Nonparametric retrospection and monitoring of predictability of financial returns

by

Stanislav Anatolyev
New Economic School
Nakhimovsky Pr., 47, Moscow, 117418 Russia
E-mail: sanatoly@nes.ru

Abstract

We develop and evaluate sequential testing tools for a class of nonparametric tests for predictability of financial returns that includes, in particular, the directional accuracy and excess profitability tests. Our sequential methods consider in a unified framework both retrospection of a historical sample and monitoring newly arriving data. To this end, we focus on linear monitoring boundaries that are continuations of horizontal lines corresponding to retrospective critical values, elaborating on both two-sided and one-sided testing. We run a simulation study and illustrate the methodology by testing for directional and mean predictability of returns in young stock markets in Eastern Europe.

Key words: Predictability testing, sequential tests, retrospection, monitoring, stock indexes.

1 Introduction

Economists have been getting more and more concerned with possible structural instabilities in economic relationships which may invalidate conclusions obtained using conventional econometric tools. More than a decade ago, econometricians revived old CUSUM-type fluctuation tests allowing one to track structural shifts in model parameters in time (e.g., Ploberger, Krämer and Kontrus, 1989). More recently, there has been a new burst of interest to developing tools of sequential testing for practitioners who make decisions in real time. This was started by Chu, Stinchcombe, and White (1996) and Chu, Hornik, and Kuan (1995), and continued in Leisch, Hornik, and Kuan (2000), Altissimo and Corradi (2003), Zeileis, Leisch, Kleiber, and Hornik (2005), Inoue and Rossi (2005), and Andreou and Ghysels (2006), among others. This resulted in a number of sequential tests designed for both static and dynamic models, for both conditional means and conditional variances. Most of this work is targeted towards parametric models.

In this paper, we develop and evaluate sequential testing tools for a certain class of nonparametric tests for predictability of financial returns. This class is quite large and allows testing for hypotheses of non-predictability of various features of a series of interest. Two representatives of this class are the directional accuracy test of Pesaran and Timmermann (1992) and the excess profitability test of Anatolyev and Gerko (2005), but also there are others. The importance of testing for stability of predictability is discussed in Pesaran and Timmermann (2004) who show that ignoring structural instability may have serious consequences for the quality of directional forecasting. Testing for predictability in a sequential manner allows one to see the evolution of predictability over time, while the nonparametric nature of the test statistic allows one to use model-free inference.

We consider both retrospective tests where a researcher wants to track predictability over time in a historical sample, and monitoring tests where a researcher conducts testing as new observations arrive. The literature does not usually consider these tasks together; considering both in a unified manner is the first novelty introduced in this paper. Underlying it is a scenario that a researcher after having carried out a retrospective test goes on to the monitoring stage. Moreover, the retrospective boundaries (horizontal lines corresponding to retrospective critical values) continuously translate into the monotonically growing monitoring boundaries. The continuity of the boundaries is an appealing property as the first

observation of the monitoring period should not affect dramatically the inference about the null. Our second novelty is that we develop both two-sided and one-sided testing, with the emphasis put on the latter as more appropriate in the context of testing for predictability (cf. Inoue and Rossi, 2005). We focus on the use of the supremum functional over the sequential statistic, which is most widely used functional in the rest of the literature.

In the monitoring context, a widely discussed issue is the shape of monitoring boundaries. Some authors follow Chu, Stinchcombe, and White (1996) who suggest complicated "parabolic" boundaries which however lead to an analytical form of critical values. Zeileis, Leisch, Kleiber, and Hornik (2005) proposed more intuitively appealing linear boundaries which tend to distribute the size throughout the monitoring period more evenly. In Monte–Carlo exercises reported in Zeileis, Leisch, Kleiber, and Hornik (2005) and Andreou and Ghysels (2006) the linear boundaries performed well. We concentrate on such linear boundaries for two reasons: first, the possibility of considering retrospection and monitoring in a unified framework, and second, because linear boundaries lead to analytical critical values. It is worth noting that the tools developed here may be applied to other contexts, parametric or non-parametric, where testing on the basis of t-statistics is performed.

Note that in most of the work on sequential stability testing, the emphasis is usually put on testing for stability rather than testing whether a particular hypothesis holds throughout the sample. In this sense, the closest to the present work is the paper by Inoue and Rossi (2005) who sequentially track deviations of parameter combinations from hypothesized values. The main consequence is that the asymptotic analog of sequential statistic paths is the Wiener process (and functions thereof) rather than the Brownian Bridge. Inoue and Rossi (2005), however, do not consider one-sided testing, adapt "parabolic" boundaries, and their framework is, as mentioned above, parametric, albeit nonlinear.

We run a number of simulation experiments to verify the size and power properties of the tests, both in terms of rejection rates and delay lags. Simulations show good size properties but indicate that sometimes the power of sequential tests may be low when the time span in which the tests operate is small, or when predictability is concentrated on short time periods far from the beginning of the historical interval. We also illustrate our methodology by testing for directional and mean predictability of returns in ten young stock markets in Eastern Europe. Such markets are an ideal polygon for applying predictability tests as it is documented using other econometric tools that the pattern of predictability there is

changing (e.g., Rockinger and Urga, 2000). It turns out that the markets differ from each other a lot by whether tests signal predictability or not, by whether one or both tests indicate predictability, and by timing of boundary crossings if any.

The paper is organized as follows. In Section 2 we review the class of one-shot tests for predictability and its special cases. Sequential tests are developed in Section 3. In Section 4, simulation evidence is discussed, while the empirical application is presented in Section 5. Proofs and other auxiliary material are collected in Appendixes. Throughout, $\lfloor a \rfloor$ denotes taking an integer part of a, and $\Rightarrow$ denotes weak uniform convergence in the space of cadlag functions.

2 One-shot predictability tests

Let y_t represent some economic variable, and denote by $\mathcal{I}_{t-1}$ the information set $\{y_{t-1}, y_{t-2}, \cdots\}$ of past values of y_t. We are interested in testing the null hypothesis

$$H_0^g : E\left[g(y_t)|\mathcal{I}_{t-1}\right] = \text{const},$$

where $g(u)$ is a given function that depends on which feature is tested for predictability. Let x_t be a forecast of y_t that depends only on the data from $\mathcal{I}_{t-1}$, and T denote the sample size. The predictability test is based on the contrast

$$A^{g,h} - B^{g,h} \equiv \frac{1}{T}\sum_t h(x_t)g(y_t) - \left(\frac{1}{T}\sum_t h(x_t)\right)\left(\frac{1}{T}\sum_t g(y_t)\right), \tag{1}$$

where $h(u)$ is an arbitrary measurable function. The motivation for basing the test on the contrast (1) is that, under H_0^g, the population analog of the constant is zero:

$$\begin{aligned} E\left[h(x_t)g(y_t)\right] - E\left[h(x_t)\right]E\left[g(y_t)\right] &= E\left[h(x_t)E\left[g(y_t)|\mathcal{I}_{t-1}\right]\right] - E\left[h(x_t)\right]E\left[g(y_t)\right] \\ &= E\left[h(x_t)c\right] - E\left[h(x_t)\right]c \\ &= 0, \end{aligned}$$

where c is the const in the null.

The function $h(u)$ is chosen by the researcher. A popular choice is $h(u) = \text{sign}(u)$, where $\text{sign}(\cdot)$ takes value -1 when its argument is negative, and value $+1$ when its argument is non-negative. In this case setting $g(u) = \text{sign}(u)$ leads to the directional accuracy (DA) test for conditional sign independence of Pesaran and Timmermann (1992), or an asymptotically

equivalent parametric test in Breen, Glosten and Jagannathan (1989). The DA test is routinely used as a predictive-failure test in constructing forecasting models, or for evaluating the quality of predictors; see, for example, Pesaran and Timmermann (1995), Franses and van Dijk (2000), and Qi and Wu (2003). Setting $h(u) = \text{sign}(u)$ and $g(u) = u$ leads to the excess profitability (EP) test for conditional mean independence of Anatolyev and Gerko (2005), or an asymptotically equivalent parametric test in Cumby and Modest (1987). When y_t is a logarithmic return on some financial asset or index, the EP statistic can be interpreted as a normalized return of the position implied by a simple trading strategy that issues a buy signal if a forecast of next period return is positive and a sell signal otherwise, over a certain benchmark (see Anatolyev and Gerko, 2005 for details). These two examples of special interest will be intensively tackled throughout, although we develop testing algorithms for the general framework.

While the choice of function $g(u)$ is driven by what feature is tested for predictability, the function $h(u)$ is pretty arbitrary. Of course, the choice of $h(u)$ affects the power of the test, and this can be taken into account in practice if the researcher has a priori beliefs about possible deviations from non-predictability. For example, setting $h(u) = u$ in the context of testing for mean predictability may well be reasonable, and in fact leads to testing for the "big hit" forecast ability from Hartzmark (1991). A similar arbitrariness applies to the choice of the predictor x_t, which may be simply set to y_{t-1}, but alternatively may be constructed as fitted values from a parametric or non-parametric autoregression.

