Nonlinearity, Nonstationarity and Spurious Forecasts[1]

Vadim Marmer[2]
Yale University

April 2004
Revised: November 2004

[1]Job market paper. I am grateful to Don Andrews, Peter Phillips, Yuichi Kitamura, Erik Hjalmarsson, Alex Maynard, Randi Pintoff, Kevin Song, and participants in the econometric seminar and workshop at Yale University for helpful comments and suggestions. I thank the Cowles Foundation for support under a Cowles Prize.

[2]Department of Economics, Yale University, New Haven, CT 06520. Email: vadim.marmer@yale.edu

Abstract

Various implications of nonlinearity, nonstationarity and misspecification are considered from a forecasting perspective. My model allows for small departures from the martingale difference sequence hypothesis by including an additive nonlinear component, formulated as a general, integrable transformation of the predictor, which is assumed to be I(1). Such a generating mechanism provides for predictability only in the extremely short run. In the stock market example, this formulation corresponds to a situation where some relevant information may escape the attention of market participants only for very short periods of time.

I assume that the true generating mechanism involving the nonlinear dependency is unknown to the econometrician and he is therefore forced to use some approximating functions. I show that the usual regression techniques lead to spurious forecasts. Improvements of the forecast accuracy are possible with properly chosen integrable approximating functions.

This paper derives the limiting distribution of the forecast MSE. In the case of square integrable approximants, it depends on the L_2-distance between the nonlinear component and the approximating function. Optimal forecasts are available for a given class of approximants. Finally, I present a Monte Carlo simulation study and an empirical example in support of the theoretical findings.

Keywords: Forecasting, integrated time series, misspecified models, nonlinear transformations, stock returns.

JEL Classification Numbers: C22, C53, G14.

1 Introduction

Nonlinear models are extensively used in econometrics (see, for example, Granger and Teräsvirta (1993) for a description and analysis of various nonlinear models). The theoretical foundation for estimation of nonlinear, nonstationary models has been developed fairly recently. Park and Phillips (1999) derived the asymptotic results for the sums of nonlinear transformations of integrated time series. They considered three classes of nonlinear functions: integrable, homogeneous and exponential. They show, for example, that partial sums of integrable functions that have a non-zero Lebesgue measure converge in distribution to local times of the Brownian motion. These methods have been applied to various nonlinear econometric models. Chang, Park, and Phillips (2001) considered nonlinear regression with separably additive regression functions. Chang and Park (2003) considered nonstationary index models, which extend switching regressions to the stochastic trends framework. Hu and Phillips (2004) studied nonstationary discrete choice models. Kasparis (2004) considered effects of functional form misspecification on estimation, when the true and the estimated models involve nonlinearity and nonstationarity. He focused on convergence of estimators to some pseudo-true values and detection of functional form misspecification.

An attractive feature of nonlinear models is flexibility that allows one to model relationships between nonstationary and seemingly stationary variables. A linear regression requires the dependent variable to have the same order of integration as the right-hand side of the regression equation. However, it is known that nonlinear transformations can change the memory properties of a process. Thus, contrary to linear regressions, properly chosen nonlinear functions can link, in a single equation, variables that appear to have different orders of integrations. Nonlinear functions that can be used to model relationships between seemingly stationary and persistent variables include Lebesgue integrable functions and asymptotically homogeneous functions of degree zero (the distribution-like functions). For example, Chang and Park (2003) modelled nonstationary switching behavior by adding a distribution type function of an integrated variable to a noise process.

There are many situations in economics that may require one to relate variables of different orders of integration. A typical example is the predictive regressions literature in empirical finance, which studies stock returns predictability. In a predictive regression, stock returns are regressed on lagged values of various financial and economic variables such as the dividend-price ratio, earnings-price ratio or interest rates. Predictability is usually concluded on the basis of t-tests for slope coefficients. Often, it is implicitly assumed that non-zero in-sample correlations found between regressors and stock returns can be used for the construction of out-of-sample forecasts. Many papers report statistically significant slope estimates. (See, for example, Fama (1991) and Cochrane (1997) for surveys of the literature). Despite the collected empirical evidence on in-sample relations between stock returns and predictors, the out-of-sample predictability is a controversial issue. Goyal and Welch (2003, 2004) report that the performance of out-of-sample forecasts based on linear regression methods can be rather poor, while Campbell and Thompson (2004) argue that there exists small but economically meaningful out-of-sample predictive power, once restrictions on the coefficients and forecasts are imposed. An additional complication arises from the fact that many potential predictors are best modelled as I(1) variables. While most researchers agree that stock market returns are not persistent, predictors such as dividend-price ratio appear to have a stochastic trend component. Naturally, it is impossible to relate such variables by a linear equation.

The results of this paper imply that significant regression slopes do not necessarily indicate usefulness of the linear regression as a forecasting equation. My model allows for small departures from the martingale difference sequence (MDS) hypothesis by including an additive nonlinear component, formulated as a general, integrable transformation of the predictor, which is assumed to be I(1). In this model, the signal coming from the nonlinear component is very weak relative to the noise. This is implied by the properties of integrable functions and I(1) variables. An integrable function approaches zero at a fast rate as the absolute value of its argument increases. At the same time, a unit root process usually takes on very large negative or positive values. As a result, the signal coming from the predictor (the nonlinear component)

is relatively strong only during rare events when the unit root process visits the neighborhood of zero. Such a generating mechanism provides for predictability only in the extremely short run, which in the stock market example corresponds to a situation where some relevant information may escape the attention of market participants only for very short periods of time.

It is natural to assume that the true data generating process (DGP) involving the nonlinear dependency is unknown to the econometrician and he is therefore forced to use some approximating functions. Furthermore, the class of approximants used by the econometrician does not necessarily include the true function. I show that a combination of nonstationarity, nonlinearity and misspecification leads to results often seen in the predictive regressions literature. Consider for example a linear regression, which is used most often as an approximating function. I show that, in this case, commonly used diagnostic tools tend to indicate predictive power despite the fact that estimated regression slopes converge to zero in probability. Moreover, I show that out-of-sample forecasts constructed from a predictive regression are dominated by the sample average in terms of the mean squared error (MSE). Hence, spurious forecasts occur: diagnostic tools may indicate usefulness of the model while, in fact, equivalent or better forecasts may be obtained if one completely ignores the information contained in the predictor.

The predictability in my model is very limited due to the nature of the generating process. Nevertheless, I show that out-of-sample forecast accuracy can be improved by using square integrable approximating functions instead of historic averages or linear regressions. I derive the limiting distribution of the out-of-sample MSE. In the case of square integrable approximants, it depends on the L_2-distance between the nonlinear component and the approximating function. I show that, for a given class of square integrable approximating functions, one can obtain the best possible forecasts in the MSE sense.

The paper is organized as follows. In Section 2, I introduce the model and the preliminary results from the asymptotic theory of nonlinear functions of integrated processes. In Section 3, I consider the class of forecasts constructed as polynomials in

the predictor. This class contains predictive regressions as a particular case. Section 4 discusses forecasting with integrable approximants. In Section 5, I consider a Wald-type predictability test based on square integrable transformations. I present some simulation results in Section 6, and Section 7 provides an empirical example. Section 8 concludes. All proofs are given in the Appendix.

2 Definitions and preliminary results

I consider a nonstationary nonlinear model described by the following equations.

$$\begin{aligned} y_t &= \mu^* + f(z_{t-1}) + u_t, \\ z_t &= z_{t-1} + C(L)\varepsilon_t, \\ z_0 &= 0. \end{aligned} \tag{2.1}$$

In the above equations, μ^* is a constant, $\{(u_t, \varepsilon_t) : t = 1, ..., n\}$ are random variables, $C(L)$ is a polynomial in the lag operator and $f : R \to R$ is a nonlinear function. Two classes of nonlinear function will be considered in this paper. The first class consists of Lebesgue integrable functions that have the Lebesgue measure different from zero. I denote this class by $\mathcal{I}$.

2.1. Definition. A function $\varphi : R \to R$ is said to belong to the class $\mathcal{I}$ if

(a) φ and φ^2 are Lebesgue integrable,

(b) $\int_{-\infty}^{+\infty} \varphi(x)dx \neq 0$.

I make the following assumption.

2.2. Assumption (Integrability). f and $f^2 \in \mathcal{I}$.

The second class consists of zero energy integrable functions (the Lebesgue integral over the entire real line is equal to zero). Such functions appear later in the paper (Section 3). This class is denoted by $\mathcal{Z}$.

2.3. Definition. A function $\varphi : R \to R$ is said to belong to the class $\mathcal{Z}$ if

(a) φ and φ^2 are Lebesgue integrable,

(b) $\int_{-\infty}^{+\infty} \varphi(x)dx = 0$,
(c) $\int_{-\infty}^{+\infty} |x\varphi(x)|\, dx < \infty$.

Assumption 2.2 implies that $f(x) \to 0$ as $x \to \pm\infty$. Since $\{z_t\}$ is an integrated process, it takes on very large negative or positive values most of the time. As a result, $\{f(z_{t-1})\}$ is arbitrarily close to zero except when $\{z_{t-1}\}$ visits the neighborhood of zero. The integrated variable z_{t-1} becomes a useful predictor for y_t only on such rare occasions. Furthermore, sample paths of $\{f(z_{t-1}) + u_t\}$ appear to be similar to the sample paths of the noise process $\{u_t\}$.

Let $\mathcal{F}_t$ be the σ-field generated by the sequence $\{(u_s, \varepsilon_s) : s \leq t\}$. I assume that the error u_t cannot be predicted from its lagged values and the lagged values of the predictor z_t.

2.4. Assumption (MDS). $\{((u_t, \varepsilon_t), \mathcal{F}_t) : t \geq 1\}$ is a stationary and ergodic martingale difference sequence.

Equation (2.1) and Assumption 2.4 define a predictive model. Suppose that the function f were known. In this case, optimal forecasts in the Mean Squared Error (MSE) sense are given by

$$\widehat{y}_t(\widehat{\mu}) = \widehat{\mu} + f(z_{t-1}),$$

where $\widehat{\mu}$ is the Least Squares (LS) estimator of μ. However, the optimal forecasts are infeasible if f is unknown to the econometrician. Such an assumption is plausible in the vast majority of situations. In this case, the econometrician is forced to use some approximating functions instead of f. The class of approximating functions considered by the econometrician may or may not include the true function f, and in general he accepts a misspecified forecasting model. I denote the approximating function by $g(\cdot, \theta)$, where θ is a vector of constants. The value of θ is chosen by the econometrician in order to obtain best forecasts given his choice of g. In this paper, I consider two alternatives for $g(x, \theta)$. The first is a polynomial function in x, which includes predictive regression and constant forecasts ($g(x, \theta) = 0$ for all $x \in R$) as particular cases. Since the true DGP involves a square integrable function, the

second type of approximating functions that I consider consists of square integrable functions in x.

Suppose that the econometrician observes the data $\{(y_t, z_{t-1}) : t = 1, ..., n_1\}$. His objective is to construct one period ahead forecasts of y_t for periods $\{n_1 + 1, ..., n\}$ using the actual values of z_{t-1}, which are observed before y_t is realized. The forecasting model is defined as

$$\widehat{y}_t(\mu, \theta) = \mu + g(z_{t-1}, \theta), \tag{2.2}$$

where $\widehat{y}_t$ is the predicted value of y_t, the function g is chosen by the econometrician, and μ and θ are scalar and vector constants respectively. I assume that μ and θ are estimated from observations $\{(y_t, z_{t-1}) : t = 1, ..., n_1\}$, which I call a training sample. Let $\left(\widehat{\mu}, \widehat{\theta}'\right)'$ be the estimates of $(\mu, \theta')'$. The econometrician uses the estimated version of (2.2) in order to construct forecasts for the observations in the forecasting sample, which consists of observations $n_1 + 1$ through n. I assume that the forecasts are evaluated with a quadratic loss function:

$$Q_n\left(\widehat{\mu}, \widehat{\theta}\right) = (n - n_1)^{-1} \sum_{t=n_1+1}^{n} \left(y_t - \widehat{\mu} - g\left(z_{t-1}, \widehat{\theta}\right)\right)^2 - (n - n_1)^{-1} \sum_{t=n_1+1}^{n} u_t^2.$$

The first term in the above expression is the MSE of the series of forecasts

$$\left\{\widehat{y}_t\left(\widehat{\mu}, \widehat{\theta}\right) : t = n_1 + 1, ..., n\right\}.$$

The second term is the MSE of the infeasible forecasts

$$\{\mu^* + f(z_{t-1}) : t = n_1 + 1, ..., n\}.$$

The second term does not depend on the choice of a forecasting function or the values of μ and θ, and therefore minimization of $Q_n(\mu, \theta)$ is achieved only through minimization of the first component. The second term is included for the derivation of the asymptotic results. I assume that $n_1 = [nr]$, where $r \in (0, 1)$, and $[\cdot]$ denotes the integer part. This setup mimics the situation where a researcher updates his forecasts when more observations become available. However, he also extends his forecasting horizon so that the relative sizes of the training and forecasting samples remain unchanged.

