Insper Instituto de Ensino e Pesquisa

*Inspirar para Transformar*

# Building portfolios of stocks in the São Paulo Stock Exchange usingRandom Matrix Theory

Leonidas Sandoval Junior

Adriana Bruscato

Maria Kelly Venezuela

**Insper Working Paper**

WPE: 270/2012

Insper | Instituto de Ensino e Pesquisa

*Inspirar para Transformar*

# Building portfolios of stocks in the São Paulo Stock Exchange using Random Matrix Theory

Leonidas Sandoval Junior *
Adriana Bruscato †
Maria Kelly Venezuela ‡

Insper, Institute of Education and Research

February 22, 2012

**Abstract**

By using Random Matrix Theory, we build covariance matrices between stocks of the BM&F-Bovespa (Bolsa de Valores, Mercadorias e Futuros de São Paulo), which is cleaned of some of the noise due to the complex interactions between the many stocks and the finiteness of available data. We also use a regression model in order to remove the market effect due to the common movement of all stocks. These two procedures are then used to build stock portfolios based on Markowitz's theory, trying to obtain better predictions of future risk based on past data. This is done for years of both low and high volatility of the Brazilian stock market, from 2004 to 2010.

**Keywords:** portfolio building; covariance matrix; random matrix theory; BM&F-Bovespa.

**JEL Codes:** G11, C02.

**Resumo**

Usando a Teoria da Matriz Aleatória, construimos matrizes de covariância entre ações da BM&F-Bovespa (Bolsa de Valores, Mercadorias e Futuros de São Paulo) que é limpa de parte do ruído devido às interações complexas entre as muitas ações e à quantidade finita de dados. Também utilizamos um modelo de regressão para remover o efeito do mercado devido ao movimento comum a todas as ações. Esses dois procedimentos são então usados na construção de carteiras baseadas na teoria de Markowitz, tentando obter melhores predições de risco futuro baseado em dados passados. Isto é feito para anos de baixa e de alta volatilidades do mercado financeiro brasileiro, de 2004 a 2010.

**Palavras-chave:** construção de carteiras; matriz de covariância; teoria da matriz aleatória; BM&F-Bovespa.

## 1 Introduction

Modern portfolio theory is largely based on Markowit's ideas, as in Markowitz (1962), Elton, Gruber, Brown, & Goetzmann (2009), and Bodie, Kane, & Marcus (2009), where a portfolio of various equities is built on the principle of minimizing risk given an expected return. Risk is assessed as the volatility of each stock that comprises the portfolio, as well as their covariance. Preference is given to stocks that have negative or low covariance between each other, which leads to diversification of the equities held in one particular portfolio.

*E-mail: LeonidasSJ@insper.edu.br (corresponding author)
†E-mail: AdrianaB@insper.edu.br
‡E-mail: MariaKV@insper.edu.br

Both volatility and covariance are integrated into the covariance matrix, which is built using the stock returns of past data. This is used in order to predict the risk of a portfolio, and it is usualy different from the realized risk of the same portfolio.

Three problems arise from this approach. The first one is that past data reflect the market as it was, and not as it will be. So, the theory assumes the hypothesis that future events shall mimic past events, which is usually not true, for it does not incorporate news releases, or the current mood of the market. There is not much that can be do about this, but to minimize effects of events that might change the behavior of a market, one cannot use past data that is too old.

That leads us to the second problem, which is the noise associated with past data that arises purely from the fact that the available data are finite. Since one cannot go back in time indefinitely, and even if one could, it wouldn't be advisable given the discussion in the preceding paragraph, there is only a limited amount of data (in our case, price quotations) from which to build a covariance matrix. The problem gets even more severe if we think that an efficient portfolio should be built from many and diverse equities. A third source of noise comes from the complex interactions between the many elements of a stock market: traders, news, foreign markets, and the very prices of stocks interact in order to guide the price of a stock. Those interactions are usually too complex to be acommodated by any econometric model.

So, all this noise is incorporated into the covariance matrix that is used in the attempt to forecast the risk of a particular portfolio, and if one can remove some of that noise from the matrix, one is then able to make better risk predictions. Frankfurter, Phillips, & Seagle (1971), Frankfurter, Phillips, & Seagle (1972), Dickinson (1994), Jobson & Korkie (1990), Michaud (1989), and Chopra & Ziemba (1993) made studies on the influence of noise and other factors on the covariance matrix in the building of portfolios. Most of the approaches for solving them involve the reduction of the dimensionality of the covariance matrix by introducing some structure into it, obtained by principal component analysis, separation of stocks into economic sectors, among other means - see Jorion (1986) and DeMiguel, Garlappi, Nogales, & Uppal (2009) for two of these approaches.

A technique first developed for the study of the nuclei of the atoms of heavier elements, called Random Matrix Theory - see Mehta (2004), compares the eigenvalues of a correlation matrix with those of a correlation matrix built from a purely random matrix. From such a comparison, one may then discern elements which are clearly not random, and study them separately. Such technique has been applied to a number of complex systems, and, particularly, to financial markets. Of the many results that were obtained, the building of portfolios that most closely resemble the realized risk of the future market, based on past data, is one of them, as in Laloux, Cizeau, Bouchaud, & Potters (2000) and Rosenow, Plerou, Gopikrishnan, & Stanley (2002), and it has been successfuly applied to stocks by Plerou, Gopikrishnan, Rosenow, Amaral, Guhr, & Stanley (2002) and Sharifi, Crane, Shamaie, & Ruskin (2004), and to mutual funds by Conlon, Ruskin, & Crane (2007).

Besides the cleaning of the correlation matrix, we used a regression model to remove the market effect on the asset returns. This procedure makes it possible to estimate the correlation matrix with greater precision, for there is just a part of the dependence which is due to the assets, which generates more reliable forecasts for the risk of a portfolio.

The contribution of this article is the use of a method which is capable of ameliorating the risk forecasts of a portfolio built with Brazilian stock market assets, based on past data. This method involves three steps: (1) the removal of the market effect of the assets; (2) the cleaning of the correlation matrix, which encodes the structure of the dependence of the assets being considered, based on Random Matrix Theory, and (3) the construction of portfolios using Markowitz's theory and the cleaned correlation matrix. In this article, we calculate portfolios of stocks with and without the removal of the market effect so as to compare both results.

In order to analyze the suitability of the proposed method, we shall use the daily returns of BM&F-Bovespa stocks with 100% liquidity, whith pairs of years ranging from 2004 to 2010. For each year being analyzed, we build a portfolio using data from the previous year in order to make a forecast of the risk for a determined year, and that forecasted risk is then compared with the realized risk in that year. As data used in this article include periods of both low and high volatility in the BM&F-Bovespa, in particular the data collected during the Subprime Mortgage Crisis of 2007 and 2008, we are able to study how this technique of cleaning the correlation matrix applies to times of high volatility.

The article is organized as follows: Section 2 introduces the basic concepts of Random Matrix Theory, which are then applied to the building of portfolios (according to Markowitz) with and without cleaning the

correlation matrix for short selling allowed, as well as the regression for the removal of the market effect. Section 3 shows the results of portfolio forecasts using data from 2004 to 2010 and compares them to the realized risks. The article ends with a conclusion and comments on years of high volatility.

# 2 Methodology

In this section, we briefly describe the method proposed for the construction of portfolios by cleaning the correlation matrix and removing the market effect, aiming at better forecasting of risk based on the previous behaviors of the assets. We use the year 2004 as an example of the application of such method in this section.

## 2.1 Random matrix theory

Random matrix theory had its origins in 1953, in the work of the German physicist Eugene Wigner (1955) (1958). He was studying the energy levels of complex atomic nuclei, such as uranium, and had no means of calculating the distance between those levels. He then assumed that those distances between energy levels should be similar to the ones obtained from a random matrix which expressed the connections between the many energy levels. Surprisingly, he could then be able to make sensible predictions about how the energy levels related to one another.

The theory was later developed, with many and surprising results arising. Today, random matrix theory is applied to quantum physics, nanotechnology, quantum gravity, the study of the structure of crystals, and may have applications in ecology, linguistics, and many other fields where a huge amount of apparently unrelated information may be understood as being somehow connected. The theory has also been applied to finance in a series of works dealing with the correlation matrices of stock prices, as well as with risk management in portfolios, as in Pafka & Kondor (2002), Onnela, Chakraborti, & Kaski (2003), Tola, Lillo, Gallegati, & Mantegna (2008), and Pantaleo, Tumminello, Lillo, & Mantegna (2011). For a recent review on the subject, see Bouchaud & Potters (2011).

The first result of the theory that we shall mention is that, given an $L \times N$ matrix with random numbers from a Gaussian distribution with zero mean and standard deviation $\sigma$, then, in the limit $L \to \infty$ and $N \to \infty$ such that $Q = L/N$ remains finite and greater than one, the eigenvalues $\lambda$ of such a matrix will have the following probability density function, called a Marčenku-Pastur distribution, developed in Marěnko & Pastur (1967):

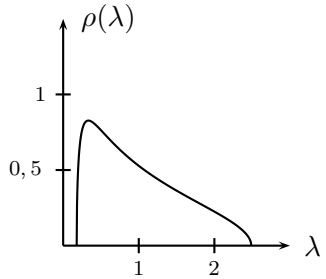
$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \ , \tag{1}$$

where

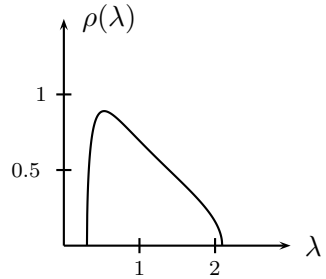
$$\lambda_- = \sigma^2 \left(1 + \frac{1}{Q} - 2\sqrt{\frac{1}{Q}}\right) \quad , \quad \lambda_+ = \sigma^2 \left(1 + \frac{1}{Q} + 2\sqrt{\frac{1}{Q}}\right) \ , \tag{2}$$

and $\lambda$ is restricted to the interval $[\lambda_-, \lambda_+]$.