Let us impose

Assumption 1 *Under* H_0^g,
(i) the series y_t *and its forecast* x_t *are strictly stationary and strongly mixing with mixing coefficients* $\alpha(j)$ *satisfying* $\sum_{j=1}^{\infty} \alpha(j)^{1-1/\nu} < \infty$ *for some* $\nu > 1$*;*
(ii) the functions $g(u)$ *and* $h(u)$ *are measurable, and* $E\left[|g(y_t)|^{2\nu q}\right]$ *and* $E\left[|h(x_t)|^{2\nu p}\right]$ *exist and are finite for* ν *from (i), and for some* q *and* p *such that* $q^{-1} + p^{-1} = 1$.

Absolute regularity, which is a stronger notion than strong mixing posited in assumption 1(i), is shown to hold for various GARCH and stochastic volatility models often fit to financial returns (Carrasco and Chen, 2002). The moment condition in assumption 1(ii) is sufficient, but not necessary. With a choice of bounded $h(u)$, as is the case for the DA and EP tests, it is possible to set $p = \infty$ and $q = 1$, so that the moment condition on $g(y_t)$ is quite mild.

Let us introduce the following notation for future use:

$$\begin{aligned} M_g &= E\left[g(y_t)\right], \quad V_g = \text{var}\left[g(y_t)\right], \\ M_h &= E\left[h(x_t)\right], \quad V_h = \text{var}\left[h(x_t)\right], \end{aligned}$$

and

$$m_y = E\left[\text{sign}(y_t)\right], \quad m_x = E\left[\text{sign}(x_t)\right], \quad V_y = \text{var}\left[y_t\right].$$

We will base our tests on the following result which we will generalize to the context of sequential testing in the next Section.

Lemma 1 *Suppose $g(u)$ and $h(u)$ satisfy the regularity conditions specified in Assumption 1. Consider the contrast (1). Under $H_0^g : E\left[g(y_t)|\mathcal{I}_{t-1}\right] = \text{const}$,*

$$\sqrt{T}\left(A^{g,h} - B^{g,h}\right) \xrightarrow{d} \text{N}(0, V^{g,h})$$

as $T \to \infty$, where

$$V^{g,h} = V_h V_g + C_1 - 2M_h C_2,$$

where $C_1 = \text{cov}\left[h(x_t)^2, g(y_t)^2\right]$ and $C_2 = \text{cov}\left[h(x_t), g(y_t)^2\right]$.

Specialization of Lemma 1 to the two special cases of DA and EP tests yields

Corollary 1

(i) Under the null of conditional sign independence, i.e. $H_0^{DA} : E\left[\text{sign}(y_t)|\mathcal{I}_{t-1}\right] = \text{const}$,

$$\sqrt{T}\left(A^{DA} - B^{DA}\right) \xrightarrow{d} \text{N}(0, V^{DA})$$

as $T \to \infty$, where

$$V^{DA} = \left(1 - m_x^2\right)\left(1 - m_y^2\right).$$

(ii) Under the null of conditional mean independence, i.e. $H_0^{EP} : E\left[y_t|\mathcal{I}_{t-1}\right] = \text{const}$,

$$\sqrt{T}\left(A^{EP} - B^{EP}\right) \xrightarrow{d} \text{N}(0, V^{EP})$$

as $T \to \infty$, where

$$V^{EP} = \left(1 - m_x^2\right) V_y - 2m_x \,\text{cov}\left[\text{sign}(x_t), y_t^2\right].$$

To construct the test statistic, the contrast (1) may be pivotized using

$$\hat{V}^{g,h} = \hat{V}_h \hat{V}_g + \hat{C}_1 - 2\hat{M}_h \hat{C}_2,$$

where $\hat{M}_h$, $\hat{V}_h$, $\hat{V}_g$, $\hat{C}_1$ and $\hat{C}_2$ are empirical analogs of corresponding population quantities. For example, for the DA and EP tests,

$$\hat{V}^{DA} = \left(1 - \hat{m}_x^2\right)\left(1 - \hat{m}_y^2\right),$$

$$\hat{V}^{EP} = \left(1 - \hat{m}_x^2\right)\hat{V}_y - 2\hat{m}_x \hat{C},$$

where

$$\begin{aligned} m_y &= \frac{1}{T}\sum_t \text{sign}(y_t), \quad m_x = \frac{1}{T}\sum_t \text{sign}(x_t), \\ \hat{V}_y &= \frac{1}{T}\sum_t y_t^2 - \left(\frac{1}{T}\sum_t y_t\right)^2, \\ \hat{C} &= \frac{1}{T}\sum_t \left(\text{sign}(x_t) - \hat{m}_x\right) y_t^2. \end{aligned}$$

The analogs of the DA and EP tests in the case $h(u) = u$ are derived in Appendix C.

3 Sequential tests

3.1 Sequential testing and boundaries

In the sequential context, the null hypothesis of interest is the conditional independence of $g(y_t)$ throughout the entire period, i.e. that

$$H_0^g : E\left[g(y_t)|\mathcal{I}_{t-1}\right] = \text{const for all } t. \tag{2}$$

Note that we do not require that the const in (2) be the same across time; all we want to test is that $g(y_t)$ cannot be predicted by information at $t-1$ at all times. Thus, the emerging tests may not be able to detect deviations of the risk premium from a constant value.

Let us continue denoting the size of the historical sample by T. Then, if we do retrospective testing of H_0^g on the historical sample, t in (2) runs from 1 to T. If we monitor H_0^g further, t in (2) runs from $T+1$ to infinity. We manage to set the boundaries so that horizontal lines corresponding to retrospective critical values continuously translate into linear monitoring boundaries going upward (see Fig.1).

The underlying scenario is the following: a researcher has a historical sample in hands and carries out a retrospective test; then he/she goes on to the monitoring stage as new observations begin to arrive. The continuity of the boundaries makes sense as first several observations in the monitoring period should not affect dramatically the inference about the null. With linear boundaries, this continuity is possible to impose provided that the test sizes are equal in the retrospective and monitoring stages. Technically, this happens due to the property

$$\Pr\left\{\sup_{r\geq 1}\left(w\left(r\right)-\lambda r\right)\geq 0\right\}=\Pr\left\{\sup_{r\geq 1}\frac{w\left(r\right)}{r}\geq\lambda\right\}=\Pr\left\{\sup_{0<r\leq 1}w\left(r\right)\geq\lambda\right\},$$

and to a similar property for $\left|w\left(r\right)\right|$, where $\lambda>0$ is a constant, and $w\left(r\right)$ is a univariate standard Wiener process on $[0,+\infty)$, a limiting process for the sequential test statistic to be developed below.

3.2 Asymptotics for partial contrasts

For a generic series a_t, $t=1,2,\cdots,T,T+1,\cdots$, let us introduce the notation for a sequence of partial averages

$$\bar{a}_\tau=\frac{1}{\lfloor T\tau\rfloor}\sum_{t=1}^{\lfloor T\tau\rfloor}a_t,$$

where $\tau\geq 0$. When a_t is a product of several series, $a_t=b_tc_t$, say, then we write $\bar{a}_\tau$ also as $\overline{bc}_\tau$.

Kuan and Chen (1994) discovered that fluctuation tests are better sized in finite samples when variance estimators use the data from the same window over which the contrast is computed (rather than all available data), and we follow this strategy throughout. To this end, let us denote by $\hat{V}_\tau^{g,h}$ the value of $\hat{V}^{g,h}$ computed using the data from 1 to $\lfloor T\tau\rfloor$:

$$\hat{V}_\tau^{g,h}=\left(\overline{h^2}_\tau-\bar{h}_\tau^2\right)\left(\overline{g^2}_\tau-\bar{g}_\tau^2\right)+\overline{h^2g^2}_\tau-\overline{h^2}_\tau\overline{g^2}_\tau-2\bar{h}_\tau\left(\overline{hg^2}_\tau-\bar{h}_\tau\overline{g^2}_\tau\right).$$

In particular,

$$\hat{V}_\tau^{DA}=\left(1-\overline{\mathrm{sign}(x)}_\tau^2\right)\left(1-\overline{\mathrm{sign}(y)}_\tau^2\right)$$

and

$$\hat{V}_\tau^{EP}=\left(1-\overline{\mathrm{sign}(x)}_\tau^2\right)\left(\overline{y^2}_\tau-\bar{y}_\tau^2\right)-2\overline{\mathrm{sign}(x)}_\tau\left(\overline{\mathrm{sign}(x)y^2}_\tau-\overline{\mathrm{sign}(x)}_\tau\overline{y^2}_\tau\right).$$

The sequence of partial contrasts corresponding to the one-shot test based on $g(u)$ and $h(u)$, is

$$P_{t/T} = \frac{t}{\sqrt{T\hat{V}^{g,h}_{t/T}}} \left(\overline{gh}_{t/T} - \bar{g}_{t/T}\bar{h}_{t/T}\right). \tag{3}$$

The following theorem describes the asymptotic distribution of the sequence of partial contrasts which will serve as a basis for constructing the sequential tests. Recall that the conventional time t is related to τ by $t = \lfloor T\tau \rfloor$.

Theorem 1 *Suppose the null hypothesis*

$$H_0^g : E\left[g(y_t)|\mathcal{I}_{t-1}\right] = \text{const } \textit{for all } t$$

holds, and $h(u)$ and $g(u)$ satisfy the regularity conditions specified in Assumption 1. Then we have that as $T \to \infty$,

$$P_\tau \Rightarrow w(\tau),$$

where $w(r)$ is a univariate standard Wiener process on $[0, +\infty)$.