In addition to the MDS assumption, I make the following assumptions concerning the innovations process $\{(u_t, \varepsilon_t) : t \geq 1\}$.

2.5. Assumption. (a) $E\left((u_1, \varepsilon_1)'(u_1, \varepsilon_1) \,|\mathcal{F}_{t-1}\right) = \Sigma > 0$ a.s. for all t.

(b) $\sup_{t\geq 1} E\left(|u_t|^h \,|\mathcal{F}_{t-1}\right) < \infty$ for some $h > 2$.

(c) $\{\varepsilon_t : t \geq 1\}$ are iid with $E\,|\varepsilon_1|^h < \infty$ for some $h > 8$.

(d) $C(1) \neq 0$ and $\sum_{k=0}^{\infty} k\,|c_k| < \infty$.

(e) The distribution of ε_1 is absolutely continuous with respect to the Lebesgue measure and $\left|Ee^{ik\varepsilon_1}\right| = o\left(k^{-\beta}\right)$ for some $\beta > 0$.

Write

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_{zu} \\ \sigma_{zu} & \sigma_u^2 \end{pmatrix}.$$

Under Assumptions 2.4 and 2.5 in large samples, the distribution of a process

$$\left(n^{-1/2} z_{[nr]}, n^{-1/2} \sum_{t=1}^{[nr]} u_t\right)$$

can be approximated by the distribution of a two-dimensional Brownian motion with the covariance matrix

$$\Omega = \begin{pmatrix} \omega_z^2 & \omega_{zu} \\ \omega_{zu} & \sigma_u^2 \end{pmatrix},$$

where $\omega_z^2 = C(1)^2 \sigma_z^2$ is the long-run variance of the innovations of z_t, and ω_{zu} is the long-run covariance of u_t and the innovations of z_t : $\omega_{zu} = \sum_{h=1}^{\infty} E\left(\varepsilon_h u_1\right)$. Note that the correlations between u_t and the lagged values of ε_t are equal to zero due to Assumption 2.4.

I assume that the parameters in (2.2) are estimated by LS, which leads to expressions of the form $\sum_t f\left(z_{t-1}\right)$. Asymptotically, partial sums of integrable transformations of I(1) variables are approximated by local times of a Brownian motion. The local time (L) of the Brownian motion B at $x \in R$ is defined as follows:

$$L\left(t, x\right) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t 1_{(x-\varepsilon, x+\varepsilon)}\left(B(s)\right) ds,$$

where 1_A is an indicator function of a set $A \subset R$. The local time measures the amount of time that the Brownian motion B spends in the neighborhood of the point x (see, for example, Chung and Williams (1990) for an introduction to local time). The following result, due to Park and Phillips (1999) and Jeganathan (2003), is the foundation of the subsequent discussion.

2.6. Lemma. Let $\varphi_h : R \to R$ be a homogeneous function of degree $h > 0$, $\varphi_I \in \mathcal{I}$ and $\varphi_0 \in \mathcal{Z}$. Let w_u and w_z be two independent standard Brownian motions. Define $\ell(t) = \omega_z^{-1} L_z(t,0)$, where L_z is the local time of w_z. If Assumptions 2.4 and 2.5 hold, then for fixed $0 < s < r < 1$ the following results hold jointly:

(a) $\left(n^{-1/2}\sum_{t=[ns]}^{[nr]} \varepsilon_t, n^{-1/2}\sum_{t=[ns]}^{[nr]} u_t\right) \to_d \left(\int_s^r db(t), \int_s^r du(t)\right)$, where $(b,u)' = \Omega^{1/2}(w_z, w_u)'$.

(b) $n^{-1-h/2}\sum_{t=[ns]}^{[nr]} \varphi_h(z_{t-1}) \to_d \int_s^r \varphi_h(b(t))\, dt$.

(c) $n^{-1/2-h/2}\sum_{t=[ns]}^{[nr]} \varphi_h(z_{t-1})\, u_t \to_d \int_s^r \varphi_h(b(t))\, du(t)$.

(d) $n^{-1/2}\sum_{t=[ns]}^{[nr]} \varphi_I(z_{t-1}) \to_d \left(\int_{-\infty}^{\infty} \varphi_I(x)\, dx\right) \int_s^r d\ell(t)$.

(e) $n^{-1/4}\sum_{t=[ns]}^{[nr]} \varphi_I(z_{t-1})\, u_t \to_d \sigma_u \left(\int_{-\infty}^{\infty} \varphi_I^2(x)\, dx \int_s^r d\ell(t)\right)^{1/2} \int_s^r dW(t)$, where W is a Brownian motion independent of (w_z, w_u).

(f) $n^{-1/4}\sum_{t=[ns]}^{[nr]} \varphi_0(z_{t-1}) \to_d \left(\lambda \int_s^r d\ell(t)\right)^{1/2} \int_s^r dV(t)$, provided that there exist constants a and $0 < b < 1$ such that $\left|\widehat{\varphi}_0(t) E e^{it\varepsilon_1}\right| \leq a\,|t|^{-(2+b)}$ as $t \to \infty$, where V is a Brownian motion independent of W and (w_z, w_u), $\lambda = (2\pi)^{-1} \int |\widehat{\varphi}_0(t)|^2 \frac{1+Ee^{it\varepsilon_1}}{1-Ee^{it\varepsilon_1}} dt$, and $\widehat{\varphi}_0$ is the Fourier transform of φ_0.

Remark. Using the Cramer-Wold device, the result in this lemma can be extended to cover functions $\varphi : R \to R^p$; see Lemma 9.2 in the Appendix.

3 Forecasting with polynomials

In this section, I consider forecasts constructed as polynomials in lagged values of the predictor. In this case,

$$g(z_{t-1}, \theta) = \sum_{j=1}^{p} \theta_j z_{t-1}^j, \tag{3.3}$$

where p is a positive integer chosen by the econometrician. For $p = 1$, equation (3.3) reduces to a simple linear regression considered in the predictive regression literature. In addition to Assumption 2.2, I assume that the function f in (2.1) satisfies the following condition.

3.1. Assumption. $f(x)x^p \in \mathcal{I}$.

It is assumed that the parameters in the forecasting equation are estimated by LS using the training sample. Collecting powers of z_t in a single vector, I define for a fixed value $r \in (0,1)$:

$$\begin{aligned} Z_t &= (z_t, \ldots, z_t^p)', \\ \underline{Z}_t &= Z_t - [nr]^{-1} \sum_{s=1}^{[nr]} Z_{s-1}. \end{aligned} \tag{3.4}$$

The LS estimator of μ and θ in (3.3) is given by

$$\begin{aligned} \widehat{\theta}_n &= \left(\sum_{t=1}^{[nr]} \underline{Z}_{t-1} \underline{Z}_{t-1}' \right)^{-1} \sum_{t=1}^{[nr]} \underline{Z}_{t-1} y_t, \\ \widehat{\mu}_n &= [nr]^{-1} \sum_{t=1}^{[nr]} y_t - [nr]^{-1} \sum_{t=1}^{[nr]} Z_{t-1}' \widehat{\theta}_n. \end{aligned} \tag{3.5}$$

Suppose that the econometrician draws a conclusion regarding the predictability of y_t from a test based on the usual F-statistic for θ:

$$\begin{aligned} F_n &= \widehat{\theta}_n' \left(\sum_{t=1}^{[nr]} \underline{Z}_{t-1} \underline{Z}_{t-1}' \right) \widehat{\theta}_n / s_n^2, \text{ where} \\ s_n^2 &= [nr]^{-1} \sum_{t=1}^{[nr]} \left(y_t - \widehat{\mu}_n - Z_{t-1}' \widehat{\theta}_n \right)^2. \end{aligned} \tag{3.6}$$

I compare the forecasts constructed according to model (3.3) with a baseline model that assumes no predictability:

$$\widehat{y}_t(\mu_0) = \mu_0. \tag{3.7}$$

Equation (3.7) is a particular case of a polynomial forecasting function. It depends on a single parameter μ_0, which is estimated by the average value of $\{y_t : t = 1, \ldots, [nr]\}$:

$$\widehat{\mu}_{0,n} = [nr]^{-1} \sum_{t=1}^{[nr]} y_t.$$

Similarly to equation (3.4), I define

$$\begin{aligned} B(t) &= (b(t), \ldots, b(t)^p)', \\ \underline{B}(t) &= B(t) - r^{-1}\int_0^r B(s)ds, \text{ and} \\ \underline{u}(t) &= u(t) - r^{-1}\int_0^r u(s)ds, \end{aligned}$$

where the Brownian motions b and u are introduced in Lemma 2.6. Let $D_n = \operatorname{diag}(n, \ldots, n^p)$. The following theorem describes the asymptotic behavior of the estimators $\widehat{\theta}_n$ and $\widehat{\mu}_n$, the test statistic F_n, and the loss function Q_n.

3.2. Theorem. Under Assumptions 2.2, 2.4, 2.5 and 3.1:

(a) $n^{1/2}D_n^{1/2}\widehat{\theta}_n \to_d \Psi$, where

$$\begin{aligned} \Psi &= \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{-1}\int_0^r \underline{B}(s)du(s) \\ &\quad -r^{-1}\ell(r)\left(\int_{-\infty}^{\infty} f(x)dx\right)\left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{-1}\int_0^r B(s)ds. \end{aligned} \tag{3.8}$$

(b) $n^{1/2}\left(\widehat{\mu}_n - \mu^*\right) \to_d r^{-1}\left(\ell(r)\int_{-\infty}^{\infty} f(x)dx + u(r) - \Psi'\int_0^r B(s)ds\right)$.

(c) $F_n \to_d \left\|\left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{1/2}\Psi/\sigma_u\right\|^2$.

(d) For any $p \geq 0$, $n^{1/2}Q_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right) \to_d (1-r)^{-1}\left(\ell(1) - \ell(r)\right)\int_{-\infty}^{\infty} f^2(x)dx$.

(e) For any $p > 0$, $n\left(Q_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right) - Q_n\left(\widehat{\mu}_{0,n}, 0\right)\right) \to_d \Delta$, where the random variable Δ is given by

$$\begin{aligned} (1-r)\Delta &= \Psi'\left(\int_r^1 \underline{B}(s)\underline{B}(s)'ds\right)\Psi \\ &\quad +2r^{-1}\left(\int_{-\infty}^{\infty} f(x)dx\right)\Psi'\left(\ell(r)\int_r^1 \underline{B}(s)d(s) + \int_1^r d\ell(s)\int_0^r B(s)ds\right) \\ &\quad -2\Psi'\int_r^1 \underline{B}(s)d\underline{u}(s). \end{aligned}$$

Part (a) of the theorem implies that $\widehat{\theta}_n$ converges in probability to zero. In large samples, its distribution can be approximated by the distribution of the random variable defined in (3.8). The first term on the right-hand side of (3.8) is the usual expression obtained in the limit when one regresses an I(0) variable on I(1) regressors.

This term has a mixed normal distribution when $\omega_{zx} = 0$. The second term in equation (3.8) comes from the nonlinear part of the DGP. It depends both on the integral of f over the entire real line and on the local time at zero of the limiting process of z_t. The second component gives the mean of the mixed normal distribution when u_s and ε_t are uncorrelated for all s and t.

Despite the fact that $\widehat{\theta}_n$ converges to zero in probability, part (c) of the theorem implies that a test based on F_n tends to reject the hypothesis of no predictability. (In the current context, the hypothesis of no predictive power is equivalent to $\theta = 0$). For example, consider the case $\omega_{zx} = 0$. In this case, F_n has a mixed noncentral χ^2_p distribution with the noncentrality parameter given by

$$\left(\sigma_u^{-1}\int_{-\infty}^{\infty} f(x)dx\right)^2 \left\| \left(r^{-1}\ell(r)\int_0^r \underline{B}(s)\underline{B}'(s)ds\right)^{-1/2} \int_0^r B(s)ds \right\|^2 .$$

Consequently, the test that rejects the null of no predictability when $F_n > \chi^2_{p,1-\alpha}$, where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$ quantile of a central χ^2_p distribution, rejects the null of no predictive power with probability greater than the nominal probability α. Actual rejection probabilities depend on the ratio of the integral of f to the standard deviation of the noise process u_t. Note that the shape of the nonlinear function f has no effect on the rejection rates, since f appears in the expression for the noncentrality parameter only through its integral over the entire real line.

Finally, part (d) of the theorem shows that Q_n has the same limiting distribution regardless of the value of p. In particular, the baseline model ($p = 0$) asymptotically yields the same loss function as a model with $p > 0$. Moreover, the asymptotic distribution of the loss function does not depend on the information contained in the predictor. Therefore, the inclusion of powers of the predictor in a forecasting equation does not improve the forecast accuracy. Part (e) of the theorem describes the asymptotic distribution of the difference of the loss functions for the polynomial and baseline forecasting models. The support of the limiting distribution includes both positive and negative parts of the real line. Later in the paper, I show, using Monte Carlo simulations, that this difference tends to be positive in finite samples and, therefore, that the baseline model dominates polynomials in the MSE sense.

