Regime Specific Predictability in Predictive Regressions*

Jesùs Gonzalo
Universidad Carlos III de Madrid
Department of Economics
Calle Madrid 126
28903 Getafe (Madrid) - Spain

Jean-Yves Pitarakis
University of Southampton
Economics Division
Southampton SO17 1BJ, U.K

December 24, 2010

Abstract

Predictive regressions are linear specifications linking a noisy variable such as stock returns to past values of a more persistent regressor such as valuation ratios, interest rates etc with the aim of assessing the presence or absence of predictability. Key complications that arise when conducting such inferences are the potential presence of endogeneity, the poor adequacy of the asymptotic approximations amongst numerous others. In this paper we develop an inference theory for uncovering the presence of predictability in such models when the strength or direction of predictability, if present, may alternate across different economically meaningful episodes. This allows us to uncover economically interesting scenarios whereby the predictive power of some variable may kick in solely during particular regimes or alternate in strength and direction (e.g. recessions versus expansions, periods of high versus low stock market valuation, periods of high versus low term spreads etc). The limiting distributions of our test statistics are free of nuisance parameters and some are readily tabulated in the literature. Finally our empirical application reconsiders the literature on Dividend Yield based stock return predictability and contrary to the existing literature documents a strong presence of predictability that is countercyclical, occurring solely during bad economic times.

Keywords: Endogeneity, Persistence, Return Predictability, Threshold Models.

*Gonzalo wishes to thank the Spanish Ministerio de Ciencia e Innovacion, grant SEJ-2007-63098 and CONSOLIDER 2010 (CSD 2006-00016) and the DGUCM (Community of Madrid) grant EXCELECON S-2007/HUM-044 for partially supporting this research. Pitarakis wishes to thank the ESRC for partially supporting this research through an individual research grant RES-000-22-3983. Both authors are grateful to Grant Hillier, Tassos Magdalinos, Peter Phillips and Peter Robinson for very useful suggestions and helpful discussions. Detailed comments and feedback from the Editor, Associate Editor and two anonymous Referees are also gratefully acknowledged. Last but not least we also thank seminar participants at Queen-Mary, LSE, Southampton, Exeter, Manchester and Nottingham, the ESEM 2009 meetings in Barcelona, the SNDE 2010 meeting in Novara and the 2010 CFE conference in London for useful comments. All errors are our own responsability.
Address for correspondence: Jean-Yves Pitarakis, University of Southampton, School of Social Sciences, Economics Division, Southampton SO17 1BJ, UK. Email: j.pitarakis@soton.ac.uk

1 Introduction

Predictive regressions with a persistent regressor (e.g. dividend yields, interest rates, realised volatility) aim to uncover the ability of a slowly moving variable to predict future values of another typically noisier variable (e.g. stock returns, GDP growth) within a bivariate regression framework. Their pervasive nature in many areas of Economics and Finance and their importance in the empirical assessment of theoretical predictions of economic models made this particular modelling environment an important and active area of theoretical and applied research (see for instance Jansson and Moreira (2006) and references therein).

A common assumption underlying old and new developments in this area involves working within a model in which the persistent regressor enters the predictive regression linearly, thus not allowing for the possibility that the strength and direction of predictability may themselves be a function of some economic factor or time itself. Given this restriction, existing work has focused on improving the quality of estimators and inferences in this environment characterised by persistence and endogeneity amongst other econometric complications. These complications manifest themselves in the form of nonstandard asymptotics, distributions that are not free of nuisance parameters, poor finite sample approximations etc. Important recent methodological breakthroughs have been obtained in Jansson and Moreira (2006), Campbell and Yogo (2006), Valkanov (2003), Lewellen (2004) while recent applications in the area of financial economics and asset pricing can be found in Cochrane (2008), Lettau and Nieuwerburgh (2008), Bandi and Perron (2008) amongst others.

The purpose of this paper is to instead develop an econometric toolkit for uncovering the presence of predictability within regression models with highly persistent regressors when the strength or direction of predictability, if present, may alternate across different economically meaningful episodes (e.g. periods of rapid versus slow growth, period of high versus low stock market valuation, periods of high versus low consumer confidence etc). For this purpose, we propose to expand the traditional linear predictive regression framework to a more general environment which allows for the possibility that the strength of predictability may itself be affected by observable economic factors. We have in mind scenarios whereby the predictability induced by some economic variable kicks in under particular instances such as when the magnitude of the variable in question (or some other variable) crosses a threshold but is useless in terms of predictive power otherwise. Alternatively, the predictive impact of a variable may alternate in sign/strength across different regimes. Ignoring such phenomena by proceeding within a linear framework as it has been done in the literature may mask the forecasting ability of a particular variable and more generally mask the presence of interesting and economically meaningful dynamics. We subsequently apply our methodology to the prediction of stock returns with Dividend Yields. Contrary to what has been documented in the linear predictability literature our findings strongly point towards the presence of regimes in which Dividend Yield (DY) based predictability kicks in solely during bad economic times. More

importantly, our analysis also illustrates the fact that the presence of regimes may make predictability appear as nonexistent when assessed within a linear model.

The plan of the paper is as follows. Section 2 introduces our model and hypotheses of interest. Section 3 develops the limiting distribution theory of our test statistics. Section 4 explores the finite sample properties of the inferences developed in Section 3, Section 5 proposes an application and Section 6 concludes. All proofs are relegated to the appendix. Due to space considerations additional Monte-Carlo simulations and further details on some of the proofs are provided as a supplementary appendix.

2 The Model and Hypotheses

We will initially be interested in developing the limiting distribution theory for a Wald type test statistic designed to test the null hypothesis of a linear relationship between y_{t+1} and x_t against the following threshold alternative

$$y_{t+1} = \begin{cases} \alpha_1 + \beta_1 x_t + u_{t+1} & q_t \leq \gamma \\ \alpha_2 + \beta_2 x_t + u_{t+1} & q_t > \gamma \end{cases} \tag{1}$$

where x_t is parameterized as the nearly nonstationary process

$$x_t = \rho_T x_{t-1} + v_t, \qquad \rho_T = 1 - \frac{c}{T} \tag{2}$$

with $c > 0$, $q_t = \mu_q + u_{qt}$ and u_t, u_{qt} and v_t are stationary random disturbances. The above parameterisation allows x_t to display local to unit root behaviour and has become the norm for modelling highly persistent series for which a pure unit root assumption may not always be sensible. The threshold variable q_t is taken to be a stationary process and γ refers to the unknown threshold parameter. Under $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ our model in (1)-(2) coincides with that in Jansson and Moreira (2006) or Campbell and Yogo (2006) and is commonly referred to as a predictive regression model while under $\alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ we have a constant mean specification.

