Stock Movement Predictability and Classifier Induction

Pedro N. Rodriguez[a]
(Comments are welcome)

Latest version: March 29, 2005

Abstract

This paper uses classifier induction to categorize the predictable components in stock returns according to the particular movements they can actually predict. We document empirical results that suggest past returns can be used to *(a)* discriminate either absolute, or negative, or positive large returns from the rest of stock movements regardless of whether or not they exhibit low-order serial correlation, and *(b)* discriminate up from down movements only when such returns are serially correlated.

[a] Universidad Complutense de Madrid; E-mail: pedro_nahum@ucm.universia.es; I thank Arnulfo Rodriguez, Sandra Rottensteiner, Simon Sosvilla-Rivero, and seminar participants at the XII Foro de Finanzas in Barcelona and the discussant, Gábor Lugosi, for helpful comments. I also thank CONACYT (Mexico) for financial support (Fellowship: 170328). All errors are solely mine.

Both academic finance and industry practice have long been interested in predicting future stock returns by using publicly available information.[1] Although recent research indicates that short-horizon returns are predictable from past returns (see, e.g., Fama (1965); Lo and MacKinlay (1988); Conrad and Kaul (1988); Jegadeesh (1990); and Kaul (1996)), it is not clear what type of movements one is able to predict.

The purpose of this study is to try to categorize the predictable components in stock returns according to the particular movements they can actually predict. We examine the (CRSP) portfolio of firms with market values in the largest NYSE-AMEX quintile in the context of classifier induction, which provides us with several advantages over previous work. First, it allows us to explicitly evaluate the predictability of large price changes. Second, classification techniques enable us to assess the forecastability of absolute large returns. Third, classifier induction is a well-suited tool to test whether or not large U.S. stocks exhibit direction-of-change predictability, which is found in emerging market indices.[2]

Our evidence suggests that past returns can be used to *(a)* discriminate either absolute, or negative, or positive large returns from the rest of stock movements regardless of whether or not they exhibit low-order serial correlation,[3] and *(b)* discriminate up from down movements only when such returns are serially correlated. Even though these results do not necessarily imply that the stock market is inefficient or that stock prices are not rational appraisals of "fundamental" values, they improve our ability to describe the time-series behavior of security returns.

The remainder of the paper proceeds as follows. In Section I, we provide a brief review of the classification technique and accuracy measures employed in this study.

Movement codification and data sets are described in Section II. We apply the classification technique described in Section I to the daily returns of the (CRSP) portfolio of firms with market values in the largest NYSE-AMEX quintile, and report the out-of-sample results in Section III. We summarize briefly and conclude in Section IV.

I. Classification Techniques

In the function approximation problem one has a system consisting of a random response variable y and a set of random explanatory variables $\mathbf{x} = \{x_1, ..., x_n\}$. Given a training sample $\{y_i, \mathbf{x}_i\}_1^N$ of known $(y, \mathbf{x})$-values, the goal is to find a function $F^*(\mathbf{x})$ that maps $\mathbf{x}$ to y, such that over the joint distribution of all $(y, \mathbf{x})$-values, the expected value of some loss function is minimized.

Regression and classification problems can be viewed as a task in function approximation. In this paper we will focus on classification problems, which allow us to concentrate on movements of interest, such as large price changes or returns' signs. In a classification problem, the goal is to discriminate between two (or more) populations, given a set of explanatory variables. Are lagged returns capable of providing useful information for discriminating price changes solely under the presence of serially correlated returns?

To answer this question empirically, we must test the out-of-sample discriminatory accuracy of the classifiers trained to understand specific movements. In addition, the out-of-sample performance of the classifiers should be evaluated with a technique invariant to a priori class probabilities and independent of a decision threshold (or cut-off value), and the statistical significance of such performance must be assessed

through re-sampling techniques to avoid misleading inference due to a possible underestimation of parameter uncertainty.

Evaluating the out-of-sample predictability accuracy in a large test sample—e.g., one greater than one thousand observations (see, for instance, Henery (1994))—allows us to evade the poor approximation of the large-sample theory to the actual finite-sample distribution of test statistics when explanatory variables are persistent (see, e.g., Stambaugh (1999); and Elliot and Stock (1994)). In other words, we will not conclude that there is strong evidence of predictability of returns based on *t*-statistics or other non-parametric measures of predictor significance, such as relative contribution, partial dependence plots, or predictor's survival. In fact, we will confirm or refute the existence of predictable behavior of security returns through the comparison of the out-of-sample accuracy of the trained classifier versus the one of a random classifier.

Furthermore, assessing the forecastability with a technique invariant to a priori class probabilities and independent of a decision threshold (or cut-off value) permits the robust assessment of a classifier's accuracy in the presence of unbalanced data bases, a characteristic which the success rate (or classification error) criterion does not possess (see, e.g., Hand (1997)). Additionally, it eliminates the necessity of selecting a cut-off value with an ad hoc approach.

In Section I.A we provide a brief review of tree-based models, which are the cornerstone of the Gradient Boosting Machine, which is the classification technique used in this study. The Gradient Boosting Machine is described in Section I.B. The receiver operating characteristic (ROC) curve, used to evaluate the discriminatory accuracy, is discussed in Section I.C.

A. Tree-based models

The origins of classification trees or hierarchical classification come from two areas of investigation. In the field of statistical pattern recognition, Breiman, Friedman, Olshen, and Stone (1984) developed a technique named CART (Classification and Regression Trees). The Machine Learning community provided a computer program called ID3, which evolved into a new system named C4.5 (Quinlan (1986, 1993)).

Tree-based techniques partition the explanatory variables space into a set of rectangles and then fit a simple model to each one. A tree-based model tries to find the split that maximizes the decrement in an impurity function (or loss function) in order to make a tree grow. This is done iteratively until a certain amount of observations is reached or no further decrements in impurity functions are found. More formally, a tree may be expressed as

$$T(\mathbf{x};\Theta) = \sum_{j=1}^{J} \gamma_j I(\mathbf{x} \in R_j), \tag{1}$$

with parameters $\Theta = \{R_j, \gamma_j\}_1^J$. Where γ_j (a constant) is assigned to a region (R_j). The constant can be a value, a probability or a class label assigned to an element in the region R_j. J is usually treated as a meta-parameter and can be interpreted as the maximum amount of admissible interactions among explanatory variables less one, and $I(\bullet)$ is an indicator function. Here the parameters $\Theta = \{R_j, \gamma_j\}_1^J$ are found by minimizing the empirical risk, like in the following equation:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} L(y_i, \gamma_j) \tag{2}$$

where $L(\bullet)$ denotes a loss function. This is an extraordinary combinatorial optimization problem, so we must relay on sub-optimal solutions. The aforementioned optimization problem can be divided into two parts. The first one, finding γ_j given R_j, is typically trivial, where $\hat{\gamma}_j$ is the modal class of observations falling in region R_j. The difficulty of this combinatorial optimization problem is based on finding R_j. A helpful solution is to employ heuristic methods.