4 Forecasting with integrable functions

The previous section illustrates that polynomials can be poor predictors when the DGP involves nonstationarity and nonlinearities of a certain type. In fact, a simple average can dominate a polynomial forecasting model in terms of the MSE. The reason for this lies in the global nature of the LS approximation in the current framework. Consider minimization of the L_2-distance between $f(x)$ and some approximating function $g(x,\theta)$:

$$\inf_{\theta\in\Theta}\int_{-\infty}^{\infty}\left(f(x)-g(x,\theta)\right)^2 dx, \tag{4.9}$$

where $\Theta \subset R^p$ is a compact set. Suppose that $g(x,\theta)$ is unbounded and diverges to $\pm\infty$ as $x \to \pm\infty$, which is true for polynomials. In this case, a solution to (4.9) demands a choice of θ such that $g(x,\theta) = 0$ for all $x \in R$. This illustrates why, in the previous section, the limit of the loss function Q_n is proportional to $\int_{-\infty}^{\infty} f^2(x)dx$ and does not depend on the information contained in the predictor. The situation changes if one uses square integrable approximating functions instead of polynomials. If $g(x,\theta)$ is square integrable, then a non-trivial solution to (4.9) exists, which leads to improvements in forecast accuracy. This occurs since

$$\int_{-\infty}^{\infty} f^2(x)dx \geq \inf_{\theta\in\Theta}\int_{-\infty}^{\infty}\left(f(x)-g(x,\theta)\right)^2 dx$$

whenever there exists $\widetilde{\theta} \in \Theta$ such that $g\left(x,\widetilde{\theta}\right)$ is equal to zero almost everywhere.

In this section, I consider forecasting with square integrable (with respect to x) functions $g(x,\theta)$ in the forecasting equation (2.2). The function g has to be nonlinear in x due to the integrability assumption; however, it may depend on θ in a linear or nonlinear way. I assume that the econometrician restricts θ to a compact subset of R^p, denoted by Θ. The dimension of θ is chosen by the econometrician together with the functional form of g. Linear (in θ) forecasting functions are of greatest interest:

$$g(x,\theta) = \sum_{i=1}^{p}\theta_i\phi_i(x), \tag{4.10}$$

where $\phi_i \in \mathcal{I}$ for $1 \leq i \leq p$. An example of a nonlinear function in θ is the class of

extended rational polynomials (ERP's):

$$\phi(x)\frac{a_0 + a_1 x + ... + a_p x^p}{1 + b_1 x + ... + b_p x^p}, \tag{4.11}$$

where $\phi \in \mathcal{I}$ and $\theta = (a_0, a_1 ..., a_p, b_1, ..., b_p)$. Functions of this type were used by Phillips (1983) for density approximation. I make the following assumption concerning the approximation function g :

4.1. Assumption. (a) $g(x, \theta)$ is differentiable with respect to θ.

(b) $g^2(\cdot, \theta) \in \mathcal{I}$ for all $\theta \in \Theta$.

(c) $\sup_{\theta \in \Theta} |g(\cdot, \theta)| \in \mathcal{I}$.

(d) $\sup_{\theta \in \Theta} \left\| \frac{\partial}{\partial \theta} g(\cdot, \theta) \right\|, \sup_{\theta \in \Theta} \left\| \frac{\partial}{\partial \theta} g(\cdot, \theta) \right\|^2 \in \mathcal{I}$.

The solution to the problem described in (4.9) depends on the choice of g. In some cases, such as (4.10) and (4.11), there exists a unique θ that solves (4.9). However, in general, multiple solutions may exist. Let Θ^* be the set of solutions to (4.9):

$$\begin{aligned} \Theta^* &= \left\{ \theta^* \in \Theta : \int_{-\infty}^{\infty} (f(x) - g(x, \theta^*))^2 dx = M^* \right\}, \text{ where} \\ M^* &= \inf_{\theta \in \Theta} \int_{-\infty}^{\infty} (f(x) - g(x, \theta))^2 dx. \end{aligned}$$

Note that different choices of g and p lead to different M^* and Θ^*. Any value $\theta^* \in \Theta^*$ can be treated as a pseudo-true value of θ for a given choice of g.

Similarly to Section 3, I assume that μ and θ are estimated by LS from the training sample (nonlinear LS if g is nonlinear in θ). In the second step, the estimates of μ and θ are used to compute predicted values of y in the forecasting sample. Let Θ_n be the set of values of $\theta \in \Theta$ that solve the in-sample LS problem:

$$\min_{\theta \in \Theta, \mu} \sum_{t=1}^{[nr]} (y_t - \mu - g(z_{t-1}, \theta))^2.$$

For each $\widehat{\theta}_n \in \Theta_n$, the corresponding estimate of μ is given by

$$\widehat{\mu}_n \left(\widehat{\theta}_n \right) = [nr]^{-1} \sum_{t=1}^{[nr]} \left(y_t - g \left(z_{t-1}, \widehat{\theta}_n \right) \right). \tag{4.12}$$

I define the distance between θ and the set $A \subset R^p$ as

$$d(\theta, A) = \inf_{a \in A} \|a - \theta\|.$$

The following result describes the behavior of the LS estimators of θ and μ and the error function Q_n as the sample size approaches infinity.

4.2. Theorem. Under Assumptions 2.2, 2.4, 2.5 and 4.1:

(a) $\sup_{\widehat{\theta}_n \in \Theta_n} d\left(\widehat{\theta}_n, \Theta^*\right) \to_p 0$.

(b) $\sup_{\widehat{\theta}_n \in \Theta_n} \left|\widehat{\mu}_n\left(\widehat{\theta}_n\right) - \mu^*\right| \to_p 0$.

(c) $\sup_{\widehat{\theta}_n \in \Theta_n} n^{1/2} Q_n\left(\widehat{\mu}_n\left(\widehat{\theta}_n\right), \widehat{\theta}_n\right) \to_d (1-r)^{-1} (\ell(1) - \ell(r)) M^*$.

Theorem 4.2(a) implies that the LS estimator of θ is consistent for its pseudo-true value. Further, it follows from part (b) of the Theorem that $\widehat{\mu}_n\left(\widehat{\theta}_n\right)$ is a consistent estimator of μ^*. Next, part (c) of the Theorem shows that in the case of square integrable approximating functions, Q_n is proportional to the least distance between the true nonlinear function and its approximant: $M^* = \inf_{\theta \in \Theta} \int_{-\infty}^{\infty} (f(x) - g(x, \theta))^2 dx$. Thus, asymptotically one achieves the lowest possible out-of-sample MSE for a given class of functions g. Finally, comparison of the results of Theorem 3.2(d) and 4.2(c) implies that integrable functions yield a non-trivial improvement of the forecast accuracy over polynomials and the baseline model.

For certain choices of g, the solution to (4.9) is unique. The approximating functions in (4.10) and (4.11) are two such examples. In this case, stronger statements than in Theorem 4.2(a)-(b) can be made. Since θ^* is unique, $\widehat{\theta}_n$, the LS estimate of θ, is unique with probability approaching 1. I define

$$\begin{aligned} S(x, \theta) &= \frac{1}{2} \frac{\partial}{\partial \theta} (f(x) - g(x, \theta))^2, \\ H(x, \theta) &= \frac{1}{2} \frac{\partial^2}{\partial \theta \partial \theta'} (f(x) - g(x, \theta))^2. \end{aligned}$$

In addition to Assumption 4.1, I assume:

4.3. Assumption. (a) $\Theta^* = \{\theta^*\}$, where θ^* lies in the interior of Θ.

(b) $g(x, \theta)$ has three derivatives with respect to θ.

(c) $H(\cdot,\theta) \in \mathcal{I}$ element-by-element for all $\theta \in \Theta$.

(d) $\sup_{\theta\in\Theta} \left\| \frac{\partial}{\partial\theta} H_{i,j}(\cdot,\theta) \right\| \in \mathcal{I}$ for all $1 \leq i,j \leq p$, where $H_{i,j}$ is the element (i,j) of the $p \times p$-matrix H.

(e) $\int_{-\infty}^{\infty} H(s,\theta^*)\,dx$ is invertible.

Assumption 4.3(a) implies that $S(x,\theta^*)$ is a zero energy function:

$$\int_{-\infty}^{\infty} S(x,\theta^*)\,dx = 0. \tag{4.13}$$

I assume that $S(x,\theta^*)$ satisfies all the additional requirements of the $\mathcal{Z}$ class. Let $\widehat{S}(\cdot,\theta)$ denote the Fourier transform of $S(x,\theta)$ and $\widehat{S}^*(\cdot,\theta)$ denote the complex conjugate of $\widehat{S}(\cdot,\theta)$.

4.4. Assumption. (a) $S(x,\theta^*) \in \mathcal{Z}$ element-by-element.

(b) $\left|\widehat{S}(t,\theta^*)Ee^{it\varepsilon_1}\right| \leq a\,|t|^{-(2+b)}$ for some constants a and $0 < b < 1$, as $t \to \infty$.

The theorem below describes the asymptotic distribution of $\widehat{\mu}_n$ and $\widehat{\theta}_n$.

4.5. Theorem. Under Assumptions 2.2, 2.4, 2.5, 4.1, 4.3 and 4.4

(a) $n^{1/2}(\widehat{\mu}_n - \mu^*) \to_d r^{-1}\left(\ell(r)\int_{-\infty}^{\infty}(f(x) - g(x,\theta^*))\,dx + u(r)\right)$.

(b) $n^{1/4}\left(\widehat{\theta}_n - \theta^*\right) \to_d \ell(r)^{-1/2}A^{1/2}W(r)$, where W is a standard Brownian motion independent of ℓ and u, and A is $p \times p$ matrix of constants such that

$$\begin{aligned} & \left(\int_{-\infty}^{\infty} H(x,\theta^*)\,dx\right) A \left(\int_{-\infty}^{\infty} H(x,\theta^*)\,dx\right) \\ = \; & (2\pi)^{-1}\int_{-\infty}^{\infty} \widehat{S}(t,\theta^*)\,\widehat{S}^*(t,\theta^*)\,\frac{1+Ee^{it\varepsilon_1}}{1-Ee^{it\varepsilon_1}}dt \\ & +\sigma_u \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} g(x,\theta^*)\,\frac{\partial}{\partial\theta'} g(x,\theta^*)\,dx. \end{aligned} \tag{4.14}$$

According to part (a) of the theorem, $\widehat{\mu}_n$ has the usual $1/n^{1/2}$ rate of convergence to μ^*. Its limiting distribution is normal conditional on ℓ and centered around $\ell(r)M^*$, which depends on the unknown function f. Part (b) of the theorem shows that the rate of convergence of $\widehat{\theta}_n$ is $1/n^{1/4}$, which is slower than the usual $1/n^{1/2}$. Its asymptotic distribution is mixed normal. Further, the covariance matrix depends on

the unknown function f through the Fourier transform of S. Nevertheless, estimation of the covariance is possible in some cases. For example, this is true if $f(x) = 0$ for all $x \in R.$; such a situation occurs if y_t does not depend on z_{t-1}.

5 Testing predictability

In the previous section, I show that the LS estimator of θ converges in probability to its pseudo-true value θ^*, which minimizes the L_2-distance between $f(x)$ and the approximating function $g(x, \theta)$. Consider a situation where, for a given choice of the square integrable function g, z_{t-1} has no predictive power over y_t, and $M^* = \int_{-\infty}^{\infty} f^2(x)dx$. In this case, it follows from the results of Theorem 4.2 that, in large samples, one should expect the estimator $\widehat{\theta}_n$ to be close to some $\theta^* \in \Theta$ that satisfies $g(x, \theta^*) = 0$ for all $x \in R$. For example, $\widehat{\theta}_n$ converges in probability to zero if y_t and z_{t-1} are unrelated and g is linear in θ.

The purpose of this section is to construct a testing procedure that rejects the null of no predictability only if the corresponding model possesses out-of-sample predictive power superior to that of the baseline model. In view of the results presented in the previous sections, I propose a modification of predictive regressions based on integrable transformations of the predictor. It is convenient to consider the class of linear approximants defined by equation (4.10). In this case, the hypothesis of interest is whether $\theta = 0$. The linear forecasting model is computationally simple. However, its advantages are not limited to computational convenience. For nonlinear (in θ) approximants, some parameters may be unidentified under the null. For example, in the case of ERP's described in equation (4.11), under the null of no predictability $a_0 = a_1 = ... = a_p = 0$. However, the coefficients in the denominator $(b_1, ..., b_p)$ are not identified under the null. Any value of $(b_1, ..., b_p)$ would give asymptotically the same result.