The motivation underlying our specification in (1)-(2) is its ability to capture phenomena such as regime specific predictability within a simple and intuitive framework. We have in mind scenarios whereby the slope corresponding to the predictor variable becomes significant solely in one regime. Alternatively, the strength of predictability may differ depending on the regime determined by the magnitude of q_t. The predictive instability in stock returns that has been extensively documented in the recent literature and the vanishing impact of dividend yields from the 90s onwards in particular (see Ang and Bekaert (2007) and also Table 9 below) may well be the consequence of the presence of regimes for instance. Among the important advantages of a threshold based parameterisation are the rich set of dynamics it allows to capture despite its mathematical simplicity, its estimability via a simple least squares based approach and the observability of the variable triggering regime switches which may help attach a "cause" to the underlying predictability. Following Petruccelli (1992) it is also useful to recall that the piecewise linear

structure can be viewed as an approximation to a much wider family of nonlinear functional forms. In this sense, although we do not argue that our chosen threshold specification mimics reality we believe it offers a realistic approximation to a wide range of more complicated functional forms and regime specific behaviour in particular. It is also interesting to highlight the consequences that a behaviour such as (1)-(2) may have if ignored and predictability is assessed within a linear specifications instead, say $y_t = \beta x_{t-1} + u_t$. Imposing zero intercepts for simplicity and assuming (1)-(2) holds with some γ_0 it is easy to establish that $\hat{\beta} \xrightarrow{p} \beta_1 + (\beta_2 - \beta_1)P(q_t > \gamma_0)$. This raises the possibility that $\hat{\beta}$ may converge to a quantity that is very close to zero (e.g. when $P(q_t > \gamma_0) \approx \beta_1/(\beta_1 - \beta_2)$) so that tests conducted within a linear specification may frequently and wrongly suggest absence of any predictability.

Our choice of modelling x_t as a nearly integrated process follows the same motivation as in the linear predictive regression literature where such a choice for x_t has been advocated as an alternative to proceeding with conventional Gaussian critical values which typically provide poor finite sample approximations to the distribution of t statistics. In the context of a stationary AR(1) for instance, Chan (1988) demonstrates that for values of $T(1-\rho) \geq 50$ the normal distribution offers a good approximation while for $T(1-\rho) \leq 50$ the limit obtained assuming near integratedness works better when the objective involves conducting inferences about the slope parameter of the AR(1) (see also Cavanagh, Elliott and Stock (1995) for similar points in the context of a predictive regression model). Models that combine persistent variables with nonlinear dynamics as (1)-(2) offer an interesting framework for capturing stylised facts observed in economic data. Within a univariate setting (e.g. threshold unit root models) recent contributions towards their theoretical properties have been obtained in Caner and Hansen (2001) and Pitarakis (2008).

In what follows the threshold parameter γ is assumed unknown with $\gamma \in \Gamma = [\gamma_1, \gamma_2]$ and γ_1 and γ_2 are selected such that $P(q_t \leq \gamma_1) = \pi_1 > 0$ and $P(q_t \leq \gamma_2) = \pi_2 < 1$ as in Caner and Hansen (2001). We also define $I_{1t} \equiv I(q_t \leq \gamma)$ and $I_{2t} \equiv I(q_t > \gamma)$ but replace the threshold variable with a uniformly distributed random variable making use of the equality $I(q_t \leq \gamma) = I(F(q_t) \leq F(\gamma)) \equiv I(U_t \leq \lambda)$. Here $F(.)$ is the marginal distribution of q_t and U_t denotes a uniformly distributed random variable on $[0, 1]$. Before proceeding further it is also useful to reformulate (1) in matrix format. Letting y denote the vector stacking y_{t+1} and X_i the matrix stacking $(I_{it} \;\; x_t I_{it})$ for $i = 1, 2$ we can write (1) as $y = X_1\theta_1 + X_2\theta_2 + u$ or $y = Z\theta + u$ with $Z = (X_1 \;\; X_2)$, $\theta = (\theta_1, \theta_2)$ and $\theta_i = (\alpha_i, \beta_i)'$ $i = 1, 2$. For later use we also define $X = X_1 + X_2$ as the regressor matrix which stacks the constant and x_t. It is now easy to see that for given γ or λ the homoskedastic Wald statistic for testing a general restriction on θ, say $R\theta = 0$ is given by $W_T(\lambda) = \hat{\theta}'R'(R(Z'Z)^{-1}R')^{-1}R\hat{\theta}/\hat{\sigma}_u^2$ with $\hat{\theta} = (Z'Z)^{-1}Z'y$ and $\hat{\sigma}_u^2 = (y'y - \sum_{i=1}^{2} y'X_i(X_i'X_i)^{-1}X_i'y)/T$ is the residual variance obtained from (1). In practice since the threshold parameter is unidentified under the null hypothesis inferences are conducted using the SupWald formulation expressed as $\sup_{\lambda \in [\pi_1, \pi_2]} W_T(\lambda)$ with $\pi_1 = F(\gamma_1)$ and $\pi_2 = F(\gamma_2)$.

In the context of our specification in (1)-(2) we will initially be interested in the null hypothesis of

linearity given by $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$. We write the corresponding restriction matrix as $R_A = [I \ -I]$ with I denoting a 2×2 identity matrix and the SupWald statistic $\sup_\lambda W_T^A(\lambda)$. At this stage it is important to note that the null hypothesis given by H_0^A corresponds to the linear specification $y_{t+1} = \alpha + \beta x_t + u_{t+1}$ and thus does not test predictability per se since x_t may appear as a predictor under both the null and the alternative hypotheses. Thus we also consider the null given by $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ with the corresponding SupWald statistic written as $\sup_\lambda W_T^B(\lambda)$ where now $R_B = [1 \ 0 \ -1 \ 0, 0 \ 1 \ 0 \ 0, 0 \ 0 \ 0 \ 1]$. Under this null hypothesis the model is given by $y_{t+1} = \alpha + u_{t+1}$ and the test is expected to have power against departures from both linearity and predictability. Finally, our framework will also cover the case whereby one wishes to test the hypothesis $H_0^C : \beta_1 = \beta_2 = 0$ without restricting the intercept parameters so that the null is compatible with both $\alpha_1 = \alpha_2$ and $\alpha_1 \neq \alpha_2$. We will refer to the corresponding Wald statistic as $W_T^C(\lambda)$ with the restriction matrix given by $R_C = [0 \ 1 \ 0 \ 0, 0 \ 0 \ 0 \ 1]$.