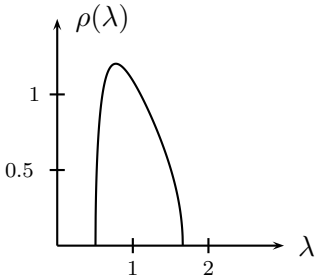
Figure 1 shows some of these distributions for diverse values of $Q$ and $\sigma$.



$Q = 3$ and $\sigma = 1$



$Q = 5$ and $\sigma = 1$



$Q = 12$ and $\sigma = 1$

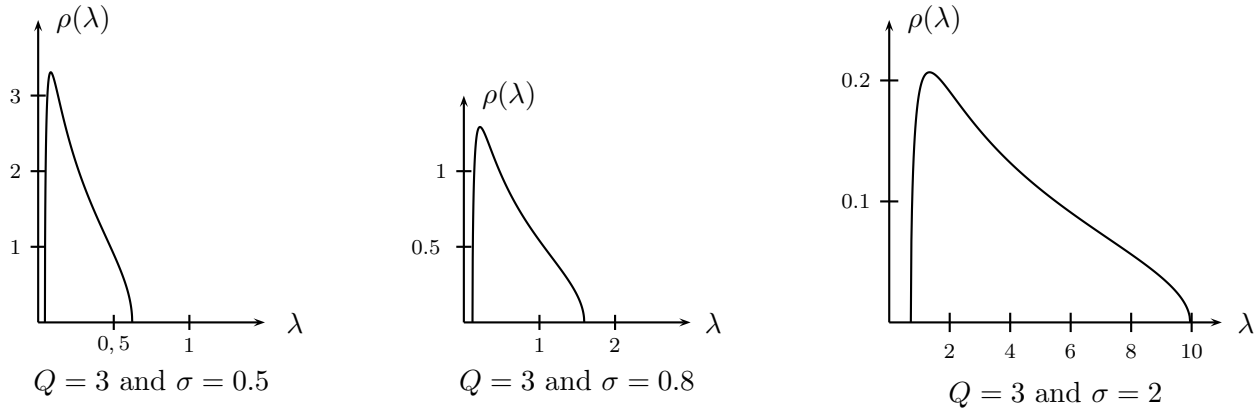



**Fig. 1.** Marčenku-Pastur distribution for fixed values of $Q$ and $\sigma$.

Since the distribution (1) is only valid for the limit $L \to \infty$ and $N \to \infty$, finite distributions will present differences from this behavior. Another source of deviations is the fact that financial time series are better described by non-Gaussian distributions, such as t Student or Tsallis distribution.

In Figure 2, we compare the theoretical distribution for $Q = 10$ and $\sigma = 1$ to distributions of the eigenvalues of three correlation matrices generated from finite $L \times N$ matrices such that $Q = L/M = 10$, and the elements of the matrices are random numbers with zero mean and standard deviation one.

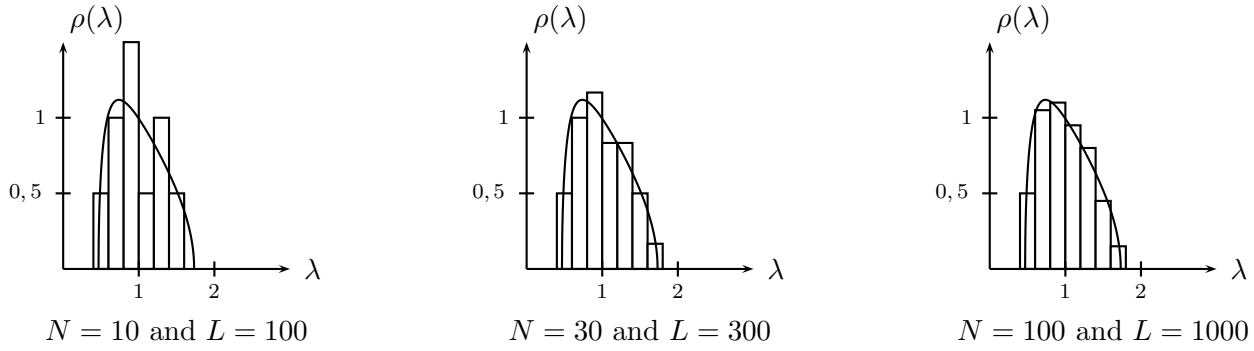



**Fig. 2.** Histogram of eigenvalues for generated correlation matrix and Marčenku-Pastur theoretical distribution (solid line) $Q = 10$ and $\sigma = 1$.

Consequently, real data will deviate from the theoretical probability distribution. Nevertheless, the theoretical result may serve as a parameter to the results obtained experimentally.

We shall now explain why random matrix theory is useful for portfolio building, starting by clarifying how it can be used to remove part of the noise from the correlation matrix. In order to do that we shall consider the data concerning the year 2004, the first of the years considered here in our study. For this period we chose Bovespa stocks (by then, Bovespa had not yet joined with BM&F) which were negotiated every trading day during the years 2004 and 2005 (2004 will be the past data that will be used to predict the risk in 2005). Those stocks are listed in Appendix A, totalizing 61 stocks.

For each stock, we calculated the returns, more precisely the log-returns, given by

$$R_t = \ln(P_t) - \ln(P_{t-1}) \approx \frac{P_t - P_{t-1}}{P_t} \ , \tag{3}$$

where $P_t$ is the closing price of one stock at the trading day $t$. Those returns were then normalized in order to obtain zero mean and standard deviation one by using the formula

$$X_t = \frac{R_t - \mu_R}{\sigma_R} \ , \tag{4}$$

where $\mu_R$ is the average of the time series used, and $\sigma_R$ is its standard deviation. This is done in order to best compare the resulting correlation matrix with the theoretical one, which is a random matrix with zero mean and standard deviation $\sigma$ chosen to be equal to one. The correlation matrix (a $61 \times 61$ matrix) between the variables $X_t$ for the year 2004 was then calculated.

The distribution density of the eigenvalues of the correlation matrix thus obtained is shown in Figure 3 (left picture). Also, the eigenvalues are plotted in order of magnitude in the right picture of Figure 3. The shaded area indicates the region predicted by the theory for the data related to a purely random behavior of the normalized returns.

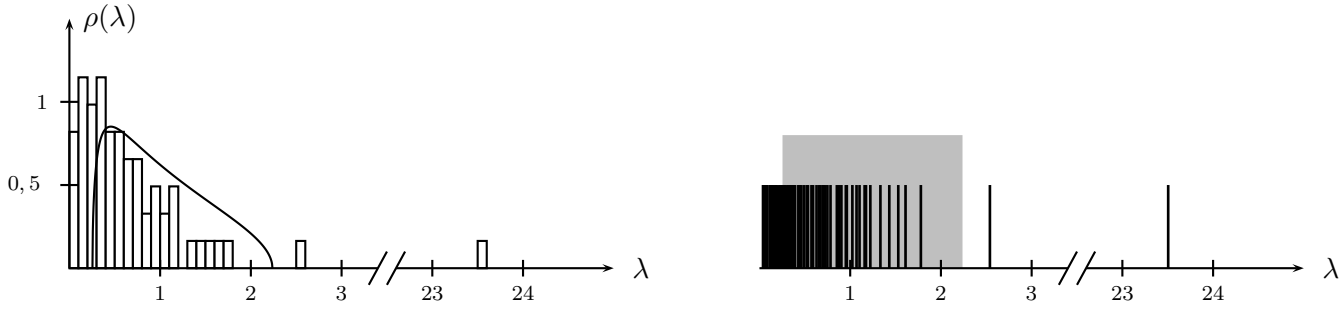



**Fig. 3.** Left: histogram of eigenvalues for the correlation matrix of 62 stocks in 2004 and Marčenku-Pastur theoretical distribution (solid line). Right: eigenvalues for the correlation matrix of 62 stocks in 2004 and purely random region.