Safavian and Landgrebe (1991) provide a survey on heuristic methods proposed for designing decision trees. The most popular heuristic method in tree-based models is the top-down recursive partitioning, which starts with a single region covering the entire space of all joint input values. This is partitioned into two regions by choosing an optimal splitting input variable x_j and a corresponding optimal split point *s*. Values in $\mathbf{x}$ for which $x_j \le s$ are defined to be the left daughter region, and those for which $x_j > s$ denote the right daughter region. Then each of this two daughter regions is optimally partitioned with the same strategy, and so forth.

In this article we replaced the loss function with the Gini index, given by

$$\text{Gini index}: \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}). \tag{3}$$

where in a node *m*, representing a region R_j, let $\hat{p}_{mk}$ be the proportion of class *k* observations in the node *m*, and *K* represents the total number of classes or populations in the study.

B. The Gradient Boosting Machine

Boosting was created from the desire to transform a collection of weak classifiers into a strong ensemble or weighted committee. It is a general method for improving the performance of any learning algorithm. Boosting was proposed in the computational learning theory literature by Schapire (1990) and Freund (1995). Freund and Schapire (1997) solved many practical difficulties of earlier boosting algorithms with the creation of AdaBoost.M1.

Much has been written about the success of AdaBoost.M1 in producing accurate classifiers. In fact, one of the main characteristic of this procedure is that the test error seems to consistently decrease and then level off as more classifiers (trees) are added, without having an ultimately increase.

Recent explanations of AdaBoost.M1's performance have focused on Boosting as a gradient descent algorithm that minimizes some loss functions (Friedman, Hastie, and Tibshirani (2000)). Ridgeway (1999) traces the developments of boosting methodology and its applications to the exponential family and proportional hazards regression models.

Friedman (2001) applies Boosting to a variety of prediction settings, which includes non-linear and robust non-linear regression problems via a Gradient Boosting Machine (hereafter GBM). In this technique, function approximation is viewed from the perspective of numerical optimization in the function space, rather than in the parameter space. The generic algorithm of the GBM is shown in Figure 1. Hastie, Tibshirani, and Friedman (2002, p. 345) provide the specific algorithm for *k*-class classification problems.

****

Insert Figure 1 about here

****

With the purpose of increasing execution speed, approximation accuracy, and robustness against over-fitting we incorporated randomness in the procedure as described in Friedman (2002). At each iteration, a sub-sample consisting of 50 percent of the total observations is drawn at random (without replacement) from the training sample. This sub-sample is then used to fit the tree and compute the model output for the current iteration. The loss function employed in all our experiments with the GBM was the Bernoulli function, and as mentioned in Section I.A, the loss function used for the tree-based models was the Gini index.

The GBM has three tuning parameters: the total number of iterations M, the learning rate (shrinkage parameter v) and the level of interaction among explanatory variables J. It is widely known that fitting the data too well can lead to over-fitting, which degrades the accuracy power on independent data bases. However, M and v do not operate independently—i.e., smaller values of v lead to larger values of the "optimal" M.

Friedman (2001) finds that low values of v ($v < 1\%$) favor better accuracy on test samples. To get reliable estimates, we opted for five hundred trees in the forest (M) and fixed the shrinkage parameter (v) to one percent. Clearly, better results may be obtained if we monitor such parameters either in a validation set or with bootstrap's by product—i.e., out-of-bag estimates. However, these strategies were not carried out to avoid possible data-snooping effects and as a way to illustrate the effectiveness of the GBM as a tool for prediction.

Hastie, Tibshirani, and Friedman (2001) indicate that $4 \leq J \leq 8$ works well in the context of boosting with results being fairly insensitive to particular choices in this range. Consequently, we fixed *J* equal to five in all our experiments. It is worth mentioning that *J* also represents the stopping criteria of the top-down algorithm of the tree-based models.

Note that we do not use any forward information in estimating the GBM. In other words, fixing the GBM's parameters ensures that, under the presence of a test sample, we are computing (and evaluating) our probabilistic forecasts completely out-of-sample and without any 'look-ahead' or 'peeking' bias. The implementation was carried out in R: Environment for Statistical Computing and Graphics with the following add-on package: gbm (developed by Greg Ridgeway).

Admittedly, the use of predictive (machine) learning techniques in the domain of financial time series remains an important challenge for future research to develop a procedure capable of incorporating temporal dependence, e.g., how should one model the movements of a pure ARIMA(*p*,*d*,*q*) process? One promising direction of future investigation is to consider alternatives to the traditional bootstrap in the election of the training cases. Although the traditional bootstrap is useful for its simplicity, it suffers from a well-known deficiency, for instance, inappropriateness for time series data. A popular alternative that overcomes this particular deficiency is the moving block bootstrap (Künch (1989)), which divides the data into overlapping blocks of cases and sampling the blocks randomly with replacement. Another tempting line of future work is to consider alternatives in the tree's voting power, as the GBM use the "one tree one vote" system. Gaining ground alternatives include error diversity, variance-optimized and out-of-bag predictability-based voting systems. Such alternatives may yield

improvements in predictive (machine) learning techniques when applied to financial time series.

C. Assessing the discriminatory accuracy

Instead of analyzing the profitability of trading strategies based on the estimated probabilities, we will focus on predictability. In this way we eliminate the necessity of specifying an asset-pricing model, which is necessary for determining the economic source of trading profits. However, obtaining discriminatory accuracy in independent test samples does not guarantee a profitable trading strategy.

The receiver operating characteristic (ROC) curve is a well-established method for summarizing performances of diagnostic tests (see, e.g., Hanley (1999); and Zhou, McClish, and Obuchowski (2002)). Lately, the ROC curve has been gaining more ground in evaluating machine learning algorithms (see, e.g., Provost, Fawcett, and Kohavi (1998); Bradley (1998); Weiss and Provost (2003); and Rodriguez and Rodriguez (2005)). One of the most important applications of the ROC curve analysis is to evaluate the 'overall' performance of learning classifiers. The term 'overall' is emphasized, since the interest is in the global picture of a test and not in the accuracy performance at a particular cut-off value.