3 Large Sample Inference

Our objective here is to investigate the asymptotic properties of Wald type tests for detecting the presence of threshold effects in our predictive regression setup. We initially obtain the limiting distribution of $W_T^A(\lambda)$ under the null hypothesis $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$. We subsequently turn to the joint null hypothesis of linearity and no predictability given by $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ and explore the limiting behaviour of $W_T^B(\lambda)$. This is then followed by the treatment of the null given by $H_0^C : \beta_1 = \beta_2 = 0$ via $W_T^C(\lambda)$ and designed to explore potential predictability induced by x regardless of any restrictions on the intercepts.

Our operating assumptions about the core probabilistic structure of (1)-(2) will closely mimic the assumptions imposed in the linear predictive regression literature but will occasionally also allow for a greater degree of generality (e.g. Campbell and Yogo (2006), Jansson and Moreira (2006), Cavanagh, Elliott and Stock (1995) amongst others). Specifically, the innovations v_t will be assumed to follow a general linear process we write as $v_t = \Psi(L)e_t$ where $\Psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$, $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ and $\Psi(1) \neq 0$ while the shocks to y_t, denoted u_t, will take the form of a martingale difference sequence with respect to an appropriately defined information set. More formally, letting $\widetilde{w}_t = (u_t, e_t)'$ and $\mathcal{F}_t^{\widetilde{w}q} = \{\widetilde{w}_s, u_{qs} | s \leq t\}$ the filtration generated by $(\widetilde{w}_t, u_{qt})$ we will operate under the following assumptions

Assumptions. A1: $E[\widetilde{w}_t|\mathcal{F}_{t-1}^{\widetilde{w}q}] = 0$, $E[\widetilde{w}_t\widetilde{w}_t'|\mathcal{F}_{t-1}^{\widetilde{w}q}] = \widetilde{\Sigma} > 0$, $\sup_t E\widetilde{w}_{it}^4 < \infty$; **A2:** *the threshold variable* $q_t = \mu_q + u_{qt}$ *has a continuous and strictly increasing distribution* $F(.)$ *and is such that* u_{qt} *is a strictly stationary, ergodic and strong mixing sequence with mixing numbers* α_m *satisfying* $\sum_{m=1}^{\infty} \alpha^{\frac{1}{m}-\frac{1}{r}} < \infty$ *for some* $r > 2$.

One implication of assumption A1 and the properties of $\Psi(L)$ is that a functional central limit theorem holds for the joint process $w_t = (u_t, v_t)'$ (see Phillips (1987)). More formally $\sum_{t=1}^{[Tr]} w_t/\sqrt{T} \Rightarrow B(r) =$

$(B_u(r), B_v(r))'$ with the long run variance of the bivariate Brownian Motion $B(r)$ being given by $\Omega = \sum_{k=-\infty}^{\infty} E[w_0 w_k'] = [(\omega_u^2, \omega_{uv}), (\omega_{vu}, \omega_v^2)] = \Sigma + \Lambda + \Lambda'$. Our notation is such that $\widetilde{\Sigma} = [(\sigma_u^2, \sigma_{ue}), (\sigma_{ue}, \sigma_e^2)]$ and $\Sigma = [(\sigma_u^2, \sigma_{uv}), (\sigma_{uv}, \sigma_v^2)]$ with $\sigma_v^2 = \sigma_e^2 \sum_{j=0}^{\infty} \psi_j^2$ and $\sigma_{uv} = \sigma_{ue}$ since $E[u_t e_{t-j}] = 0 \ \forall j \geq 1$ by assumption. Given our parameterisation of v_t and the m.d.s assumption for u_t we have $\omega_{uv} = \sigma_{ue}\Psi(1)$ and $\omega_v^2 = \sigma_e^2 \Psi(1)^2$. For later use we also let $\lambda_{vv} = \sum_{k=1}^{\infty} E[v_t v_{t-k}]$ denote the one sided autocovariance so that $\omega_v^2 = \sigma_v^2 + 2\lambda_{vv} \equiv \sigma_e^2 \sum_{j=0}^{\infty} \psi_j^2 + 2\lambda_{vv}$. At this stage it is useful to note that the martingale difference assumption in A1 imposes a particular structure on Ω. For instance since serial correlation in u_t is ruled out we have $\omega_u^2 = \sigma_u^2$. It is worth emphasising however that while ruling out serial correlation in u_t our assumptions allow for a sufficiently general covariance structure linking (1)-(2) and a general dependence structure for the disturbance terms driving x_t and q_t. The martingale difference assumption on u_t is a standard assumption that has been made throughout all recent research on predictive regression models (see for instance Jansson and Moreira (2006), Campbell and Yogo (2005) and references therein) and appears to be an intuitive operating framework given that many applications take y_{t+1} to be stock returns. Writing $\Lambda = \sum_{k=1}^{\infty} E[w_t w_{t-k}'] = [(\lambda_{uu}, \lambda_{uv}), (\lambda_{vu}, \lambda_{vv})]$ it is also useful to explicitly highlight the fact that within our probabilitic environment $\lambda_{uu} = 0$ and $\lambda_{uv} = 0$ due to the m.d.s property of the $u_t's$ while λ_{vv} and λ_{vu} may be nonzero.

Regarding the dynamics of the threshold variable q_t and how it interacts with the remaining variables driving the system, assumption A1 requires q_{t-j}'s to be orthogonal to u_t for $j \geq 1$. Since q_t is stationary this is in a way a standard regression model assumption and is crucial for the development of our asymptotic theory. We note however that our assumptions allow for a broad level of dependence between the threshold variable q_t and the other variables included in the model (e.g. q_t may be contemporaneously correlated with both u_t and v_t). At this stage it is perhaps also useful to reiterate the fact that our assumption about the correlation of q_t with the remaining components of the system are less restrictive than what is typically found in the literature on marked empirical processes or functional coefficient models such as $y_{t+1} = f(q_t)x_t + u_{t+1}$ which commonly take q_t to be independent of u_t and x_t.