We have $L = 248$ days of data for each of the $N = 61$ stocks, so that $Q = 248/61 \approx 4,06$. The probability distribution function for a random matrix with $L \to \infty$ and $N \to \infty$ with $Q \approx 4,06$ is also plotted in Figure 3, so that we may compare the result of pure noise with the one obtained for our data. The minimum ($\lambda_-$) and maximum ($\lambda_+$) values of the probability distribution function are given by

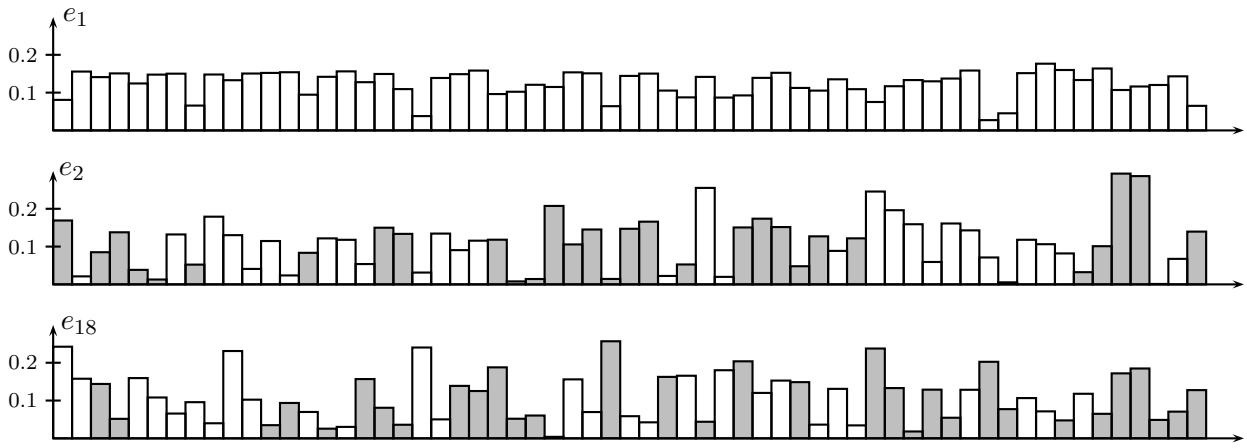
$$\lambda_- = 0.254 \quad \text{and} \quad \lambda_+ = 2.238 \; .$$

The first striking feature is that the largest eigenvalue is more than ten times larger than the maximum value predicted for a purely random correlation matrix. About 72% of the eigenvalues fall within the shaded region associated with pure noise, 15 of them fall below this region, and another one is above it.

The eigenvectors $e_1$ and $e_2$ for the two largest eigenvalues, $\lambda_1 = 23.505$, and $\lambda_2 = 2.540$, are represented in Figure 4 (first two graphs). The white bars represent positive values and the gray bars represent negative ones.

The distribution of individual values of eigenvector $e_1$ is very similar for all the stocks considered, showing that all stocks contribute to this mode, which is considered "the market mode". For eigenvalue $e_2$, one can see the prevalence of some stocks over others. In comparison, eigenvectors corresponding to the shaded region (Wishart region) do not show any preference for particular stocks.

The third and fourth graphs of Figure 4 show the eigenvectors distributions for the eigenvalues of two eigenvectors that are inside the region considered as noise, $\lambda_{18} = 0.853$, and $\lambda_{37} = 0.393$. Note that there are no clearly defined stock structures.



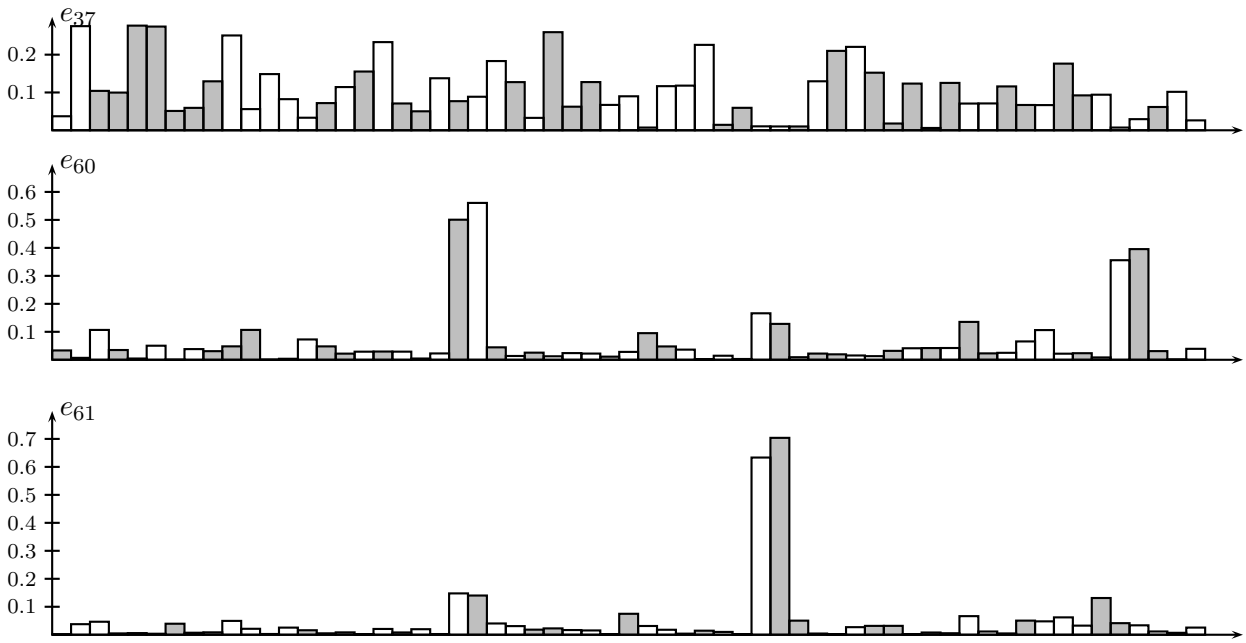



**Fig. 4.** Eigenvectors of some fixed eigenvalues: $\lambda_1$, $\lambda_2$ (largest), $\lambda_{18}$, $\lambda_{37}$ (noise region), and $\lambda_{60}$, $\lambda_{61}$ (lowest eigenvalues).

We also show the eigenvectors corresponding to the two lowest eigenvalues of the correlation matrix, $\lambda_{60} = 0,046$ and $\lambda_{62} = 0,039$ (last two graphs in Figure 4). These eigenvectors corresponding to low eigenvalues represent "portfolios" of low risk, in opposition to the eigenvectors of the largest two eigenvetors, which represent the oscillations of the market and the common behavior of a cluster of stocks that behave in a similar way. Eigenvector $e_{61}$ represents a portfolio from which the investor buys PETR4 and short-sells PETR3, which are stocks belonging to the same company, Petrobras, and buys ELET3 and short-sells ELET6, which also belong to the same company, Eletrobras. Eigenvector $e_{60}$, in its turn, represents a portfolio from which the investor buys VALE3 and ELET6 and short-sells in VALE5 and ELET3, which again are two pairs of stocks of the same companies, and also buys PETR3 and short-sells PETR4.

Figure 5 shows the daily log-returns of portfolio $P_1$ built with eigenvalue $e_1$, plotted against the log-returns of the Ibovespa, which is an index that describes the general behavior of the São Paulo Stock Exchange. The correlation between the two vectors is 0.9865, which is a very strong indication that the portfolio $P_1$ corresponds to a combination of stocks that behave much like the market, although with a much larger volatility: the standard deviation of the returns of $P_1$ is 12.51%, an the standard deviation for the Ibovespa is 1.80%.

The situation changes if we consider a portfolio built with eigenvector $e_{37}$, which corresponds to the noisy part of the eigenvalue spectrum: the correlation between this portfolio, $P_{37}$, and the Ibovespa is 0.1824, and it has a standard deviation 1.72%, very close to the standard deviation of the Ibovespa. For the portfolio $P_{61}$, built with eigenvector $e_{61}$, which corresponds to the lowest eigenvalue, the correlation with the Ibovespa is 0.0932, and its standard deviation is 0.44%. This portfolio presents the lowest correlation with the Ibovespa.

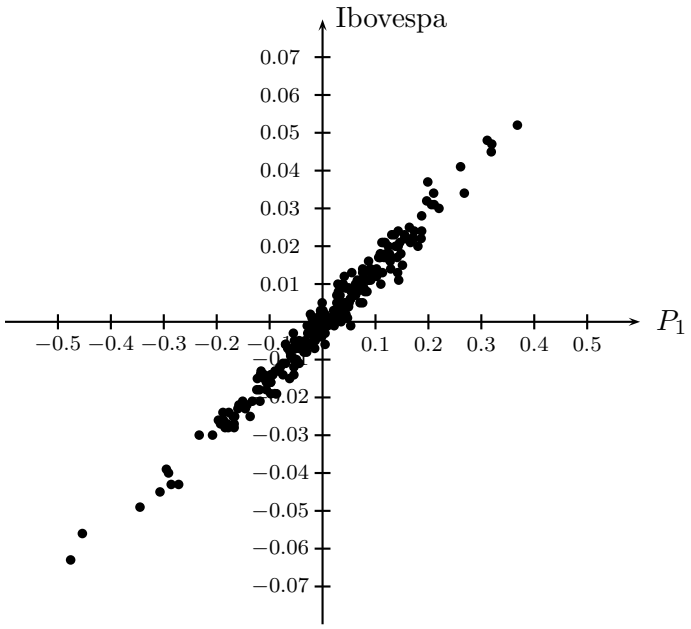



**Fig. 5.** Scatter plot of the portfolio built with $e_1$ and Ibovespa returns.

## 2.2 Building portfolios using Markowitz's Theory

In this section, we shall build portfolios using the $N = 61$ stocks we are considering based on the correlation matrix of their returns in the year 2004. According to the usual portfolio theory, we can obtain $w$, the vector of weights of the portfolio due to each stock, by fixing the portfolio return (RE) and minimizing the risk (RI) of the portfolio, as in Markowitz (1952).

The return of the portfolio is given by

$$RE = w^T R \ , \tag{5}$$

where $R$ is the vector of average returns of each stock in 2004.