Thus, in large samples, deviations of partial contrasts $P_{t/T}$ from zero may be classified as statistically significant evidence of predictability if the associated path is unusual for the standard Wiener process.

Note that the methods proposed in this paper can be applied to other contexts, parametric or non-parametric. One may want to track the importance of a particular variable of interest in a time series parametric model (for example, unemployment or output gap in the Phillips curve, or some instrument in the monetary policy rule). One may alternatively want to track the significance of some nonparametric measure of time series data (for example, an autoregressive coefficient, a skewness coefficient, or a BDS statistic). In a more complex setting, one may want to track which of two competing models better fits the data as time progresses (for example, using the weighted likelihood ratio statistic of Amisano and Giacomini, 2007). In order to perform any of these exercises, one has to establish a functional central limit theorem similar to Theorem 1 for the sequential version of the t-statistic (i.e. the sequence of suitably scaled partial estimates) for the parameter or measure of interest, and then to apply the sequential tools developed in the present paper.

3.3 Retrospective tests

By computing the EP statistic from data in a moving window, Anatolyev and Gerko (2005) track the historical evolution of mean predictability in the American stock market. However, the comparison of the statistic with conventional critical values is an invalid testing procedure because the overall size of the sequential test has little to do with the intended nominal size (see Inoue and Rossi, 2005, for illustrations of this point). In this subsection we develop a formal sequential procedure that correctly controls the overall size of the test using historical data.

Using the supremum functional, we obtain the asymptotic size α one-sided test

$$\text{Reject if } \max_{t=2,\ldots,T} P_{t/T} \geq q_{\alpha}^{(1)},$$

and the asymptotic size α two-sided test

$$\text{Reject if } \max_{t=2,\ldots,T} \left|P_{t/T}\right| \geq q_{\alpha}^{(2)},$$

where $q_{\alpha}^{(j)}$ is a critical value for the j-sided test with significance level α.

It is widely known (e.g., Karatzas and Shreve (1988, problem 8.2)) that for $\lambda > 0$,

$$\Pr\left\{\sup_{0\leq r\leq 1} w\left(r\right) \geq \lambda\right\} = 2\left(1 - \Phi\left(\lambda\right)\right),$$

where $\Phi\left(\circ\right)$ is the CDF of the standard normal distribution. Hence, $q_{\alpha}^{(1)}$ can be easily found as a solution to the equation

$$\Phi\left(q_{\alpha}^{(1)}\right) = 1 - \frac{\alpha}{2},$$

so the α-level critical values for the one-sided test are equal to conventionally used α-level critical values for two-sided one-shot tests. Next, from Erdös and Kac (1946),

$$\Psi(\lambda) \equiv \Pr\left\{\sup_{0\leq r\leq 1} \left|w\left(r\right)\right| \leq \lambda\right\} = \frac{4}{\pi}\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left(-\frac{\pi^2\left(2k+1\right)^2}{8\lambda^2}\right).$$

Hence, $q_{\alpha}^{(2)}$ can be easily found as a solution to the equation

$$\Psi\left(q_{\alpha}^{(2)}\right) = 1 - \alpha.$$

In Table 1 we list the critical values for popular levels of significance. Note that for small α, as typically is the case, $q_{\alpha}^{(1)} \approx q_{2\alpha}^{(2)}$. This reflects a low probability of the standard Wiener process' hitting both $-\lambda$ and $+\lambda$ when λ is large enough (i.e. when α is small enough).

3.4 Monitoring tests

In this subsection we proceed to testing using newly arriving data. Using the supremum functional, we obtain the asymptotic size α one-sided test

$$\text{Reject if } \max_{t=T+1,T+2,\ldots} \left(P_{t/T} - b_\alpha^{(1)}(t)\right) \geq 0,$$

and the asymptotic size α two-sided test

$$\text{Reject if } \max_{t=T+1,T+2,\ldots} \left(\left|P_{t/T}\right| - b_\alpha^{(2)}(t)\right) \geq 0,$$

where $b_\alpha^{(1)}(t)$ and $b_\alpha^{(2)}(t)$ are upper boundaries.

We base our recursive monitoring one-sided tests on the boundaries of the type

$$b_\alpha^{(j)}(t) = \lambda_\alpha^{(j)} \frac{t}{T},$$

which tend to distribute the size throughout the monitoring period more evenly when the underlying process has growing variance that the so called "parabolic" boundaries (Zeileis, Leisch, Kleiber, and Hornik, 2005), see Appendix B. From the results of Robbins and Siegmund (1970, example 1) we obtain that for $\lambda > 0$,

$$\lim_{T\to\infty} \Pr\left\{\max_{t=T+1,T+2,\ldots} \left(P_{t/T} - \lambda\frac{t}{T}\right) \geq 0\right\} = 2\left(1 - \Phi(\lambda)\right).$$

Hence, $\lambda_\alpha^{(1)}$ can be found as a solution of the equation

$$\Phi\left(\lambda_\alpha^{(1)}\right) = 1 - \frac{\alpha}{2}.$$

For the two-sided test,

$$b_\alpha^{(2)}(t) = \lambda_\alpha^{(2)} \frac{t}{T},$$

and $\lambda_\alpha^{(2)}$ solves

$$\Psi\left(\lambda_\alpha^{(2)}\right) = 1 - \alpha.$$

Note that the same equations are used by the retrospective tests in the previous subsection. Hence, we can consult Table 1 to get values of $\lambda_\alpha^{(j)}$. The property $b_\alpha^{(j)}(T) = q_\alpha^{(j)}$, $j = 1, 2$, provides the continuity of the boundaries.

4 Simulation evidence

In this Section, we use Monte–Carlo simulations to check on actual sizes of the developed tests in finite samples and to study their power properties. Throughout we set the predictor x_t to the total return from two previous periods, i.e. $x_t = y_{t-2} + y_{t-1}$. This is an easy way to construct a predictor, and it is always available. In what follows, we report actual rejection frequencies for the sequential DA and EP tests corresponding to the nominal size of 5%. The simulation results are collected in Tables 2a and 2b for the retrospective tests, and in Tables 3a, 3b and 3c for the monitoring tests. All experiments are based on $10,000$ draws of time series of y_t according to the data generating processes (DGPs) described below; in each experiment we read off whether the sequential statistic crosses the boundary or not, and in which period the crossing occurred if this is informative. We consider a few values of T that approximately match sample sizes in the empirical application in Section 5, and one relatively big T. Namely, when we study the size, T is set to 50, 100, 150, or 500, and when we investigate the power, T is either 100 or 500. The parameters in DGPs A, B below that are used to verify actual sizes are calibrated using the Hungarian stock index returns, while all other DGPs are created artificially using as benchmarks the parameters calibrated to the Ukrainian stock index returns. This choice is motivated by the fact that no test has detected predictability in the Hungarian market, while almost all tests agree that there is high level of predictability in the Ukrainian data.

The first two DGPs are

$$\begin{aligned} \text{DGP A} \quad y_t &= 0.00289 + \varepsilon_t,\ \varepsilon_t \sim iid\ \mathrm{N}(0, 0.00166), && (4)\\ \text{DGP B} \quad y_t &= 0.00481 + \varepsilon_t,\ \varepsilon_t = \sigma_t \eta_t,\ \eta_t \sim iid\ \mathrm{N}(0,1), && (5)\\ \sigma_t^2 &= 8.70 \cdot 10^{-5} + 0.129\, \varepsilon_{t-1}^2 + 0.820\, \sigma_{t-1}^2. \end{aligned}$$

DGP A possesses no sign or mean predictability, while DGP B contains only sign predictability: Christoffersen and Diebold (2006) recently showed that conditional heteroskedasticity alone can induce sign predictability. This means that rejection rates in simulations will describe the actual size in the cases A/DA, A/EP and B/EP, and as a by-product we will have evidence in the B/DA case of how big directional predictability can be induced by (quite strong) conditional heteroskedasticity.

The following DGPs C, D and E in several variations calibrated to the Ukrainian stock

index data are used to investigate power of retrospective tests. While in DGP C there is the same non-zero amount of predictability throughout the sample, in DGPs D it is observed only during subperiods in the middle or towards the beginning or the end of the sample. Finally, in DGPs E there is a continuous transition from no predictability to higher and even higher predictability, or vice versa. Extra factors 3 and 2 attached to the autoregressive parameter serve to equalize the "amount" of predictability across the DGPs.

$$\text{DGP C} \quad y_t = 0.00300 + 0.192\, y_{t-1} + \varepsilon_t, \tag{6}$$

$$\text{DGP D}_k \quad y_t = 0.00300 + 3 \cdot 0.192\, \mathbb{I}_{\{t \in T_k\}}\, y_{t-1} + \varepsilon_t,\ k = 1, 2, 3, \tag{7}$$

$$\text{DGP E}_k \quad y_t = 0.00300 + 2 \cdot 0.192\, \frac{t\mathbb{I}_{\{k=1\}} + (T-t)\, \mathbb{I}_{\{k=2\}}}{T}\, y_{t-1} + \varepsilon_t,\ k = 1, 2, \tag{8}$$

where in all cases

$$\varepsilon_t = \sigma_t \eta_t,\ \eta_t \sim iid\ \mathrm{N}(0,1),\ \sigma_t^2 = 4.35 \cdot 10^{-5} + 0.092\, \varepsilon_{t-1}^2 + 0.894\, \sigma_{t-1}^2, \tag{9}$$

$\mathbb{I}_{\{\circ\}}$ is an indicator function, and T_k contains time periods from the k's third of the sample. That is, T_1 contains observations from the first third of the sample, T_2 – those from the second third of the sample, and T_3 – those from the last third of the sample. Hence, in DGPs D_1 through D_3 the predictability is observed during one of the three periods, and is not observed during the other two. In contrast, the predictability is continuously escalating as time passes in DGP E_1, but is vanishing as time passes in DGP E_2.