I consider a local to zero alternative DGP:

5.1. Assumption. $y_t = \mu + n^{-1/4} f(z_{t-1}) + u_t$.

Scaling by $n^{-1/4}$ instead of usual $n^{-1/2}$ in the alternative DGP follows from

the convergence results described in Lemma 2.6. I make the following assumption regarding the basis functions ϕ_i in (4.10):

5.2. Assumption. ϕ_i and $\phi_i^2 \in \mathcal{I}$ for all $1 \leq i \leq p$.

Let $\Phi(z_t) = \left(\phi_1(z_t), ..., \phi_p(z_t)\right)'$. The LS estimator of θ is given by

$$\widehat{\theta}_n = \left(\sum_{t=1}^{n} \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right) \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right)' \right)^{-1} \sum_{t=1}^{n} \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right) y_t,$$

where $\overline{\Phi}_n = n^{-1} \sum_{t=1}^{n} \Phi_{t-1}$. The Wald test statistic for $H_0 : \theta = 0$ is defined as

$$T_n = \widehat{\theta}_n' \left(\sum_{t=1}^{[nr]} \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right) \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right)' \right) \widehat{\theta}_n / \widehat{\sigma}_{u,n}^2, \tag{5.15}$$

where $\widehat{\sigma}_{u,n}^2 = n^{-1} \sum_{t=1}^{n} \left(y_t - \widehat{\mu}_n - \Phi'(z_{t-1})\widehat{\theta}_n\right)^2$. Suppose that one rejects the null hypothesis if $T_n > \chi_{p,1-\alpha}^2$, where $\chi_{p,1-\alpha}^2$ is the $(1-\alpha)$ quantile of the χ_p^2 distribution. The following theorem describes the asymptotic size and power of the test.

5.3. Theorem. Under Assumptions 2.2, 2.4, 2.5, 5.1 and 5.2, T_n has an asymptotically noncentral χ_p^2 distribution with the noncentrality parameter given by

$$\left\| \left(\ell(1)\sigma_u^2 \int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1/2} \int_{-\infty}^{\infty} f(x)\Phi(x)dx \right\|^2.$$

Under the null hypothesis, $f(x) = 0$ for all $x \in R$. In this case, the test statistic T_n asymptotically has a χ_p^2 distribution regardless of the value of the long-run covariance ω_{zu}. In large samples, the test detects alternatives approaching the null at the rate slower than $n^{-1/4}$. Furthermore, Theorems 4.2 and 5.3 together imply that, in contrast to the test based on F_n, the test based on T_n does not tend to reject the null hypothesis of no predictive power, unless the forecasting model has a better out-of-sample fit than the baseline forecasting equation.

6 Simulation

The theoretical results of the previous sections suggest that nonlinear processes described by equation (2.1) may be confused with MDS's. In this case, usual linear regression methods will lead to spurious forecasts, while integrable approximants will provide a better out-of-sample fit. This section presents a series of Monte Carlo experiments motivated by these findings. First, I would like to illustrate similarities between a MDS and a process generated according to equation (2.1). In this and the next section, I use ϕ to denote a standard normal density function.

[Figure 1 about here.]

Figure 1 describes a typical sample path of a process generated according to (2.1) with $f(x) = 10\phi(x)$, independent standard normal errors $\{u_t\}$, and a random walk $\{z_t\}$ with standard normal increments independent of $\{u_t\}$. The random walk was initialized at zero. The top graph plots the errors $\{u_t\}$, the graph in the middle shows the sample path of the nonlinear component $\{f(z_t)\}$, and the graph at the bottom shows the sum of the two components. The figure shows that the signal generated by the nonlinear part is strong relative to the noise only during the first 15 periods. After that, the sum of the nonlinear component and the noise cannot be distinguished from a MDS.

The results in Section 3 suggest that a test based on F_n will tend to indicate predictive power, if equation (2.1) is a good approximation of the true DGP. It is important to see rejection rates in finite samples. I simulate the data according to the following equations:

$$\begin{aligned} y_t &= af(z_{t-1}) + u_t, \\ (u_t, \Delta z_t)' &\sim \text{iid N}(0, I_2), \\ z_0 &= 0, \end{aligned}$$

where the constant a allows one to vary the strength of the signal coming from the nonlinear part. I consider the following alternative functions for f :

$$\begin{aligned} f_1(x) &= 1\{0 < x < 1\}, \\ f_2(x) &= (1 - 0.5x)1\{0 < x < 1\}, \\ f_3(x) &= x^2 1\left\{0 < x < 3^{1/3}\right\}, \\ f_4(x) &= 2\phi(x - 0.25) - \phi(x - 0.75), \text{ and} \\ f_5(x) &= 2\phi(x + 2) - \phi(x - 1). \end{aligned}$$

The above functions have different shapes, however, they all have the same Lebesgue measure: $\int_{-\infty}^{\infty} f_i(x)dx = 1$ for all $i = 1, 2, ..., 5$. Thus, according to the results in Section 3, all five functions should provide similar rejection rates. This is due to the fact that the asymptotic distribution of F_n depends on the integral of f and not on the shape of the function. I construct F_n using the LS estimates of $(\theta_1, ..., \theta_p)'$ in the forecasting model below:

$$\widehat{y}_t(\mu, \theta) = \mu + \sum_{i=1}^{p} \theta_i z_{t-1}^i.$$

[Table 1 about here.]

Table 1 reports the simulated rejection rates for the sample size of 100 observations. The number of simulations is 1,000, the nominal size is set to 5%, $a \in \{1, 2, 4, 8\}$, and $p \in \{1, 2, 3, 4\}$. As one can see from Table 1, the actual rejection rates are higher than the nominal 5%. For example, in the case of a usual predictive regression ($p = 1$), the rejection rates are around 20% for $a = 2$, and they exceed 50% for most of the models when $a = 8$. Thus, the econometrician will tend to conclude that the polynomial forecasting function has predictive power.

Simulations confirm that the shape of f does not have an effect on the rejection rates. The power appears to be the same for all functions except f_5. The difference between the results corresponding to the first four functions and f_5 follows from a difference in the location of the functions. The limiting behavior of sums of the form $\sum_{i=1}^{n} f(z_t)$ differs from that of $\sum_{i=1}^{n} f(z_t + \sqrt{n}c)$. In the first case, it converges to

the Brownian local time at zero, while in the second case, it converges to the local time at c. The first four functions are concentrated on $(0,1)$ interval, while f_5 puts relatively more mass on the values outside $(0,1)$ interval. The shift in the location of f_5 results in higher rejection rates.

Next, I compare the out-of-sample performance of various forecasting models. I consider constant, polynomial, integrable linear (in parameters) and ERP forecasting equations:

$$\widehat{y}_t(\mu,0) = \mu, \tag{6.16}$$

$$\widehat{y}_t(\mu,\theta) = \mu + \sum_{i=1}^{p} \theta_i z_{t-1}^i, \tag{6.17}$$

$$\widehat{y}_t(\mu,\theta) = \mu + \phi(z_{t-1}) \sum_{i=1}^{p} \theta_i z_{t-1}^i, \tag{6.18}$$

$$\widehat{y}_t(\mu,\theta) = \mu + \phi(z_{t-1}) \frac{\theta_1 + \theta_2 z_{t-1} + ... + \theta_{p+1} z_{t-1}^p}{1 + \theta_{p+2} z_{t-1} + ... + \theta_{2p+1} z_{t-1}^p}. \tag{6.19}$$

Since, $f_1, ..., f_4$ give similar results, only f_1 and f_5 are used for the next set of simulations. First, I simulate 200 observations and use observations $\{1, ..., 100\}$ to estimate the parameters in (6.16)-(6.19). The parameters are estimated by LS (nonlinear LS in the case of equation (6.19)). In the second step, I construct one period ahead forecasts for observations $\{101, ..., 200\}$ using the estimated versions of (6.16)-(6.19). Finally, using predicted values of y_t, I compute the out-of-sample MSE's for four forecasting equations.

[Table 2 about here.]

Table 2 reports the proportion of cases in which forecasting functions (6.17)-(6.19) have smaller MSE's than that of the historic average (model (6.16)). The number of simulations is 1,000, $a \in \{1, 10\}$, and $p \in \{1, 2, 3, 4\}$. Consider the case $p = 1$. The numbers corresponding to the polynomial forecasting model show that the historic average provides a better out-of-sample fit than predictive regression in approximately 63%-70% of the repetitions. Increasing the value of a from 1 to 10 leads only to a marginal improvement in the performance of the polynomials. Thus, by ignoring the information contained in the predictor, one can obtain better forecasts despite

the fact that F_n indicates predictive power. Performance of the integrable forecasting functions depends on the strength of the signal coming from the nonlinear component. In the case of $a = 1$, the historic average provides a better out-of-sample fit in 51%-56% of the repetitions. However, in the case of $a = 10$, the integrable forecasting models perform better than the baseline model in 63%-90% of the repetitions. Thus, the result of Theorem 4.2(c) holds in finite samples, provided that the signal-to-noise ratio is large enough.

Table 2 shows that the performance of (6.17)-(6.19) deteriorates as p increases. This can be explained by the slow rates of convergence in the case of integrable transformations of I(1) processes. Evidently, larger sample sizes are required when $p > 1$ in order to obtain better approximations.

Finally, I evaluate finite sample size and power properties of the test proposed in Section 5. I consider the DGP's f_1 and f_5, and values of $a \in \{0, 0.5, 1, 2\}$. The test statistic is constructed using the estimates of θ in (6.18). I set the number of observations equal to 100, and the number of simulations to 5,000.

[Table 3 about here.]

Table 3 reports the results for the nominal size 5%. First, consider $a = 0$, which corresponds to the case of no predictive power. As one can see from the table, the actual rejection rates are close to the nominal, especially when $p = 1$. I conclude that, under the null, the χ^2_p distribution provides a reasonable approximation to the actual distribution of T_n in finite samples. Next, the results for $a > 0$ show that the test has non-trivial power. The test attains 50%-80% rejection rates for $a = 2$.

7 Empirical Example

The dividend-price ratio (dividend yield) has received much attention in the literature as a potential predictor for stock returns. In a recent study Lewellen (2005) considered the regression of stock returns on the natural log of the dividend yield and reported strong predictive power. Goyal and Welch (2003) approached the same problem from a different perspective. They focused on out-of-sample fit and arrived at an

opposite conclusion. This section evaluates the predictive power of the natural log of the dividend-price ratio (LDP) in view of the theoretical findings of the previous sections. I consider the same data as Goyal and Welch (2003): monthly observations, for the period 1946-2000, of value-weighted NYSE stock returns.

[Figure 2 about here.]

Figure 2 plots stock returns and LDP. While the stock returns show fluctuations around a constant level, the behavior of the LDP resembles a stochastic trend. Next, I proceed to formal unit root tests.

[Table 4 about here.]

Table 4 shows the results of unit root tests for LDP. I consider two alternative autoregressive specifications for LDP: with and without a linear deterministic trend. The first line of the table shows that the estimated autoregressive coefficient is very close to unity in both cases. Furthermore, Phillips-Perron Z_t is unable to reject the null hypothesis of a unit root for either specification. Finally, the strongest evidence in support of the I(1) hypothesis for LDP comes from KPSS tests (see Kwiatkowski, Phillips, Schmidt, and Shin (1992)). The KPSS test assumes stationarity under the null hypothesis. Rejection of the null suggests that there exists strong evidence in favor of the nonstationary alternative. As one can see from Table 4, this is the case with LDP. At 1% significance level, the null hypothesis is rejected for both models, with or without the deterministic trend. I conclude that a unit root model is a reasonable approximation for LDP.

Next, I look at in-sample predictability. I consider three alternative testing procedures. The first procedure is based on the usual OLS regression of stock returns on the lagged value of LDP. The second test is based on the fully modified OLS (FM-OLS) estimator of the regression slope, which is corrected for endogeneity of errors. While the OLS based t-test is invalid if errors and the predictor are correlated, the FM-OLS t-statistic has a mixed normal distribution regardless of correlations between errors and the regressor (see Phillips and Hansen (1990)). Finally, I consider the statistic proposed in Section 5, equation (5.15). For that purpose, I use $\Phi(x) = x\phi(x)$.

[Table 5 about here.]

Table 5 reports the results of the tests. The OLS based t-statistic is large; however, it is not significant at 5% significance level. The statistic based on FM-OLS estimates and the test statistic introduced in Section 5 are both significant. Hence, in-sample evidence indicates possible predictive power of LDP.

Lastly, I compare the out-of-sample performance of the four predictive models given in equations (6.16)-(6.19). I set $p = 1$ in (6.17)-(6.19). Equation (6.16) corresponds to the assumption that stock returns cannot be predicted from historic values of LDP, and provides constant forecasts. Equation (6.17) is the usual predictive regression model. Equation (6.18) corresponds to the integrable forecasting function linear in the parameters. Finally, equation (6.19) describes the ERP forecasting model. For the purpose of this exercise, I select $r \in (0,1)$ and divide the sample into two parts: observations $\{1, ..., [nr]\}$ and observations $\{[nr]+1, ..., n\}$. In the first step, I use observations $\{1, ..., [nr]\}$ to estimate the unknown parameters in (6.16)-(6.19). The parameters are estimated by LS. In the case of equation (6.16), the historic average is used to estimate μ. In the case of ERP, parameters are estimated by nonlinear LS. I used zeros as starting values for $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$ during numerical optimization. This choice follows from the fact that $\theta_1 = \theta_2 = 0$ in (6.19) implies that LDP has no predictive power. In the second step, estimated versions of (6.16)-(6.19) and the actual values of LDP in the forecasting sample are used to construct one period ahead forecasts for observations $\{[nr]+1, ..., n\}$.

[Table 6 about here.]