The risk is defined by the variance of the portfolio

$$RI = w^T \Sigma_R w \ , \tag{6}$$

where $\Sigma_R$ is the estimated covariance matrix of the $N$ stocks.

The risk is then minimized with the constraint that the sum of all weights in the portfolio should be equal to one,

$$\sum_{i=1}^{N} w_i = 1 \ . \tag{7}$$

One can do that for several values of the average return, leaving the coordinates of $w$ free to assume negative values, as well as positive ones, so that short selling is allowed. In Finance, this is not always possible, or sometimes it is limited, and so we shall consider the case of no short selling later on.

In order to build a portfolio, the covariance matrix of a period of time (usually some months) prior to the period of investment is used together with a forecast of the expected returns. Those returns, which are unknown, may be approximated by many means, with relative degrees of success. There is a vast literature on the forecasting of returns, as in Elton,Gruber, Brown, & Goetzmann (2009) and bibliography therein, but this does not concerns us in our study of how to improve the prediction of risk. So, in order to restrict ourselves to the analysis of the correlation matrix, we shall consider that our prediction of returns is the best one possible, which is a perfect forecast of returns. Of course, if we had a perfect forecast of returns, and we knew it was a perfect forecast, we would not need to make any portfolio analysis. We use here the perfect forecast of returns in order to compare different ways of calculating risk in a fashion that is independent of the way one tries to forecast returns.

So, we first use the covariance matrix from 2004, together with the average returns of 2005 (perfect forecast of returns), in order to build minimum risk portfolios for 2005. Doing so, we build an *efficient frontier*, which is a curve whose coordinates are the minimum risk for a given return. We also use the data form 2005, which means perfect forecasts of risk and return, in order to build an efficient frontier for the realized risk.

In Figure 6, we have the predicted return and risk of portfolios using 2004 covariances of stocks (dashed line) and the realized return and risk using data from 2005 (solid line). Note that, for a given return, the predicted risk is smaller than the realized one. This may lead to a false perception of how risky an investment truly is, and may cause wrong decisions by the portfolio manager. The agreement of the curves (predicted and realized risk) can be measured by

$$AG = \frac{1}{n}\sum_{i=1}^{n}\frac{RI_i^{real} - RI_i^{pred}}{RI_i^{pred}} \ , \tag{8}$$

where $RI_i^{real}$ is the realized risk and $Ri_i^{pred}$ is the predicted risk, both for $i = 1, \cdots, n$ values of fixed returns. In our case, this number is $AG = -0.176$ ($n = 100$), which means that the predicted risk is, on average, 18% smaller than the realized risk.

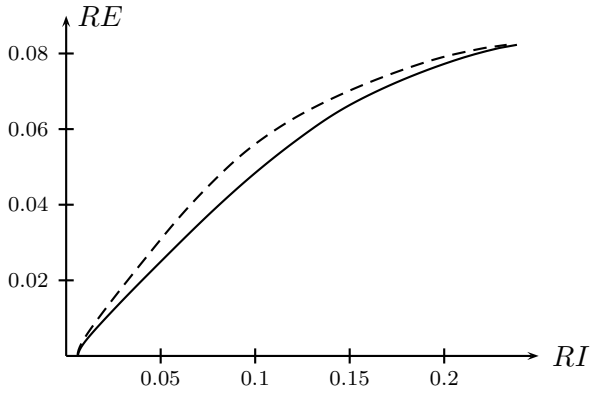



**Fig. 6.** Realized (solid line) and predicted (dashed line) returns and risks of portfolios for 2005.

### 2.3 Building portfolios with a cleaned correlation matrix

The situation may be improved by trying to remove some of the noise of the correlation matrices of 2004 and 2005 returns. One way this can be done is by building a diagonal matrix $D$ where the elements of the diagonal are the eigenvalues of the original correlation matrix, but now with all eigenvalues corresponding to noise (those between $\lambda_-$ and $\lambda_+$) replaced by their average, as it is done in Laloux, Cizeau, Bouchaud, & Potters (2000), Rosenow, Plerou, Gopikrishnan, & Stanley (2002), Plerou, Gopikrishnan, Rosenow, Amaral, Guhr, & Stanley (2002) and Sharifi, Crane, Shamaie, & Ruskin (2004), and Conlon, Ruskin, & Crane (2007). In our present case, this average is $\bar{\lambda} = 0.748$ for the eigenvalues based on data from 2004 and $\bar{\lambda} = 0.790$ for the eigenvalues based on data from 2005. The cleaned correlation matrix is then built using the formula

$$\mathbb{C}_{\text{clean}} = PDP^{-1} \ , \tag{9}$$

where $P$ are matrices whose columns are the eigenvectors of the original correlation matrix.

Calculating now the efficient frontier built with the covariance matrix obtained from the cleaned correlation matrix of 2004, together with the average returns of 2005 (perfect forecast of returns), dashed line, and comparing with the real curve calculated with the covariance matrix obtained from the cleaned correlation matrix of 2005, solid line, we obtain the results represented in Figure 7.

The ratio between predicted and realized risk has now gone from $AG = -0.176$ (Figure 6) to $AG = -0.102$ (Figure 7), which means that the predicted risk is, in average, 10% smaller than the realized risk. This is an improvement on the previous result and shows how the cleaning of the correlation matrix may help building portfolios which account best for the realized risk based on previous data.

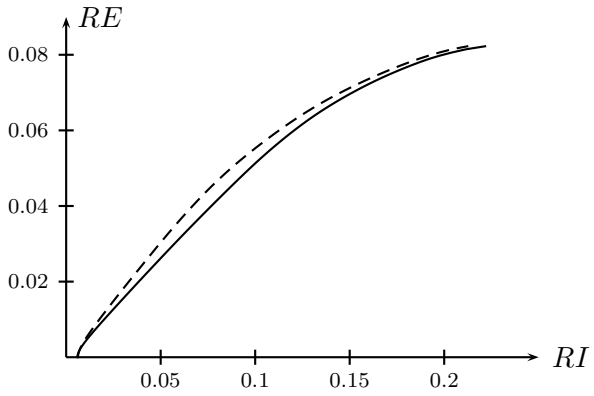



**Fig. 7.** Realized (solid line) and predicted (dashed line) returns and risks of portfolios for 2005 based on the cleaned correlation matrices.

### 2.4 Systemic risk

One of the stylized facts that remains, concerning the returns of assets, is the existence of volatility clustering. Generaly, those volatility clusterings occur due to patterns of the stock market, and they make it difficult to forecast the risk of portfolios, since they show a structure of dependence between the assets and the market, and not solely the dependence between assets. As an example, the prediction for 2008 using data from 2007 grossly underestimates the risk of 2008, for 2007 was a year with relatively low volatility while 2008 witnessed the height of the USA Subprime Mortgage Crisis. Similarly, risk prediction for 2009 using data from 2008 overestimates the risk for 2009. Figure 8a shows the volatility of the Ibovespa (the index for the São Paulo Stock Exchange) over the years considered in this article. The volatility is calculated as the absolute value of the log returns of the index. It clearly shows regions with high and low volatilities.

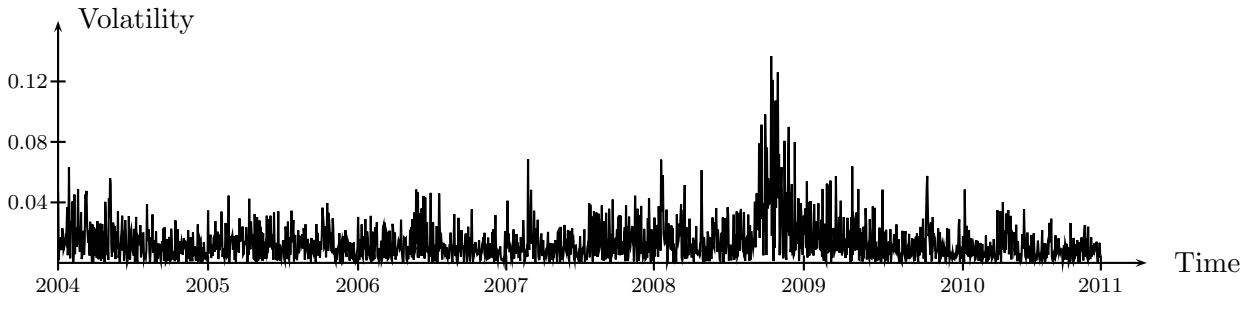
The most common way to remove this so called systemic risk is to use a sinlge index model, where all log returns $R_t$ are written in terms of a first degree function of a market index $I_t$, as, for example, the Ibovespa, plus an error $E_t$:

$$R_t = a + bI_t + E_t \ . \tag{10}$$

The coefficients $a$ and $b$ are estimated for each equity using simple linear regression.

As an alternative to the use of the Ibovespa as the market index, one may use the index obtained by the log returns of the portfolio of stocks that may be built using the eingenvector corresponding to the highest eigenvalue of the correlation matrix of those same stocks. As we have shown for the data concerning the year 2004, both this index and the Ibovespa are very highly correlated, so the results should not be substantially altered by using any of those two indices.

Figures 8b and 8c show, respectively, the volatility of the log returns and of the residuals of PETR4, stocks of Petrobras, a gas and oil company that is one of the largest assets of the BM&F-Bovespa in terms of negotiated volume. Note that the volatility of returns is less prone to effects due to market fluctuations, although those effects are still present in its variability. This remaining nonstationarity shall have its effects on the results in the next section.