The following DGPs F and G are used to investigate power of monitoring tests. In DGP F there is temporary predictability lasting for T periods from the start of the monitoring interval. In DGPs G predictability is permanent: in DGP G_1 it is constant, while in DGP G_2 it is escalating.

$$\text{DGP F} \quad y_t = 0.00300 + 0.192\, \mathbb{I}_{\{T+1 \le t \le 2T\}}\, y_{t-1} + \varepsilon_t, \tag{10}$$

$$\text{DGP G}_k \quad y_t = 0.00300 + 0.192\, \frac{T\mathbb{I}_{\{k=1\}} + (t-T)\, \mathbb{I}_{\{k=2\}}}{T} \mathbb{I}_{\{t \ge T+1\}}\, y_{t-1} + \varepsilon_t,\ k = 1, 2, \tag{11}$$

and in all cases ε_t follows (9).

First let us look at panel A of Table 2a which contains actual sizes when data are serially independent. One can immediately see that the sequential tests are very well-sized, especially for larger samples. The DA tests tend to be a little undersized in very short samples, but distortions quickly diminish as T grows. Next consider panel B corresponding to the GARCH

process which is mean non-predictable. While the EP test is very well sized, the DA tests display practically the same sizes as for DGP A, even though under DGP B the series is, in contrast to DGP A, sign predictable. It follows that directional predictability induced by volatility clustering alone is very weak. Next, if we compare size distortions across alternatives, those for one-sided tests a bit exceed those for two-sided tests for both DA and EP.

Let us now turn to power figures in Table 2b which reports, along with rejection frequencies, boundary crossing dates averaged over those experiments where crossings did take place. Overall, the EP tests are more powerful than the DA analogs in detecting predictability in all DGPs, which is in line with the analytical results in Anatolyev and Gerko (2005, section 3). Naturally, power increases quickly with the sample size. Also, power figures are significantly higher for one-sided tests than for two-sided ones, and tend to detect predictability earlier. For DGPs D_1, D_2 and E_2 testing for mean predictability allows one to detect predictability much earlier than does testing for sign predictability. In DGP C where predictability is uniformly distributed, the tests detect predictability on average after $\frac{3}{4}$ of the historical period are over when $T = 100$, but this measure changes to about $\frac{2}{3}$ when $T = 500$. Of course, the detection dates shift significantly in DGPs D and E where the same "amount" of predictability is unevenly distributed. For DGP D_1, for example, the tests detect predictability on average after a half of the historical period is over when $T = 100$, but only after its first $\frac{1}{3}$ is over when $T = 500$; in the latter case detection is practically inevitable, especially by the EP test. When predictability is concentrated towards the end of a historical interval (D_2 and D_3 in contrast to D_1), the tests detect it relatively faster after the moment when predictability becomes in effect. On average, the power figures tend to get larger along the following sequences of DGPs: $E_1 \prec E_2$ and $D_3 \prec D_2 \prec D_1$, showing comparatively better performance against DGPs where predictability is concentrated towards the beginning of a historical sample.

Tables 3a, 3b and 3c provide analogous information for monitoring tests. The numbers in the tables are rejection frequencies during the periods $[T+1, \tau T]$ with τ equaling $\frac{3}{2}$, $\frac{5}{2}$, and 5, i.e. from the beginning of the monitoring period up to the point where a half, or two and a half, or four times the length of the historical interval pass. One can see from Table 3a that the monitoring tests are very well sized, with sizes at $\tau = 5$ strictly smaller than 5%; the total size is exhausted pretty rapidly. Other observations confirm conclusions

for retrospective tests: there is slight underrejection for small samples, the directional predictability induced by conditional heteroskedasticity alone is very weak (panel B/DA), size distortions for one-sided tests a little exceed those for two-sided tests. Table 3b presents rejection frequencies together with how much time passes before the monitoring procedure detects predictability provided that it does detect it. From panel F one can infer that a short (up to $\tau = 2$) period of temporary predictability is very rarely detected when $T = 100$, especially when the alternative is two-sided, and is moderately frequently detected when $T = 500$. On the other hand, when detection does take place, it does so quite quickly, at τ smaller or around $\frac{3}{2}$ when $T = 500$. When predictability is ever-lasting or even escalating (panel G), the power figures are naturally much higher. For DGP G_1, the power numbers when $T = 100$ are not great either, but when $T = 500$, the detection occurs pretty frequently (especially with the EP test), and it does so at τ around or a bit higher than $\frac{5}{2}$. For DGP G_2, tests detect predictability quite often even when $T = 100$, and with certainty when $T = 500$; the corresponding detection times are concentrated around $\tau = 4$ and τ from 3 to $\frac{7}{2}$, respectively. Overall, one-sided tests turn out to be more powerful, sometimes quite appreciably, than two-sided tests, and EP tests appear more powerful than DA analogs, just as in the case of retrospective testing. Finally, Table 3c contains information on test performance using "parabolic" boundaries; the two-sided version was previously considered in Chu, Stinchcombe, and White (1996) and Inoue and Rossi (2005), and the one-sided version is similarly derived from the results in Robbins and Siegmund (1970); see Appendix B. The upper linear boundary lies below the parabolic one for smaller values of τ and above it for larger values of τ; hence, early instances of predictability are easier to detect with a linear boundary. Panel A reflects the shape of "parabolic" boundaries: the size is "eaten up" more slowly than in the case of linear boundaries; even accounting for possible slight underrejection, there is more size remaining after $\tau = 5$. From panels F and G it is clear that "parabolic" boundaries allow one to detect more successfully predictability that goes on permanently, but temporary episodes of predictability are sensed less successfully. Comparison of corresponding rows in Tables 3b and 3c reveals that linear boundaries tend to detect predictability earlier than "parabolic" ones; the exception is DGP G_2 with escalating predictability for which the results on boundary crossing dates are comparable.

The experiments with sequential tests indicate that the power of such tests may be quite low when the time span in which the tests operate is small. Further, retrospective tests may

miss predictability concentrated towards the end of the historical interval; however, this predictability may be captured by the subsequent monitoring test. Monitoring tests in turn may have low power when predictability is limited to a relatively short time period provided that there is no predictability in the historical interval. Also, we would like to again draw attention to the attractive property of the sequential tests (see remark under (2)) of not being able to detect structural breaks not associated with predictability. For example, an introduction of a big change in the intercept term ("risk premium") of DGP A in the middle of the historical interval leads to the same rejection frequencies by the sequential EP test as without the break.

5 Application to returns from Eastern European stock markets

In this Section we apply the developed methodology to the analysis of predictability of weekly stock market indexes in ten former communist countries in Eastern Europe. The indexes are listed in Table 4. They start on January of the year 1997 (7 series), 1998 (2 series), or 1999 (1 series), and end on January 2005. The data are taken from Bloomberg. The literature has documented a significant amount of predictability in such markets at the end of 20th century when these markets were very young, but one could observe a movement towards non-predictability in most of them; see Zalewska-Mitura and Hall (1999) and Rockinger and Urga (2000, 2001), subsequently R&U. Note that in the literature, predictability is sometimes referred to as the absence of (weak form) "efficiency". However, in these markets there are a few market limitations that do not support such interpretation; see Pesaran and Timmermann (1995) and Timmermann and Granger (2004).

Our empirical strategy is the following. We consider a virtual researcher who at the beginning of the year 2000 is given the data available at that moment and who is interested in retrospectively testing the ten returns for predictability and in further monitoring their predictability in the new millennium. Hence, T is different for different series (it corresponds to one, two, or three years of historical data), while the monitoring interval is the same for all series; precise subsample sizes are listed in Table 4. In Tables 5a and 5b we document the dates when the sequential statistic paths hit the boundaries if they do. Our main strategy remains using the two-week return $y_{t-2} + y_{t-1}$ for the predictor x_t and the sign function $\text{sign}(u)$ for $h(u)$, but we also give complementary results from using other choices of x_t and

$h(u)$, as well as from using the "parabolic" type of monitoring boundaries. All boundaries correspond to one-sided testing, and all conclusions are drawn based on 5% test sizes.

Figure 2 presents graphs of evolution of the sequential DA and EP statistics in four selected markets: Russian, Polish, Czech, and Estonian. Superimposed are horizontal retrospective and linear monitoring boundaries. The columns "Linear" in Table 5a contain information on whether the sequential statistics hit the boundaries and when if they do, relative to the beginning of the monitoring interval (i.e. Jan 2000). One can see that only the Ukrainian and Estonian stock indices show a clear indication of strong predictability of both types. For Ukraine, both retrospective statistics sense that directional and mean predictability appeared at least half a year ago, in the middle of 1999. For Estonia, the EP-statistic felt mean predictability about a year and a half ago, while the DA-statistic did not detect sign predictability in the historical period but does detect it after 7 weeks of monitoring. For the Polish stock index, mean predictability was sensed a year ago, at the beginning of 1999, although no sign predictability is found neither in the historical nor in the monitoring period. Finally, the Czech and Slovenian stock indices display the presence of predictability of only one type in about 3 to 4 months of monitoring, mean predictability in the Czech market and sign predictability in the Slovenian market.

