Index Tracking with Feature Selection

Paulo Ferreira Naibert[a,1], João F. Caldeira[b]

[a] *PhD Candidate at the Economics Department of the Federal University of Rio Grande do Sul*
[b] *Department of Economics, Universidade Federal de Santa Catarina*

Abstract

We compare the performance of passive investment portfolio with a small number of assets (sparse index-tracking portfolios) using different feature selection algorithms. To isolate the effect of the selection methods, we separate the asset selection and the asset allocation phase. In the asset allocation phase, we follow Liu (2009), and select minimum tracking error volatility portfolios. The selection methods used are the backward stepwise selection, forward stepwise selection and the lasso. Our results show that the lasso selection method outperforms the other methods, in the brazilian case. It presents similar tracking error volatility and higher mean return, which leads to a better risk-adjusted performance. In the american case, the lasso presents better risk-adjusted performance, but this is due to higher mean returns, not lower volatility. This is undesirable in our case. One highlight of this paper is that the forward and backward iteration algorithms (simple methods that receive little attention in the literature) perform well in selecting assets for index tracking.

Keywords: Index-Tracking;, Portfolio Theory;, Feature Selection;, Optimal Portfolios
JEL classification: G11; G12; G17.

[1] Corresponding author. Department of Economics, Universidade Federal do Rio Grande do Sul, Porto Alegre, R.S. 90040-000, Brazil. Telephone: +55-51. Fax: +55-51. e-mail: paulo.naibert@gmail.com

1. Introduction

Tu & Zhou (2011) and Liu (2009) point out that the majority of institutional investors are benchmarked by an index, e.g. the SP500. So, their performance has to be evaluated in relation to that index. Consequently, having to choose index tracking portfolio is closer to the practical life of a institutional investor. Index tracking can be defined as a quantitative method of trading and of passive portfolio management (McWilliams & Montana, 2010; Wu *et al.* , 2014). The most popular way to use this strategy is to form a portfolio of assets to mimic the performance of an externally specified index that, usually, is a market index such as the SP500. This performance takes into account the risk-return profile of the specified benchmark.

There are two main ways of maintaining a fund that tracks an index; one is the full replication and the other is the partial replication. Full replication is the simplest one; it consists on maintaining all the assets with the same relative weights of the index being tracked. However, despite being possible to achieve a perfect replication with it, the full replication presents some disadvantages: (i) it results in small weights, which increase transaction costs; (ii) it presents instability of the weights, which means frequent revisions of them; (iii) its portfolio rebalancing can become complicated; and (iv) price fluctuation bring too many additions and subtractions of assets to the index.

It is also possible to try to replicate an index with a portfolio that contains only a small proportion of all the available assets, this is the partial replication. But this method also presents some problems: (i) the initial selection of the small number of assets to be included in the portfolio can be hard, and (ii) the need to estimate optimal weights (allocation) complicate the problem.

To implement an index tracking strategy, the investor has to solve two interconnected problems associated with it. The first is the asset selection problem, where the investor selects p assets from a universe of N assets. The second is the asset allocation problem, where the investor decides the relative weights of each selected asset.

The main goal of this paper is to compare the performance of different feature selection algorithms applied to index tracking and portfolio optimization. More specifically, we study the case where an investor has to track an Index and he rebalances his portfolio monthly, bimonthly and quarterly (20, 40, and 60 days rebalancing). First, we select the assets with the following strategies: backward iteration, forward iteration, lasso regression. Here, we highlight the use of the backward and forward iteration algorithms that receive very little attention in the literature. With the selected assets, we form portfolios by minimizing the tracking error volatility. Then, we take the out of sample portfolio returns and analyze those returns in comparison to the selected Index. We use two differente datasets. The first is a brazilian dataset of asset prices with daily frequency from January 1999 to December 2010 (2970 days). The second is an american dataset of asset prices with daily frequency from January 2010 to October 2017 (1950 days).

The benefits of this two phase method is twofold. First, it isolates the effect of the selection methods, so the differences in performance may be attributed to differences in asset selection methods. Second, we make use of the Global Minimum Variance Portfolio (GMVP) in the second phase. The GMVP has several desirable properties discussed in Jagannathan & Ma (2003) and Clarke *et al.* (2006). As the goal of this paper is to compare the performance of the selection algorithms applied to index tracking and portfolio selection, this separation of phases is very suitable to our needs.

Next, we list papers that evaluated different methods of asset selection and portfolio optimization to implement index tracking and we compare their works to ours. Liu (2009) selects portfolios that track the SP500 by minimizing tracking error volatility without regard for the size of mean of the tracking error. In that paper, the asset selection phase is solved by only using the 30 assets in the Dow Jones Industrail Average (DJIA). Jansen & van Dijk (2002) use index tracking with small portfolios by asset selection methods. McWilliams & Montana (2010) mention the possibility that the problems of asset selection and portfolio optimization can be put in the form of variable selection. They implement the asset selection with methods of penalized regression. Thus, we can use other methods of variable selection to choose the assets that will enter our portfolio. Here, we will use some methods listed in James *et al.* (2014).

Santos (2015) selects minimum variance portfolios using cardinality constraints to form portfolios with few assets and assess their performance. Sant'Anna *et al.* (2014) selects Index-Tracking portfolios with cardinality constraints to form portfolios with few assets and track the Bovespa Index. This paper resembles

those once we will form portfolios with few assets; however we will use a different method. We will not use cardinality constraints, but feature selection algorithms. This can be justified by the lower computational cost of those algorithms. We will use only regressions to select the portfolio assets, the optimization of portfolios weights will occur in a different phase.

Other related works are Wu *et al.* (2014) e Wu & Yang (2014), which use non-negative lasso and elastic net regressions to select assets. Those works use non-negative least-squares for portfolio optimization (weight selection). Yet another related work is Medeiros *et al.* (2015), which use lasso regression to select variables to forecast inflation. This work differs from that in terms of the goal for which we use the lasso regression. Here we use it to form portfolios to track a financial Index.