Table 6 reports out-of-sample square-root MSE's. I consider $r \in \{1/4, 1/2, 3/4\}$. It appears that the linear regression is the worst performer, since it is dominated by the historic average and integrable functions at all forecasting horizons considered. The MSE's of integrable functions is comparable to that of the historic mean. In fact, the ERP forecasting function is the best performer in the given sample.

I offer the following interpretation of the results reported in Table 6. It is reasonable to assume that LDP contains an autoregressive unit root. Consequently, LDP

cannot be a good predictor for stock returns during the periods which exhibit apparent trending behavior. Integrable transformations of LDP improve out-of-sample fit because they filter out large values of LDP. This allows one to ignore LDP during the trending periods and extract useful information during the other times. This interpretation is consistent with the idea that stock returns are predictable only on rare occasions, when the predictor does not show clear patterns immediately observable by all market participants.

It is important to emphasize that above comparison of the MSE's does not prove the out-of-sample predictability. For that purpose, one has to test whether the difference between the out-of-sample MSE's of nonlinear models and the historic average is statistically significant. It is possible to derive the asymptotic distribution of the difference of MSE's for the DGP described in Assumption 5.1, which can be used for developing a testing procedure based on a out-of-sample criterion similar to the tests proposed by Diebold and Mariano (1995). Such a procedure will allow one to compare MSE's of various forecasting functions. However, since an econometrician has a freedom to choose the nonlinear function g and the value of p, there is a danger of the data snooping bias. The testing procedure has to take into account the search for the best forecasting model. This problem can be approached along the lines described by White (2000). Further, note that testing with ERP's is complicated by the fact that θ_3 is not identified when $f = 0$. Development of such tests is a part of the ongoing research project.

8 Conclusion

In this paper, I consider the forecasting time series that contain a nonstationary, nonlinear component. The nonlinear component is modeled as an integrable transformation of the predictor, which is assumed to be a I(1) variable. I assume that the true form of the nonlinear component is unknown to the econometrician and that he is forced to use some approximating functions. I show that standard tools such as t-type tests and linear regressions lead to spurious forecasts. The diagnostic tests

tend to indicate predictive ability, while the forecasts based on the usual linear regression perform worse in terms of the MSE than constant forecasts, which ignore the information contained in the predictor. I derive general approximating results which allow one to improve the forecast accuracy with properly chosen forecasting functions. I show that one can obtain non-trivial improvements in forecast accuracy over polynomials and historic averages by using square integrable forecasting functions.

The results of this paper are supported by a Monte Carlo study and an empirical example. In the empirical application, I consider forecasting the NYSE stock returns using dividend-price ratio. I show that some integrable transformations of the dividend-price ratio provide a better out-of-sample fit than the forecasts constructed from the typical linear models. The accuracy of the forecasts is improved because nonlinear transformations filter out irrelevant information.

In conclusion, I would like to emphasize the importance of nonstationarity in the current context. The paper shows that in the case of a nonstationary predictor z_t, the loss function converges to the L_2-distance between the true function f and the approximating function g (multiplied by the local time process):

$$\int_{-\infty}^{\infty} \left(f(x) - g\left(x, \theta\right)\right)^2 dx.$$

In contrast, in the case of a strictly stationary and ergodic predictor z_t, the loss function converges to the L_2-distance weighted by the density of the predictor:

$$\int_{-\infty}^{\infty} \left(f(x) - g\left(x, \theta\right)\right)^2 pdf_z(x) dx.$$

In the second case, polynomials can provide a good approximation since $pdf_z(x) \to 0$ as $x \to \pm\infty$, while in the nonstationary case, approximation with polynomials is impossible due to unweighted integration over the entire real line.

This paper studies implications of nonlinearity and nonstationarity in the forecasting framework. These results can be extended to nonparametric estimation of the nonlinear component, which I postpone for further study.

9 Appendix

The following three lemmas are used in the proofs of the results in this paper. Lemma 9.1 extends some of the results of Lemma 2.6 to the weak convergence of stochastic processes indexed by the parameters $\theta \in \Theta$. Lemma 9.2 uses the Cramer-Wold device to extend some of the results of Lemma 2.6 to random vectors. Finally, Lemma 9.3 considers the case of extremum estimators with multiple optimal points. Below, the symbol "$\Rightarrow$" denotes weak convergence.

9.1. Lemma. Let $\varphi : R \times \Theta \to R$ be a differentiable function with respect to θ such that $\varphi(\cdot, \theta) \in \mathcal{I}$ for all $\theta \in \Theta \subset R^p$. Furthermore, assume that $\sup_{\theta \in \Theta} \left\| \frac{\partial}{\partial \theta} \varphi(\cdot, \theta) \right\|$ is in $\mathcal{I}$. Fix s and r such that $0 < s < r < 1$. Then, under Assumptions 2.4 and 2.5

(a) $n^{-1/2} \sum_{t=[ns]}^{[nr]} \varphi(z_{t-1}, \theta) \Longrightarrow (\ell(r) - \ell(s)) \int_{-\infty}^{\infty} \varphi(x, \theta)\, dx$, a stochastic process indexed by θ.

(b) $\sup_{\theta \in \Theta} \left| n^{-1/2} \sum_{t=[ns]}^{[nr]} \varphi(z_{t-1}, \theta)\, u_t \right| \to_p 0$.

Proof of Lemma 9.1. I prove part (a) of the lemma. The convergence of finite dimensional distributions follows from Lemma 2.6(d). Thus, it suffices to show that $n^{-1/2} \sum_{t=[ns]}^{[nr]} \varphi(z_{t-1}, \theta)$ is stochastically equicontinuous. The result will follow from Pollard (1990, Theorem (10.2)). Applying the mean value theorem, one obtains the expansion

$$\varphi(x, \theta_1) - \varphi(x, \theta_2) = (\theta_1 - \theta_2)' \frac{\partial}{\partial \theta} \varphi(x, \theta(x)),$$

where $\theta(x)$ lies between θ_1 and θ_2 and depends on the value of x. Next, write

$$\begin{aligned} & \left| n^{-1/2} \sum_{t=[ns]}^{[nr]} \varphi(z_{t-1}, \theta_1) - n^{-1/2} \sum_{t=[ns]}^{[nr]} \varphi(z_{t-1}, \theta_2) \right| \\ \leq\ & n^{-1/2} \sum_{t=[ns]}^{[nr]} \left\| \frac{\partial}{\partial \theta} \varphi(z_{t-1}, \theta(z_{t-1})) \right\| \|\theta_1 - \theta_2\| \\ \leq\ & n^{-1/2} \sum_{t=[ns]}^{[nr]} \sup_{\theta \in \Theta} \left\| \frac{\partial}{\partial \theta} \varphi(z_{t-1}, \theta) \right\| \|\theta_1 - \theta_2\|. \end{aligned}$$

By the assumption, $\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\theta}\varphi(\cdot,\theta)\right\|\in\mathcal{I}$. Consequently,

$$n^{-1/2}\sum_{t=[ns]}^{[nr]}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\theta}\varphi(z_{t-1},\theta)\right\|=O_p(1). \tag{9.20}$$

It follows from Lemma 2(a) of Andrews (1992) that $n^{-1/2}\sum_{t=[ns]}^{[nr]}\varphi(z_{t-1},\theta)$ is stochastically equicontinuous.

Next, I prove part (b). It is sufficient to show convergence for all $\theta\in\Theta$ and stochastic equicontinuity of $n^{-1/2}\sum_{t=[ns]}^{[nr]}\varphi(z_{t-1},\theta)u_t$. For a fixed value of θ, we have that $n^{-1/2}\sum_{t=1}^{[nr]}\varphi(z_{t-1},\theta)u_t=o_p(1)$ by Lemma 2.6(e). Now, write

$$\begin{aligned}&\left|n^{-1/2}\sum_{t=[ns]}^{[nr]}\varphi(z_{t-1},\theta_1)u_t-n^{-1/2}\sum_{t=[ns]}^{[nr]}\varphi(z_{t-1},\theta_2)u_t\right|\\ \leq\;& n^{-1/2}\sum_{t=[ns]}^{[nr]}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\theta}\varphi(z_{t-1},\theta)\right\||u_t|\,\|\theta_1-\theta_2\|. \end{aligned}\tag{9.21}$$

By the assumptions of the Lemma and by Lemma 2.6(e) we have that

$$n^{-1/2}\sum_{t=[ns]}^{[nr]}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\theta}\varphi(z_{t-1},\theta)\right\|(|u_t|-E(|u_t||\mathcal{F}_{t-1}))=o_p(1).$$

Further, by Assumption 2.5(b), there exists $K<\infty$ such that $\sup_t E(|u_t||\mathcal{F}_{t-1})<K$. Hence,

$$\begin{aligned}n^{-1/2}\sum_{t=[ns]}^{[nr]}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\theta}\varphi(z_{t-1},\theta)\right\|E(|u_t||\mathcal{F}_{t-1}) &\leq Kn^{-1/2}\sum_{t=[ns]}^{[nr]}\sup_{\theta\in\Theta}\left\|\frac{\partial}{\partial\theta}\varphi(z_{t-1},\theta)\right\|\\ &= O_p(1).\end{aligned}$$

The result of part (b) follows from Pollard (1990, Theorem (10.2)) and the continuous mapping theorem. □

9.2. Lemma. Let $\varphi_I:R\to R^p$ be a function that belongs to $\mathcal{I}$ element-by-element. Let $\varphi_0:R\to R^p$ be a function that belongs to $\mathcal{Z}$ element-by-element. Fix s and r such that $0<s<r<1$. Then, under Assumptions 2.4 and 2.5

(a) $n^{-1/4}\sum_{t=[ns]}^{[nr]}\varphi_I(z_{t-1})u_t$ converges in distribution to

$$\sigma_u\left((\ell(r)-\ell(s))\int_{-\infty}^{\infty}\varphi_I(x)\varphi(x)'\,dx\right)^{1/2}\int_s^r dW(t),$$

where W is a p-vector standard Brownian motion independent of ℓ and u.

(b) $n^{-1/4}\sum_{t=[ns]}^{[nr]}\varphi_0(z_{t-1})$ converges in distribution to

$$\left((2\pi)^{-1}(\ell(r)-\ell(s))\int_{-\infty}^{\infty}\widehat{\varphi}_0(t)\,\widehat{\varphi}_0^*(t)\,\frac{1+Ee^{it\varepsilon_1}}{1-Ee^{it\varepsilon_1}}dt\right)^{1/2}\int_s^r dV(t),$$

where $\widehat{\varphi}_0$ is the Fourier transform of φ_0, $\widehat{\varphi}_0^*$ is its complex conjugate, and V is another p-vector standard Brownian motion independent of ℓ, u, and W.

Convergence in (a) and (b) is joint.

Proof of Lemma 9.2. (a) Let c be a p-vector of constants. Lemma 2.6(e) implies that

$$\begin{aligned}
&n^{-1/4}\sum_{t=[ns]}^{[nr]}c'\varphi_I(z_{t-1})\,u_t\\
\to_d\ &\sigma_u\left((\ell(r)-\ell(s))\int_{-\infty}^{\infty}\left|c'\varphi_I(x)\right|^2dx\right)^{1/2}\int_s^r dw(t)\\
=^d\ &\sigma_u c'\left((\ell(r)-\ell(s))\int_{-\infty}^{\infty}\varphi_I(x)\,\varphi_I(x)'\,dx\right)^{1/2}\int_s^r dW(t),
\end{aligned}$$

where $=^d$ means equal in distribution, and $w(r)$ and $W(r)$ are scalar and p-vector standard Brownian motion respectively, both independent of ℓ and u. The result follows from the Cramer-Wold device.

(b) Again, let c be a p-vector of constants. It follows from Lemma 2.6(f) that

$$\begin{aligned}
&n^{-1/4}\sum_{t=[ns]}^{[nr]}c'\varphi_0(z_{t-1})\\
\to_d\ &\left((2\pi)^{-1}(\ell(r)-\ell(s))\int_{-\infty}^{\infty}\left|c'\widehat{\varphi}_0(t)\right|^2\frac{1+Ee^{it\varepsilon_1}}{1-Ee^{it\varepsilon_1}}dt\right)^{1/2}\int_s^r dv(t),\\
=^d\ &c'\left((2\pi)^{-1}(\ell(r)-\ell(s))\int_{-\infty}^{\infty}\widehat{\varphi}_0(t)\,\widehat{\varphi}_0^*(t)\,\frac{1+Ee^{it\varepsilon_1}}{1-Ee^{it\varepsilon_1}}dt\right)^{1/2}\int_s^r dV(t),
\end{aligned}$$

where v and V are a scalar and a p-vector standard Brownian motions respectively, both independent of ℓ, u, W. The result follows from the Cramer-Wold device. The joint convergence of (a) and (b) is implied by the joint convergence in Lemma 2.6.□

9.3. Lemma. Suppose $(Q_{1n}(\theta), Q_{2n}(\theta)) \Longrightarrow (Q_1(\theta), Q_2(\theta))$, some stochastic processes indexed by $\theta \in \Theta$, where Θ is a compact subset of R^p. Define

$$\Theta^* = \left\{ \theta^* \in \Theta : P\left(Q_1(\theta^*) = \inf_{\theta \in \Theta} Q_1(\theta) \right) = 1 \right\}.$$

Let Θ_n be the set of values of θ that minimize $Q_{1n}(\theta)$ on Θ. Then,

(a) $\sup_{\theta_n \in \Theta_n} d(\theta_n, \Theta^*) \to_p 0$ as $n \to \infty$.