For comparison purposes, we also consider "parabolic" monitoring boundaries. Recall that unlike the approach of this paper monitoring schemes usually considered rely on the stability, or non-predictability in the present context, during the historical period. The columns "Parabolic" in Table 5a contain dates of crossings the parabolic boundaries. One can see that the alternative procedure detects the same types of predictability as our procedure in the Ukrainian, Polish and Estonian stock markets, in four out of five cases it senses early the predictability which presumably appeared in the historical period; in one case (DA/Estonia) it feels sign predictability only after two years while our procedure feels it after less than two months. For the Czech and Slovenian markets, the alternative procedure does not detect predictability during the whole monitoring period while our procedure does it quite early. Note that in no case the alternative procedure senses predictability not detected by our procedure, or senses it earlier.

In turn, Table 5b contains similar information in cases when one uses alternative predictors in place of x_t or an alternative popular choice of function $h(u)$. Namely, x_t is set to the

one-week return y_{t-1} ("shortest" predictor), or to the four-week (nearly one-month) return $\sum_{j=1}^{4} y_{t-j}$ ("long" predictor), or to the to the "OLS" predictor $z_t'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{4} \hat{\beta}_j y_{t-j}$, instead of the two-week return $y_{t-2} + y_{t-1}$ ("short" predictor); an alternative choice for $h(u)$ is natural $h(u) = u$ (see Appendix C). One can see that the DA and EP tests continue to detect predictability of both types in the Ukrainian stock market, albeit the EP test with the "long" predictor finds it only during the monitoring stage, and both tests detect it much later when the alternative $h(u)$ is used. Moreover, when $h(u) = u$, neither test senses deviations from the null for all other markets. Alternative predictors, however, sometimes make the tests more sensitive to the presence of predictability. For example, the DA test with the "shortest" predictor in the Estonian market finds retrospectively that sign predictability appeared in the same early week as the mean predictability, while it detected sign predictability only during the monitoring stage when based on the "short" predictor; the EP test based on either the "shortest" or "long" predictor detects retrospectively mean predictability in the stock market in Slovakia, while it fails to do so during both retrospective and monitoring stages when based on the "short" predictor. However, there are opposite examples, like the EP/Czech case.

It is interesting to compare our empirical evidence to studies that describe evolution of mean predictability by means of regression models with time-varying parameters. R&U (2000, 2001) consider Russian, Polish, Czech, and Hungarian stock markets, the former article using the data from early 1994 to mid-1999, the latter – from early 1994 to mid-1997. (The results in the two studies are somewhat contradictory which can be explained by different regression specifications reflecting different focus of investigation.) Zalewska-Mitura and Hall (1999) study, in particular, the Hungarian market from early 1990s till late 1997. For the Russian market, R&U (2000, 2001) find strong predictability shortly after 1994 for which one can blame institutional imperfections in early stages and low liquidity, then one observes convergence towards no predictability; there is a peak in predictability in late 1998, however. From the bottom panel of Figure 2a one can see that this peak did affect the EP sequential statistic, but the (1998 Russian financial crisis) period was short, and the statistic touched the 10% boundary at the end of 1999. After that, the market became much more liquid because of a rise in oil prices, so predictability did evaporate as is clear from the further path of the EP empirical process. Regarding the Polish market, R&U (2000)

state that by the end of 1994 predictability caused by panic sells and initial restrictions on foreign participation vanished; however, R&U (2001) document strong mean predictability in the Polish market up to the end of their sample (mid-1997). The latter evidence is more in agreement with the behavior of the sequential EP statistic which shows strong mean predictability from presumably the second half of 1998; however, a subsequent upward trend in its path may be entirely due to a temporary period of predictability rather than its being permanent. For the Czech market, R&U (2000, 2001) document temporary periods of mean predictability between June 1996 and March 1997 and between March 1998 and mid-1999, and no predictability at other times. The behavior of the sequential EP statistic clearly reflects the presence of both periods of predictability, exhibiting a hike at early 1997 and a series of upward trends from early 1998, and even touching the 10% level boundary in the first half of 1999. It is presumably the second predictability period that caused the crossing the 5% boundary during the monitoring stage. As long as the Hungarian stock market is concerned, Zalewska-Mitura and Hall (1999) find no movement towards non-predictability throughout their sample that end in late 1997. However, R&U (2000, 2001) claim that this market could not be predicted from at least 1994, the reasons being that this market was founded earlier than the others, and it had independent supervisory authority. It seems that the behavior of our sequential EP statistic is more consistent with the former view, as its upward trend during 1997 leads to crossing the 10% boundary and nearly touching the 5% boundary in the second half of 1998.

Of other countries, Ukraine and Estonia exhibit persistently high predictability in their stock markets. In Ukraine, this may be caused by initially poor legislation concerning financial markets, which subsequently lead to premature centralization and inefficient stock market functioning accompanied by heavy insider trading and low liquidity. Among other things, the market turned out to be unattractive for foreign investors (and closed to them until 2001), in contrast to, for example, the historically related Russian market. In Estonia, for a long time the stock market was spontaneous with no formal rules and no market infrastructure, and only in June 1996 the Tallinn stock exchange started functioning, after which the legislation strengthening the regulatory framework continued being polished up. The causes of high predictability apparently lie in the absence of stimuli for stock issuers to be listed at the exchange and of attractiveness for foreign investors, and as a consequence, low liquidity and thin trading.

Conclusion

We have developed tools for nonparametrically testing predictability of financial returns (or, for that matter, of any stationary variable) in a sequential context, where both retrospection of a historical sample and monitoring newly arriving data are conducted in a unified framework. The size and power of sequential tests for mean and directional predictability are analyzed. The technique is illustrated using stock return indexes from developing Eastern European markets.

Acknowledgments

The comments of the Editor Serena Ng, Associate Editor and two anonymous referees helped greatly improve the presentation. The author also thanks audiences of relevant sessions at the 2006 North American and European meetings of Econometric Society in Minneapolis and Vienna.

References

Altissimo, F. and Corradi, V. (2003), "Strong rules for detecting the number of breaks in a time series," *Journal of Econometrics*, 117, 207–244.

Amisano, G. and Giacomini, R. (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business & Economic Statistics*, 25, 177–190

Anatolyev, S. and Gerko, A. (2005), "A trading approach to testing for predictability," *Journal of Business & Economic Statistics*, 23, 455–461.

Andreou, E. and E. Ghysels (2006), "Monitoring disruptions in financial markets," *Journal of Econometrics*, 135, 77–124.

Breen, W., L.R. Glosten and R. Jagannathan (1989), "Economic significance of predictable variations in stock index returns," *Journal of Finance*, 44, 1177–1189.

Carrasco, M. and X. Chen (2002), "Mixing and moment properties of various GARCH and stochastic volatility models," *Econometric Theory*, 18, 17–39.

Christoffersen, P.F. and F.X. Diebold (2006), "Financial asset returns, direction-of-change forecasting, and volatility dynamics," *Management Science*, 52, 1273–1288.

Chu, C.S.J., K. Hornik, and C.M. Kuan (1995), "The moving-estimates test for parameter

stability," *Econometric Theory*, 11, 669–720.

Chu, C.S.J., M. Stinchcombe, and H. White (1996), "Monitoring structural change," *Econometrica*, 64, 1045–1065.

Cumby, R.E., and D.M. Modest (1987), "Testing for market timing ability: a framework for forecast evaluation," *Journal of Financial Economics*, 19, 169–89.

Erdös, P. and M. Kac (1946), "On certain limit theorems of the theory of probability," *Bulletin of American Mathematical Society*, 52, 292–302.

Franses, P. and D. van Dijk (2000), *Nonlinear Time Series Models in Empirical Finance*, Cambridge: Cambridge University Press.

Hartzmark, M.L. (1991), "Luck versus forecast ability: determinants of trader performance in futures markets," *Journal of Business*, 64, 49–74.

Inoue, A. and B. Rossi (2005), "Recursive predictability tests for real time data," *Journal of Business & Economic Statistics*, 23, 336–345.

Karatzas, I. and S.E. Shreve (1988), *Brownian motion and stochastic calculus*, New York: Springer-Verlag.

Kuan, C.M. and M.Y. Chen (1994), "Implementing the fluctuation and moving-estimates tests in dynamic econometric models," *Economics Letters*, 44, 235–239.

Leisch, F., K. Hornik, and C.M. Kuan (2000), "Monitoring structural changes with the generalized fluctuation test," *Econometric Theory*, 16, 835–854.

Pesaran, M.H. and A. Timmermann (1992), "A simple nonparametric test of predictive performance," *Journal of Business & Economic Statistics,* 10, 561–565.

Pesaran, M.H. and A. Timmermann (1995), "Predictability of stock returns: robustness and economic significance," *Journal of Finance*, 50, 1201–1228.

Pesaran, M.H. and A. Timmermann (2002), "Market timing and return prediction under model instability," *Journal of Empirical Finance*, 9, 495–510.