Our results show that, in the brazilian case, the lasso selection method outperforms the other methods by presenting similar tracking error volatility and higher mean returns. Note that this overperformance is due only to asset selection, not due to the shrinkage of the portfolio weights that results from the consstraints of the problem. This happens because the asset allocation is the same for all portfolios. More on the shrinkage of portfolio weights can be found in Fan *et al.* (2012); Jagannathan & Ma (2003); Brodie *et al.* (2009).

However, in the american case, the lasso method presents more volatility, but this is more than offset by its superior mean returns, which ultimately delievers better risk-adjusted performance. But, because our primary goal is to minimize tracking error volatility, this kind of risk-return tradeoff is not desirable in our case. Even though the backward and forward selection methods might be outperformed by the lasso, there are some instances where these methods are superior or comparable to the benchmark index. This has to be highlighted, because those simple methods of feature selection receive too little attention in the portfolio literature, even though they might bring good results.

Beyond this introduction, this paper is organized as follows. Section 2 presents the basics of portfolio selection that will be used throughout the paper. Section 3 presents the algorithms used to apply feature selection in the dataset. Section 4 presents the methodology of the empirical study. Section 5 shows the results obtained in the empirical study. Finally, the section 6 concludes.

2. Basic Definitions

In this section, we present the basics of portfolio selection that we will use throughout this paper. We start by considering an investment universe with N risky assets. Each asset has a price in time t denoted by $P_{t,i}$. By holding asset i from $t-1$ to t, the investor earns the return

$$X_{t,i} = \frac{P_{t,i}}{P_{t-1,i}} - 1. \tag{1}$$

Let's denote the $N \times 1$ vector of future and uncertain returns of those assets in time t by X_t, where

$$X_t = [X_{t1}, X_{t2}, \ldots, X_{tN}]'.$$

We will assume that the returns have mean vector $E[X_t] = \mu_X$ and covariance matrix $V[X_t] = \Sigma$. We also denote the future and uncertain return of the index by y_t. We assume that the index return have mean $E[y_t] = \mu_y$ and variance $V[y_t] = \sigma_y^2$.

With those definitions, we can define the **excess returns** of the risky assets on the index as:

$$R_t = X_t - ey_t.$$

where e is a $N \times 1$ vector of ones. We assume that R_t has mean vector and covariance matrix

$$\begin{aligned} E[R_t] &= \mu_X - e\mu_y = \mu \\ V[R_t] &= \Sigma + \sigma_y^2 ee' - 2Cov(X_t, y_t)e' = \Omega. \end{aligned}$$

Porfolio Returns

Next, we define portfolio and portfolio returns. A portfolio of the N risky assets is represented by a $N \times 1$ vector w, where

$$w = [w_1, w_2, \ldots, w_N]',$$

and w_i is the fraction of total wealth invested in asset i. The vector w can also be called the allocation vector. If we constrain the weights in w to sum up to one ($e'w = 1$), we have the **fully bought condition**.

Holding the portfolio w_t from t to $t+1$ yields the out-of-sample return in $t+1$,

$$X_{p,t+1} = w_t' X_{t+1}. \tag{2}$$

$X_{p,t+1}$ is a weighted average of the asset returns selected to the portfolio with weights w_i for $i = 1, \ldots, N$.

$$X_{p,t+1} = w_{t,1} X_{t+1,1} + \cdots + w_{t,N} X_{t+1,N} = \sum_{i=1}^{N} w_{t,i} X_{t+1,i} = w_t' X_{t+1}.$$

Evolution of weights

Liu (2009) points out that in the moment prior to rebalancing, each dollar invested in asset i in the portfolio has changed its value from $w_{t,i}$ to $w_{t,i}(1 + X_{t+1,i})$, where $w_{t,i}$ is the i-th element of w_t and the i-th element of X_t is denoted as $X_{t,i}$. Also, in general, each dollar invested in the whole portfolio has changed from w_t to $w_t(1 + X_{p,t+1})$. Therefore, prior to rebalancing, the weight on asset i has changed from $w_{i,t}$ to

$$w_{t,i}^{+1} = w_{t,i} \frac{1 + X_{t+1,i}}{1 + w_t' X_{p,t+1}}.$$

More generally, we can express equation the changed vector of weights as

$$w_t^{+1} = w_t \odot \frac{e + X_{t+1}}{1 + X_{p,t+1}}, \tag{3}$$

where $\odot$ denotes the Hadamard (direct) product. Compounding another period, the portfolio w_t in $t+2$ will have changed to

$$w_t^{+2} = w_t \odot \frac{(e + X_{t+1}) \odot (e + X_{t+2})}{(1 + X_{p,t+1}) \times (1 + X_{p,t+2})}.$$

Following this logic, we reach a equation for h periods:

$$w_t^{+h} = w_t \odot \frac{(e + X_{t+1}) \odot \cdots \odot (e + X_{t+h})}{(1 + X_{p,t+1}) \times \cdots \times (1 + X_{p,t+h})}. \tag{4}$$

Tracking Error

The tracking error may be defined as a measure of the difference between the index tracking portfolio return, $X_{p,t}$ and the index return, y_t. So, we have the following expression for the tracking error:

$$TE_{t+1} = w_t' X_{t+1} - y_{t+1} = X_{p,t+1} - y_{t+1}, \tag{5}$$

where w_t is the allocation vector of our portfolio.

With the expression of the tracking error above, we can see that, if the weights are restricted to sum up to one ($e'w = 1$), then the excess return of a portfolio on a benchmark is the same as the tracking error:

$$\begin{aligned} R_{p,t+1} &= w_t' R_{t+1} = w'(X_{t+1} - ey_{t+1}) \\ &= w_t' X_{t+1} - w' e y_{t+1} \\ R_{p,t+1} &= X_{p,t+1} - y_{t+1} = TE_{t+1}. \end{aligned} \tag{6}$$

Hence, if the weights sum up to one ($e'w = 1$), minimizing the variance of the Tracking Error is equivalent to minimizing the variance of the excess returns of a portfolio on a benchmark. Note that the restriction is important.