A ROC curve depicts the relationship of Sensitivity and 1 - Specificity of a learning classifier at various cut-off values used to distinguish population one cases (e.g., up-movements) from population two instances (e.g., down-movements). The points on a ROC curve are either joined by line segments (non-parametric approach) or smooth curves (parametric procedure).

The area under a ROC curve (AUC) is equal to the probability that a randomly selected observation from population one scores higher than a randomly selected observation from population two (Hanley and McNeil (1982)). This is true if and only if population one was codified with ones and population two with zeros, otherwise the reverse is true. Formally, the AUC can be expressed as:

$$\text{AUC} = P(y > x) + \frac{1}{2}P(y = x) \tag{8}$$

where y and x denote the classifier output (i.e. population one posterior probability) for a randomly selected observation from population one and two, correspondingly. A ranking probability equal to 1 and ½ would imply a perfect classifier and a random classifier, respectively.

Both parametric and nonparametric approaches can be used to derive an AUC index of accuracy. In this article, the Mann-Whitney-U Statistic, a nonparametric approach, was chosen to obtain the AUC. The Mann-Whitney-U Statistic is given by:

$$\text{AUC} = \widehat{\oplus} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij} \tag{9}$$

where $Z_{ij} = I(y_i > x_j) + \frac{1}{2}I(y_i = x_j)$ and $I(\bullet)$ is an indicator function. The variables m and n are the total number of observations for population one and two, respectively.

The AUC has a number of desired properties when compared to global accuracy (see, e.g., Bradley (1998)). However, an accurate estimate of AUC uncertainty is essential to avoid misleading inference. This uncertainty is usually summarized by standard errors. In this paper, we assess the AUC's standard error with three non-

parametric techniques. First, we use the Obuchowski and Lieber's (1998) version of DeLong et al.'s (1988) method given by

$$Var(\widehat{\oplus}) = \frac{\sum_{i=0}^{m}\left[\left[\sum_{j=0}^{n}\frac{Z_{ij}}{n}\right]-\widehat{\oplus}\right]^2}{m(m-1)} + \frac{\sum_{i=0}^{n}\left[\left[\sum_{j=0}^{m}\frac{Z_{ij}}{m}\right]-\widehat{\oplus}\right]^2}{n(n-1)}. \quad (10)$$

Second, we employ re-sampling techniques. Re-sampling techniques are computer-intensive statistical tools for estimating the distribution of a parameter that in other ways would be difficult to obtain.[4] The traditional bootstrap algorithm is: (1) draw a sample of size *m* with replacement from the observed sample of values from population one cases and a separate sample of size *n* with replacement from the observed sample of population two instances, (2) calculate the AUC, (3) repeat steps 1 and 2 thousand times to obtain a set of bootstrap replications, and (4) compute the variance of the set of bootstrap replications.

To incorporate into the analysis possible dependences in tests results, we employed the Block Bootstrap (Künch (1989)). The Block bootstrap (hereafter Bb) is the best-known method for implementing the bootstrap with time series data. It consists of dividing the data into overlapping blocks of cases and sampling the blocks randomly with replacement. The steps for the Bb go as follows: Given a weekly dependent time series $\mathbf{y} = \{y_1, ..., y_N\}$, (1) choose block's length *l*, (2) block one is $\mathbf{y}_j = \{y_j, ..., y_{j+l}\}$, block two is observations $\mathbf{y}_{j+1} = \{y_{j+1}, ..., y_{j+1+l}\}$, and so forth, (3) create a bootstrap sample by concatenating blocks that are drawn at random with replacements from the set of blocks, (4) compute the AUC, (5) repeat steps 3 and 4 thousand times to obtain a set of bootstrap replications, and (6) calculate the variance of the set of bootstrap replications.

In this paper we will determine the block's length with a heuristic approach when establishing the 95% confidence interval for the AUC. In this conservative heuristic approach, we will search for the parameter that maximizes the difference between upper and lower bounds. In doing so, we are able to find a parameter that is large enough to capture the dependences, since it is expected that the confidence interval widens as the parameter is increased. However, a plateau region will eventually be encountered as larger parameter values are analyzed, because large values of block's length will generate small variability in the re-sampled blocks. Hence, we will report the variance where the difference is maximized.

II. Movements codification

Regarding the dependent variable, four conditionals were evaluated and the class labels generated by each conditional were used as the response variable. In other words, for the CRSP index we have four *2*-class classification problems. Figure 2 shows the visualization of the movement codification algorithm per conditional evaluated. Population one and two will be represented by the observations codified with ones and zeros, respectively.

****

Insert Figure 2 about here

****

The first conditional evaluated was set to obtain daily up-and-down movements. To codify the response variable we used a logical expression. Specifically,

$$if\,(\text{return}_t > \text{Value})\{y_t \leftarrow 1;\}else\{y_t \leftarrow 0;\}. \tag{11}$$

The value of this conditional was fixed to zero to analyze the direction-of-change predictability. The second conditional tested was employed to obtain large positive price

changes. We used Equation (11) to codify the output. Nevertheless, the value was adjusted to obtain a population that analogously represented the right-tail of the distribution of returns. The value of the second conditional was tweaked to obtain a population represented by 20%, given by the complement of the accumulated probability of the distribution of returns.

The third conditional used was utilized from the yearning to study large negative price changes. The value of this conditional was modified to generate a population that is comparable to the left-tail of the distribution of returns. Thus, the value was calibrated to obtain a population with 20%, given by the accumulated probability of the distribution of returns. However, to properly monitor large negative price changes we must replace the "Greater than" operator in Equation (11) with the "Less than" operator.

To assuage the need of analyzing movements relevant to option traders, who usually implement volatility trading strategies (e.g., strangles, strips and straps), we modeled volatility-like price changes. The logical expression used to obtain the fourth response variable is given by

$$if\,(\text{return}_t < \text{Value}(3) \,\|\, \text{return}_t > \text{Value}(2))\{y_t \leftarrow 1;\}else\{y_t \leftarrow 0;\}. \qquad (12)$$

Where $\|$ denotes the logical operator OR, and Value(2) and (3) are given by those values obtained in the second and third conditional, respectively. The well-known asymmetry in the distribution in stock returns (see, e.g., Cheng, Hong, and Stein (2001) and references therein) will eventually force us to analyze asymmetric volatility-like movements, which are nonetheless useful for asymmetric option trading strategies, e.g., strips and straps. Regarding the set of explanatory variables, for any given time *t*, we used twenty-four return lags.