Since our assumptions also satisfy Caner and Hansen's (2001) framework, from their Theorem 1 we can write $\sum_{t=1}^{[Tr]} u_t I_{1t-1}/\sqrt{T} \Rightarrow B_u(r, \lambda)$ as $T \to \infty$ with $B_u(r, \lambda)$ denoting a two parameter Brownian Motion with covariance $\sigma_u^2(r_1 \wedge r_2)(\lambda_1 \wedge \lambda_2)$ for $(r_1, r_2), (\lambda_1, \lambda_2) \in [0, 1]^2$ and where $a \wedge b \equiv \min\{a, b\}$. Noting that $B_u(r, 1) \equiv B_u(r)$ we will also make use of a particular process known as a Kiefer process and defined as $G_u(r, \lambda) = B_u(r, \lambda) - \lambda B_u(r, 1)$. A Kiefer process on $[0, 1]^2$ is Gaussian with zero mean and covariance function $\sigma_u^2(r_1 \wedge r_2)(\lambda_1 \wedge \lambda_2 - \lambda_1 \lambda_2)$. Finally, we introduce the diffusion process $K_c(r) = \int_0^r e^{(r-s)c} dB_v(s)$ with $K_c(r)$ such that $dK_c(r) = cK_c(r) + dB_v(r)$ and $K_c(0) = 0$. Note that we can also write $K_c(r) = B_v(r) + c\int_0^r e^{(r-s)c} B_v(s) ds$. Under our assumptions it follows directly from Lemma 3.1 in Phillips (1988) that $x_{[Tr]}/\sqrt{T} \Rightarrow K_c(r)$.

3.1 Testing $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$

Having outlined our key operating assumptions we now turn to the limiting behaviour of our test statistics. We will initially concentrate on the null hypothesis given by $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and the behaviour of $\sup_\lambda W_T^A(\lambda)$ which is summarised in the following Proposition.

Proposition 1: *Under the null hypothesis $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$, assumptions A1-A2 and as $T \to \infty$ the limiting distribution of the SupWald statistic is given by*

$$\begin{aligned} \sup_\lambda W_T^A(\lambda) \quad &\Rightarrow \quad \sup_\lambda \frac{1}{\lambda(1-\lambda)\sigma_u^2} \left[\int_0^1 \overline{K}_c(r) dG_u(r,\lambda)\right]' \left[\int_0^1 \overline{K}_c(r)\overline{K}_c(r)'\right]^{-1} \\ &\times \quad \left[\int_0^1 \overline{K}_c(r) dG_u(r,\lambda)\right] \end{aligned} \tag{3}$$

where $\overline{K}_c(r) = (1, K_c(r))'$, $G_u(r,\lambda)$ is a a Kiefer process and $K_c(r)$ an Ornstein-Uhlenbeck process.

Although the limiting random variable in (3) appears to depend on unknown parameters such as the correlation between B_u and B_v, σ_u^2 and the near integration parameter c a closer analysis of the expression suggests instead that it is equivalent to a random variable given by a quadratic form in normalised Brownian Bridges, identical to the one that occurs when testing for structural breaks in a purely stationary framework. We can write it as

$$\sup_\lambda \frac{BB(\lambda)' BB(\lambda)}{\lambda(1-\lambda)} \tag{4}$$

with $BB(\lambda)$ denoting a standard bivariate Brownian Bridge (recall that a Brownian Bridge is a zero mean Gaussian process with covariance $\lambda_1 \wedge \lambda_2 - \lambda_1\lambda_2$). This result follows from the fact that the processes $K_c(r)$ and $G_u(r,\lambda)$ appearing in the stochastic integrals in (3) are uncorrelated and thus independent since Gaussian. Indeed

$$\begin{aligned} E[G_u(r_1,\lambda_1)K_c(r_2)] &= E[(B_u(r_1,\lambda_1) - \lambda_1 B_u(r_1,1))(B_v(r_2) + \\ &\quad c\int_0^{r_2} e^{(r_2-s)c} B_v(s)ds)] \\ &= E[B_u(r_1,\lambda_1)B_v(r_2)] - \lambda_1 E[B_u(r_1,1)B_v(r_2)] + \\ &\quad c\int_0^{r_2} e^{(r_2-s)c} E[B_u(r_1,\lambda_1)B_v(s)]ds - \\ &\quad \lambda_1 c\int_0^{r_2} e^{(r_2-s)c} E[B_u(r_1,1)B_v(s)]ds \\ &= \omega_{uv}(r_1 \wedge r_2)\lambda_1 - \lambda_1\omega_{uv}(r_1 \wedge r_2) \\ &+ c\lambda_1\int_0^{r_2} e^{(r_2-s)c}(r_1 \wedge s)ds - c\lambda_1\int_0^{r_2} e^{(r_2-s)c}(r_1 \wedge s)ds = 0. \end{aligned}$$

Given that $K_c(r)$ is Gaussian and independent of $G_u(r,\lambda)$ and also $E[G_u(r_1,\lambda_1)G_u(r_2,\lambda_2)] = \sigma_u^2(r_1 \wedge r_2)((\lambda_1 \wedge \lambda_2) - \lambda_1\lambda_2$ we have $\int K_c(r)dG_u(r,\lambda) \equiv N(0, \sigma_u^2\lambda(1-\lambda)\int K_c(r)^2)$ conditionally on a realisation of $K_c(r)$. Normalising by $\sigma_u^2\int K_c^2(r)$ as in (3) gives the Brownian Bridge process in (4) which is also the unconditional distribution since it is not dependent on a realisation of $K_c(r)$ (see also Lemma 5.1

in Park and Phillips (1988)). Obviously the discussion trivially carries through to $\overline{K}_c$ and G_u since $E[\overline{K}_c(r_2)G_u(r_1,\lambda_1)]' = E[G_u(r_1,\lambda_1) \;\; K_c(r_2)G_u(r_1,\lambda_1)]' = [0 \;\; 0]'$.

The result in Proposition 1 is unusual and interesting for a variety of reasons. It highlights an environment in which the null distribution of the SupWald statistic no longer depends on any nuisance parameters as it is typically the case in a purely stationary environment and thus no bootstrapping schemes are needed for conducting inferences. In fact, the distribution presented in Proposition 1 is extensively tabulated in Andrews (1993) and Hansen (1997) also provides p-value approximations which can be used for inference purposes. More recently, Estrella (2003) provides exact p-values for the same distribution. Finally and perhaps more importantly the limiting distribution does not appear to depend on c the near integration parameter which is another unusual specificity of our framework.

All these properties are in contrast with what has been documented in the recent literature on testing for threshold effects in purely stationary contexts. In Hansen (1996) for instance the author investigated the limiting behaviour of a SupLM type test statistic for detecting the presence of threshold nonlinearities in purely stationary models. There it was established that the key limiting random variables depend on numerous nuisance parameters involving unknown population moments of variables included in the fitted model. From Theorem 1 in Hansen (1996) it is straightforward to establish for instance that under stationarity the limiting distribution of a Wald type test statistic would be given by $S^*(\lambda)'M^*(\lambda)^{-1}S^*(\lambda)$ with $M^*(\lambda) = M(\lambda) - M(\lambda)M(1)^{-1}M(\lambda)$, and $S^*(\lambda) = S(\lambda) - M(\lambda)M(1)^{-1}S(1)$. Here $M(\lambda) = E[X_1'X_1]$ and $S(\lambda)$ is a zero mean Gaussian process with variance $M(\lambda)$. Since in this context the limiting distribution depends on the unknown model specific population moments the practical implementation of inferences is through a bootstrap style methodology.