3. Feature Selection Algorithms

Here, we present the methods through which we will select the assets, they are: (i) forward stepwise selection, (ii) backward stepwise selection, and (iii) lasso regression, More information about this methods can be found in James *et al.* (2014). The methods of asset selection used here depend heavily on the Ordinary Least Squares (OLS) Regression. Below, we offer more details about these methods, beggining with some notes on the OLS Regression.

3.1. Ordinary Least Squares (OLS) Regression

By regressing the index returns against the asset returns with the portfolio weights being parameters to estimate, we have a regression problem with stochastic regressors, more details about this problem can be found in Rao *et al.* (2008) and Rencher & Schaalje (2007). We may represent this problem as:

$$y_t = \alpha + X_t \beta + \varepsilon_t.$$

Note that the intercept α is necessary to let $E[\varepsilon_t] = 0$, so we can have $V[\varepsilon_t] = E[\varepsilon_t^2]$.

The OLS regression has the goal to minimize the sum of squared residuals, which is expressed as:

$$RSS(\beta) = \sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t - \alpha - X_t \beta)^2. \tag{7}$$

where ε_t is the residual of the regression.

By minimizing $RSS(\beta)$ in equation (7), we find β_{ols}:

$$\beta_{ols} = (X'X)^{-1}(X'y) \tag{8}$$

where X is the $J \times N$ matrix of asset returns: $X = [X_{t-J}, X_{t-J+1}, \ldots, X_{t-1}]'$ and y is the $J \times 1$ vector of index returns $y = [y_{t-J}, y_{t-J+1}, \ldots, y_{t-1}]'$.

3.2. Forward Stepwise Selection

The forward stepwise selection uses a series of OLS regressions to select the assets that will enter the investor's portfolio, i.e. the columns that will enter in the matrix X. It starts with a null model with no predictor, then we add one predictor at a time to the model until all predictors are added to the model. More specifically, at each step the variable that provides the best fit is added to the model.

3.3. Backward Stepwise Selection

In opposition to the forward stepwise selection, the backward stepwise selection starts with the full OLS model, which contains all N predictors available, then the least useful predictors are removed, one at a time, until we reach a model with the desired p predictors.

The forward and backward stepwise selections result in a set of models, each of which contains a subset of p predictors. Usually, the researcher determines which of the models in the set of models is the best. Here, however, we want a predefined number of assets in our portfolios. So we just choose the model with the predefined number p of assets with the smaller RSS and bigger R^2. To implement both of those algorithms, we use the `R` software with the `leaps` package.[2]

3.4. LASSO

To fit the least squares regression, we estimate the values of $\alpha, \beta_1, \cdots, \beta_p$ that minimize the RSS, as in equation (7). The lasso regression has a similar form, with its coefficients minimizing the RSS plus a penalty:

$$\sum_{t=1}^{T}\left(y_t - \alpha - \sum_{j=1}^{p}\beta_j x_{tj}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{9}$$

that is, the lasso is a penalized regression. In the equation above, $\lambda \geq 0$ controls the amount of regularization in the regression. When $\lambda = 0$, we have the least squares regression. The bigger the value of λ, the bigger the amount of regularization and tighter the variable selection will be. Thus, we have to fit λ in a convenient way to our objectives. In this paper we use the `glmnet` package of the `R` software.[3]

Because the betas in the algorithms above might not sum to one, we only use the algorithms to select the assets. Wu *et al.* (2014) and Wu & Yang (2014) also segregate the asset selection phase and the asset allocation phase by doing different regressions for each phase. Another reason to do this is that by using the same equation for all the portfolios' allocation, we can isolate the effect of the asset selection strategies.

4. The Experiment

In this section, we present the methodology of our experiment. The goal of such experiment is to evaluate the performance of the alternative asset selection techniques against each other and the benchmark index. It is done in three steps. The first step is the asset selection phase, which is done by using the algorithms from section 3. The second step is asset allocation phase, where we form optimal portfolios in pseudo real time. The third step is the out of sample performance evaluation of those portfolios against each other and the benchmark.

4.1. Data

In this empirical analysis we have two different datasets. The first is the Brazilian dataset, it consists of daily closing prices for 61 stocks traded in the Bovespa and a series of daily closing prices for the Bovespa Index (IBovespa). The sample period used here is from January 1999 to December 2010 (2970 days). The second dataset is the American dataset, it consists of daily closing prices for 462 stocks and a series of daily closing prices for the SP500 index. The sample period used here is from January 2010 to October 2017 (1950 days).

From the price and index series, we take the simple daily raw return of each asset, according equation (1). That will leave us with a total of $N + 1$ (N stocks plus one index) series of raw returns, with length of T, each. With the raw return series (X_t), we use the algorithms from Section 3 to select which assets will enter in our portfolios (asset selection phase).

[2]More information about the leaps package can be found in https://cran.r-project.org/web/packages/leaps/
[3]More information about the glmnet package can be found in https://cran.r-project.org/web/packages/glmnet/

Allocation

After that, we enter the asset allocation phase. To that end, we follow Liu (2009) and select portfolios that minimize the volatility of the tracking error, without any regard for the size of the mean of the tracking error. To achieve this, we find the portfolio that minimizes the variance of the excess returns on the benchmark with a fully bought contraint ($e'w = 1$). The answer to the problem is:

$$w_t = \frac{\hat{\Omega}_t^{-1} e}{e' \hat{\Omega}_t^{-1} e} \tag{10}$$

where e is a $N \times 1$ vector of ones and $\hat{\Omega}_t$ is the covariance matrix estimator of Ω using information untill time t.