The virtual impossibility of control experimentation in economics and finance shows that care must be taken to avoid data snooping biases (Lo and MacKinlay (1990)) or model over-fitting (Bossaerts and Hillion (1999)). Indeed, conducting uncontrolled tests in the context of prediction entails that the estimated model, without tuning the model's parameters contingent on the error provided by the test data, should provide good out-of-sample predictability to assuage data snooping effects or model over-fitting biases.

Our out-of-sample predictability tests are based on a rolling window scheme. The fixed-size rolling window prunes the oldest observations at each update and recalibrates the estimated model. In the analysis, we fixed the window's size to twelve years of daily data. This scheme usually forecasts one step-ahead at each up-date. However, to get unbiased accuracy measures' estimates, it is necessary to run the estimated function in a large test sample. Consequently, we predicted the next one thousand and one observations with one step-ahead forecast and pruned the first thousand and one instances at each update. Clearly, this slight modification of the traditional rolling window scheme resembles a train-and-test procedure with multiple evaluation periods. Moreover, this strategy allows us to analyze whether or not predictable components have diminished over time. The empirical examinations use daily closing prices from July 2, 1962 to December 31, 2003.

Although the use of daily data in empirical tests is a contentious issue owing to the fact that it spawns well-known biases (e.g., bid-ask spread, non-trading, and non-synchronous price quotations), in recent years the price efficiency of the CRSP index has improved substantially. To illustrate, we performed a fixed-size rolling window analysis

of Lo and MacKinlay's (1988) variance ratio test from 1962 to 2003. In the analysis, we fixed the window's size to (approximate) twelve years of data, employed overlapping observations with aggregation ratio q of two, and measured weekly returns from Wednesday close to the following Wednesday close. Therefore, the first variance ratio test is estimated using data spanning from July 2, 1962 to July 5, 1974, the second using data from July 3, 1962 to July 8, 1974, and so forth. Figures 3A-B display the variance ratios (first y-axis) and heteroscedasticity-robust statistics (second y-axis) for daily- and weekly-holding period returns, respectively.

****

Insert Figure 3 about here

****

Note that under the random walk hypothesis, the value of the variance ratio is one and the test statistic has a standard normal distribution (asymptotically). Figures 3A-B show that the stochastic behavior of daily-holding stock returns in the post-1987 era is very similar to the one of weekly-holding returns—i.e., the random walk hypothesis is generally not rejected. It is worth noting that three days after the Black Monday of October 1987 large-capitalization stocks became generally linearly unpredictable. Thus, with post-1987 daily data, the biases associated with daily sampling are to some extent minimized whereas its virtue is maintained—i.e., large test samples.

III. Empirical Results

This section is divided into two parts. Direction-of-change predictability is analyzed in Section III.A. Section III.B shows the large price movement predictability accuracy of the estimated nonparametric function approximation technique.

A. Direction-of-change predictability

To answer whether or not linear correlatedness of price changes is a necessary condition to foresee stocks up-and-down movements, we applied the GBM technique described in Section I.B to CAP10's data. The results of the rolling window analysis are shown in Table I.

****

Insert Table I about here

****

Table I illustrates that coin-toss classification may be rejected at the usual significance level solely for the first two subperiods. Unsurprisingly, these periods are characterized by high positive auto-correlation. For example, from 1974 to 1978 the variance ratio test with an aggregation ratio q of two, displayed in Figure 3A, implies that the first-order auto-correlation in daily returns was higher than 24 per cent. Moreover, the aforementioned tests entail that from 1978 to 1982 a linear relationship with the lagged price change was able to explain between 3.16 and 6.76% of the variation of the current price change.

Evidently, the variance ratio's downward trend after 1982 had an effect on the GBM's ability to identify up-and-down movements. Particularly, during the period of the highest positive auto-correlation, the GBM was able to discriminate with 57.74% of accuracy positive from negative returns, and following 1978, its predictive power decreased to 54.78%. However, with post-1982 data, it mutated into a simple random classifier—past price changes did not provide useful information to foresee up-and-down movements! Hence, high degree of correlatedness of price changes is a necessary condition to obtain sign predictability.

Our finding of market timing ability under the presence of serially correlated returns is consistent with Allen and Karjalainen's (1999) finding of usefulness of daily prices, at the presence of positive low-order serial correlation in S&P 500 returns, to identify periods to be in the S&P 500 index when returns are positive and volatility is low and out when the reverse is true.

B. Large movement predictability

The lack of autocorrelation observed in some stock prices may not necessarily imply unforecastability, as Granger (1981) illustrates in a time series context. Can data-intensive techniques find a predictable behavior that the variance ratio, which is (approximately) a linear combination of autocorrelation coefficients, is not able to detect?

When we applied the GBM to the data to analyze whether or not past returns provide incremental information to forecast large movements, we had to consider the time-varying definition of large movements. Specifically, we sought a return value which represented a large price change in the training and testing sample (or rolling window's iteration). For example, the first train-and-test period spans from August 7, 1962 to July 25, 1978. Thus, we searched in that period for the necessary value that enabled us to codify either absolute, or negative, or positive large price changes. Table II reports the overall accuracy of the Gradient Boosting Machine when forecasting large positive price changes.

****

Insert Table II about here

****

In contrast to the results reported in Table I, the out-of-sample performance is satisfactory in all the time periods considered, even in periods where variance ratio tests

are not able to reject the random walk hypothesis. Although the highest accuracy is found in the first test period (August 9, 1974 to July 25, 1978), the predictable components have not diminished over time. Indeed, the external validity of the estimated model, in terms of discriminatory accuracy, is above 58 per cent after May 26, 1994, reaching to 61.65% from May 13, 1998 to May 7, 2002.

****

Insert Table III about here

****

Table III provides the out-of-sample results of the GBM when discriminating large negative price changes. Random classifications are rejected at the 5 per cent level in five evaluation periods. The second evaluation period from July 26, 1978 to July 12, 1982 ratifies the importance of incorporating into the analysis possible dependence's structures, since the AUC's standard errors were underestimated when assuming independence in test results. Table IV reports the external validity of the GBM when predicting large absolute price changes.