One interesting instance worth pointing out however is the fact that this limiting random variable simplifies to a Brownian Bridge type of limit when the threshold variable is taken as exogenous in the sense $M(\lambda) = \lambda M(1)$. Although the comparison with the present context is not obvious since x_t is taken as near integrated and we allow the innovations in q_t to be correlated with those of x_t the force behind the analogy comes from the fact that x_t and q_t have variances with different orders of magnitude. In a purely stationary setup, taking x_t as stationary and the threshold variable as some uniformly distributed random variable leads to results such as $\sum x_t^2 I(U_t \leq \lambda)/T \xrightarrow{p} E[x_t^2 I(U_t \leq \lambda)]$ and if x_t and U_t are independent we also have $E[x_t^2 I(U_t \leq \lambda)] = \lambda E[x_t^2]$. It is this last key simplification which is instrumental in leading to the Brownian Bridge type of limit in Hansen's (1996) framework. If now x_t is taken as a nearly integrated process and regardless of whether its shocks are correlated with U_t or not we have $\sum x_t^2 I(U_t \leq \lambda)/T^2 \Rightarrow \lambda \int K_c^2(r)$ which can informally be viewed as analogous to the previous scenario. Heuristically this result follows by establishing that asymptotically, objects interacting $x_t/\sqrt{T}$ and $(I_{1t} - \lambda)$ such as $\frac{1}{T}\sum(\frac{x_t}{\sqrt{T}})^2(I_{1t} - \lambda)$ or $\frac{1}{T}\sum(\frac{x_t}{\sqrt{T}})(I_{1t} - \lambda)$ converge to zero (see also Caner and Hansen (2001, page 1585) and Pitarakis (2008)). This would be similar to arguing that $x_t/\sqrt{T}$ and I_{1t} are asymptotically uncorrelated in the sense that their sample covariance (normalised by T) is zero

in the limit.

3.2 Testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$

We next turn to the case where the null hypothesis of interest tests jointly the absence of linearity and no predictive power i.e. we focus on testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ using the supremum of $W_T^B(\lambda)$. The following Proposition summarises its limiting behaviour.

Proposition 2: *Under the null hypothesis* $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$*, assumptions A1-A2 and as* $T \to \infty$*, the limiting distribution of the SupWald statistic is given by*

$$\sup_{\lambda} W_T^B(\lambda) \quad \Rightarrow \quad \frac{\left[\int K_c^*(r) dB_u(r,1)\right]^2}{\sigma_u^2 \int K_c^*(r)^2} + \sup_{\lambda} \frac{1}{\lambda(1-\lambda)\sigma_u^2} \left[\int \overline{K}_c^*(r) dG_u(r,\lambda)\right]' \left[\int \overline{K}_c^* \overline{K}_c^*(r)'\right]^{-1} \left[\int \overline{K}_c^*(r) dG_u(r,\lambda)\right]' \quad (5)$$

where $\overline{K}_c^*(r) = (1, K_c^*(r))'$*,* $K_c^*(r) = K_c(r) - \int_0^1 K_c(r)dr$ *and the remaining variables are as in Proposition 1.*

Looking at the expression of the limiting random variable in (5) we note that it consists of two components with the second one being equivalent to the limiting random variable we obtained under Proposition 1. The first component in the right hand side of (5) is more problematic in the sense that it does not simplify further due to the fact that $K_c^*(r)$ and $B_u(r,1)$ are correlated since ω_{uv} may take nonzero values. However, if we were to rule out endogeneity by setting $\omega_{uv} = 0$ then it is interesting to note that the limiting distribution of the SupWald statistic in Proposition 2 takes the following simpler form

$$\sup_{\lambda} W_T^B(\lambda) \quad \Rightarrow \quad W(1)^2 + \sup_{\lambda} \frac{BB(\lambda)' BB(\lambda)}{\lambda(1-\lambda)} \quad (6)$$

where $BB(\lambda)$ is a Brownian Bridge and $W(1)$ a standard normally distributed random variable. The first component in the right hand side of either (5) or (6) can be recognised as the $\chi^2(1)$ limiting distribution of the Wald statistic for testing $H_0 : \beta = 0$ in the linear specification

$$y_{t+1} \quad = \quad \alpha + \beta x_t + u_{t+1} \quad (7)$$

and the presence of this first component makes the test powerful in detecting deviations from the null (see Rossi (2005) for the illustration of a similar phenomenon in a different context).

Our next concern is to explore ways of making (5) operational since as it stands the first component of the limiting random variable depends on model specific moments and cannot be universally tabulated. For this purpose it is useful to notice that the problems arising from the practical implementation of (5) are partly analogous to the difficulties documented in the single equation cointegration testing literature where the goal was to obtain nuisance parameter free chisquare asymptotics for Wald type tests on β

in (7) despite the presence of endogeneity (see Phillips and Hansen (1990), Saikkonen (1991, 1992)). As shown in Elliott (1998) however inferences about β in (7) can no longer be mixed normal when x_t is a near unit root process. It is only very recently that Phillips and Magdalinos (2009) (PM09 thereafter) reconsidered the issue and resolved the difficulties discussed in Elliott (1998) via the introduction of a new Instrumental Variable type estimator of β in (7). Their method is referred to as IVX estimation since the relevant IV is constructed solely via a transformation of the existing regressor x_t. It is this same method that we propose to adapt to our present context.