**(a)**

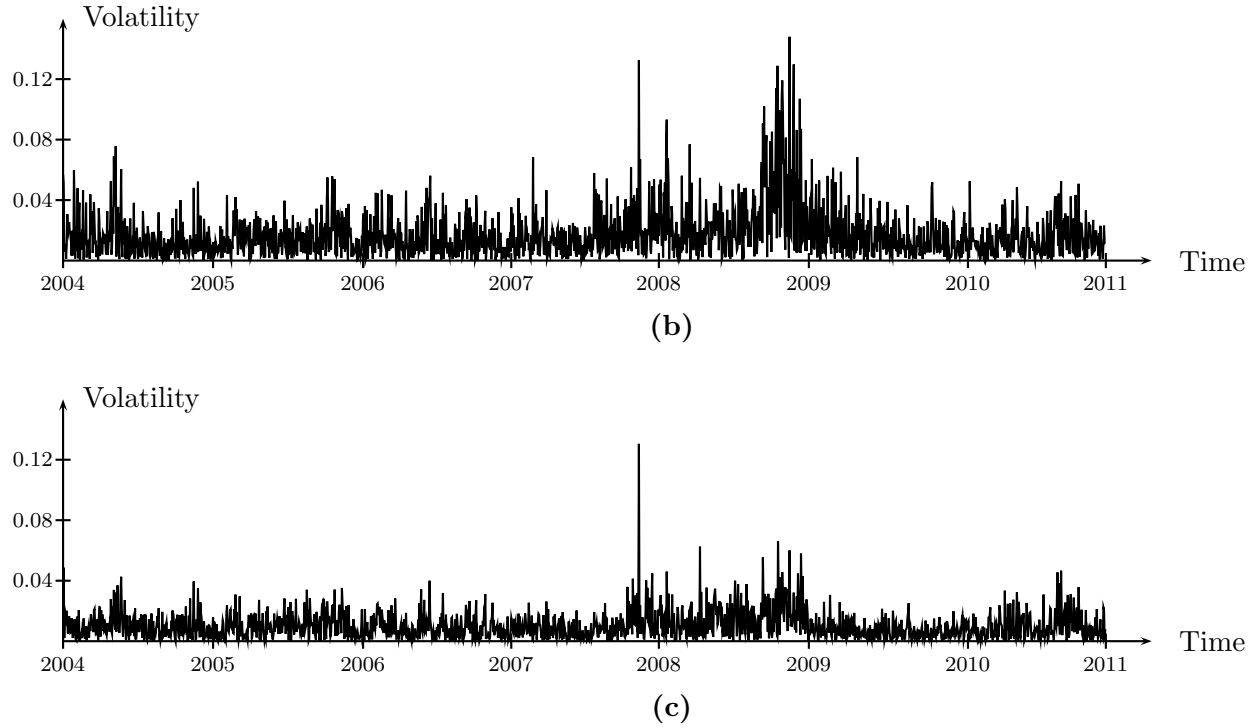



**Fig. 8.** Volatility of Ibovespa (a), PETR4 (b) and the residuals of PETR4 (c) from 2004 to 2010.

So, in what follows, we calculate the residuals for all stocks being considered for each pair of years using the Ibovespa as the market index. We then proceed to building portfolios using the correlation matrices between those residuals and also the cleaned correlation matrices.

## 3 Results

In this section, we shall describe the results obtained with the procedure presented in Section 2 for each of the years considered in this article. We shall make a comparison of the predicted risk and the realized risk for the years 2005 to 2010, using graps and calculating $AG$ in (8).

Table 1 presents the results for the measure $AG$ calculated for this method with and without the cleaning of the correlation matrix and with and wihtout the regression for the removal of the market effect, as well as the volatility of the Ibovespa in the predicted year.

| | Without Cleaning | | With Cleaning | | Ibovespa |
|---|---|---|---|---|---|
| Previous-Predicted | No Regression | Regression | No Regression | Regression | Volatility |
| $2004-2005$ | $-0,176$ | $-0,105$ | $-0.102$ | $0,021$ | $1,57\%$ |
| $2005-2006$ | $-0.212$ | $-0,079$ | $-0,203$ | $0,064$ | $1,53\%$ |
| $2006-2007$ | $-0.071$ | $0,018$ | $-0.056$ | $-0,101$ | $1,73\%$ |
| $2007-2008$ | $-1.011$ | $-0,220$ | $-0.841$ | $-0,283$ | $3,32\%$ |
| $2008-2009$ | $0.259$ | $0,290$ | $0.169$ | $0,263$ | $1,93\%$ |
| $2009-2010$ | $-0,056$ | $-0,218$ | $-0,116$ | $-0,148$ | $1,28\%$ |

**Table 1.** Agreement measure of the curves (AG) and the volatility of the Ibovespa in the predicted year.

One could notice that in Table 1 the majority of the forecast results were better with the use of regression in order to eliminate the effect of the market movements, and with the cleaning of the correlation matrix. Moreover, in every year in which there was a volatility drop from the previous year to the forecasted year (2005, 2006, 2009 and 2010), the cleaning of the correlation matrix showed better results, in other words, when volatility drops with respect to the previous year, it is best to use the cleaning of the correlation matrix, and

when volatility goes up with respect to the previous year, it is best not to use the cleaning of the correlation matrix. So, an investor may judge which would be the best method based on his expectations in terms of the volatility of the market or use both procedures, with and without the cleaning the correlation matrix, in order to decide his strategy.

Figure 9 contains the graphs of the return and realized risk (solid line), and predicted (dashed line) with the original correlation matrix (left plots) and the cleaned correlation matrix (right plots), always using regression in order to remove the market effect. Based on these graphs, one may notice that the cleaning procedure was best for years of low volatility of the Ibovespa.

The year 2005 was relatively quiet for the Bovespa (by then, it hadn't yet merged with the BM&F). So, there was low volatility for 2005, and the predictions of risk are better using the residuals of data concerning this year. For this period, we chose stocks of the Bovespa which were negotiated every trading day during the years 2004 and 2005. Those stocks are listed in the Appendix, totalizing 61 stocks. For the correlation matrix of 2004 and the returns of 2005, we obtain $AG = -0.105$, which means that the predicted risk is 10% smaller than the realized risk. For the cleaned correlation matrices, $AG = 0.021$, which means that the predicted risk is 2% larger than the realized risk (Table 1).

The year of 2006 was also a low volatility year for the Bovespa. For this period, we chose stocks of the Bovespa which were negotiated every trading day during the years 2005 and 2006. Those stocks are listed in the Appendix, totalizing 72 stocks. The predicted risk is 8% smaller than the realized risk for the original correlation matrix, and for the cleaned correlation matrix, the predicted risk is 6% larger than the realized risk.

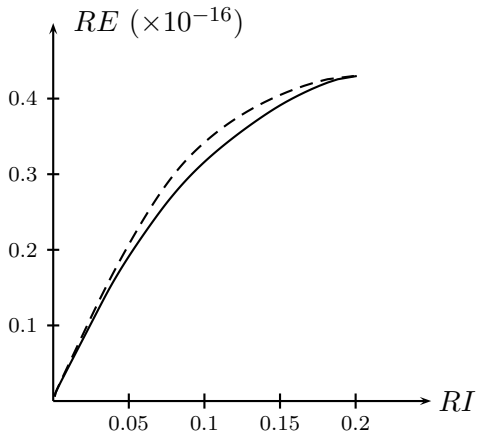
Although 2006 was a year of low volatility, 2007 witnessed the beginning of the USA Subprime Mortage Crisis, so there is a strong difference in volatility between those two years even when removing the market effect. For this period we chose stocks of the Bovespa which were negotiated every trading day during the years 2006 and 2007, listed in the Appendix, totalizing 86 stocks. For the original correlation matrix the predicted risk is 2% larger than the realized risk and for the cleaned correlation matrix, the predicted risk is 10% smaller than the realized risk.

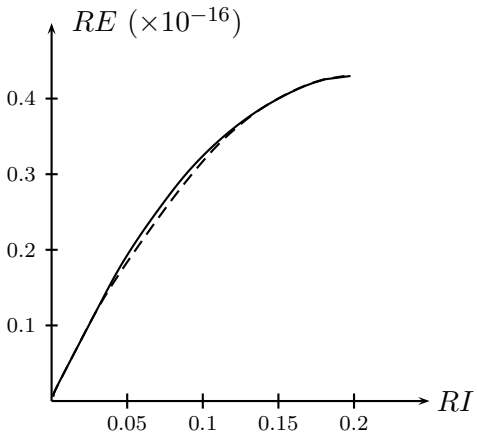
The year 2008 was the most turbulent year in Brazil and in stock exchanges worldwide since the Black Monday crisis of 1987. The USA Subprime Mortgage Crisis has spread to become a crisis of trust in financial institutions and a worldwide credit crisis. For this period, we chose stocks of the Bovespa (BM&F-Bovespa from 2008 onwards) which were negotiated every trading day during the years 2007 and 2008, also listed in the Appendix, totalizing 105 stocks. For the original correlation matrix, the predicted risk is 22% smaller than the realized risk, and for the cleaned correlation matrix, the predicted risk is 28% smaller than the realized risk. From all procedures, this was the year that resulted in the worst forecasts.