One clarification has to be made here, what the formula (10) finds is the portfolio with the least variance of excess returns, without any regard for the mean. So, following the index exactly and zeroing the tracking error is not a concern, only minimizing the volatility of the tracking error. Using the equation (10) can be defended on the grounds that it is the global minimum variance portfolio (GMVP). This portfolio achieves just what we want and it offers several desirable properties as presented by Jagannathan & Ma (2003) and Clarke *et al.* (2006). If we wanted to zero the tracking error, we would have to optimize a different objective function and use different restrictions in our portfolio. However, this portfolio would not have the properties of the GMVP highlighted above.

Rolling Windows

To estimate the $\hat{\Omega}_t$ we adopt a recursive estimation approach based on a *rolling window time series of returns* $\{R_j\}_{j=t}^{t+J-1}$ with t varying from $t = 1$ to $t = T - J - 1$. For this experiment, we use window length equal to $J = 250$ for the Brazilian dataset and $J = 500$ for the American dataset, because the American dataset has more assets, and we need $J > N$ for the covariance matrix to be invertible. Other way to see that is by stacking J vectors of excess returns on top of each other and form the $J \times N$ matrix

$$R = \begin{bmatrix} R_{t-J} \\ R_{t-J+1} \\ \cdots \\ R_{t-1} \end{bmatrix}$$

With the rolling windows' scheme, we form the vector of expected means and the covariance matrices of excess returns with the formulas:

$$\hat{\mu}_t = J^{-1} \sum_{j=t}^{t+J-1} R_t, \tag{11}$$

$$\hat{\Omega}_t = (J-1)^{-1} \sum_{j=t}^{t+J-1} (R_t - \hat{\mu})(R_t - \hat{\mu})'. \tag{12}$$

This results in $T - J - 1$ estimates of expected return vectors and covariance matrices. Using them, we find the allocation vector using equation (10). By the end of this process, we generate $T - J - 1$ allocation vectors for each asset selection algorithm, in each dataset. With the allocation vectors, we compute the *out-of-sample portfolio returns* with equation (2), which generates the time series $\{X_{p,t+1}\}_{t=J}^{T-1}$. This time series is the subject of our analysis.

4.2. Evaluation Measures

What we will evaluate are the out-of-sample portfolio returns in relation to the benchmark index. Since the goal of index tracking portfolio is to closely follow the series of interest, our evaluation has to consider

how our portfolio performs in relation to the benchmark. Note that, as Tu & Zhou (2011) and Liu (2009) point out, the majority of institutional investors are benchmarked by an index, e.g. the SP500. So, their performance has to be evaluated in relation to that index. Consequently, having to choose index tracking portfolio is closer to the practical life of a institutional investor.

The statistics that we want to evaluate are the Sharpe Ratio (SR), which is the mean of portfolio returns divided by the standard deviation of the portfolio returns; and the Information Ratio (IR), which is the mean of the tracking error divided by the standard deviation of the tracking error.

Transaction Costs

To get a sense of the amount of trading required to implement each portfolio strategy, we compute the portfolio turnover and report its average through time. Turnover in time t has the following formula

$$TO_t = |w_{t+1} - w_t^{+1}|, \tag{13}$$

where w_t^{+1} is the portfolio prior to rebalancing as defined in equation (3).

Associated with the turnover there is the associated transaction costs of rebalancing a portfolio. If the rebalancing is too frequent or too extreme, it can lead to overtrading. According to Barber & Odean (2000), overtrading has negative effects on portfolio performance due to its associated transaction costs. To take into account the impact of such costs we also evaluate the returns net of transaction costs, which have formula:

$$r_{p,t} = (1 + X_{p,t})(1 - cTO) - 1, \tag{14}$$

where c is the fee that must be paid for each transaction. We use $c = 50bp$ like in DeMiguel *et al.* (2009) and Kirby & Ostdiek (2012).

Weight Statistics

To gauge the amount of short selling in the portfolios, we report their **Short Interest**. It expresses the average size of the short positions (weights smaller than zero) in the portfolio, which has the following formula:

$$\text{SI}_t = \frac{\sum_{i=1}^{N} |w_{i,t}| - 1}{2} \tag{15}$$

To assess how many assets our strategies select, we also report the **average number of assets in the portfolio**.

5. Results

In this section we present the results of our empirical analysis. Each row of the Tables 1 through 4 presents the statistics of a portfolio, and the first column of the rows presents the name of the portfolios. In the Tables, **IBOV** is the Bovespa Index and **SP500** is the SP500 index. Also, in each Table, **bwd** stands for the portfolio formed using the Backward Selection Algorithm, **fwd** stands for the portfolio formed using the Forward Selection Algorithm, and **las** stands for the portfolio formed using the Lasso Algorithm. The numbers after the letters denote the number of active positions allowed in each portfolio.

Tables 1 and 3 shows statistics without considering transaction costs. The presented statistics are: annualized Mean and Standard Deviation (SD) of the benchmark index and of the portfolio returns. The index is show in absolute terms and the portfolios are shown in relation to the index (Mean and SD of the portfolios are divided by the Mean and the SD of the index). SR denotes the annualized Sharpe Ratio. The Table also shows the Mean and SD of the tracking error of the portfolios (portfolio return minus the index return). And the IR denotes the annulized Information Ratio which is the Mean of the tracking error divided by its SD. The last three columns of the Table shows the alfa, beta, and R^2 (intercept in annualized terms, the slope, and the coefficient of determination) of a univariate regression of the out of sample portfolio returns against the benchmark Index in the out of sample period.

The Figures **??** through **??** and **??** through **??**, shows cumulative returns of each portfolio with a specified rebalancing period and a number of asssets. Those Figures also show in pointed lines the difference between those cumulative returns and the index. The Figures **??** through **??** and **??** through **??**, shows rolling window (250 days) annualized standard deviation. Those Figures also show in pointed lines the tracking error standard deviation.