Pesaran, M.H. and A. Timmermann (2004), "How costly is it to ignore breaks when forecasting the direction of a time series?" *International Journal of Forecasting*, 20, 411–425.

Phillips, P.C.B. and S.N. Durlauf (1986), "Multiple time series regression with integrated processes," *Review of Economic Studies*, 53, 473–495.

Ploberger, W., W. Krämer, and K. Kontrus (1989), "A new test for structural stability in the linear regression model," *Journal of Econometrics*, 40, 307–318.

Qi, M. and Y. Wu (2003), "Nonlinear prediction of exchange rates with monetary fundamentals," *Journal of Empirical Finance*, 10, 623–640.

Robbins, H. and D. Siegmund (1970), "Boundary Crossing Probabilities for the Wiener Process and Sample Sums," *Annals of Mathematical Statistics*, 41, 1410–1429.

Rockinger, M. and G. Urga (2000), "The evolution of stock markets in transition economies," *Journal of Comparative Economics*, 28, 456–472.

Rockinger, M. and G. Urga (2001), "A time varying parameter model to test for predictability and integration in the stock markets of transition economies," *Journal of Business & Economic Statistics*, 19, 73–84.

Timmermann, A. and C. Granger (2004), "Efficient market hypothesis and forecasting," *International Journal of Forecasting*, 20, 15–27.

Zalewska-Mitura, A. and S.G. Hall (1999), "Examining the first stages of market performance: a test for evolving market efficiency," *Economics Letters*, 64, 1–12.

Zeileis A., F. Leisch, C. Kleiber, and K. Hornik (2005), "Monitoring structural change in dynamic econometric models," *Journal of Applied Econometrics*, 20, 99–121.

A Appendix: proofs

Proof. [of Lemma 1] Follows as a special case of Theorem 1 by setting $\tau_1 = 0$ and $\tau_2 = 1$. ■

Lemma 2 *Suppose $h(u)$ and $g(u)$ satisfy the regularity conditions specified in Assumption 1. Then under*

$$H_0^g : E\left[g(y_t)|\mathcal{I}_{t-1}\right] = \text{const},$$

as $T \to \infty$, we have

$$\frac{1}{\sqrt{T}}\mathbf{V}^{-1/2}\sum_{t=1}^{[\tau T]}\begin{pmatrix} h_t g_t - M_h M_g \\ g_t - M_g \\ h_t - M_h \end{pmatrix} \Rightarrow \mathbf{W}(\tau)$$

where $\mathbf{W}(\tau)$ *is a trivariate standard Brownian motion, and the elements of* $\mathbf{V}$ *are given by*

$$\begin{aligned}
V_{11} &= \operatorname{var}\left[h(x_t)g(y_t)\right] + 2M_g \sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), h(x_{t+j})\right], \\
V_{22} &= V_g, \\
V_{33} &= V_h + 2\sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t), h(x_{t+j})\right], \\
V_{12} &= \operatorname{cov}\left[h(x_t), g(y_t)^2\right] + M_h V_g + M_g \sum_{j=1}^{+\infty} \operatorname{cov}\left[g(y_t), h(x_{t+j})\right], \\
V_{13} &= M_g V_h + M_g \sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t), h(x_{t+j})\right] + \sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), h(x_{t+j})\right], \\
V_{23} &= \sum_{j=1}^{+\infty} \operatorname{cov}\left[g(y_t), h(x_{t+j})\right].
\end{aligned}$$

Proof. The conclusion follows directly from Phillips and Durlauf (1986, corollary 2.2), with the elements of the long-run covariance $\mathbf{V}$ given by

$$\begin{aligned}
V_{11} &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), h(x_{t+j})g(y_{t+j})\right] \\
&= \operatorname{var}\left[h(x_t)g(y_t)\right] + 2M_g \sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), h(x_{t+j})\right], \\
V_{22} &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left[g(y_t), g(y_{t+j})\right] = V_g, \\
V_{33} &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left[h(x_t), h(x_{t+j})\right] = V_h + 2\sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t), h(x_{t+j})\right], \\
V_{12} &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), g(y_{t+j})\right] \\
&= \operatorname{cov}\left[h(x_t), g(y_t)^2\right] + M_h V_g + M_g \sum_{j=1}^{+\infty} \operatorname{cov}\left[g(y_t), h(x_{t+j})\right], \\
V_{13} &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), h(x_{t+j})\right] \\
&= M_g V_h + M_g \sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t), h(x_{t+j})\right] + \sum_{j=1}^{+\infty} \operatorname{cov}\left[h(x_t)g(y_t), h(x_{t+j})\right], \\
V_{23} &= \sum_{j=-\infty}^{+\infty} \operatorname{cov}\left[g(y_t), h(x_{t+j})\right] = \sum_{j=1}^{+\infty} \operatorname{cov}\left[g(y_t), h(x_{t+j})\right],
\end{aligned}$$

where the law of iterated expectations and the statement of the null hypothesis are intensively used. ■

Proof. [of Theorem 1] Let us denote

$$\boldsymbol{\mu} = (1, -M_h, -M_g)'.$$

From Lemma 2, it follows that

$$\sqrt{T}\left(\overline{gh}_\tau - \bar{g}_\tau \bar{h}_\tau\right) \Rightarrow \frac{\boldsymbol{\mu}'\mathbf{V}^{1/2}\mathbf{W}(\tau)}{\tau}.$$

When pivotized,

$$P_\tau \Rightarrow \frac{\boldsymbol{\mu}'\mathbf{V}^{1/2}\mathbf{W}(\tau)}{\sqrt{\boldsymbol{\mu}'\mathbf{V}\boldsymbol{\mu}}} \overset{d}{=} w(\tau),$$

because $\hat{V}_\tau^{g,h} \xrightarrow{p} V^{g,h} = \boldsymbol{\mu}'\mathbf{V}\boldsymbol{\mu}$. ■

B Appendix: "parabolic" boundaries

For one-sided testing we deduce from Robbins and Siegmund (1970, example 2) that

$$\Pr\left\{w(r) \geq \sqrt{r}\delta^{-1}\left(\delta(\varrho) + \log r\right) \text{ for some } r \geq 1\right\} = 1 - \Phi(\varrho) + \varphi(\varrho)\left(\varrho + \frac{\varphi(\varrho)}{\Phi(\varrho)}\right),$$

where $\delta(u) = u^2 + 2\log\Phi(u)$. Using numerical methods, we obtain that the 5% level of significance of the monitoring test corresponds to $\varrho = 2.503$.

For two-sided testing we deduce from Robbins and Siegmund (1970, example 3) that

$$\Pr\left\{|w(r)| \geq \sqrt{r\left(\varrho^2 + \log r\right)} \text{ for some } r \geq 1\right\} = 2\left(1 - \Phi(\varrho) + \varphi(\varrho)\,\varrho\right).$$

From Inoue and Rossi (2005, Table 1), the 5% level of significance of the monitoring test corresponds to $\varrho = 2.796$.

C Appendix: testing with $h(u) = u$

Using Lemma 1, we set for the DA-analog

$$\hat{V}^{DA'} = \hat{V}_x\left(1 - \hat{m}_y^2\right)$$

with

$$\begin{aligned}
\hat{V}_x &= \frac{1}{T}\sum_t x_t^2 - \hat{M}_x^2, \\
\hat{M}_x &= \frac{1}{T}\sum_t x_t,
\end{aligned}$$

because $C_1 = \text{cov}\left[h(x_t)^2, g(y_t)^2\right] = 0$ and $C_2 = \text{cov}\left[h(x_t), g(y_t)^2\right] = 0$. For the EP-analog,

$$\hat{V}^{EP'} = \hat{V}_x \hat{V}_y + \hat{C}_1 - 2\hat{M}_x \hat{C}_2$$

with

$$\begin{aligned} \hat{C}_1 &= \frac{1}{T}\sum_t x_t^2 y_t^2 - \left(\frac{1}{T}\sum_t x_t^2\right)\left(\frac{1}{T}\sum_t y_t^2\right), \\ \hat{C}_2 &= \frac{1}{T}\sum_t \left(x_t - \hat{M}_x\right) y_t^2. \end{aligned}$$

Table 1. Critical values for the retrospective and monitoring tests.

Test type (j)	One-sided			Two-sided		
Test size (α)	10%	5%	1%	10%	5%	1%
Critical values	1.645	1.960	2.576	1.960	2.241	2.807

Notes: The table shows the parameters of boundaries: the height of horizontal lines $q_\alpha^{(j)}$ for retrospective tests and the slope of monotonically increasing lines $\lambda_\alpha^{(j)}$ for monitoring tests.

Table 2a. Size for retrospective tests with nominal size 5%.

		Directional accuracy		Excess profitability	
DGP	T	One-sided	Two-sided	One-sided	Two-sided
A	50	3.0	4.4	3.9	4.8
	100	3.8	4.6	4.1	4.6
	150	4.0	4.7	4.2	5.1
	500	4.7	5.0	4.4	4.7
B	50	3.4	4.4	4.6	5.0
	100	4.0	4.7	4.4	4.9
	150	4.2	4.6	4.4	4.8
	500	4.9	4.7	4.4	4.8

Notes: Figures in the table show actual rejection (boundary crossing) frequencies corresponding to test nominal size of 5%.

Table 2b. Power for retrospective tests with nominal size 5%.