****

Insert Table IV about here

****

Table IV shows that for all evaluation periods, and as in Table II, in contrast to the direction-of-change predictability, random classifications are rejected at the 5 per cent of significance. However, the highest discriminatory accuracy, 69.14%, is found for the most recent data, rejecting, as with all large movements analyzed, the hypothesis that predictable components have been diminishing over time.[5]

IV. Conclusion

This article extents the work of Fama (1965), Lo and MacKinlay (1988), Conrad and Kaul (1988, 1989), Jegadeesh (1990), and many others concerning the ability of past prices to foresee future returns. The prior work points out a size asymmetry in the auto-correlation patterns—value-weighted portfolios of stocks with market values in the largest NYSE-AMEX quintile are not serially correlated.

We document a new empirical characteristic of the data—large movement predictability without auto-correlation patterns—which poses a new challenge to those seeking to explain the predictability patterns of short-horizon stock returns. Specifically, past prices of both correlated and uncorrelated financial time series may be used to classify better than random either absolute, or negative, or positive large price changes from the rest of price movements. Additionally, we find that a direction-of-change predictability is attributable to the low-order serial correlation in stock returns.

Although our results shed new light on how price changes can be predicted by past returns, and indicate that larger capitalization stocks' returns are predictable, it is a more difficult task to determine precisely whether or not information is transmitted from larger to smaller stocks, and this will pursued in subsequent research.

Notes

1. For a review of the literature on stock return predictability see Fama (1970, 1991).
2. Studies by Apte and Hong (1995), Tsaih, Hsu, and Lai (1998), Zemke (1999), Chen, Leung, and Daouk (2003), Kim (2003), and Rodriguez and Rodriguez (2004) provide evidence in support of direction-of-change predictability in short-horizon returns through classification techniques.
3. Nevertheless, the empirical evidence is somewhat weaker when distinguishing large negative returns from the remainder of price changes.
4. Re-sampling techniques are described in more technical detail in Hall (1992) and Davison and Hinkley (1997). Practical examples of confidence interval construction are given by Efron and Tibshirani (1993). Guide for choosing a bootstrap confidence method when using nonparametric or parametric simulation is given by Carpenter and Bithell (2000).
5. Patently, we cannot report all the estimated committees (or ensembles) with either their relative importance measures or partial dependence plots for brevity's sake. Nevertheless, they are available upon request to the authors.

References

Allen, Franklin, and Risto Karjalainen, 1999, Using genetic algorithms to find technical trading rules, *Journal of Financial Economics* 51, 245-271.

Apte, Chidanand, and Se June Hong, 1995, Predicting equities returns from securities data with minimal rule generation, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, eds.: *Advances in Knowledge Discovery and Data Mining* (AAAI Press).

Bossaerts, Peter, and Pierre Hillion, 1999, Implementing statistical criteria to select return forecasting models: What do we learn?, *Review of Financial Studies* 12, 405-428.

Bradley, Andrew P., 1998, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition* 30, 1145-1159.

Breiman, Leo, 1998, Arcing classifiers, *Annals of Statistics* 26, 801-849.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, 1984, *Classification and Regression Trees* (Wadsworth, Belmont, California).

Carpenter, James, and John Bithell, 2000, Bootstrap confidence intervals: when?, which?, what? A practical guide for medical statisticians, *Statistics in Medicine* 19, 1141-1164.

Chen, An-Sing, Mark T. Leung, and Hazem Daouk, 2003, Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan stock index, *Computers and Operations Research* 30, 901-923.

Chen, Joseph, Harrison Hong, and Jeremy C. Stein, 2001, Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices, *Journal of Financial Economics* 61, 345-381.

Conrad, Jennifer, and Gautam Kaul, 1988, Time varying expected returns, *Journal of Business* 61, 409-425.

Conrad, Jennifer, y Gautam Kaul, 1989, Mean reversion in short-horizon expected returns, Review of Financial Studies 2, 225-240.

Davison, Anthony C., and David V. Hinkley, 1997, *Bootstrap Methods and their Applications* (Cambridge University Press, USA).

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson, 1988, Comparing the areas under two or more correlated receiver operating characteristics curves: A nonparametric approach, *Biometrics* 44, 837-845.

Efron, Bradley, and Robert Tibshirani, 1993. *An Introduction to the Bootstrap* (Chapman and Hall, London).

Elliot, Graham, and James H. Stock, 1994, Inferences in time series regressions when the order of integration of a regressor is unknown, *Econometric Theory* 10, 672-700.

Fama, Eugene F., 1965, The behavior of stock market prices, *Journal of Business* 30, 34-105.

Fama, Eugene F., 1970, Efficient capital markets: A review of theory and empirical work, *Journal of Finance* 25, 383-417.

Fama, Eugene F., 1991, Efficient capital markets: II, *Journal of Finance* 46, 1575-1617.

Freund, Yoav, 1995, Boosting a weak learning algorithm by majority, *Information and Computation* 121, 256-285.

Freund, Yoav, and Robert E. Schapire, 1997, A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and Systems Sciences* 55, 119-139.

Friedman, Jerome H., 2001, Greedy function approximation: A gradient boosting machine, *Annals of Statistics* 29, 1189-1232.

Friedman, Jerome H., 2002, Stochastic gradient boosting, *Computational Statistics and Data Analysis* 38, 367-378.

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani, 2000, Additive logistic regression: A statistical view of boosting, *Annals of Statistics* 28, 337-374.

Granger, Clive W.J., 1981, Forecasting white noise, in Ghysels, Eric, Norman R. Swanson, and Mark W. Watson, eds.: *Essays in Econometrics: Collected Papers of Clive W.J. Granger* (Cambridge University Press, USA).

Hanley, James A., 1999, Receiver operating characteristic (ROC) curves. *Encyclopedia of Biostatistics* (Wiley).

Hanley, James A., and Barbara J. McNeil, 1982, The meaning and use of the area under a ROC curve, *Radiology* 143, 29-36.

Hall, Peter, 1988, Theoretical comparison of bootstrap confidence intervals (with Discussion), *Annals of Statistics* 16, 927-985.

Hall, Peter, 1992. *The Bootstrap and Edgeworth Expansion* (Springer-Verlag, London).

Hand, David J., 1997, *Construction and Assessment of Classification Rules* (John Wiley & Sons, Chichester).

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman, 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer-Verlag, New York).

Henery, R.J., 1994, Methods for comparison, in Donald Michie, David J. Spiegelhalter, and Charles C. Taylor, eds.: *Machine Learning, Neural and Statistical Classification* (Ellis Horwood, Hemel Hempstead, UK.).