(b) Suppose that $Q_{2n}(\theta)$ is stochastically equicontinuous on Θ. Suppose further that the following condition is satisfied for all $\varepsilon > 0$

$$P\left(\sup_{\theta_1^*, \theta_2^* \in \Theta^*} |Q_2(\theta_1^*) - Q_2(\theta_2^*)| \geq \varepsilon \right) = 0. \tag{9.22}$$

Then, as $n \to \infty$

$$\sup_{\theta_n \in \Theta_n} \sup_{\theta^* \in \Theta^*} |Q_{2n}(\theta_n) - Q_{2n}(\theta^*)| \to_p 0.$$

Proof of Lemma 9.3. (a) As $n \to \infty$,

$$\overline{\lim} P\left(\sup_{\theta_n \in \Theta_n} d(\theta_n, \Theta^*) \geq \delta \right) \leq \overline{\lim} P\left(\inf_{\theta \in \Theta : d(\theta, \Theta^*) \geq \delta} Q_{1n}(\theta) \leq \inf_{\theta^* \in \Theta^*} Q_{1n}(\theta) \right).$$

Define

$$h(Q) = 1\left\{ \inf_{\theta \in \Theta : d(\theta, \Theta^*) \geq \delta} Q(\theta) \leq \inf_{\theta^* \in \Theta^*} Q(\theta) \right\}.$$

The definition of Θ^* implies that $h(Q_1) = 0$ with probability one. Next, for all $\varepsilon > 0$,

$$\begin{aligned} \overline{\lim} P\left(\inf_{\theta \in \Theta : d(\theta, \Theta^*) \geq \delta} Q_{1n}(\theta) \leq \inf_{\theta^* \in \Theta^*} Q_{1n}(\theta) \right) &= \overline{\lim} P\left(h(Q_{1n}) \geq \varepsilon \right) \\ &\leq P\left(h(Q_1) \geq \varepsilon \right) \\ &= 0. \end{aligned} \tag{9.23}$$

The inequality in (9.23) follows from weak convergence and the continuous mapping theorem.

(b) As $n \to \infty$,

$$
\begin{aligned}
& \overline{\lim} P\left(\sup_{\theta_n\in\Theta_n}\sup_{\theta^*\in\Theta^*}\left|Q_{2n}\left(\theta_n\right)-Q_{2n}\left(\theta^*\right)\right|\geq\varepsilon\right) \\
\leq\ & \overline{\lim} P\left(\sup_{\theta_n\in\Theta_n}\sup_{\theta^*\in\Theta^*}\left|Q_{2n}\left(\theta_n\right)-Q_{2n}\left(\theta^*\right)\right|\geq\varepsilon,\ \sup_{\theta_n\in\Theta_n} d\left(\theta_n,\Theta^*\right)<\delta\right) \\
& +\overline{\lim} P\left(\sup_{\theta_n\in\Theta_n} d\left(\theta_n,\Theta^*\right)\geq\delta\right). \qquad (9.24)
\end{aligned}
$$

Next, the condition $\sup_{\theta_n\in\Theta_n} d\left(\theta_n,\Theta^*\right)<\delta$ implies that for all $\theta_n\in\Theta_n$ there exists $\theta_n^*\in\Theta^*$ such that $\|\theta_n^*-\theta_n\|<\delta$. The first summand on the right-hand side of (9.24) is bounded by

$$
\begin{aligned}
& \overline{\lim} P\left(\sup_{\theta_n\in\Theta_n}\sup_{\theta^*\in\Theta^*}\left|Q_{2n}\left(\theta_n\right)-Q_{2n}\left(\theta_n^*\right)\right|+\left|Q_{2n}\left(\theta_n^*\right)-Q_{2n}\left(\theta^*\right)\right|\geq\varepsilon\right) \\
\leq\ & \overline{\lim} P\left(\sup_{\theta^*\in\Theta^*}\sup_{\|\theta^*-\theta\|<\delta}\left|Q_{2n}\left(\theta\right)-Q_{2n}\left(\theta^*\right)\right|+\sup_{\theta_1^*,\theta_2^*\in\Theta^*}\left|Q_{2n}\left(\theta_1^*\right)-Q_{2n}\left(\theta_2^*\right)\right|\geq\varepsilon\right).
\end{aligned}
$$

Now, weak convergence, the continuous mapping theorem and (9.22) imply that

$$
\sup_{\theta_1^*,\theta_2^*\in\Theta^*}\left|Q_{2n}\left(\theta_1^*\right)-Q_{2n}\left(\theta_2^*\right)\right|\rightarrow_p 0.
$$

The second summand on the right-hand side of (9.24) is $o(1)$ due to part (a) of the lemma. The desired result follows from the stochastic equicontinuity of Q_{2n} and Slutsky's Lemma. □

Next, I present the proofs of the main results.

Proof of Theorem 3.2. I prove part (a) first. Write $\widehat{\theta}_n=A_{1n}-A_{2n}+A_{3n}$, where

$$
\begin{aligned}
A_{1n} &= \left(\sum_{t=1}^{[nr]}\underline{Z}_{t-1}\underline{Z}_{t-1}'\right)^{-1}\sum_{t=1}^{[nr]} f\left(z_{t-1}\right)Z_{t-1}, \\
A_{2n} &= \left(\sum_{t=1}^{[nr]} f\left(z_{t-1}\right)\right)\left(\sum_{t=1}^{[nr]}\underline{Z}_{t-1}\underline{Z}_{t-1}'\right)^{-1}[nr]^{-1}\sum_{t=1}^{[nr]} Z_{t-1}, \\
A_{3n} &= \left(\sum_{t=1}^{[nr]}\underline{Z}_{t-1}\underline{Z}_{t-1}'\right)^{-1}\sum_{t=1}^{[nr]} u_t\underline{Z}_{t-1}.
\end{aligned}
$$

It follows from Lemma 2.6(b) and the Cramer-Wold device that

$$
n^{-1}\sum_{t=1}^{[nr]} D_n^{-1/2}\underline{Z}_{t-1}\underline{Z}_{t-1}'D_n^{-1/2}\rightarrow_d\int_0^r\underline{B}(s)\underline{B}(s)'ds. \qquad (9.25)
$$

Assumption 3.1 and Lemma 2.6(d) imply that

$$n^{-1/2}\sum_{t=1}^{[nr]} f\left(z_{t-1}\right) Z_{t-1} = O_p(1). \tag{9.26}$$

Therefore, it follows from (9.25) and (9.26) that

$$n^{1/2} D_n^{1/2} A_{1n} = o_p(1). \tag{9.27}$$

Next, joint convergence in Lemma 2.6(b),(d) and the continuous mapping theorem imply that

$$n^{1/2} D_n^{1/2} A_{2n} \to_d r^{-1}\ell(r) \int_{-\infty}^{\infty} f(x)dx \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{-1} \int_0^r \underline{B}(s)ds. \tag{9.28}$$

Finally, Lemma 2.6(c) implies that

$$n^{1/2} D_n^{1/2} A_{3n} \to_d \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{-1} \int_0^r \underline{B}(s)d\underline{u}(s). \tag{9.29}$$

The result in part (a) follows from (9.27)-(9.29) and the joint convergence in Lemma 2.6.

The result in part (b) of the theorem follows immediately from the definition of $\widehat{\mu}_n$, Lemma 2.6 and part (a) of the theorem.

For part (c) of the theorem, it is sufficient to show that s_n^2 in (3.6) converges in probability to σ_u^2. The result will follow from part (a) and the continuous mapping theorem. Define the averages $\overline{f}_n = [nr]^{-1}\sum_{t=1}^{[nr]} f\left(z_{t-1}\right)$, and $\overline{u}_n = [nr]^{-1}\sum_{t=1}^{[nr]} u_t$. Write

$$\sum_{i=1}^{[nr]} \left(y_t - \widehat{\mu}_n - Z_{t-1}'\widehat{\theta}_n\right)^2 = \sum_{i=1}^{[nr]} \left(\left(f\left(z_{t-1}\right) - \overline{f}_n\right) + \left(u_t - \overline{u}_n\right) - \underline{Z}_{t-1}'\widehat{\theta}_n\right)^2. \tag{9.30}$$

We have the following results:

$$\begin{aligned}\sum_{i=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)^2 &= \sum_{i=1}^{[nr]} f^2\left(z_{t-1}\right) - \left([nr]^{-1/2}\sum_{i=1}^{[nr]} f\left(z_{t-1}\right)\right)^2 \\ &= O_p\left(n^{1/2}\right),\end{aligned} \tag{9.31}$$

$$\widehat{\theta}_n' \sum_{t=1}^{[nr]} \underline{Z}_{t-1} \underline{Z}_{t-1}' \widehat{\theta}_n = O_p(1), \tag{9.32}$$

$$\sum_{i=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)\left(u_t - \overline{u}_n\right) = O_p\left(n^{1/4}\right), \tag{9.33}$$

$$\sum_{i=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right) \underline{Z}_{t-1}' \widehat{\theta}_n = O_p\left(1\right), \tag{9.34}$$

$$\sum_{i=1}^{[nr]} \left(u_t - \overline{u}_n\right) \underline{Z}_{t-1}' \widehat{\theta}_n = O_p\left(1\right). \tag{9.35}$$

Equation (9.31) follows from Assumption 2.2 and Lemma 2.6(d); (9.32) is implied by part (a) of the Theorem and Lemma 2.6(b); (9.33) is due to Lemma 2.6(e); (9.34) follows from part (a) of the Theorem, Lemma 2.6(d) and Assumption 3.1; and, finally, (9.35) follows from 2.6(c) and part (a) of the theorem. Next, (9.30)-(9.35) together imply that

$$\sum_{i=1}^{[nr]} \left(y_t - \widehat{\mu}_n - Z_{t-1}' \widehat{\theta}_n\right)^2 = \sum_{i=1}^{[nr]} u_t^2 + \sum_{i=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)^2 + O_p\left(n^{1/4}\right).$$

The result follows from Assumption 2.5 and Slutsky's lemma.

The proof of part (d) is similar to the derivation of the probability limit of s_n^2 in (b). Using the same arguments as in (9.30)-(9.35), one can write

$$\begin{aligned} nQ_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right) &= \sum_{i=[nr]+1}^{n} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)^2 + O_p\left(n^{1/4}\right) \\ &= \sum_{i=[nr]+1}^{n} f^2\left(z_{t-1}\right) \\ &\quad + \frac{n-[nr]}{[nr]^2}\left(\sum_{i=1}^{[nr]} f\left(z_{t-1}\right)\right)^2 - \frac{1}{[nr]} \sum_{i=1}^{[nr]} f\left(z_{t-1}\right) \sum_{i=[nr]+1}^{n} f\left(z_{t-1}\right) \\ &\quad + O_p\left(n^{1/4}\right) \\ &= \sum_{i=[nr]+1}^{n} f^2\left(z_{t-1}\right) + O_p\left(n^{1/4}\right). \end{aligned} \tag{9.36}$$

The result follows from (9.36), Assumption 2.2 and Lemma 2.6(d).