For the brazilian dataset, the Figures show that the portfolios attain similiar return and tracking error volatility. As for the cumulative returns, it shows that the lasso presents the higher returns. This translates into a bigger Sharpe Ratio (SR) for the lasso method. This higher mean is a double-edged knife. On the one hand, higher returns is a good to any investor. On the other hand, the higher returns detaches the portfolio from the index, and might increase the portfolio volatility. However, because we are only minimizing tracking error volatility, it seems that the lasso outperform the other methods. In terms of the Information Ratio (IR), again, the lasso method outperforms the other methods, in all rebalancing frequencies, and for the same reasons as the ones reported for the SR.

For the american dataset, the Figures show that the lasso portfolios attain higher tracking error volatitlity than the other methods. And the forward and backward iteration show very similar tracaking error volatility. As for the cumulative returns, it seems that the lasso presents the higher returns. In the risk-adjusted performance, the lasso presents the higher Sharpe Ratio (SR), but this comes in exchange for higher volatility. Once again, we remind the reader that we are only minimizing the tracking error volatility, so the lasso, in the american case, does not attain the goal we expected, even though it presents a higher sharpe ratio. In terms of the Information Ratio (IR), the story is the same as in the SR, and the lasso attains higher values of IR by exchanging more mean returns for more volatility, an undesirable transfer in this kind of portfolio.

Tables 2 and 4 shows statistics of the out of sample portfolio returns net of transaction costs. We used transaction costs equal to $50bp$ as in DeMiguel *et al.* (2009). The statistics shown are annualized Mean and Standard Deviation (SD). SR is the annualized Sharpe Ratio of those returns. TO is the turnover and is presented as a daily percentage average. Short interest is the size of the short position and is expressed in percentage. Active is the average number of active positions in the portfolio. The last three columns of the Table shows the alfa, beta, and R^2 (intercept in annualized terms, the slope, and the coefficient of determination) of a univariate regression of the out of sample portfolio returns net of transactions costs against the benchmark Index in the out of sample period.

For both the brazilian and the american dataset, the the lasso method does not always select the maximum number of assets. This might be the reason why the lasso presents lower turnover, which ultimately leads to lower transaction costs. Next, we inspect the average short positions of the portfolios. Here we can see that while the short positions increase with the active positions, for the lasso method this increase is much less pronounced than in the other methods. This, again, lead to lower turnover and transaction costs. Another observation is that the lasso penalty diminish the portfolio turnover. But for the penalty to have this effect, we should use the lasso weights, which we don't, as we separate the asset selection and the asset allocation phase. What might be happening here is that even without the penalty, the lasso still selects assets that will present less turnover.

Examining the risk-adjusted performance of the portfolios, we attest that the lasso, once again, has the better Sharpe Ration (SR). And once again this is due more to the higher mean return, than to lower volatility. We already expounded why the higher mean return has two sides to it. Other noteworthy fact is that once we consider transaction costs, the bwd and the fwd methods might present lower SR than the index itself. One last remark to clarify those results is that the portfolios only differ on its asset selection methods, the asset allocation phase is the same for all of them, so what we are examining is how well each method selects the assets that will enter our portfolio.

6. Concluding Remarks

Here, we presented portfolios formed by methods of variable selection to track a benchmark index. We used two different datasets for two different countries. One is an american dataset to track the SP500 index; the other is a brazilian dataset to track the Ibovespa index. In the emprical exercise, we segregated the asset

selection and the asset allocation phases to isolate the effect of asset selection for each method. In the asset allocation phase we used the Global Minimum Variance Portfolio with returns in excess of the benchmark to minimize the tracking error volatility of the portfolio, as in Liu (2009).

In the results section we can observe that, in terms of risk-adjusted performance, the lasso method outperforms the backward and forward selection. In the brazilian case, the lasso presents similar return and tracking error volatility, but with higher mean return. This translates into higher risk-adjusted performance for the lasso. This evidence is robust to rebalancing frequency and transaction costs. In the american case, the lasso also presents higher mean return, but it presents higher volatility, which ultimately leads to a higher sharpe ratio. However, because our primary goal is to minimize tracking error volatility, that exchange is undesirable to us.

Even though the backward and forward selection are outperformed by the lasso in some cases, there are some instances in which they can track the benchmark index pretty well. The positive performance of the portfolios, lead us to believe that simple variable selection methods can add value to index tracking portfolios. By simple methods we mean methods that are not computationally costly as Integer Quadratic Programming like in Santos (2015); Sant'Anna *et al.* (2014).

Table 1: Out of Sample Results for the Index Tracking Portfolios without Transaction Costs (IBOV)

