additional margin requirements imposed by brokerage firms and market regulations, as a result of the use of short selling.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007. doi: https://doi.org/10.1214/009053606000001523.

A. Consiglio and S. A. Zenios. Integrated simulation and optimization models for tracking international fixed income indices. *Mathematical Programming*, 89:311–339, 2001. doi: https://doi.org/10.1007/PL00011401.

B. Do and R. Faff. Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2):261–287, 2012. doi: https://doi.org/10.1111/j.1475-6803.2012.01317.x.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, 1993.

R. F. Engle and C. Granger. Cointegration and error correction: representation, estimation and testing. *Econometrica*, 55: 251–276, 1987. doi: https://doi.org/10.2307/1913236.

C. W. J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121–130, 1981.

G. Guastaroba and M. G. Speranza. Kernel search: an application to the index tracking problem. *European Journal of Operational Research*, 217:54–68, 2012. doi: https://doi.org/10.1016/j.ejor.2011.09.004.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, New Jersey, USA, 1994.

Y. Han. Asset allocation with a high dimensional latent factor stochastic volatility model. *The Review of Financial Studies*, 19 (1):237–271, 2005. doi: https://doi.org/10.1093/rfs/hhj002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd Edition. Springer, New York, NY, 2nd ed. edition, 2009.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*. Stanford, CA, 1995.

H. Konno and A. Wijayanayake. Minimal cost index tracking under nonlinear transactions costs and minimal transactions unit constraints. *International Journal of Theoretical and Applied Finance*, 4:939–958, 2001. doi: https://doi.org/10.1142/S0219024901001292.

E. Konzen and F. A. Ziegelmann. Lasso-type penalties for covariate selection and forecasting in time series. *Journal of Forecasting*, 35(7):592–612, 2016. doi: https://doi.org/doi.org/10.1002/for.2403.

Q. Li and L. Bao. Enhanced index tracking with multiple time-scale analysis. *Economic Modelling*, 39:282–292, 2014. doi: https://doi.org/10.1016/j.econmod.2014.03.009.

J. G. MacKinnon. Critical values for cointegration tests. Queen's Economics Department Working Paper 1227, Kingston, Ont., 2010. URL http://hdl.handle.net/10419/67744.

H. Mezali and J. E. Beasley. Quantile regression for index tracking and enhanced indexation. *Journal of the Operational Research Society*, 64(11):1676–1692, 2013. doi: https://doi.org/10.1057/jors.2012.18.

S. Nasekin. High-dimensional lasso quantile regression applied to hedge funds' portfolio. Master's thesis, Humboldt-Universität zu Berlin, 2013.

R. Roll. A mean/variance analysis of tracking error. *The Journal of Portfolio Management*, 18(4):13–22, 1992. doi: https://doi.org/10.3905/jpm.1992.701922.

A. Scozzari, F. Tardella, S. Paterlini, and T. Krink. Exact and heuristic approaches for the index tracking problem with UCITS constraints. *Annals of Operations Research*, 205:235–250, 2013. doi: https://doi.org/10.1007/s10479-012-1207-1.

P. Stucchi. A unified approach to portfolio selection in a tracking error framework with additional constraints on risk. *The Quarterly Review of Economics and Finance*, 56:165–174, 2015. doi: https://doi.org/10.1016/j.qref.2014.09.008.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. URL http://www.jstor.org/stable/2346178.

L. Wu, Y. Yang, and H. Liu. Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, 70: 116–126, 2014. doi: https://doi.org/10.1016/j.csda.2013.08.012.

Y. Yang and L. Wu. 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, 2016. doi: https://doi.org/10.1016/j.jspi.2016.01.011.

P. Zeng, T. He, and Y. Zhu. A lasso-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, 21(1):92–109, 2012. doi: https://doi.org/10.1198/jcgs.2011.09156.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. URL http://www.jmlr.org/papers/v7/zhao06a.html.