I prove part (e) of the theorem now. Define $\overline{y}_n = [nr]^{-1}\sum_{t=1}^{[nr]} y_t$, and write

$$\begin{aligned}
& (n-[nr])\left(Q_n\left(\widehat{\mu}_n,\widehat{\theta}_n\right) - Q_n\left(\widehat{\mu}_{0,n},0\right)\right) \\
= & \sum_{i=[nr]+1}^{n}\left(y_t - \overline{y}_n - \underline{Z}'_{t-1}\widehat{\theta}_n\right)^2 - \sum_{i=[nr]+1}^{n}(y_t-\overline{y}_n)^2 \\
= & B_{1n} - 2B_{2n} - 2B_{3n},
\end{aligned}$$

where

$$\begin{aligned}
B_{1n} &= \widehat{\theta}'_n \sum_{i=[nr]+1}^{n} \underline{Z}_{t-1}\underline{Z}'_{t-1}\widehat{\theta}_n, \\
B_{2n} &= \sum_{i=[nr]+1}^{n}\left(f\left(z_{t-1}\right) - \overline{f}_n\right)\underline{Z}'_{t-1}\widehat{\theta}_n, \\
B_{3n} &= \sum_{i=[nr]+1}^{n}\left(u_t - \overline{u}_n\right)\underline{Z}'_{t-1}\widehat{\theta}_n.
\end{aligned}$$

Due to the result in part (a),

$$B_{1n} \to_d \Psi'\left(\int_r^1 \underline{B}(s)\underline{B}(s)'ds\right)\Psi, \tag{9.37}$$

$$\begin{aligned}
B_{2n} = & -[nr]^{-1}\sum_{i=1}^{[nr]} f\left(z_{t-1}\right)\sum_{i=[nr]+1}^{n}\underline{Z}'_{t-1}\widehat{\theta}_n - \sum_{i=[nr]+1}^{n} f\left(z_{t-1}\right)[nr]^{-1}\sum_{i=1}^{[nr]} Z'_{t-1}\widehat{\theta}_n \\
& + \sum_{i=[nr]+1}^{n} f\left(z_{t-1}\right)Z'_{t-1}\widehat{\theta}_n \\
= & -[nr]^{-1}\sum_{i=1}^{[nr]} f\left(z_{t-1}\right)\sum_{i=[nr]+1}^{n}\underline{Z}'_{t-1}\widehat{\theta}_n - \sum_{i=[nr]+1}^{n} f\left(z_{t-1}\right)[nr]^{-1}\sum_{i=1}^{[nr]} Z'_{t-1}\widehat{\theta}_n \\
& + O_p\left(D_n^{-1/2}\right) \\
\to_d & -r^{-1}\left(\int_{-\infty}^{\infty} f(x)dx\right)\times \\
& \left(\ell(r)\int_r^1 \underline{B}(s)d(s) + \int_1^r d\ell(s)\int_0^r B(s)ds\right)'\Psi,
\end{aligned} \tag{9.38}$$

and

$$B_{3n} \to_d \Psi'\int_r^1 \underline{B}(s)d\underline{u}(s). \tag{9.39}$$

The desired result follows from (9.37)-(9.39). □

Proof of Theorem 4.2. Define $\overline{g}_n(\theta) = [nr]^{-1} \sum_{t=1}^{[nr]} g(z_{t-1}, \theta)$. Concentrating out μ as in (4.12), write the in-sample MSE as

$$\begin{aligned} n^{1/2} MSE_n(\theta) &= n^{-1/2} \sum_{t=1}^{[nr]} (y_t - \overline{y}_n - g(z_{t-1}, \theta) + \overline{g}_n(\theta))^2 \\ &= n^{-1/2} \sum_{t=1}^{[nr]} (f(z_{t-1}) - g(z_{t-1}, \theta))^2 + n^{-1/2} \sum_{t=1}^{[nr]} (u_t - \overline{u}_n)^2 \\ &\quad + R_n(\theta), \end{aligned}$$

where $R_n(\theta) = R_{1n}(\theta) + 2R_{2n}(\theta) + 2R_{3n}(\theta) + 2R_{4n}(\theta)$, and

$$\begin{aligned} R_{1n}(\theta) &= n^{1/2} \left(\overline{f}_n - \overline{g}_n(\theta)\right)^2, \\ R_{2n}(\theta) &= \left(\overline{f}_n - \overline{g}_n(\theta)\right) n^{-1/2} \sum_{t=1}^{[nr]} (f(z_{t-1}) - g(z_{t-1}, \theta)), \\ R_{3n}(\theta) &= n^{-1/2} \sum_{t=1}^{[nr]} (f(z_{t-1}) - g(z_{t-1}, \theta))(u_t - \overline{u}_n), \\ R_{4n}(\theta) &= \left(\overline{f}_n - \overline{g}_n(\theta)\right) n^{-1/2} \sum_{t=1}^{[nr]} (u_t - \overline{u}_n). \end{aligned}$$

Each of the components of $R_n(\theta)$ is $o_p(1)$ uniformly in θ. For $R_{1n}(\theta)$, write $R_{1n}(\theta) = n^{-1/2} \left(n^{-1/2} \sum_{t=1}^{[nr]} (f(z_{t-1}) - g(z_{t-1}, \theta))\right)^2$. Then, Assumptions 2.2 and 4.1, Lemma 9.1(a) and continuous mapping theorem imply that $\sup_{\theta \in \Theta} |R_{1n}(\theta)| = o_p(1)$. In a similar way, one can show that $R_{2n}(\theta)$ and $R_{4n}(\theta)$ converge to zero in probability uniformly in θ. Finally, $\sup_{\theta \in \Theta} R_{3n}(\theta) = o_p(1)$ by Lemma 9.1(b).

Now, it follows from Assumption 4.1 and Lemma 9.1(a)

$$n^{1/2} MSE_n(\theta) - n^{-1/2} \sum_{t=1}^{[nr]} (u_t - \overline{u}_n)^2 \Longrightarrow \ell(r) \int_{-\infty}^{\infty} (f(x) - g(x, \theta))^2 \, dx. \tag{9.40}$$

Equation (9.40) and Lemma 9.3(a) together imply that

$$\sup_{\theta \in \Theta_n} d(\theta_n, \Theta^*) \to_p 0, \tag{9.41}$$

which completes the proof of part (a) of the Theorem.

For part (b), write

$$\sup_{\theta_n\in\Theta_n}\left|\widehat{\mu}_n\left(\widehat{\theta}_n\right)-\mu^*\right| \leq \left|[nr]^{-1}\sum_{t=1}^{[nr]}\left(f\left(z_{t-1}\right)-g\left(z_{t-1},\theta^*\right)\right)\right|+\left|[nr]^{-1}\sum_{t=1}^{[nr]}u_t\right| + \sup_{\theta_n\in\Theta_n}\left|R_{5n}\left(\theta^*,\widehat{\theta}_n\right)\right|, \tag{9.42}$$

where $\theta^*\in\Theta^*$, and

$$R_{5n}\left(\theta^*,\widehat{\theta}_n\right)=[nr]^{-1}\sum_{t=1}^{[nr]}\left(g\left(z_{t-1},\theta^*\right)-g\left(z_{t-1},\widehat{\theta}_n\right)\right).$$

The first two summands on the right-hand side of (9.42) are $o_p(1)$, as it follows from Lemma 2.6(a) and (d). Next, due to the Assumptions of the theorem, and by Lemma 9.1(a)

$$n^{1/2}R_{5n}\left(\theta^*,\theta\right)\Rightarrow r^{-1}\ell(r)\int_{-\infty}^{\infty}\left(g\left(x,\theta^*\right)-g\left(x,\theta\right)\right)dx, \tag{9.43}$$

where $\sup_{\theta\in\Theta}\left|\int_{-\infty}^{\infty}\left(g\left(x,\theta^*\right)-g\left(x,\theta\right)\right)dx\right|<\infty$ by Assumption 4.1(c). Hence, the remainder term in (9.42), $R_{5n}\left(\theta^*,\theta_n\right)$, is $o_p(1)$ uniformly in θ_n. The result of part (b) of the Theorem follows.

For part (c) of the theorem, note that $Q_n\left(\widehat{\mu}_n\left(\widehat{\theta}_n\right),\widehat{\theta}_n\right)$ depends on the out-of-sample MSE. Hence, similarly to part (a), one can show that

$$n^{1/2}Q_n\left(\widehat{\mu}_n\left(\theta\right),\theta\right)\Rightarrow\left(1-r\right)^{-1}\left(\ell(1)-\ell(r)\right)\int_{-\infty}^{\infty}\left(f(x)-g\left(x,\theta\right)\right)^2dx. \tag{9.44}$$

Next, fix $\theta^*\in\Theta^*$. Define $R_{6,n}\left(\theta_n,\theta^*\right)=Q_n\left(\widehat{\mu}_n\left(\theta_n\right),\theta_n\right)-Q_n\left(\widehat{\mu}_n\left(\theta^*\right),\theta^*\right)$. It follows from equation (9.44), the definition of Θ^* and Lemma 9.3(b) that

$$\sup_{\theta_n\in\Theta_n,\theta^*\in\Theta^*}\left|n^{1/2}R_{6,n}\left(\widehat{\theta}_n,\theta^*\right)\right|\rightarrow_p 0.$$

Hence,

$$\begin{aligned} n^{1/2}Q_n\left(\widehat{\mu}_n\left(\widehat{\theta}_n\right),\widehat{\theta}_n\right) &= n^{1/2}Q_n\left(\widehat{\mu}_n\left(\theta^*\right),\theta^*\right)+o_p(1)\\ &\rightarrow_d\left(1-r\right)^{-1}\left(\ell(1)-\ell(r)\right)M^*, \end{aligned}$$

where the last result holds uniformly in $\widehat{\theta}_n$. □

Proof of Theorem 4.5. For part (a) write

$$\begin{aligned} n^{1/2}\left(\widehat{\mu}_n - \mu^*\right) &= n^{1/2}[nr]^{-1}\sum_{t=1}^{[nr]}\left(f\left(z_{t-1}\right) - g\left(z_{t-1},\theta^*\right)\right) + n^{1/2}[nr]^{-1}\sum_{t=1}^{[nr]} u_t \\ &\quad + n^{1/2}R_{5n}\left(\theta^*, \widehat{\theta}_n\right). \end{aligned}$$

Theorem 4.2(a) and (9.43) together imply that $n^{1/2}R_{5n}\left(\theta^*, \widehat{\theta}_n\right) = o_p(1)$. The result of part (a) follows from 2.6(a) and (d).

For part (b), one can write

$$\begin{aligned} \frac{1}{2}n^{1/2}\frac{\partial}{\partial\theta}MSE_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right) &= n^{-1/2}\sum_{t=1}^{[nr]} S\left(z_{t-1}, \widehat{\theta}_n\right) + n^{-1/2}\sum_{t=1}^{[nr]}\frac{\partial}{\partial\theta}g\left(z_{t-1}, \widehat{\theta}_n\right) u_t \\ &\quad - \left(\widehat{\mu}_n - \mu^*\right) n^{-1/2}\sum_{t=1}^{[nr]}\frac{\partial}{\partial\theta}g\left(z_{t-1}, \widehat{\theta}_n\right). \end{aligned} \tag{9.45}$$

It follows from part (a) of the theorem, Lemma 9.1(a) and Theorem 4.2 that the last summand on the right-hand side of equation (9.45) is $O_p(n^{-1/2})$. Now, using the mean value expansion of $n^{-1/2}\sum_{t=1}^{[nr]} S\left(z_{t-1}, \widehat{\theta}_n\right)$ and $n^{-1/2}\sum_{t=1}^{[nr]}\frac{\partial}{\partial\theta}g\left(z_{t-1}, \widehat{\theta}_n\right) u_t$ around θ^*, one obtains

$$\begin{aligned} & n^{1/4}\left(\widehat{\theta}_n - \theta^*\right) \\ = & \left(n^{-1/2}\sum_{t=1}^{[nr]} H\left(z_{t-1}, \widetilde{\theta}_n\right) + o_p(1)\right)^{-1} \\ & \times\left(n^{-1/4}\sum_{t=1}^{[nr]} S\left(z_{t-1},\theta^*\right) + 2n^{-1/4}\sum_{t=1}^{[nr]}\frac{\partial}{\partial\theta}g\left(z_{t-1},\theta^*\right) u_t\right) + o_p(1). \end{aligned}$$

In the above expression $\widetilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ^* and, therefore, converges in probability to θ^*. Next, under Assumptions 2.2 and 4.3(c)-(d), and by applying Lemma 9.1(a) element-by-element, one obtains

$$n^{-1/2}\sum_{t=1}^{[nr]} H\left(z_{t-1}, \widetilde{\theta}_n\right) \to_d \ell(r)\int_{-\infty}^{\infty} H\left(x,\theta^*\right) dx. \tag{9.46}$$

Let W_1 and W_2 be two p-vectors standard Brownian motions independent of ℓ, u and

each other. Under Assumption 4.4, it follows from Lemma 9.2(b) that

$$n^{-1/4}\sum_{t=1}^{[nr]} S\left(z_{t-1},\theta^*\right)$$
$$\rightarrow_d \left((2\pi)^{-1}\ell(r)\int_{-\infty}^{\infty}\widehat{S}\left(t,\theta^*\right)\widehat{S}^*\left(t,\theta^*\right)\frac{1+Ee^{it\varepsilon_1}}{1-Ee^{it\varepsilon_1}}dt\right)^{1/2} W_1(r). \quad (9.47)$$

Under Assumption 4.1(d), Lemma 9.2(a) implies that

$$n^{-1/4}\sum_{t=1}^{[nr]}\frac{\partial}{\partial\theta}g\left(z_{t-1},\theta^*\right)u_t$$
$$\rightarrow_d \sigma_u\left(\ell(r)\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}g\left(x,\theta^*\right)\frac{\partial}{\partial\theta}g\left(x,\theta^*\right)'dx\right)^{1/2} W_2(r). \quad (9.48)$$

Part (b) of the theorem follows from (9.46), (9.47) and (9.48). □

Proof of Theorem 5.3. Under the alternative, the re-scaled slope coefficient $n^{1/4}\widehat{\theta}_n$ is given by

$$\left(n^{-1/2}\sum_{t=1}^{n}\left(\Phi\left(z_{t-1}\right)-\overline{\Phi}_n\right)\left(\Phi\left(z_{t-1}\right)-\overline{\Phi}_n\right)'\right)^{-1}\times$$
$$\left(n^{-1/2}\sum_{t=1}^{n}f\left(z_{t-1}\right)\left(\Phi\left(z_{t-1}\right)-\overline{\Phi}_n\right)+n^{-1/4}\sum_{t=1}^{n}u_t\left(\Phi\left(z_{t-1}\right)-\overline{\Phi}_n\right)\right).$$

Similarly to (9.31), the average $\overline{\Phi}_n$ has no effect on the asymptotics of the above expression. The Cramer-Wold device, Lemma 2.6(d) and Assumption 5.2 imply that

$$n^{-1/2}\sum_{t=1}^{n}\Phi\left(z_{t-1}\right)\Phi\left(z_{t-1}\right)' \rightarrow_d \ell(1)\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx, \quad (9.49)$$
$$n^{-1/2}\sum_{t=1}^{n}f\left(z_{t-1}\right)\Phi\left(z_{t-1}\right) \rightarrow_d \ell(1)\int_{-\infty}^{\infty}f(x)\Phi(x)dx. \quad (9.50)$$

Next, Assumption 5.2 and Lemma 9.2(b) imply that

$$n^{-1/4}\sum_{t=1}^{n}u_t\Phi\left(z_{t-1}\right)\rightarrow_d\left(\sigma_u^2\ell(1)\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx\right)^{1/2}W(1), \quad (9.51)$$

where $W(r)$ is a Brownian motion independent of $\ell(r)$. In fact, convergence in (9.49)-(9.51) is joint. Hence, $n^{1/4}\widehat{\theta}_n$ converges in distribution to

$$\left(\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx\right)^{-1}\int_{-\infty}^{\infty}f(x)\Phi(x)dx$$
$$+\sigma_u\left(\ell(1)\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx\right)^{-1/2}W(1). \quad (9.52)$$

Finally, $\widehat{\sigma}^2_{u,n} \to_p \sigma^2_u$, which can be shown similarly to $s^2_n \to_p \sigma^2_u$ in the proof of Theorem 3.2(c). Therefore, it follows from (9.52) that

$$T_n \to_d \left\| \left(\ell(1) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1/2} \int_{-\infty}^{\infty} \frac{f(x)\Phi(x)}{\sigma_u} dx + W(1) \right\|^2 . \square$$

References

Andrews, D. W. K. (1992): "Generic Uniform Convergence," *Econometric Theory*, 8, 241–257.