	Mean	SD	SR	Mean TE	SD TE	IR	alpha	beta	R2
IBOV	18.38	31.52	0.58	0.00	0.00	NaN	0.00	1.00	1.00
Panel A: Monthly Rebalancing									
bwd.10	0.99	1.02	0.57	−0.23	9.14	−0.03	0.24	0.97	0.92
fwd.10	1.04	1.02	0.60	0.77	9.16	0.08	1.18	0.98	0.92
las.10	1.14	1.05	0.63	2.55	9.51	0.27	2.42	1.01	0.92
bwd.15	1.09	0.99	0.64	1.58	8.40	0.19	2.50	0.95	0.93
fwd.15	1.07	0.99	0.63	1.37	8.36	0.16	2.23	0.95	0.93
las.15	1.13	1.02	0.65	2.37	8.48	0.28	2.70	0.98	0.93
bwd.20	1.07	0.97	0.64	1.38	8.14	0.17	2.49	0.94	0.93
fwd.20	1.09	0.98	0.65	1.62	8.06	0.20	2.67	0.94	0.93
las.20	1.10	1.00	0.64	1.91	8.00	0.24	2.49	0.97	0.94
Panel B: Bimonthly Rebalancing									
bwd.10	1.03	1.02	0.59	0.53	9.07	0.06	1.00	0.97	0.92
fwd.10	1.07	1.02	0.61	1.21	9.20	0.13	1.56	0.98	0.92
las.10	1.17	1.05	0.65	3.05	9.62	0.32	2.97	1.00	0.92
bwd.15	1.12	0.98	0.66	2.19	8.38	0.26	3.17	0.95	0.93
fwd.15	1.10	0.99	0.65	1.79	8.39	0.21	2.65	0.95	0.93
las.15	1.20	1.02	0.69	3.65	8.55	0.43	4.02	0.98	0.93
bwd.20	1.09	0.97	0.65	1.58	8.15	0.19	2.73	0.94	0.93
fwd.20	1.12	0.97	0.67	2.14	8.18	0.26	3.25	0.94	0.93
las.20	1.14	1.00	0.67	2.62	7.97	0.33	3.23	0.97	0.94
Panel C: Quarterly Rebalancing									
bwd.10	1.03	1.02	0.59	0.53	9.10	0.06	1.00	0.97	0.92
fwd.10	0.98	1.02	0.56	−0.36	9.23	−0.04	0.14	0.97	0.92
las.10	1.15	1.05	0.64	2.74	9.52	0.29	2.73	1.00	0.92
bwd.15	1.08	0.98	0.64	1.49	8.44	0.18	2.51	0.94	0.93
fwd.15	1.07	0.98	0.64	1.36	8.40	0.16	2.34	0.95	0.93
las.15	1.13	1.01	0.65	2.39	8.65	0.28	2.83	0.98	0.93
bwd.20	1.07	0.97	0.65	1.29	8.19	0.16	2.50	0.93	0.93
fwd.20	1.08	0.97	0.65	1.49	8.19	0.18	2.69	0.93	0.93
las.20	1.10	1.00	0.64	1.76	8.06	0.22	2.43	0.96	0.94

Note for Table 1: In this Table the benchmark index is the Bovespa Index (IBOV). The Table shows annualized Mean and Standard Deviation (SD) of the Index and of the portfolio returns. The index is shown in absolute terms and the portfolios are shown in relation to the index (Mean and SD of the portfolios are divided by the Mean and the SD of the index). SR denotes the annualized Sharpe Ratio. The Table also shows the Mean and SD of the tracking error of the portfolios (portfolio return minus the index). And the IR denotes the annulized Information Ratio which is the Mean of the tracking error divided by its SD. The last three columns of the Table shows the alfa, beta, and R^2 (intercept in annualized terms, the slope, and the coefficient of determination) of a univariate regression of the out of sample portfolio returns against the benchmark Index in period from january 2000 to december of 2010.

Table 2: Out of Sample Results for the Index Tracking Portfolios with Transaction Costs (IBOV)

	Mean	SD	SR	TO	Short	Active	alpha	beta	R2
MKT	18.38	31.52	0.58	*NA*	*NA*	*NA*	0.00	1.00	1.00
Panel A: Monthly Rebalancing									
bwd.10	0.84	1.02	0.48	2.13	−0.08	10.00	−2.43	0.97	0.92
fwd.10	0.89	1.02	0.51	2.18	−0.24	10.00	−1.57	0.98	0.92
las.10	1.04	1.05	0.58	1.40	−0.74	9.99	0.66	1.01	0.92
bwd.15	0.95	0.99	0.56	1.99	−2.50	15.00	−0.00	0.95	0.93
fwd.15	0.93	0.99	0.55	2.08	−1.65	15.00	−0.38	0.95	0.93
las.15	1.03	1.02	0.59	1.41	−0.55	14.93	0.93	0.98	0.93
bwd.20	0.94	0.97	0.56	1.96	−8.04	20.00	0.02	0.94	0.93
fwd.20	0.95	0.98	0.57	2.00	−6.59	20.00	0.16	0.94	0.93
las.20	1.00	1.00	0.59	1.44	−1.14	19.95	0.67	0.97	0.94
Panel B: Bimonthly Rebalancing									
bwd.10	0.93	1.02	0.53	1.46	0.00	10.00	−0.83	0.97	0.92
fwd.10	0.97	1.02	0.55	1.41	−0.48	10.00	−0.21	0.98	0.92
las.10	1.10	1.05	0.61	1.01	−0.97	9.99	1.70	1.00	0.91
bwd.15	1.03	0.98	0.61	1.36	−2.72	15.00	1.46	0.95	0.93
fwd.15	1.00	0.99	0.59	1.41	−2.08	15.00	0.88	0.95	0.93
las.15	1.13	1.02	0.65	0.98	−0.59	14.97	2.78	0.98	0.93
bwd.20	0.99	0.97	0.60	1.35	−8.59	20.00	1.03	0.94	0.93
fwd.20	1.02	0.97	0.61	1.34	−7.12	20.00	1.56	0.94	0.93
las.20	1.07	1.00	0.63	1.01	−0.98	19.93	1.96	0.97	0.94
Panel C: Quarterly Rebalancing									
bwd.10	0.95	1.02	0.54	1.17	0.00	10.00	−0.46	0.97	0.92
fwd.10	0.90	1.01	0.52	1.11	0.00	10.00	−1.24	0.97	0.92
las.10	1.09	1.04	0.61	0.81	−0.80	10.00	1.72	1.00	0.92
bwd.15	1.00	0.98	0.60	1.11	−2.20	15.00	1.12	0.94	0.93
fwd.15	1.00	0.98	0.59	1.06	−1.68	15.00	1.01	0.95	0.93
las.15	1.08	1.01	0.62	0.79	−0.44	15.00	1.84	0.98	0.93
bwd.20	0.99	0.97	0.60	1.11	−7.99	20.00	1.11	0.93	0.93
fwd.20	1.01	0.97	0.61	1.05	−6.59	20.00	1.37	0.93	0.93
las.20	1.04	1.00	0.61	0.77	−1.08	19.96	1.47	0.96	0.94

Note for Table 2: In this Table the benchmark index is the Bovespa Index (IBOV). The Table shows annualized Mean and Standard Deviation (SD) of the out of sample portfolio returns net of transaction costs. We used $c = 50bp$ as in DeMiguel *et al.* (2009). SR is the annualized Sharpe Ratio of those returns. TO is the turnover and is presented as a daily percentage average. Short interest is the size of the short position and is expressed in percentage. Active is the average number of active positions in the portfolio. The last three columns of the Table shows the alfa, beta, and R^2 (intercept in annualized terms, the slope, and the coefficient of determination) of a univariate regression of the out of sample portfolio returns net of transaction errors against the benchmark Index in period from january 2000 to december of 2010.