		Directional accuracy		Excess profitability	
DGP	T	One-sided	Two-sided	One-sided	Two-sided
C	100	17.8 (75)	11.3 (79)	25.7 (73)	17.5 (76)
	500	59.9 (339)	47.7 (362)	76.1 (320)	66.6 (347)
D_1	100	34.0 (56)	22.5 (62)	52.8 (47)	40.1 (51)
	500	92.8 (143)	85.6 (165)	99.0 (113)	97.5 (130)
D_2	100	28.9 (70)	19.1 (73)	43.3 (67)	32.2 (70)
	500	85.1 (293)	76.9 (307)	94.7 (269)	91.2 (280)
D_3	100	19.8 (87)	13.7 (87)	32.2 (87)	24.1 (88)
	500	68.8 (434)	59.8 (441)	85.7 (422)	80.3 (430)
E_1	100	17.5 (81)	11.6 (83)	23.4 (80)	15.9 (83)
	500	57.9 (395)	46.4 (412)	74.4 (386)	65.4 (405)
E_2	100	21.4 (68)	13.6 (73)	30.2 (63)	20.9 (67)
	500	70.6 (263)	58.7 (291)	86.4 (228)	78.2 (259)

Notes: Unbracketed figures in the table show actual rejection (boundary crossing) frequencies corresponding to test nominal size of 5%. Figures in brackets show average observation numbers where boundary crossing takes place, conditional on crossings.

Table 3a. Size for monitoring tests with nominal size 5%.

			Directional accuracy		Excess profitability	
DGP	T	τ	One-sided	Two-sided	One-sided	Two-sided
A	50	$\frac{3}{2}$	2.9	3.8	2.7	3.5
		$\frac{5}{2}$	3.0	3.9	2.8	3.6
		5	3.0	3.9	2.8	3.6
	100	$\frac{3}{2}$	3.6	4.1	3.3	3.9
		$\frac{5}{2}$	3.8	4.3	3.5	4.1
		5	3.8	4.3	3.5	4.1
	150	$\frac{3}{2}$	3.5	4.0	3.3	4.0
		$\frac{5}{2}$	3.7	4.2	3.5	4.1
		5	3.8	4.2	3.5	4.1
	500	$\frac{3}{2}$	3.8	4.3	4.1	4.5
		$\frac{5}{2}$	4.0	4.4	4.3	4.8
		5	4.0	4.4	4.3	4.8
B	50	$\frac{3}{2}$	2.9	3.8	2.8	3.4
		$\frac{5}{2}$	3.0	4.0	2.9	3.5
		5	3.1	4.0	2.9	3.5
	100	$\frac{3}{2}$	3.6	4.2	3.2	3.8
		$\frac{5}{2}$	3.9	4.3	3.4	4.0
		5	3.9	4.3	3.4	4.0
	150	$\frac{3}{2}$	3.8	4.2	3.4	4.2
		$\frac{5}{2}$	4.1	4.4	3.7	4.4
		5	4.1	4.5	3.7	4.4
	500	$\frac{3}{2}$	4.3	4.7	4.0	4.5
		$\frac{5}{2}$	4.4	4.8	4.3	4.7
		5	4.4	4.8	4.3	4.7

Notes: Figures in the table show actual rejection (boundary crossing) frequencies corresponding to nominal test size of 5%.

Table 3b. Power for monitoring tests with nominal size 5%.

			Directional accuracy		Excess profitability	
DGP	T	τ	One-sided	Two-sided	One-sided	Two-sided
F	100	$\frac{3}{2}$	4.9	4.5	5.3	4.7
		$\frac{5}{2}$	6.3	5.0	7.4	5.4
		5	6.3	5.1	7.5	5.4
			(29)	(15)	(36)	(19)
	500	$\frac{3}{2}$	10.2	6.9	13.5	9.0
		$\frac{5}{2}$	17.7	10.8	26.1	16.9
		5	17.5	10.9	26.4	17.0
			(230)	(196)	(257)	(236)
G_1	100	$\frac{3}{2}$	4.9	4.5	5.3	4.7
		$\frac{5}{2}$	6.7	5.1	8.1	5.8
		5	7.6	5.3	10.1	6.3
			(58)	(25)	(82)	(43)
	500	$\frac{3}{2}$	10.2	6.9	13.5	9.0
		$\frac{5}{2}$	22.0	13.2	34.2	22.3
		5	41.0	23.5	66.6	46.6
			(774)	(733)	(807)	(841)
G_2	100	$\frac{3}{2}$	3.8	4.1	3.5	4.0
		$\frac{5}{2}$	4.7	4.4	4.8	4.5
		5	32.5	17.4	69.4	51.1
			(299)	(267)	(312)	(318)
	500	$\frac{3}{2}$	4.8	4.6	4.7	4.6
		$\frac{5}{2}$	9.3	6.5	13.9	9.3
		5	100	100	100	100
			(1179)	(1290)	(1023)	(1115)

Notes: Unbracketed figures in the table show actual rejection (boundary crossing) frequencies corresponding to nominal test size of 5%. Figures in brackets show how much time on average passes before boundary crossing takes place.

Table 3c. Size and power for monitoring tests with nominal size 5% and $T = 100$, alternative parabolic boundaries.

		Directional accuracy		Excess profitability	
DGP	τ	One-sided	Two-sided	One-sided	Two-sided
A	$\frac{3}{2}$	1.1	1.3	1.1	1.3
	$\frac{5}{2}$	1.8	2.2	1.7	1.9
	5	2.5	2.9	2.2	2.5
F	$\frac{3}{2}$	2.2	1.7	2.3	1.6
	$\frac{5}{2}$	5.5	3.4	6.8	4.1
	5	7.2	4.5	8.5	5.3
		(107)	(100)	(107)	(107)
G_1	$\frac{3}{2}$	2.2	1.6	2.3	1.6
	$\frac{5}{2}$	6.7	4.1	8.7	5.6
	5	20.9	14.1	31.2	22.7
		(215)	(219)	(223)	(232)
G_2	$\frac{3}{2}$	1.4	1.3	1.3	1.2
	$\frac{5}{2}$	3.8	2.8	4.6	3.3
	5	83.6	77.3	95.8	94.1
		(303)	(314)	(280)	(293)

Notes: See notes to Table 3b.

Table 4. Series of stock indexes.

Country	Series name	Historical period begins	Historical period size	Total sample size
Russia	RUX	Jan 1998	103	369
Ukraine	PFTS	Jan 1998	103	367
Poland	WIG	Jan 1997	155	423
Czech Rep	PX50	Jan 1997	155	422
Slovakia	SKSM	Jan 1997	151	404
Hungary	BUX	Jan 1997	155	422
Croatia	CROBEX	Jan 1997	154	418
Slovenia	SBI	Jan 1997	155	422
Romania	ROL	Jan 1999	53	309
Estonia	TALSE	Jan 1997	155	423

Notes: Irrespective of the initial date in a sample, the first week of 2000 is the end of the historical period and the start of the monitoring period.

Table 5a. Dates of hitting different boundaries.

Statistic	DA		EP	
Boundary	Linear	Parabolic	Linear	Parabolic
Russia				
Ukraine	−26	1	−26	1
Poland			−48	7
Czech Republic			16	
Slovakia				
Hungary				
Croatia				
Slovenia	14			
Romania				
Estonia	7	103	−81	2

Notes: Dates are measured with respect to the beginning of the monitoring period. Thus, if the figure is negative, it denotes the number of weeks before the monitoring period starts when the sequential statistic path crosses the 5% retrospective boundary; if it is positive, it denotes the number of weeks after the monitoring period starts when the sequential statistic path crosses the 5% monitoring boundary provided that it does not cross the 5% retrospective boundary.

Table 5b. Dates of hitting the linear boundary for various predictors and their functions.

x_t or $h(u)$	$x_t = y_{t-1}$		$x_t = \sum_{j=1}^{4} y_{t-j}$		$x_t = z_t'\hat{\beta}$		$h(u) = u$	
Statistic	DA	EP	DA	EP	DA	EP	DA	EP
Russia								
Ukraine	−24	−26	−25	11			−16	−5
Poland		0	0					
Czech Republic								
Slovakia		−51		−81				
Hungary								
Croatia		3						
Slovenia			−33					
Romania								
Estonia	−80	−80	2	−120				

Notes: See notes to Table 5a.