Before proceeding further it is useful to note that $W_T^B(\lambda)$ can be expressed as the sum of the following two components

$$W_T^B(\lambda) \equiv \frac{\hat{\sigma}_{lin}^2}{\hat{\sigma}_u^2} W_T(\beta = 0) + W_T^A(\lambda) \tag{8}$$

where $W_T(\beta = 0)$ is the standard Wald statistic for testing $H_0 : \beta = 0$ in (7). Specifically,

$$W_T(\beta = 0) = \frac{1}{\hat{\sigma}_{lin}^2} \frac{[\sum x_{t-1} y_t - T\bar{x}\bar{y}]^2}{[\sum x_{t-1}^2 - T\bar{x}^2]} \tag{9}$$

with $\bar{x} = \sum x_{t-1}/T$ and $\hat{\sigma}_{lin}^2 = (y'y - y'X(X'X)^{-1}X'y)/T$ is the residual variance obtained from the same linear specification. Although not of direct interest this reformulation of $W_T^B(\lambda)$ can simplify the implementation of the IVX version of the Wald statistic since the setup is now identical to that of PM09 and involves constructing a Wald statistic for testing $H_0 : \beta = 0$ in (7) i.e we replace $W_T(\beta = 0)$ in (8) with its IVX based version which is shown to be asymptotically distributed as a $\chi^2(1)$ random variable and independent of the noncentrality parameter c. Note that although PM09 operated within a model without an intercept, in a recent paper Kostakis, Magdalinos and Stamatogiannis (2010) (KMS10) have also established the validity of the theory in models with a fitted constant term.

The IVX methodology starts by choosing an artifical slope coefficient, say

$$R_T = 1 - \frac{c_z}{T^\delta} \tag{10}$$

for a given constant c_z and $\delta < 1$ and uses the latter to construct an IV generated as $\tilde{z}_t = R_T \tilde{z}_{t-1} + \Delta x_t$ or under zero initialisation $\tilde{z}_t = \sum_{j=1}^t R_T^{t-j} \Delta x_j$. This IV is then used to obtain an IV estimator of β in (7) and to construct the corresponding Wald statistic for testing $H_0 : \beta = 0$. Through this judicious choice of instrument PM09 show that it is possible to clean out the effects of endogeneity even within the near unit root case and to subsequently obtain an estimator of β which is mixed normal under a suitable choice of δ (i.e. $\delta \in (2/3, 1)$) and setting $c_z = 1$ (see PM09, pp. 7-12).

Following PM09 and KMS10 and letting y_t^*, x_t^* and $\tilde{z}_t^*$ denote the demeaned versions of y_t, x_t and $\tilde{z}_t$ we can write the IV estimator as $\tilde{\beta}^{ivx} = \sum y_t^* \tilde{z}_{t-1}^* / \sum x_{t-1}^* \tilde{z}_{t-1}^*$. Note that contrary to PM09 and KMS10 we do not need a bias correction term in the numerator of $\tilde{\beta}^{ivx}$ since we operate under the assumption that $\lambda_{uv} = 0$. The corresponding IVX based Wald statistic for testing $H_0 : \beta = 0$ in (7) is now written as

$$W_T^{ivx}(\beta = 0) = \frac{(\tilde{\beta}^{ivx})^2 (\sum x_{t-1}^* \tilde{z}_{t-1}^*)^2}{\tilde{\sigma}_u^2 \sum (\tilde{z}_{t-1}^*)^2} \tag{11}$$

with $\tilde{\sigma}_u^2 = \sum(y_t^* - \tilde{\beta}^{IVX} x_{t-1}^*)^2/T$. Note that this latter quantity is also asymptotically equivalent to $\hat{\sigma}_{lin}^2$ since the least squares estimator of β remains consistent. Under the null hypothesis H_0^B we also have that these two residual variances are in turn asymptotically equal to $\hat{\sigma}_u^2$ so that $\hat{\sigma}_{lin}^2/\hat{\sigma}_u^2 \approx 1$ in (8).

We can now introduce our modified Wald statistic, say $W_T^{B,ivx}(\lambda)$ for testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ in (1) as

$$W_T^{B,ivx}(\lambda) \quad = \quad W_T^{ivx}(\beta = 0) + W_T^A(\lambda). \tag{12}$$

Its limiting behaviour is summarised in the following Proposition.

Proposition 3: *Under the null hypothesis* $H_0^{(B)} : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$*, assumptions A1-A2,* $\delta \in (2/3, 1)$ *in (10) and as* $T \to \infty$*, we have*

$$\sup_\lambda W_T^{B,ivx}(\lambda) \quad \Rightarrow \quad W(1)^2 + \sup_\lambda \frac{BB(\lambda)' BB(\lambda)}{\lambda(1-\lambda)} \tag{13}$$

with $BB(\lambda)$ *denoting a standard Brownian Bridge.*

Our result in (13) highlights the usefulness of the IVX based estimation methodology since the resulting limiting distribution of the SupWald statistic is now equivalent to the one obtained under strict exogeneity (i.e. under $\omega_{uv} = 0$) in (6). The practical implementation of the test is also straightforward, requiring nothing more than the computation of an IV estimator.

3.3 Some Remarks on Testing Strategies and Further Tests

So far we have developed the distribution theory for two sets of hypotheses that we explicitly did not attempt to view as connected since both may be of interest and considered individually depending on the context of the research question. The implementation of hypotheses tests in a sequence is a notoriously difficult and often controversial endeavor which we do not wish to make a core objective of this paper especially within the nonstandard probabilistic environment we are operating under. Depending on the application in hand each of the hypotheses we have considered is useful in its own right. If one is interested in predictability coming from either x or q for instance then $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ would be a natural choice. A non rejection of this null would stop the investigation and lead to the conclusion that the data do not support the presence of any form of predictability with some confidence level. If one is solely interested in the potential presence of regimes in a general sense then a null such as $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ may be the sole focus of an investigation.

Naturally, one could also be tempted to combine H_0^B and H_0^A within a sequence and upon rejection of H_0^B and a non rejection of H_0^A argue in favour of linear predictability while a rejection of H_0^A would support the presence of nonlinear predictability in a *broad* sense. In this latter case the rejection could be compatible with a model in which only the intercepts shift and x plays no role in predicting y since a

specification such as $y_{t+1} = \alpha_1 I(q_t \leq \gamma_0) + \alpha_2 I(q_t > \gamma_0) + u_{t+1}$ in which predictability is solely driven by the threshold variable q_t is compatible with the rejection of both H_0^A and H_0^B. As in most sequentially implemented tests however one should also be aware that the overall size of such an approach would be difficult to control since the two tests will be correlated. Even under independence which would allow a form of size control the choice of individual significance levels is not obvious and may lead to different conclusions.

Given the scenarios dicussed above and depending on the application in hand it is now also interesting to explore the properties of a test that focuses solely on slope parameters with its null given by $H_0^C : \beta_1 = \beta_2 = 0$. Such a null would be relevant if one were solely interested in the linear or nonlinear predictability induced by x or if one believed on à priori grounds that $\alpha_1 \neq \alpha_2$. As in Caner and Hansen (2001) the practical difficulty here lies in the fact that H_0^C is compatible with both $\alpha_1 = \alpha_2$ and $\alpha_1 \neq \alpha_2$.