Jegadeesh, Narasimhan, 1990, Evidence of predictable behavior of security returns, *Journal of Finance* 45, 881-898.

Kaul, Gautam, 1996, Predictable components in stock returns, in G.S. Maddala and C.R. Rao, eds.: *Handbook of Statistics, Vol. 14, Statistical Methods in Finance* (Elsevier Science B.V., The Netherlands).

Kim, Kyoung-jae, 2003, Financial time series forecasting using support vector machines, *Neurocomputing* 55, 307-319.

Leung, Mark T., Hazem Daouk, and An-Sing Chen, 2000, Forecasting stock indices: a comparison of classification and level estimation models, *International Journal of Forecasting* 16, 173-190.

Lo, Andrew W., and A. Craig MacKinlay, 1988, Stocks market prices do not follow random walks: Evidence from a simple specification test, *Review of Financial Studies* 1, 41-66.

Lo, Andrew W., and A. Craig MacKinlay, 1990, Data-Snooping biases in test of financial asset pricing models, *Review of Financial Studies* 3, 431-467.

Lo, Andrew W., and A. Craig MacKinlay, 1999. *A Non-Random Walk down Wall Street* (Princeton University Press, New Jersey).

Mech, Timothy S., 1993, Portfolio return autocorrelation, *Journal of Financial Economics* 34, 307-344.

Obuchowski, Nancy A., and M.L. Lieber, 1998, Confidence intervals for the receiver operating characteristic area in studies with small samples, *Academic Radiology* 5, 561-571.

Pesaran, M. Hashem, and Allan Timmermann, 1995, Predictability of stock returns: Robustness and economic significance, *Journal of Finance* 50, 1201-1228.

Pesaran, M. Hashem, and Allan Timmermann, 2002, Market timing and return prediction under model instability, *Journal of Empirical Finance* 9, 495-510.

Provost, Foster, Tom Fawcett, and Ron Kohavi, 1998, The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning* (Morgan Kaufmann, San Francisco, CA), 445-453.

Quinlan, J. Ross, 1986, Induction of decision trees, *Machine Learning* 1, 81-106.

Quinlan, J. Ross, 1993, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, California).

Ridgeway, Greg, 1999, The state of boosting, *Computing Science and Statistics* 31, 172-181.

Rodriguez, Pedro N., and Arnulfo Rodriguez, 2004, Predicting stock market indices movements, in Marco Costantino and Carlos Brebbia, eds.: *Computational Finance and its Applications* (Wessex Institute of Technology, Southampton).

Rodriguez, Arnulfo, and Pedro N. Rodriguez, 2005, Understanding and predicting sovereign debt rescheduling: A comparison of the areas under receiver operating characteristic curves, *Journal of Forecasting*, forthcoming.

Safavian, S. Rasoul, and David Landgrebe, 1991, A survey of decision tree classifier methodology, *IEEE Transactions on Systems, Man, and Cybernetics* 21, 660-674.

Schapire, Robert E., 1990, The strength of weak learnability, *Machine Learning* 5, 197-227.

Singh, K, 1981, On the asymptotic accuracy of Efron's bootstrap, *Annals of Statistics* 9, 1187-1195.

Stambaugh, Robert F., 1999, Predictive regressions, *Journal of Financial Economics* 54, 375-421.

Tsaih, Ray, Yenshan Hsu, and Charles C. Lai, 1998, Forecasting S&P 500 stock index futures with a hybrid AI system, *Decision Support System* 23, 161-174.

Weiss, Gary M., and Foster Provost, 2003, Learning when training data are costly: The effect of class distribution on tree induction, *Journal of Artificial Intelligence Research* 19, 315-354.

Zemke, Stefan, 1999, Nonlinear index prediction, *Physica A* 269, 177-183.

Zhou, Xiao-Hua, Donna K. McClish, and Nancy A. Obuchowski, 2002. *Statistical Methods in Diagnostic Medicine* (John Wiley & Sons Inc).

Table I. Direction-of-change predictability: out-of-sample results

For the (CRSP) portfolio of firms with market values in the largest NYSE-AMEX quintile, positive returns were codified with 1's and the rest with 0's. We test if the Gradient Boosting Machine is able to discriminate the cases codified with 1's from the instances encoded with 0's using past returns as explanatory variables. To assess the discriminatory accuracy, we employed the well-known ROC-curve summary index: the area under the ROC curve (AUC). The uncertainty of the AUC was summarized by three standard errors derived from: (1) DeLong et al. (1988)'s non-parametric formula, (2) Bootstrap Simulation, and (3) Moving Block Bootstrap simulation. The AUC is reported in the third column, and the standard errors, given in parenthesis, are reported in the next three columns. Under random (or coin-toss) classifications the AUC is not statistically different from 50%. Standard errors marked with asterisks indicate that the corresponding AUC is statistically different from 50 per cent. The dates of the training and testing periods are reported along with test sample discriminatory accuracy.

			Standard Error (%)		
Training period	Test sample	AUC (%)	DeLong et al.	Bootstrap	Moving Block
19620807-19740808	19740809-19780725	57.74	(1.80)*	(1.82)*	(1.82)*
19660727-19780725	19780726-19820712	54.78	(1.82)*	(1.90)*	(1.98)*
19700826-19820712	19820713-19860625	51.13	(1.82)	(1.87)	(1.90)
19740814-19860625	19860626-19900611	53.59	(1.84)	(1.87)	(1.82)
19780801-19900611	19900612-19940525	53.04	(1.82)	(1.80)	(1.94)
19820719-19940525	19940526-19980512	51.88	(1.83)	(1.87)	(1.85)
19860702-19980512	19980513-20020507	51.13	(1.82)	(1.84)	(1.92)
19900618-20020507	20020508-20031231	46.59	(2.83)	(2.88)	(2.90)

Table II. Large positive price changes predictability: out-of-sample results

Large positive movements analogously represent the right-tail of the distribution of returns. For each train-and-test period, returns higher than the value shown in the third column were codified with 1's and the rest with 0's. We test if the Gradient Boosting Machine is able to discriminate the cases codified with 1's from the instances encoded with 0's using past returns as explanatory variables. Methodology for evaluating the out-of-sample discriminatory accuracy of Gradient Boosting Machine as in Table I.