Table 3: Out of Sample Results for the Index Tracking Portfolios without Transaction Costs (SP500)

	Mean	SD	SR	Mean TE	SD TE	IR	alpha	beta	R2
SP500	12.93	12.22	1.06	0.00	0.00	NaN	0.00	1.00	1.00
Panel A: Monthly Rebalancing									
bwd.20	1.16	1.03	1.19	2.01	2.75	0.73	2.01	1.00	0.95
fwd.20	1.05	1.04	1.07	0.66	2.90	0.23	0.48	1.01	0.95
las.20	1.26	1.06	1.26	3.39	4.10	0.83	3.38	1.00	0.90
bwd.30	1.06	1.02	1.11	0.80	2.22	0.36	0.82	1.00	0.97
fwd.30	1.09	1.03	1.12	1.16	2.39	0.49	1.05	1.01	0.96
las.30	1.20	1.03	1.23	2.61	3.39	0.77	2.67	1.00	0.93
bwd.40	1.10	1.01	1.16	1.36	1.93	0.70	1.41	1.00	0.98
fwd.40	1.12	1.02	1.16	1.55	2.08	0.75	1.46	1.01	0.97
las.40	1.12	1.03	1.15	1.52	2.93	0.52	1.55	1.00	0.95
Panel B: Bimonthly Rebalancing									
bwd.20	1.07	1.01	1.12	0.91	2.86	0.32	1.08	0.99	0.95
fwd.20	1.13	1.03	1.15	1.62	2.79	0.58	1.55	1.01	0.95
las.20	1.25	1.06	1.25	3.21	4.06	0.79	3.13	1.01	0.90
bwd.30	0.99	1.01	1.04	−0.16	2.31	−0.07	−0.02	0.99	0.96
fwd.30	1.19	1.02	1.23	2.48	2.35	1.05	2.44	1.00	0.96
las.30	1.20	1.03	1.23	2.62	3.39	0.77	2.69	0.99	0.93
bwd.40	1.04	1.00	1.10	0.47	1.96	0.24	0.62	0.99	0.97
fwd.40	1.18	1.01	1.23	2.35	2.03	1.16	2.35	1.00	0.97
las.40	1.16	1.02	1.20	2.02	2.99	0.68	2.16	0.99	0.94
Panel C: Quarterly Rebalancing									
bwd.20	1.26	1.01	1.31	3.35	2.73	1.23	3.48	0.99	0.95
fwd.20	1.16	1.04	1.18	2.06	2.85	0.72	1.87	1.02	0.95
las.20	1.22	1.06	1.22	2.84	4.18	0.68	2.84	1.00	0.90
bwd.30	1.23	1.00	1.30	2.99	2.22	1.34	3.21	0.98	0.97
fwd.30	1.20	1.02	1.24	2.65	2.33	1.14	2.55	1.01	0.97
las.30	1.19	1.04	1.22	2.45	3.44	0.71	2.49	1.00	0.93
bwd.40	1.24	1.00	1.31	3.09	1.92	1.61	3.27	0.99	0.98
fwd.40	1.18	1.01	1.23	2.34	1.93	1.21	2.31	1.00	0.98
las.40	1.16	1.03	1.20	2.06	2.93	0.70	2.10	1.00	0.95

Note for Table 3: In this Table the benchmark index is the SP500 Index (SP500). The Table shows annualized Mean and Standard Deviation (SD) of the Index and of the portfolio returns. The index is shown in absolute terms and the portfolios are shown in relation to the index (Mean and SD of the portfolios are divided by the Mean and the SD of the index). SR denotes the annualized Sharpe Ratio. The Table also shows the Mean and SD of the tracking error of the portfolios (portfolio return minus the index). And the IR denotes the annulized Information Ratio which is the Mean of the tracking error divided by its SD. The last three columns of the Table shows the alfa, beta, and R^2 (intercept in annualized terms, the slope, and the coefficient of determination) of a univariate regression of the out of sample portfolio returns against the benchmark Index in period from january 2012 to october of 2017.

Table 4: Out of Sample Results for the Index Tracking Portfolios with Transaction Costs (SP500)