We let $W_T^C(\lambda) = \hat{\theta}' R_C' (R_C (Z'Z)^{-1} R_C)^{-1} R_C \hat{\theta} / \hat{\sigma}_u^2$ denote the Wald statistic for testing H_0^C within the unrestricted specification in (1) and for some given $\lambda \in (0,1)$. When we wish to explicitly impose the constancy of intercepts in the fitted model used to calculate $W_T^C(\lambda)$ we will refer to the same test statistic as $W_T^C(\lambda|\alpha_1 = \alpha_2)$. The latter is computed from the intercept restricted model which in matrix form can be written as $y = \widetilde{Z}\phi + u$ with $\widetilde{Z} = [1 \; x_1 \; x_2]$, $\phi = (\alpha \; \beta_1 \; \beta_2)'$ and where the lower-case vectors x_i stack the elements of $x_t I_{it}$ for $i = 1, 2$. More specifically $W_T^C(\lambda|\alpha_1 = \alpha_2) = \hat{\phi}' \widetilde{R}' (\widetilde{R} (\widetilde{Z}'\widetilde{Z})^{-1} \widetilde{R}')^{-1} \widetilde{R} \hat{\phi} / \tilde{\sigma}_u^2$ with $\widetilde{R} = [0 \; 1 \; 0, 0 \; 0 \; 1]$ and $\tilde{\sigma}_u^2$ referring to the residual variance from the same intercept restricted specification. Unless explictly stated however $W_T^C(\lambda)$ will be understood to be computed within (1). When $\alpha_1 \neq \alpha_2$ we also denote by $\hat{\lambda} = F(\hat{\gamma})$ the least squares based estimator of the threshold parameter obtained from the null model $y_{t+1} = \alpha_1 I_{1t} + \alpha_2 I_{2t} + u_{t+1}$ and $\lambda_0 = F(\gamma_0)$ its true counterpart. Note that since this threshold parameter estimator is obtained within a purely stationary setting of the null model its T-consistency follows from Gonzalo and Pitarakis (2002). The following Proposition summarises the large sample behaviour of $W_T^C(\lambda)$ under alternative scenarios.

Proposition 4: *(i) Under the null hypothesis $H_0^C : \beta_1 = \beta_2 = 0$, assumptions A1-A2, and if $\alpha_1 = \alpha_2$ in the DGP, we have as $T \to \infty$,*

$$W_T^C(\lambda) \;\; \Rightarrow \;\; \frac{\left[\int K_c^*(r) dB_u(r,1)\right]^2}{\sigma_u^2 \int K_c^*(r)^2} + \chi^2(1) \tag{14}$$

for any constant $\lambda \in (0,1)$ and similarly for $W_T^C(\lambda|\alpha_1 = \alpha_2)$. (ii) If $\alpha_1 \neq \alpha_2$ the limiting result in (14) continues to hold for $W_T^C(\lambda_0)$ and $W_T^C(\hat{\lambda})$ but not for any other $\lambda \in (0,1)$. (iii) Under exogeneity the limiting random variable in (14) is equivalent to a $\chi^2(2)$.

The above results highlight a series of important facts. When $\alpha_1 = \alpha_2$, the Wald statistics $W_T^C(\lambda)$ or $W_T^C(\lambda|\alpha_1 = \alpha_2)$ evaluated at any $\lambda \in (0,1)$ are seen to converge to a random variable that does not depend on λ. This is obviously no longer the case when $\alpha_1 \neq \alpha_2$ and is intuitively due to the fact that fitting a model with the wrong choice of λ (i.e. $\lambda \neq \lambda_0$) leads to inconsistent parameter estimates. This

is why $W_T^C(\lambda)$ needs to be evaluated at $\hat{\lambda}$ or λ_0 when $\alpha_1 \neq \alpha_2$.

One practical and well known limitation of (14) comes from its first component which depends on the noncentrality parameter c in addition to other endogeneity induced nuisance parameters. As pointed out in Proposition 4(iii) however imposing exogeneity leads to the interesting and unusual outcome of a simple nuisance parameter free standard distributional result. Thus if we are willing to entertain an exogeneity assumption our result in Proposition 4 offers a simple and trivial way of conducting inferences on the $\beta' s$.

Naturally and analogously to the framework of Caner and Hansen (2001) our result in Proposition 4(i) crucially depends on the knowledge that $\alpha_1 = \alpha_2$ while the use of $W_T^C(\lambda_0)$ or $W_T^C(\hat{\lambda})$ presume knowledge that $\alpha_1 \neq \alpha_2$ so that λ_0 becomes a meaningful quantity. If $\alpha_1 \neq \alpha_2$ with the switch occurring at some λ_0, it is straightforward to show that both $W_T^C(\lambda|\alpha_1 = \alpha_2)$ and $W_T^C(\lambda)$ will be diverging to infinity with T. In the former case this will be happening because the test is evaluated at some $\lambda \neq \lambda_0$ in addition to the fact that the fitted model ignores the shifting intercepts while in the case of $W_T^C(\lambda)$ this will be happening solely because $\lambda \neq \lambda_0$. Naturally, if the ad-hoc choice of λ happens to fall close to λ_0 the use of $W_T^C(\lambda)$ may lead to more moderate distortions when $\alpha_1 \neq \alpha_2$ while continuing to be correct in the event that $\alpha_1 = \alpha_2$. For purely practical reasons therefore it may be preferable to base inferences on $W_T^C(\lambda)$ instead of $W_T^C(\lambda|\alpha_1 = \alpha_2)$ even if we believe $\alpha_1 = \alpha_2$ to be the more likely scenario.