Volatility remained high for 2009, although in Brazil the international economic crisis did not have the same strength as in other countries. So, the BM&F-Bovespa had a quieter period than some other stock exchanges, although the movement in and out of foreign investment remained high. For this period, we chose stocks of the Bovespa which were negotiated every trading day during the years 2008 and 2009. Those stocks are listed in the Appendix, totalizing 148 stocks. For the original correlation matrix the predicted risk is 29% larger than the realized risk, and for the cleaned correlation matrices the predicted risk is 26% larger than the realized risk.

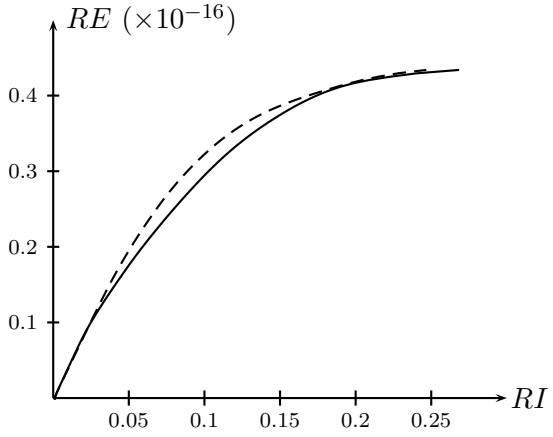
Volatility remained higher than normal for the year 2010, and apprehension, due to a succession of international financial crises, made the market unstable, but less than in the previous years. For this period we chose stocks of the Bovespa which were negotiated every trading day during the years 2009 and 2010, listed in the Appendix, totalizing 153 stocks. For the original correlation matrix the predicted risk is 22% larger than the realized risk, and for the cleaned correlation matrix the predicted risk is 15% larger than the realized risk.

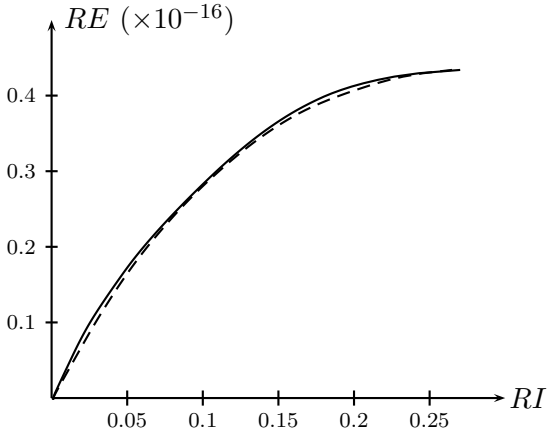
After the crisis of 2008, the forecasts have become poorer when compared with those of previous years, indicating that the higher the change in volatility in subsequent years, the worse are the forecasts obtained for portfolios.



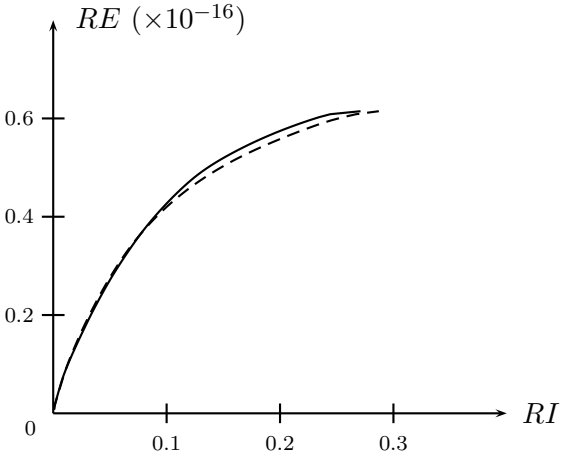



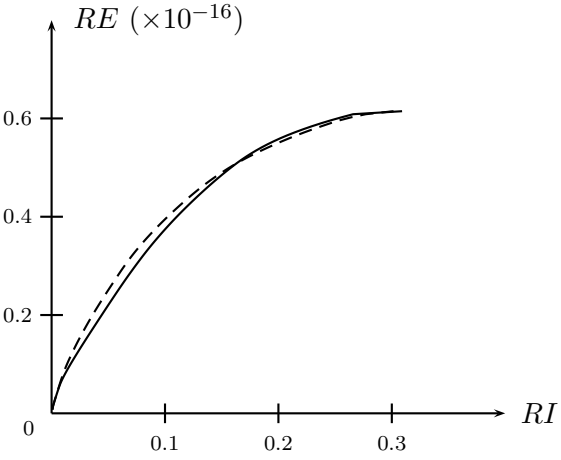

**(a)** 2005



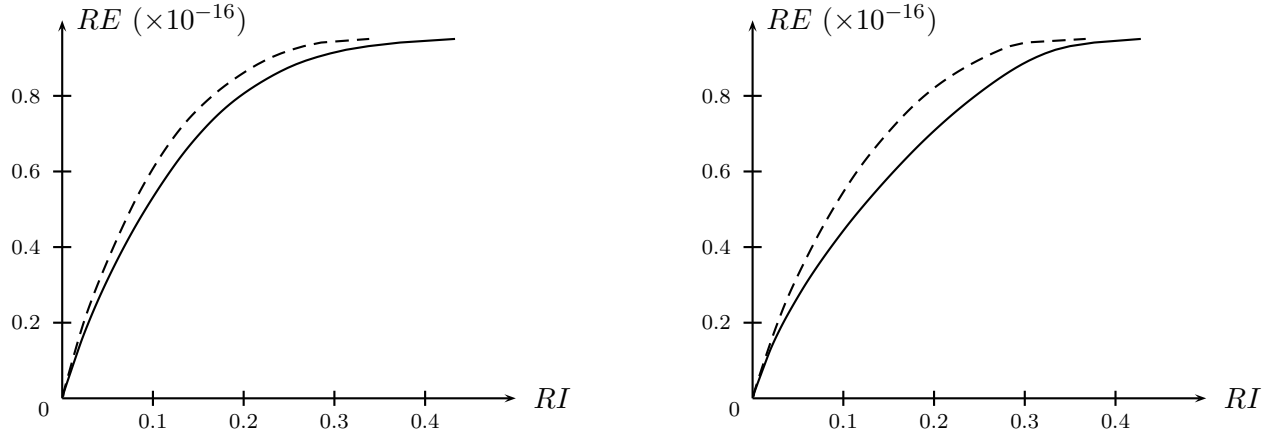




**(b)** 2006

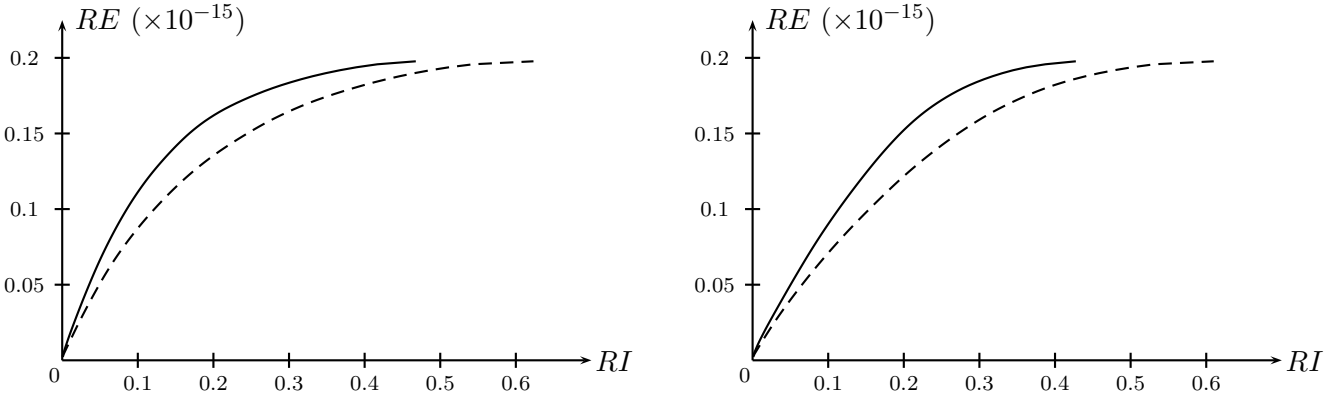


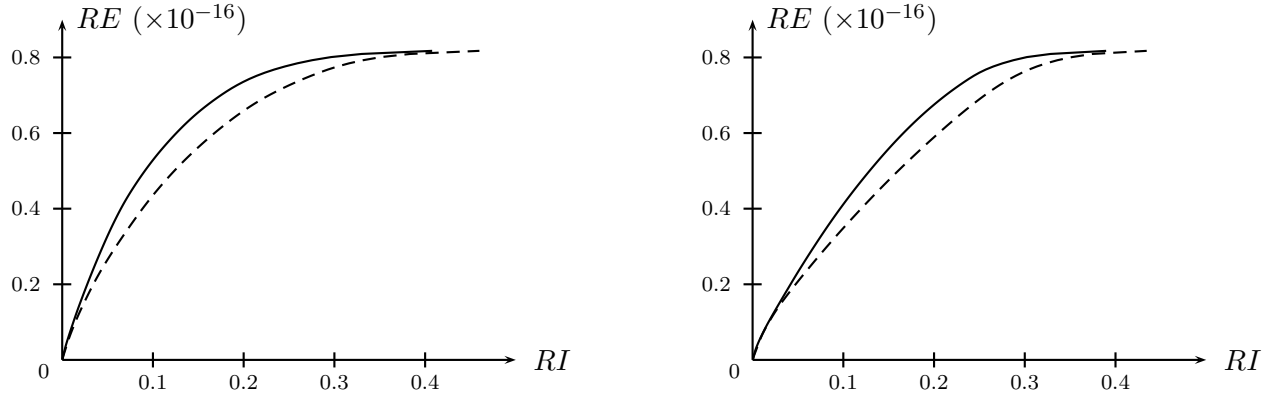

**(c)** 2007




**(d)** 2008




**(e)** 2009




**(f)** 2010

**Fig. 9.** Realized (solid lines) and predicted (dashed lines) returns and risks of portfolios using the original correlation matrices (left plots) and the cleaned correlation matrices (right plots) form 2005 to 2010.