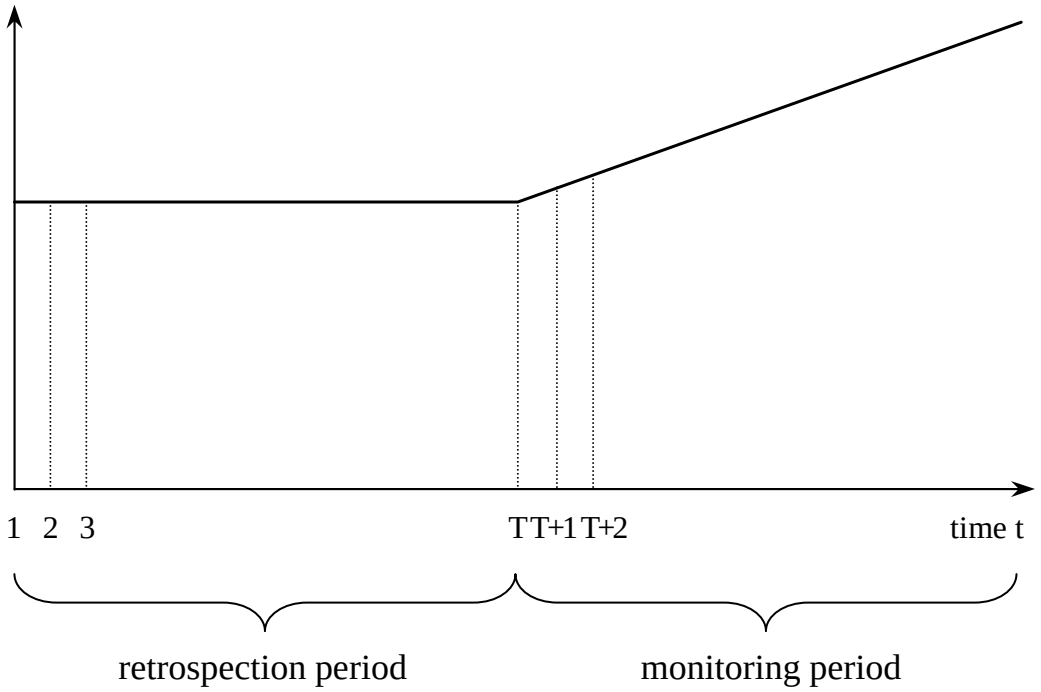

Figure 1. Sequential testing: periods and boundaries.

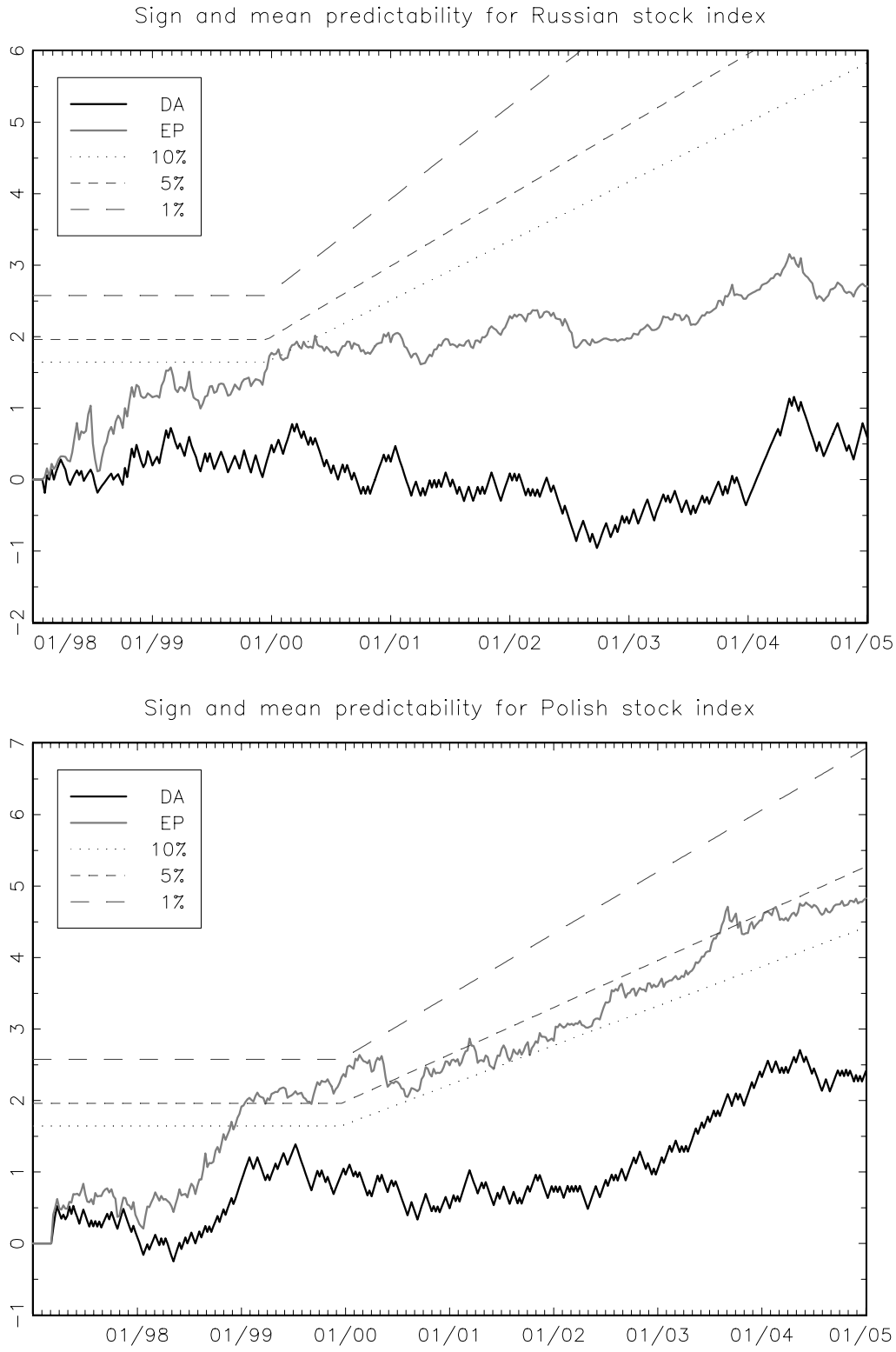

Figure 2. Sequential tests for some Eastern European stock indexes.

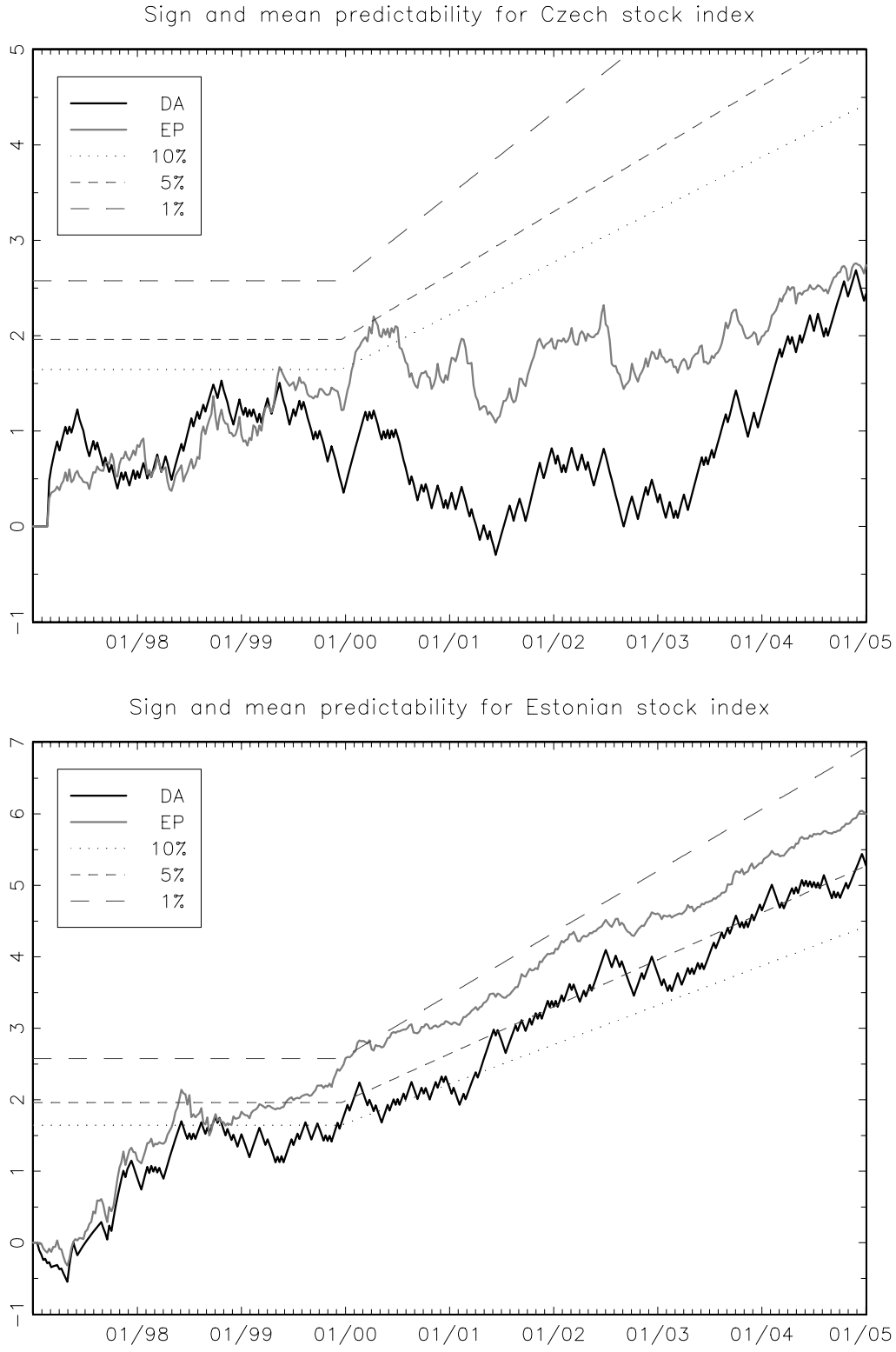

Figure 2 (continued). Sequential tests for some Eastern European stock indexes.