				Standard Error (%)		
Training period	Test sample	Value	AUC (%)	DeLong et al.	Bootstrap	Moving Block
19620807-19740808	19740809-19780725	0.51%	63.66	(1.97)*	(2.00)*	(2.03)*
19660727-19780725	19780726-19820712	0.62%	56.78	(2.06)*	(2.11)*	(2.07)*
19700826-19820712	19820713-19860625	0.67%	59.45	(2.14)*	(2.13)*	(2.21)*
19740814-19860625	19860626-19900611	0.71%	56.35	(2.31)*	(2.30)*	(2.60)*
19780801-19900611	19900612-19940525	0.67%	57.29	(2.56)*	(2.59)*	(2.88)*
19820719-19940525	19940526-19980512	0.65%	60.26	(2.22)*	(2.22)*	(2.31)*
19860702-19980512	19980513-20020507	0.68%	61.65	(1.98)*	(1.93)*	(2.06)*
19900618-20020507	20020508-20031231	0.69%	58.60	(3.13)*	(3.18)*	(3.20)*

Table III. Large negative price changes predictability: out-of-sample results

Large negative movements analogously represent the left-tail of the distribution of returns. For each train-and-test period, returns lower than the values shown in the third column were codified with 1's and the rest with 0's. We test if the Gradient Boosting Machine is able to discriminate the cases codified with 1's from the instances encoded with 0's using past returns as explanatory variables. Methodology for evaluating the out-of-sample discriminatory accuracy of Gradient Boosting Machine as in Table I.

				Standard Error (%)		
Training period	Test sample	Value	AUC (%)	DeLong et al.	Bootstrap	Moving Block
19620807-19740808	19740809-19780725	-0.48%	62.28	(2.01)*	(2.09)*	(2.16)*
19660727-19780725	19780726-19820712	-0.59%	54.51	(2.20)*	(2.19)*	(2.34)
19700826-19820712	19820713-19860625	-0.61%	52.29	(2.35)	(2.38)	(2.46)
19740814-19860625	19860626-19900611	-0.60%	59.76	(2.37)*	(2.34)*	(2.40)*
19780801-19900611	19900612-19940525	-0.57%	59.08	(2.44)*	(2.38)*	(2.54)*
19820719-19940525	19940526-19980512	-0.50%	59.73	(2.29)*	(2.34)*	(2.79)*
19860702-19980512	19980513-20020507	-0.55%	51.70	(2.03)	(2.02)	(2.12)
19900618-20020507	20020508-20031231	-0.60%	59.86	(2.86)*	(2.86)*	(3.60)*

Table IV. Large absolute price changes predictability: out-of-sample results

Large absolute movements are comparable to option volatility trading strategies, such as strips and straps. For each train-and-test period, returns lower than the value shown in Table III (third column) or higher than the value reported in Table II (third column) were codified with 1's and the rest with 0's. We test if the Gradient Boosting Machine is able to discriminate the cases codified with 1's from the instances encoded with 0's using past returns as explanatory variables. Methodology for evaluating the out-of-sample discriminatory accuracy of Gradient Boosting Machine as in Table I.

			Standard Error (%)		
Training period	Test sample	AUC (%)	DeLong et al.	Bootstrap	Moving Block
19620807-19740808	19740809-19780725	59.77	(1.78)*	(1.77)*	(2.12)*
19660727-19780725	19780726-19820712	57.04	(1.81)*	(1.82)*	(1.98)*
19700826-19820712	19820713-19860625	59.98	(1.84)*	(1.93)*	(2.30)*
19740814-19860625	19860626-19900611	57.44	(1.88)*	(1.88)*	(2.57)*
19780801-19900611	19900612-19940525	63.05	(1.92)*	(2.00)*	(2.58)*
19820719-19940525	19940526-19980512	64.07	(1.80)*	(1.74)*	(2.05)*
19860702-19980512	19980513-20020507	58.17	(1.80)*	(1.77)*	(2.12)*
19900618-20020507	20020508-20031231	69.14	(2.62)*	(2.68)*	(3.25)*

Initialize $\hat{F}_0(\mathbf{x}) = \arg\min \sum_{i=1}^{N} L(y_i, \rho)$.

For $m = 1,\ldots, M$ do:

1. Given a training sample, $\{y_i, \mathbf{x}_i\}_1^N$, perform a random permutation of the integers $\{1,...,N\}$ and obtain a sub-sample ($\hat{N} < N$)

$$\{\pi(i)\}_1^N = \text{rand\_perm}\{\text{i}\}_1^N \rightarrow \{y_{\pi(i)}, x_{\pi(i)}\}_1^{\hat{N}} \tag{4}$$

2. Compute the negative gradient as the working response/output

$$g_{\pi(i)m} = -\left.\frac{\partial L(y_{\pi(i)}, F(\mathbf{x}_{\pi(i)})}{\partial F(\mathbf{x}_{\pi(i)})}\right|_{F(\mathbf{x}_{\pi(i)})=\hat{F}_{m-1}(\mathbf{x}_{\pi(i)})} \tag{5}$$

3. Fit a regression tree, $T(\mathbf{x}_{\pi(i)}; \Theta_m)$, predicting $g_{\pi(i)m}$ from the explanatory variables $\mathbf{x}$.

4. Choose a gradient descent step size as

$$\rho = \arg\min_{\rho} \sum_{i=1}^{\hat{N}} L(y_{\pi(i)}, \hat{F}_{m-1}(\mathbf{x}_{\pi(i)}) + \rho T(\mathbf{x}_{\pi(i)}; \Theta_m)) \tag{6}$$

5. Given a learning rate *v*, update the estimate of $F_m(\mathbf{x})$, as

$$\hat{F}_m(\mathbf{x}) \leftarrow \hat{F}_{m-1}(\mathbf{x}) + v\rho T(\mathbf{x}_{\pi(i)}; \Theta_m) \tag{7}$$