Table 1: Overall results for index tracking using lasso - S&P 100 and Ibovespa[1]

S&P 100					
		Portfolios up to 15 stocks			
	S&P 100	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	11.43%	10.88%	13.93%	14.18%	12.41%
Cumulative Return	106.04%	103.08%	138.48%	140.38%	119.62%
Portfolios' Average Tracking Error	-	0.040%	0.029%	0.020%	0.008%
Annual Volatility	12.47%	14.56%	14.47%	14.49%	14.52%
Correlation	-	0.930	0.936	0.937	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	11.43%	7.49%	11.67%	13.23%	12.48%
Cumulative Return	106.04%	66.84%	109.78%	127.74%	118.35%
Portfolios' Average Tracking Error	-	0.031%	0.022%	0.016%	0.007%
Annual Volatility	12.47%	14.48%	14.17%	14.15%	14.19%
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	6.87%	6.98%	10.02%	10.04%	11.46%
Cumulative Return	25.30%	14.43%	36.80%	35.44%	37.51%
Portfolios' Average Tracking Error	-	0.083%	0.060%	0.044%	0.019%
Annual Volatility	23.05%	29.25%	29.22%	29.65%	29.33%
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	6.87%	5.06%	7.75%	9.32%	10.92%
Cumulative Return	25.30%	4.32%	20.75%	30.12%	40.76%
Portfolios' Average Tracking Error	-	0.069%	0.048%	0.035%	0.016%
Annual Volatility	23.05%	28.07%	27.87%	27.93%	28.29%
Correlation	-	0.954	0.956	0.958	0.952
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 Equation (11). 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 Equation (15).

Table 2: Overall results for index tracking using lasso - Russell 1000

Russell 1000					
		Portfolios up to 30 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	12.03%	7.63%	14.04%	14.39%	13.57%
Cumulative Return	109.86%	64.80%	133.51%	138.65%	128.43%
Portfolios' Average Tracking Error	-	0.038%	0.027%	0.019%	0.008%
Annual Volatility	12.70%	15.67%	15.12%	14.89%	14.70%
Correlation	-	0.916	0.937	0.947	0.947
Average Monthly Turnover	-	8.38%	6.25%	4.52%	0.00%
		Portfolios up to 40 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	12.03%	2.63%	11.95%	13.68%	13.57%
Cumulative Return	109.86%	23.71%	109.31%	129.61%	128.43%
Portfolios' Average Tracking Error	-	0.034%	0.025%	0.018%	0.008%
Annual Volatility	12.70%	16.17%	15.05%	14.82%	14.70%
Correlation	-	0.880	0.930	0.946	0.947
Average Monthly Turnover	-	8.53%	6.40%	4.60%	0.00%

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)

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	11.43%	10.88%	13.93%	14.18%	12.41%	7.49%	11.67%	13.23%	12.48%
Cumulative Return	106.04%	103.08%	138.48%	140.38%	119.62%	66.84%	109.78%	127.74%	118.35%
Annual Volatility	12.47%	14.56%	14.47%	14.49%	14.52%	14.48%	14.17%	14.15%	14.19%
Portfolios' Average Tracking Error	-	0.040%	0.029%	0.020%	0.008%	0.031%	0.022%	0.016%	0.007%
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	11.43%	10.43%	13.46%	15.28%	14.56%	6.88%	11.67%	13.88%	16.55%
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%	13.36%	13.69%	13.92%	13.22%	12.83%	12.79%	13.26%
Portfolios' Average Tracking Error	-	0.032%	0.023%	0.016%	0.006%	0.024%	0.017%	0.012%	0.005%
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	6.87%	6.98%	10.02%	10.04%	11.46%	5.06%	7.75%	9.32%	10.92%
Cumulative Return	25.30%	14.43%	36.80%	35.44%	37.51%	4.32%	20.75%	30.12%	40.76%
Annual Volatility	23.05%	29.25%	29.22%	29.65%	29.33%	28.07%	27.87%	27.93%	28.29%
Portfolios' Average Tracking Error	-	0.083%	0.060%	0.044%	0.019%	0.069%	0.048%	0.035%	0.016%
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			

Table 3 continued from previous page

	Ibovespa	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	6.87%	7.55%	7.68%	7.94%	12.19%	11.67%	11.85%	11.21%	13.26%
Cumulative Return	25.30%	18.66%	31.88%	31.77%	55.83%	56.59%	66.33%	64.45%	84.70%
Annual Volatility	23.05%	26.24%	25.94%	26.25%	27.89%	25.22%	24.97%	25.35%	23.45%
Portfolios' Average Tracking Error	-	0.069%	0.047%	0.034%	0.016%	0.059%	0.040%	0.030%	0.010%
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 30 stocks				Portfolios up to 40 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	12.03%	7.63%	14.04%	14.39%	13.57%	2.63%	11.95%	13.68%	13.57%
Cumulative Return	109.86%	64.80%	133.51%	138.65%	128.43%	23.71%	109.31%	129.61%	128.43%
Annual Volatility	12.70%	15.67%	15.12%	14.89%	14.70%	16.17%	15.05%	14.82%	14.70%
Portfolios' Average Tracking Error	-	0.038%	0.027%	0.019%	0.008%	0.034%	0.025%	0.018%	0.008%
Average Monthly Turnover	-	8.38%	6.25%	4.52%	0.00%	8.53%	6.40%	4.60%	0.00%