As mentioned before, short selling is not usually freely allowed in financial transactions, mainly due to the increase in risks it might bring to a portfolio. So, we also did the calculations for no short selling, which implies the addition of a constraint that specifies that all values in the weight vector $w$ must be positive or zero. The results are summarized in Table 2, and they show that the conclusions to be derived from the case with no short selling allowed are basically the same as those from the case in which short selling is allowed: there is usually a significant better result when removing the market effect through a regression and, in years of a small

variation of volatility, the cleaning of the correlation matrices is a better method for forecasting risk, and it fails for the forecasting of risk for years of high volatility.

| | Without Cleaning | | With Cleaning | | Ibovespa |
|---|---|---|---|---|---|
| Previous-Predicted | No Regression | Regression | No Regression | Regression | Volatility |
| $2004-2005$ | 0.057 | $-0,066$ | 0.069 | $0,001$ | $1,57\%$ |
| $2005-2006$ | $-0.128$ | 0.041 | $-0,104$ | $0,074$ | $1,53\%$ |
| $2006-2007$ | $-0.263$ | $0,133$ | $-0.261$ | $-0,178$ | $1,73\%$ |
| $2007-2008$ | $-0.462$ | $-0,564$ | $-0.467$ | $-0,606$ | $3,32\%$ |
| $2008-2009$ | 0.506 | $0,322$ | 0.514 | $0,337$ | $1,93\%$ |
| $2009-2010$ | $-0,149$ | $0,207$ | $-0,136$ | $0,176$ | $1,28\%$ |

**Table 2.** Agreement measure of the curves (AG) and the volatility of the Ibovespa at the predicted year without the use of short selling.

# 4 Final remarks

In this article, we used two techniques in order to clean the covariance matrix used in the building of portfolios using Markowitz's theory. The first technique is the use of Random Matrix Theory in order to clean the correlation matrix built from the time series data of stocks in the year prior to that for which the portfolio is to be built. The second technique is to use a regression model in the removal of the market effect due to the common movement of all stocks. These are used in order to forecast the risk of a portfolio in a particular year using data from its previous year with better precision. The data were the time series returns of the 100% liquid assets of the BM&F-Bovespa covering the years from 2004 to 2010. The aim was to combine these two methods in different configurations, and to compare these results in order to obtain the best risk forecasts for portfolios.

Based on a measure of the aggreement (AG) between the forecasted and the realized risks, we conclude that the forecasted risk is closer to the realized risk, depending on the volatility of the forecasted year being smaller or larger than the volatility of the year used for the forecast. In the case there is a drop in volatility from one year to another, the results of the agreement measure were best when we built a portfolio using the cleaning of the correlation matrices. In the case there is a raise in volatility, it is best to use the original correlation matrix. This result is valid when the market effect is both removed and maintained by a regression.

The best performance of the method of cleaning the correlation matrices for the building of portfolios occurs because it eliminates the volatility (noise due to the complex interactions between the many stocks) which is useless for the forecasting of risk for a subsequent year of lower volatility. We noticed the advantage, in slightly more than half of the cases, in making the forecast of risk having the market effect removed, but without the presence of a clear pattern. Last, in the year of the greatest crisis (2008), the use of regression was particularly better in terms of previsibility.

So, the use of regression methods in the removal of market effects is usually advisable, but the use of Random Matrix Theory in the removal of noise from the correlation matrices tends to fail in the forecasting for years of high volatility, which are precisely the occasions in which a reliable risk forecast is most needed.

**References**


Bodie, Z., Kane, A., & Marcus, A.J. (2009). *Investments*, Eigth Edition , McGraww-Hill/Irwin.

Bouchaud, J-P, & M. Potters (2011). Financial applications of random matrix theory: a short review. In Akemann, G., Baik, J., Di Francesco, P. (editors), *The Oxford handbook of random matrix theory*, Oxford University Press.

Chopra, V., & Ziemba, W.T. (1993). The Effect of Errors in Mean and Co-Variance Estimates on Optimal Portfolio Choice. *J. Port. Management*, 6-11.

Conlon, T., Ruskin, H.J., & Crane, M. (2007). Random Matrix Theory and Fund of Funds Portfolio Optmizisation, *Physica A*, 382:565-578.

DeMiguel, V., Garlappi, L., Nogales, F.J., & Uppal, R. (2009). A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. *Management Science*, 55:782-812.

Dickinson, J.P. (1974). The Reliability of Estimation Procedures in Portfolio Analysis. *J. Fin. Quant. Anal.*, 9:447-462.

Elton, E.J., Gruber, M.J., Brown, S.J., & Goetzmann, W. (2009). *Modern Portfolio Theory and Investment Analysis*, Eigth Edition , Wiley.

Frankfurter, G.M., Phillips, H.E., & Seagle, J.P. (1971). Portfolio Selection: The Effects of Uncertain Means, Variances, and Covariances. *J. Fin. Quant. Anal.*, 6:1251-1262.

Frankfurter, G.M., Phillips, H.E., & Seagle, J.P. (1972). Estimation Risk in the Portfolio Selection Model: A Comment. *J. Fin. Quant. Anal.*, 7:1423-1424.

Jobson, J.D., & Korkie, B.M. (1980). Estimation for Markowitz Efficient Portfolios. *J. Am. Stat. Assoc.*, 75:544-554.

Jorion, P.(1986). Bayes-Stein Estimation for Portfolio Analysis. *J. Fin. Quant. Anal.*, 21:279-292.

Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (2000). Random Matrix Theory and Financial Correlations. *Int. J. Theor. Appl. Finance.*, 3:391-397.

Marěnko, V.A., & Pastur, L.A. (1967). *USSR-Sb*, 1:457-483.

Markowitz, H.M. (1952). Portfolio Selection. *The Journal of Finance*, 7:77-91.

Mehta, M. L. (2004). *Random Matrices*, Academic Press.

Michaud, R.O. (1989). The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *Fin. Anal. Journal*, 45:31-42.

Onnela, J-P., Chakraborti, A., & Kaski, K. (2003). Dynamics of market correlations: taxonomy and portfolio analysis. *Phys. Rev. E*, 68:056110.

Pafka, S., & Kondor, I. (2002). Noisy covariance matrices and portfolio optimization. *Eur. Phys. J. B*, 27:277-280.

Pantaleo, E., Tumminello, M., Lillo, F., & Mantegna, R.S. (2011). When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators. *Quantitative Finance*, 11:1067-1080.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., & Stanley, H.E. (2002). A Random Matrix Theory Approach to Cross-Correlations in Financial Data. *Phys. Rev. E*, 65:066126.

Rosenow, B., Plerou, V., Gopikrishnan, P., & Stanley, H.E. (2002). Portfolio optimization and the random magnet problem. *Europhys. Lett.*, 59:500.

Sharifi, S., Crane, M., Shamaie, A., & Ruskin, H. (2004). Random Matrix Theory for Portfolio Optimization: A Stability Approach. *Physica A*, 335:629-643.

Tola, V., Lillo, F., Gallegati, M., & Mantegna, R.N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32:235-258.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.*, 62:548-564.

Wigner, E. P.(1958). On the distribution of the roots of certain symmetric matrices. *Ann. Math.*, 67:325-327.

## Appendix: stocks of the BM&F-Bovespa used to build the portfolios

**2004-2005**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| AMBV4 | BBAS3 | BBDC3 | BBDC4 | BRAP4 | BRKM5 | BRTO4 | CCRO3 | CESP5 | CGAS5 |
| CLSC6 | CMIG3 | CMIG4 | CNFB4 | CPLE3 | CPLE6 | CRUZ3 | CSNA3 | CTNM4 | EBTP3 |
| EBTP4 | ELET3 | ELET6 | EMBR3 | FESA4 | FFTL4 | FIBR3 | GGBR4 | GOAU4 | INEP4 |
| ITSA4 | ITUB4 | KLBN4 | LAME4 | LIGT3 | MAGG3 | PCAR5 | PETR3 | PETR4 | POMO4 |
| RAPT4 | SBSP3 | SUZB5 | TBLE3 | TCSL3 | TCSL4 | TLPP3 | TLPP4 | TMAR5 | TNCP3 |
| TNCP4 | TNLP3 | TNLP4 | TRPL4 | UNIP6 | USIM5 | VALE3 | VALE5 | VIVO3 | VIVO4 |
| WEGE3 | | | | | | | | | |