End For

Figure 1. Friedman's (Stochastic) Gradient Boosting Machine Algorithm

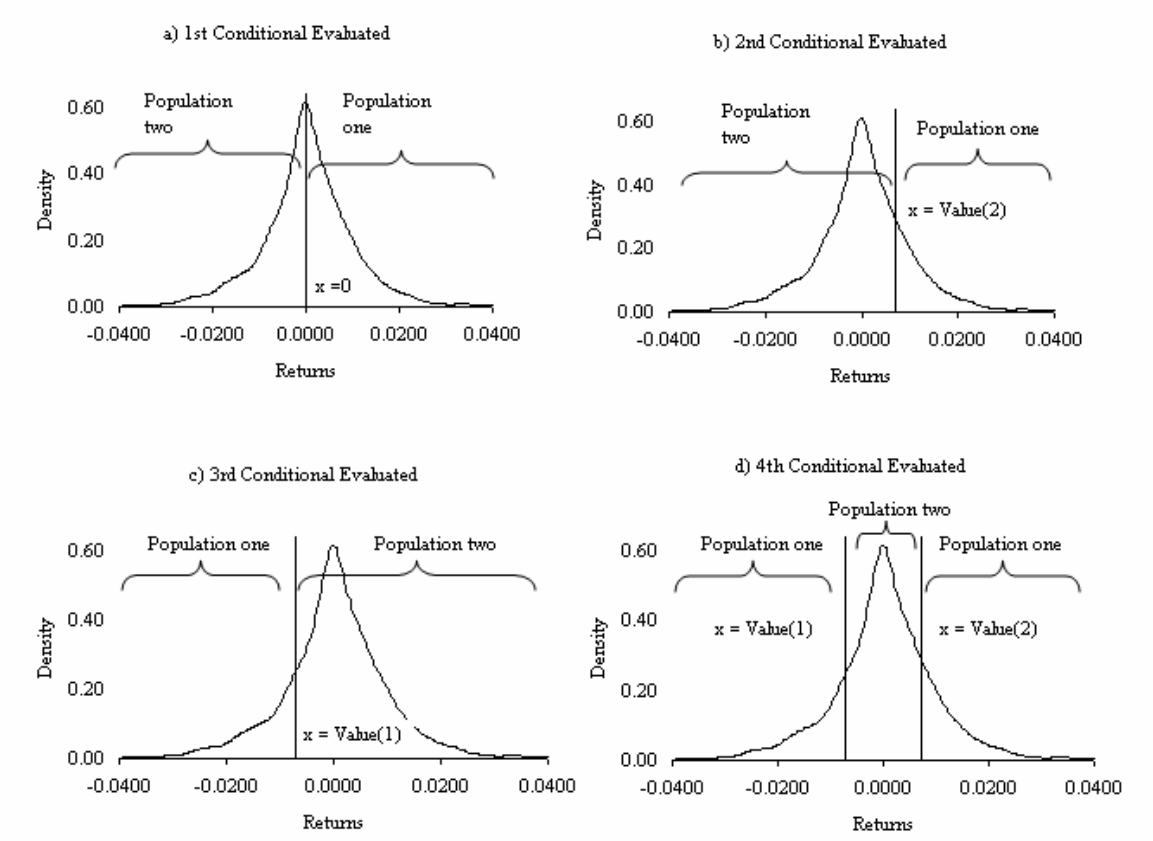

Figure 2. Visualization of the movement codification per conditional evaluated.

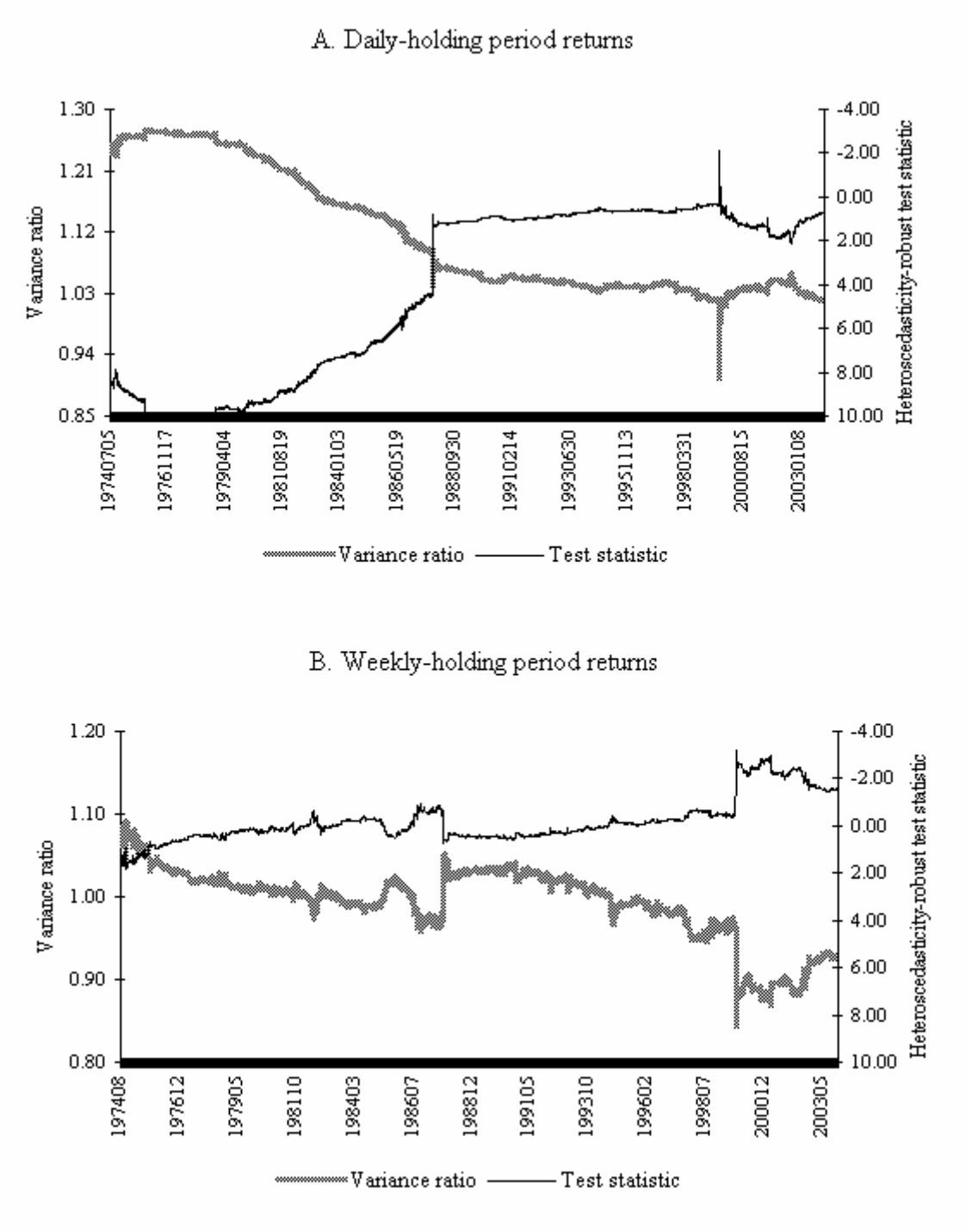

Figure 3. CRSP index's time-varying stochastic behavior