	Mean	SD	SR	TO	Short	Active	alpha	beta	R2
SP500	12.93	12.22	1.06	NA	NA	NA	0.00	1.00	1.00
Panel A: Monthly Rebalancing									
bwd.20	0.53	1.05	0.54	6.42	0.00	20.00	−6.15	1.01	0.92
fwd.20	0.53	1.06	0.53	5.39	−0.03	20.00	−6.35	1.02	0.93
las.20	1.09	1.06	1.09	1.82	−5.85	19.93	1.08	1.00	0.90
bwd.30	0.47	1.03	0.48	6.07	0.00	30.00	−6.89	1.00	0.94
fwd.30	0.56	1.04	0.57	5.43	−0.08	30.00	−5.84	1.01	0.94
las.30	1.02	1.04	1.04	1.89	−9.01	29.88	0.29	1.00	0.93
bwd.40	0.55	1.03	0.57	5.66	0.00	40.00	−5.77	1.00	0.95
fwd.40	0.59	1.04	0.60	5.48	−0.13	40.00	−5.48	1.01	0.95
las.40	0.92	1.03	0.94	2.07	−8.91	39.88	−1.07	1.00	0.94
Panel B: Bimonthly Rebalancing									
bwd.20	0.74	1.03	0.76	3.43	0.00	20.00	−3.26	0.99	0.93
fwd.20	0.85	1.04	0.87	2.82	−0.05	20.00	−2.02	1.01	0.94
las.20	1.13	1.06	1.13	1.18	−5.89	19.92	1.65	1.01	0.90
bwd.30	0.66	1.02	0.69	3.33	0.00	30.00	−4.23	0.99	0.95
fwd.30	0.92	1.03	0.94	2.84	−0.06	30.00	−1.14	1.00	0.95
las.30	1.07	1.03	1.10	1.32	−8.68	29.92	1.02	0.99	0.93
bwd.40	0.74	1.01	0.77	3.09	0.00	40.00	−3.29	0.99	0.96
fwd.40	0.91	1.02	0.94	2.81	−0.08	40.00	−1.20	1.00	0.96
las.40	1.01	1.02	1.05	1.45	−9.34	39.92	0.33	0.99	0.94
Panel C: Quarterly Rebalancing									
bwd.20	1.02	1.02	1.05	2.49	0.00	20.00	0.36	0.99	0.94
fwd.20	0.96	1.04	0.97	2.06	0.00	20.00	−0.71	1.01	0.94
las.20	1.12	1.06	1.12	1.05	−5.74	19.96	1.52	1.00	0.89
bwd.30	1.01	1.01	1.06	2.30	0.00	30.00	0.32	0.98	0.95
fwd.30	1.00	1.03	1.03	2.08	−0.03	30.00	−0.06	1.01	0.96
las.30	1.08	1.04	1.10	1.10	−9.04	29.92	1.10	1.00	0.92
bwd.40	1.03	1.00	1.09	2.11	0.00	40.00	0.62	0.99	0.96
fwd.40	0.98	1.02	1.02	2.04	−0.04	40.00	−0.25	1.00	0.97
las.40	1.05	1.03	1.08	1.17	−8.45	39.88	0.62	1.00	0.94

Note for Table 4: In this Table the benchmark index is the SP500 Index (SP500). The Table shows annualized Mean and Standard Deviation (SD) of the out of sample portfolio returns net of transaction costs. We used $c = 50bp$ as in DeMiguel *et al.* (2009). SR is the annualized Sharpe Ratio of those returns. TO is the turnover and is presented as a daily percentage average. Short interest is the size of the short position and is expressed in percentage. Active is the average number of active positions in the portfolio. The last three columns of the Table shows the alfa, beta, and R^2 (intercept in annualized terms, the slope, and the coefficient of determination) of a univariate regression of the out of sample portfolio returns against the benchmark Index in period from january 2012 to october of 2017.

References

Barber, Brad M., & Odean, Terrance. 2000. Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. *The Journal of Finance*, **55**(2), 773–806.

Brodie, Joshua, Daubechies, Ingrid, De Mol, Christine, Giannone, Domenico, & Loris, Ignace. 2009. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, **106**(30), 12267–12272.

Clarke, Roger G, de Silva, Harindra, & Thorley, Steven. 2006. Minimum-Variance Portfolios in the U.S. Equity Market. *The Journal of Portfolio Management*, **33**(1), 10–24.

DeMiguel, Victor, Garlappi, Lorenzo, & Uppal, Raman. 2009. Optimal Versus Naive Diversification: How Inefficient is the 1-N Portfolio Strategy? *Review of Financial Studies*, **22**(5), 1915–1953.

Fan, Jianqing, Zhang, Jingjin, & Yu, Ke. 2012. Vast Portfolio Selection With Gross-Exposure Constraints. *Journal of the American Statistical Association*, **107**(498), 592–606.

Jagannathan, Ravi, & Ma, Tongshu. 2003. Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *Journal of Finance*, **58**(4), 1651–1684.

James, Gareth, Witten, Daniela, Hastie, Trevor, & Tibshirani, Robert. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Jansen, Roel, & van Dijk, Ronald. 2002. Optimal Benchmark Tracking with Small Portfolios. *The Journal of Portfolio Management*, **28**(2), 33–39.

Kirby, Chris, & Ostdiek, Barbara. 2012. It's All in the Timing: Simple Active Portfolio Strategies that Outperform Naïve Diversification. *Journal of Financial and Quantitative Analysis*, **47**(2), 437–467.

Liu, Qianqiu. 2009. On portfolio optimization: How and when do we benefit from high-frequency data? *Journal of Applied Econometrics*, **24**(4), 560–582.

McWilliams, Brian, & Montana, Giovanni. 2010. Sparse partial least squares regression for on-line variable selection with multivariate data streams. *Statistical Analysis and Data Mining*, **3**(3), 170–193.

Medeiros, Marcelo C, Vasconcelos, Gabriel, & Freitas, Eduardo. 2015. Forecasting Brazilian Inflation With High-Dimensional Models. *Brazilian Review of Econometrics*, **99**(2), nil.

Rao, C. Radhakrishna, Shalabh, Toutenburg, Helge, & Heumann, Christian. 2008. *Linear Models and Generalizations*. Springer Series in Statistics. Springer.

Rencher, Alvin C., & Schaalje, G. Bruce. 2007. *Linear Models in Statistics*. Wiley Interscience.

Sant'Anna, Leonardo, Filomena, Tiago, & Borenstein, Denis. 2014. Index Tracking with Control on the Number of Assets. *Brazilian Review of Finance*, **12**(1), 89–119.

Santos, André A.P. 2015. Beating the market with small portfolios: Evidence from Brazil. *EconomiA*, **16**(1), 22 – 31.

Tu, Jun, & Zhou, Guofu. 2011. Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, **99**(1), 204 – 215.

Wu, Lan, & Yang, Yuehan. 2014. Nonnegative Elastic Net and application in index tracking. *Applied Mathematics and Computation*, **227**, 541 – 552.

Wu, Lan, Yang, Yuehan, & Liu, Hanzhong. 2014. Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, **70**, 116 – 126.