Campbell, J. Y., and S. B. Thompson (2004): "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?," mimeo.

Chang, Y., and J. Y. Park (2003): "Index models with integrated time series," *Journal of Econometrics*, 114, 73–106.

Chang, Y., J. Y. Park, and P. C. Phillips (2001): "Nonlinear Econometric Models with Cointegrated and Deterministically Trending Regressors," *Econometrics Journal*, 4, 1–36.

Chung, K., and R. Williams (1990): *Introduction to Stochastic Integration*. Birkhäuser, Boston, second edn.

Cochrane, J. H. (1997): "Where is the Market Going? Uncertain Facts and Novel Theories," *Federal Reserve Bank of Chicago - Economic Perspectives*, 21, 3–37.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economics Statistics*, 13, 253–263.

Fama, E. F. (1991): "Efficient Capital Markets: II," *Journal of Finance*, 46, 1575–1617.

Goyal, A., and I. Welch (2003): "A Note on Predicting Returns with Financial Ratios," mimeo.

——— (2004): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," NBER Working Paper 10483.

Granger, C. W. J., and T. Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press, New York.

Hu, L., and P. C. B. Phillips (2004): "Nonstationary Discrete Choice," *Journal of Econometrics*, pp. 103–138.

Jeganathan, P. (2003): "Second Order Limits of Functionals of Sums of Linear Processes that Converge to Fractional Stable Motions," .

Kasparis, I. (2004): "Functional Form Misspecification in Regression with Integrated Time Series," Ph.D. thesis, Department of Economics, University of Southampton.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root," *Journal of Econometrics*, 54, 159–178.

Lewellen, J. (2005): "Predicting Returns with Financial Ratios," *Journal of Financial Economics (forthcoming)*.

Park, J. Y., and P. C. B. Phillips (1999): "Asymptotics for Nonlinear Transformations of Integrated Time Series," *Econometric Theory*, 15, 269–298.

Phillips, P. C., and B. E. Hansen (1990): "Statistical Inference in Instrumental Variables Regression with I(1) Processes," *Review of Economic Studies*, 57, 99–125.

Phillips, P. C. B. (1983): "Best Uniform and Modified Pade Approximants to Probability Densities in Econometrics," in *Advances in Econometrics*, ed. by W. Hildenbrand, pp. 123–167. Cambridge University Press.

Pollard, D. (1990): *Empirical Processes: Theory and Applications*, vol. 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, California.

White, H. (2000): "A Reality Check For Data Snooping," *Econometrica*, 68, 1097–1126.

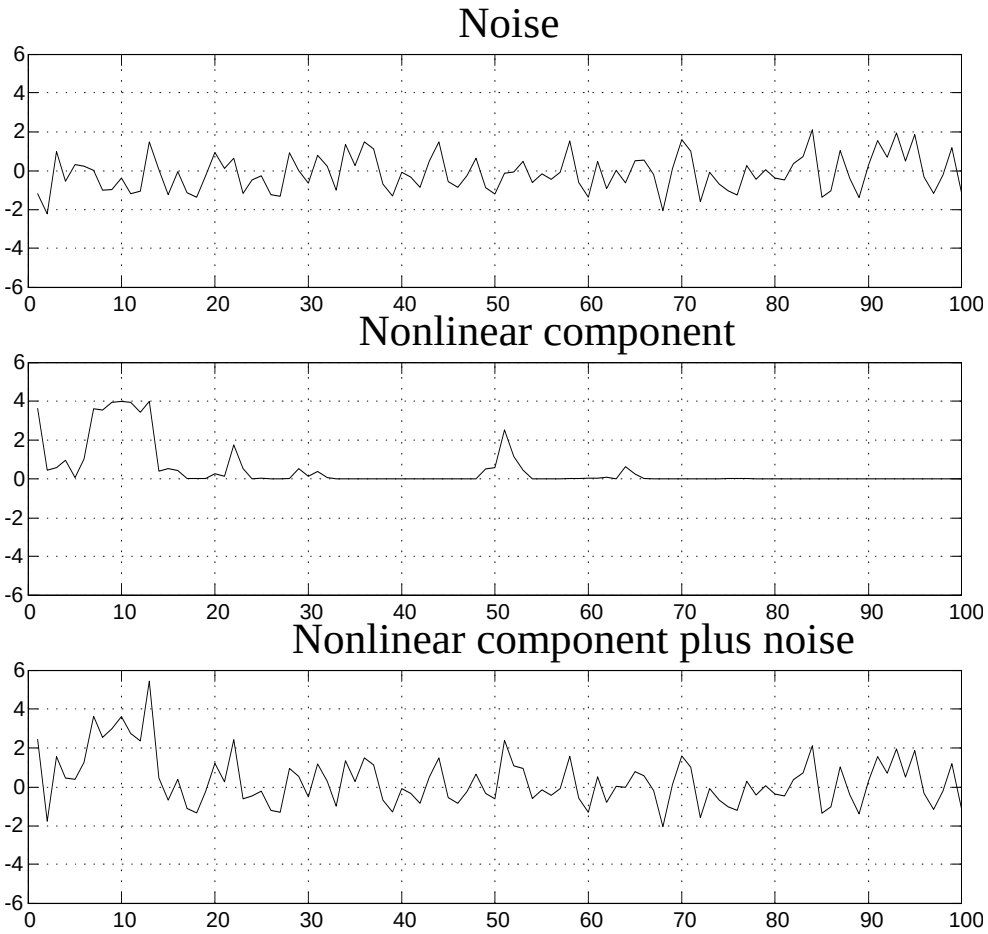

Figure 1: Simulated sample path of $y_t = f(z_{t-1}) + u_t$.

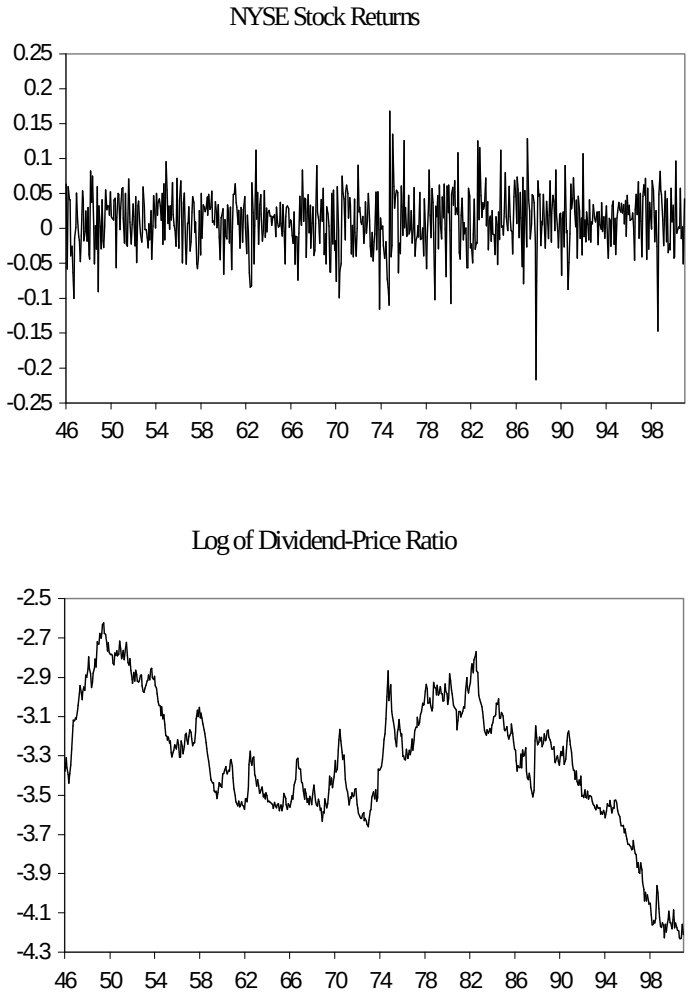

Figure 2: NYSE stock returns and dividend-price ratio, 1946-2000.

p	f_1	f_2	f_3	f_4	f_5
			$a = 1$		
1	0.115	0.113	0.112	0.103	0.184
2	0.126	0.108	0.121	0.121	0.251
3	0.142	0.136	0.136	0.133	0.315
4	0.176	0.165	0.157	0.136	0.362
			$a = 2$		
1	0.233	0.205	0.247	0.264	0.373
2	0.295	0.256	0.314	0.299	0.534
3	0.340	0.303	0.353	0.355	0.650
4	0.419	0.376	0.433	0.398	0.724
			$a = 4$		
1	0.411	0.355	0.459	0.556	0.538
2	0.536	0.457	0.603	0.782	0.693
3	0.594	0.556	0.665	0.878	0.761
4	0.691	0.629	0.748	0.906	0.789
			$a = 8$		
1	0.538	0.464	0.600	0.654	0.724
2	0.707	0.597	0.795	0.908	0.897
3	0.758	0.695	0.825	0.967	0.930
4	0.827	0.761	0.885	0.981	0.957

Table 1: Simulated rejection rates of F_n test for 0.05 significance level.

		integrable			integrable	
p	polynomial	linear	ERP	polynomial	linear	ERP
	f_1 and $a = 1$			f_5 and $a = 1$		
1	0.299	0.481	0.484	0.307	0.438	0.487
2	0.205	0.430	0.445	0.192	0.397	0.405
3	0.107	0.417	0.394	0.106	0.352	0.336
4	0.089	0.414	0.371	0.093	0.346	0.320
	f_1 and $a = 10$			f_5 and $a = 10$		
1	0.342	0.631	0.776	0.366	0.612	0.897
2	0.257	0.654	0.784	0.275	0.723	0.857
3	0.175	0.597	0.780	0.198	0.639	0.846
4	0.154	0.583	0.793	0.203	0.720	0.842

Table 2: Proportion of simulation repetitions where the out-of-sample MSE of the corresponding model is less than that of the historic average.

	$p=1$	$p=2$	$p=3$	$p=4$
		f_1		
$a=0$	0.056	0.065	0.062	0.075
$a=0.5$	0.108	0.100	0.107	0.136
$a=1$	0.226	0.205	0.255	0.315
$a=2$	0.507	0.489	0.583	0.689
		f_5		
$a=0$	0.056	0.065	0.062	0.075
$a=0.5$	0.173	0.167	0.147	0.160
$a=1$	0.476	0.442	0.463	0.474
$a=2$	0.851	0.832	0.853	0.823

Table 3: Simulated size and power of T_n test for 0.05 significance level.

	intercept	intercept and trend
autoregressive coefficient	0.997	0.990
Phillips-Perron test		
Z_t statistic	-0.702	-1.808
10% critical value	-2.569	-3.132
KPSS test		
statistic	2.969	0.500
1% critical value	0.739	0.216

Table 4: Unit root tests for the log of the dividend-price ratio.

	regression slope estimate	t-statistic
OLS	0.0092	1.92
FM-OLS	0.0065	4.37
Integrable	-0.2505	-2.58

Table 5: In-sample performance of the log of the dividend-price ratio.

r	1/4	1/2	3/4
historic average	4.2223	4.4500	4.0765
linear regression	4.2812	4.4744	4.2521
integrable linear	4.2209	4.4393	4.1243
integrable ERP	4.0779	4.0733	4.0687

Table 6: Out-of-Sample Root MSE $\times 10^2$.