**2005-2006**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| AMBV4 | BBAS3 | BBDC3 | BBDC4 | BRAP4 | BRKM5 | BRTO4 | CCRO3 | CESP5 | CGAS5 |
| CLSC6 | CMIG3 | CMIG4 | CNFB4 | CPFE3 | CPLE6 | CRUZ3 | CSNA3 | CTNM4 | DASA3 |
| EBTP3 | EBTP4 | ELET3 | ELET6 | EMAE4 | EMBR3 | FFTL4 | FIBR3 | FJTA4 | GGBR3 |
| GGBR4 | GOAU4 | GOLL4 | GRND3 | IDNT3 | ITSA4 | ITUB4 | KLBN4 | LAME4 | LIGT3 |
| MAGG3 | MGEL4 | NATU3 | PCAR5 | PETR3 | PETR4 | POMO4 | PSSA3 | RAPT4 | SBSP3 |
| SUZB5 | TBLE3 | TCSL3 | TCSL4 | TELB3 | TELB4 | TLPP3 | TLPP4 | TMAR5 | TNCP3 |
| TNCP4 | TNLP3 | TNLP4 | TRPL4 | UGPA4 | UNIP6 | USIM5 | VALE3 | VALE5 | VIVO3 |
| VIVO4 | WEGE3 | | | | | | | | |

**2006-2007**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ALLL3 | AMBV3 | AMBV4 | BBAS3 | BBDC3 | BBDC4 | BOBR4 | BRAP4 | BRKM5 | BRTO4 |
| BTOW3 | CCRO3 | CGAS5 | CLSC6 | CMIG3 | CMIG4 | CNFB4 | CPFE3 | CPLE6 | CRUZ3 |
| CSAN3 | CSNA3 | CTAX4 | CTNM4 | CYRE3 | DASA3 | ELET3 | ELET6 | EMAE4 | EMBR3 |
| ENBR3 | ENMA3B | ETER3 | FFTL4 | FIBR3 | FJTA4 | GGBR3 | GGBR4 | GOAU4 | GOLL4 |
| GRND3 | GUAR3 | IDNT3 | ITSA4 | ITUB4 | KLBN4 | LAME4 | LIGT3 | LREN3 | MAGG3 |
| MGEL4 | NATU3 | NETC4 | OHLB3 | PCAR5 | PETR3 | PETR4 | POMO4 | PSSA3 | RAPT4 |
| RENT3 | RSID3 | SBSP3 | SLED4 | SUZB5 | TAMM4 | TBLE3 | TCSL3 | TCSL4 | TELB3 |
| TELB4 | TLPP3 | TLPP4 | TMAR5 | TNLP3 | TNLP4 | TRPL4 | UGPA4 | UNIP6 | UOLL4 |
| USIM5 | VALE3 | VALE5 | VIVO3 | VIVO4 | WEGE3 | | | | |

**2007-2008**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ALLL3 | AMBV3 | AMBV4 | BBAS3 | BBDC3 | BBDC4 | BEES3 | BISA3 | BRAP4 | BRFS3 |
| BRKM5 | BRTO4 | BTOW3 | CARD3 | CCRO3 | CESP6 | CGAS5 | CLSC6 | CMIG3 | CMIG4 |
| CNFB4 | COCE5 | CPFE3 | CPLE6 | CRUZ3 | CSAN3 | CSMG3 | CSNA3 | CTNM4 | CYRE3 |
| DASA3 | ECOD3 | ELET3 | ELET6 | ELPL4 | EMBR3 | ENBR3 | EQTL3 | ESTR4 | ETER3 |
| FFTL4 | FIBR3 | FJTA4 | GETI3 | GETI4 | GFSA3 | GGBR3 | GGBR4 | GOAU4 | GOLL4 |
| GRND3 | IDNT3 | IMBI4 | ITSA4 | ITUB4 | JBDU4 | KLBN4 | LAME4 | LIGT3 | LPSB3 |
| LREN3 | LUPA3 | MAGG3 | MDIA3 | MLFT4 | NATU3 | NETC4 | ODPV3 | OHLB3 | PCAR5 |
| PETR3 | PETR4 | POMO4 | POSI3 | PSSA3 | RAPT4 | RENT3 | RNAR3 | RSID3 | SBSP3 |
| SLED4 | SUZB5 | TAMM4 | TBLE3 | TCSL3 | TCSL4 | TELB3 | TELB4 | TLPP3 | TLPP4 |
| TMAR5 | TNLP3 | TNLP4 | TOTS3 | TRPL4 | UGPA4 | UNIP6 | UOLL4 | USIM3 | USIM5 |
| VALE3 | VALE5 | VIVO4 | VLID3 | WEGE3 | | | | | |

**2008-2009**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ABCB4 | ALLL3 | AMAR3 | AMBV3 | AMBV4 | AMIL3 | BAZA3 | BBAS3 | BBDC3 | BBDC4 |
| BEEF3 | BEES3 | BEMA3 | BICB4 | BISA3 | BMTO4 | BPNM4 | BRAP4 | BRFS3 | BRKM5 |
| BRML3 | BRSR6 | BRTO4 | BTOW3 | BVMF3 | CARD3 | CCIM3 | CCRO3 | CESP6 | CGAS5 |
| CLSC6 | CMIG3 | CMIG4 | CNFB4 | COCE5 | CPFE3 | CPLE6 | CRUZ3 | CSAN3 | CSMG3 |
| CSNA3 | CYRE3 | CZRS4 | DASA3 | DAYC4 | ECOD3 | ELET3 | ELET6 | ELPL4 | EMBR3 |
| ENBR3 | EQTL3 | ESTR4 | ETER3 | EVEN3 | EZTC3 | FFTL4 | FHER3 | FJTA4 | GETI3 |
| GETI4 | GFSA3 | GGBR3 | GGBR4 | GOAU4 | GOLL4 | GPIV11 | GRND3 | IDNT3 | IGTA3 |
| INPR3 | ITSA4 | ITUB3 | ITUB4 | JBDU4 | JBSS3 | JHSF3 | KEPL3 | KLBN4 | LAME4 |
| LIGT3 | LOGN3 | LPSB3 | LREN3 | LUPA3 | MAGG3 | MILK11 | MLFT4 | MMXM3 | MRFG3 |
| MRVE3 | MULT3 | NATU3 | NETC4 | ODPV3 | OHLB3 | PCAR5 | PDGR3 | PETR3 | PETR4 |
| PINE4 | PLAS3 | PMAM3 | POMO4 | POSI3 | PRVI3 | PSSA3 | RAPT4 | RDCD3 | RDNI3 |
| RENT3 | ROMI3 | RSID3 | SBSP3 | SFSA4 | SLCE3 | SLED4 | SMTO3 | SULA11 | SUZB5 |
| TAMM4 | TBLE3 | TCSA3 | TCSL3 | TCSL4 | TELB3 | TELB4 | TGMA3 | TLPP3 | TLPP4 |
| TMAR5 | TNLP3 | TNLP4 | TOTS3 | TOYB3 | TOYB4 | TRPL4 | UGPA4 | UNIP6 | UOLL4 |
| USIM3 | USIM5 | VALE3 | VALE5 | VIVO4 | VLID3 | WEGE3 | WSON11 | | |

**2009-2010**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ABCB4 | ALLL3 | AMAR3 | AMBV3 | AMBV4 | AMIL3 | BBAS3 | BBDC3 | BBDC4 | BEEF3 |
| BEMA3 | BICB4 | BISA3 | BPNM4 | BRAP4 | BRFS3 | BRKM5 | BRML3 | BRSR6 | BRTO4 |
| BTOW3 | BVMF3 | CARD3 | CCIM3 | CCRO3 | CESP6 | CLSC6 | CMIG3 | CMIG4 | CNFB4 |
| COCE5 | CPFE3 | CPLE6 | CRDE3 | CRUZ3 | CSAN3 | CSMG3 | CSNA3 | CYRE3 | DASA3 |
| ECOD3 | ELET3 | ELET6 | ELPL4 | EMBR3 | ENBR3 | EQTL3 | ESTR4 | ETER3 | EVEN3 |
| EZTC3 | FESA4 | FFTL4 | FHER3 | FJTA4 | GETI3 | GETI4 | GFSA3 | GGBR3 | GGBR4 |
| GOAU4 | GOLL4 | GPIV11 | GRND3 | GSHP3 | HBOR3 | HYPE3 | IDNT3 | IGTA3 | INEP4 |
| INPR3 | ITSA4 | ITUB3 | ITUB4 | JBDU4 | JBSS3 | JHSF3 | KEPL3 | KLBN4 | KROT11 |
| LAME3 | LAME4 | LIGT3 | LLXL3 | LOGN3 | LPSB3 | LREN3 | LUPA3 | MAGG3 | MILK11 |
| MLFT4 | MMXM3 | MNDL4 | MPXE3 | MRFG3 | MRVE3 | MULT3 | MYPK3 | NATU3 | NETC4 |
| ODPV3 | OGXP3 | OHLB3 | PCAR5 | PDGR3 | PETR3 | PETR4 | PINE4 | PLAS3 | PMAM3 |
| POMO4 | POSI3 | PRVI3 | PSSA3 | RAPT4 | RDCD3 | RENT3 | RSID3 | SBSP3 | SFSA4 |
| SLCE3 | SLED4 | SMTO3 | SULA11 | SUZB5 | TAMM4 | TBLE3 | TCSA3 | TCSL3 | TCSL4 |
| TELB3 | TELB4 | TGMA3 | TLPP3 | TLPP4 | TMAR5 | TNLP3 | TNLP4 | TOTS3 | TOYB3 |
| TOYB4 | TPIS3 | TRPL4 | UGPA4 | UNIP6 | UOLL4 | USIM3 | USIM5 | VALE3 | VALE5 |
| VIVO4 | VLID3 | WEGE3 | | | | | | | |