OLS NON-NEGATIVE - RUSSELL 1000

		Portfolios up to 30 stocks				Portfolios up to 40 stocks			
	Russell 1000	Quarterly	Semiannual	Annual	Buy-and-Hold	Quarterly	Semiannual	Annual	Buy-and-Hold
Average Annual Return	12.03%	7.68%	11.73%	12.83%	12.35%	8.10%	12.48%	13.08%	13.08%
Cumulative Return	109.86%	64.83%	105.61%	118.99%	109.82%	71.01%	117.86%	125.17%	121.68%
Annual Volatility	12.70%	14.43%	14.01%	13.95%	13.93%	14.60%	13.97%	13.86%	13.95%
Portfolios' Average Tracking Error	-	0.033%	0.023%	0.016%	0.007%	0.029%	0.021%	0.015%	0.007%
Average Monthly Turnover	-	31.74%	15.54%	7.61%	0.00%	31.06%	15.99%	7.85%	0.00%

Table 4: Overall results for Long-Short using lasso per market benchmark[1]

S&P 100 - Portfolios up to 50 stocks			
	Monthly	Bimonthly	Quarterly
Average Annual Return	3.31%	3.85%	2.48%
Cumulative Return	23.71%	28.44%	17.66%
Annual Volatility	8.52%	8.43%	8.44%
Correlation	0.009	-0.004	-0.066
Skewness	0.265	0.206	0.284
Kurtosis	1.369	0.936	1.583
IBOVESPA - Portfolios up to 30 stocks			
	Monthly	Bimonthly	Quarterly
Average Annual Return	2.04%	2.72%	-3.66%
Cumulative Return	14.20%	18.17%	-24.98%
Annual Volatility	16.61%	17.10%	15.75%
Correlation	-0.007	0.003	0.019
Skewness	-0.133	-0.104	-0.158
Kurtosis	1.547	1.912	1.740
RUSSELL 1000 - Portfolios up to 80 stocks			
	Monthly	Bimonthly	Quarterly
Average Annual Return	4.78%	4.08%	2.44%
Cumulative Return	34.88%	29.43%	16.44%
Annual Volatility	9.91%	9.80%	9.68%
Correlation	0.115	0.074	0.016
Skewness	0.222	0.133	0.078
Kurtosis	2.627	1.913	2.492

[1] Average Annual Return, Cumulative Return, Annual Volatility, and Correlation are calculated as indicated in Table 1. 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.

Table 5: Cumulative return per year for Long-Short strategy[1]

	S&P 100			Ibovespa			Russell 1000		
	Mont.	Bimont.	Quart.	Mont.	Bimont.	Quart.	Mont.	Bimont.	Quart.
2011	2.0%	1.5%	1.5%	-0.9%	-0.9%	-0.9%	1.3%	1.0%	1.0%
2012	9.3%	7.6%	2.9%	-4.9%	-3.8%	-11.0%	24.0%	18.9%	11.4%
2013	1.8%	1.8%	-2.1%	1.0%	5.5%	-16.7%	7.9%	9.9%	6.8%
2014	6.1%	11.9%	11.8%	2.0%	-0.5%	3.0%	4.0%	5.6%	4.9%
2015	-3.1%	-2.3%	0.0%	11.1%	21.8%	8.6%	-8.1%	-9.1%	-12.3%
2016	14.4%	12.8%	7.8%	7.7%	-1.5%	-9.6%	7.3%	4.2%	3.7%
2017	-7.3%	-6.3%	-4.6%	-1.7%	-1.5%	1.1%	-3.0%	-1.9%	1.6%
Average	3.3%	3.9%	2.5%	2.0%	2.7%	-3.7%	4.8%	4.1%	2.4%

[1] The Cumulative Return per year refers to the return calculated cumulatively during each year of the out-of-sample period.