For the purpose of making our result in Proposition 4(iii) operational even under endogeneity it is again useful to note that $W_T^C(\lambda) \approx W_T(\beta = 0, \lambda) + W_T(\beta_1 = \beta_2, \lambda)$ with $W_T(\beta = 0, \lambda)$ denoting the Wald statistic for testing $\beta = 0$ in $y_{t+1} = \alpha_1 I_{1t}(\lambda) + \alpha_2 I_{2t}(\lambda) + \beta x_t + u_{t+1}$ for a given $\lambda \in (0, 1)$ and $W_T(\beta_1 = \beta_2, \lambda)$ the Wald statistic for testing $H_0 : \beta_1 = \beta_2$ in model (1). More formally, letting $Z_1 = [I_1 \ I_2 \ x]$, $\psi_1 = (\alpha_1 \ \alpha_2 \ \beta)'$ and $R_1 = [0 \ 0 \ 1]$ we have $W_T(\beta = 0, \lambda) = \hat{\psi}_1' R_1' [R_1 (Z_1' Z_1)^{-1} R_1']^{-1} R_1 \hat{\psi}_1 / \hat{\sigma}_1^2$ where $\hat{\sigma}_1^2$ is the residual variance from $y = Z_1 \psi_1 + u$, and using our notation surrounding (1), $W_T(\beta_1 = \beta_2, \lambda) = \hat{\theta}' R_2' [R_2 (Z'Z)^{-1} R_2']^{-1} R_2 \hat{\theta} / \hat{\sigma}_u^2$ for $R_2 = [0 \ 1 \ 0 \ -1]$. Naturally, if one wishes to maintain the assumption that $\alpha_1 = \alpha_2$ we could also focus on $W_T^C(\lambda|\alpha_1 = \alpha_2) \approx W_T(\beta = 0|\alpha_1 = \alpha_2) + W_T(\beta_1 = \beta_2, \lambda|\alpha_1 = \alpha_2)$ with these two components being evaluated on models with fixed intercepts (i.e. $y_{t+1} = \alpha + \beta x_t + u_{t+1}$ for $W_T(\beta = 0|\alpha_1 = \alpha_2)$ and $y_{t+1} = \alpha + \beta_1 x_t I_{1t} + \beta_2 x_t I_{2t} + u_{t+1}$ for $W_T(\beta_1 = \beta_2, \lambda|\alpha_1 = \alpha_2)$). With $\widetilde{Z}$ defined as earlier, $W_T(\beta_1 = \beta_2, \lambda|\alpha_1 = \alpha_2) = \hat{\phi}' R_3' [R_3 (\widetilde{Z}' \widetilde{Z})^{-1} R_3']^{-1} R_3 \hat{\phi} / \tilde{\sigma}_u^2$ for $R_3 = [0 \ 1 \ -1]$ while $W_T(\beta = 0|\alpha_1 = \alpha_2)$ is as in (9). Note that the above decompositions are valid asymptotically due to the omission of scaling factors adjacent to $W_T(\beta = 0, \lambda)$ that converge to 1 in probability under the null hypothesis (i.e. $\hat{\sigma}_1^2 / \hat{\sigma}_u^2 \xrightarrow{p} 1$ and $\hat{\sigma}_{lin}^2 / \tilde{\sigma}_u^2 \xrightarrow{p} 1$).

Given the above decompositions it is now possible to modify $W_T^C(\lambda)$ (and similarly for $W_T^C(\lambda|\alpha_1 = \alpha_2)$) along exactly the same lines as our treatment of $W_T^B(\lambda)$ via the IVX based modification applied to

$W_T(\beta = 0, \lambda)$ (or $W_T(\beta = 0|\alpha_1 = \alpha_2)$ when applicable). Specifically, we let

$$\begin{aligned} W_T^{C,ivx}(\lambda) &= W_T^{ivx}(\beta = 0, \lambda) + W_T(\beta_1 = \beta_2, \lambda) \\ W_T^{C,ivx}(\lambda|\alpha_1 = \alpha_2) &= W_T^{ivx}(\beta = 0|\alpha_1 = \alpha_2) + W_T(\beta_1 = \beta_2, \lambda|\alpha_1 = \alpha_2) \end{aligned} \tag{15}$$

where $W_T^{ivx}(\beta = 0|\alpha_1 = \alpha_2)$ is as in (11) while $W_T^{ivx}(\beta = 0, \lambda)$ is the IVX based Wald statistic for testing $H_0 : \beta = 0$ in $y = Z_1\psi_1 + u$. More specifically, letting $\tilde{z}$ refer to the IVX vector that stacks the $\tilde{z}_t's$ this latter Wald statistic is constructed instrumenting $Z_1 = (I_1\ I_2\ x)$ with $\overline{Z}_1 = (I_1\ I_2\ \tilde{z})$ so that $\overline{\psi}_{1,ivx} = (\overline{Z}_1'Z_1)^{-1}\overline{Z}_1'y$ and $W_T(\beta = 0, \lambda)$ is then constructed in a manner identical to equation (25) in PM09.

The purpose of this IVX based step is to ensure that the new limit corresponding to the first component in the right hand side of (14) is $\chi^2(1)$ and thus no longer depending on the noncentrality parameter c and other endogeneity induced parameters. Due to its independence from the second $\chi^2(1)$ component which arises as the limit of $W_T(\beta_1 = \beta_2, \lambda)$ or $W_T(\beta_1 = \beta_2, \lambda|\alpha_1 = \alpha_2)$ (see the proof of Proposition 4(iii)) we also have the useful outcome that $W_T^{C,ivx}(\lambda) \Rightarrow \chi^2(2)$ for some $\lambda \in (0, 1)$ when $\alpha_1 = \alpha_2$ in addition to $W_T^{C,ivx}(\hat{\lambda}) \Rightarrow \chi^2(2)$ and $W_T^{C,ivx}(\lambda_0) \Rightarrow \chi^2(2)$ when $\alpha_1 \neq \alpha_2$. Our simulation based results presented below document a remarkably accurate match of the finite sample quantiles of $W_T^{C,ivx}(\lambda)$ and $W_T^{C,ivx}(\hat{\lambda})$ with those of the $\chi^2(2)$ (see Table 7).

Although the above might suggest a unified way of testing $H_0^C : \beta_1 = \beta_2 = 0$ regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ this is not so due to the treatment of λ in the construction of $W_T^{C,ivx}(\lambda)$. When $\alpha_1 = \alpha_2$ and as in Proposition 4 above the test statistic can be evaluated at any constant $\lambda \in (0, 1)$. This is no longer true however if $\alpha_1 \neq \alpha_2$ with the switch occurring at λ_0. In this latter case we have $W_T^{C,ivx}(\hat{\lambda}) \Rightarrow \chi^2(2)$ and obviously $W_T^{C,ivx}(\lambda_0) \Rightarrow \chi^2(2)$. When $\alpha_1 \neq \alpha_2$ evaluating $W_T^{C,ivx}(.)$ at some $\lambda \neq \lambda_0$ would lead to wrong inferences and similarly when $\alpha_1 = \alpha_2$, evaluating the same test statistic at $\hat{\lambda}$ would be misleading since $\hat{\lambda}$ is not a well defined quantity when $\alpha_1 = \alpha_2$. Indeed under $\alpha_1 = \alpha_2$, $\hat{\lambda}$ does not converge in probability to a constant and the consequences of this on the bahaviour of the test statistic are unclear.

4 Finite Sample Analysis

4.1 Testing $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$

Having established the limiting properties of the SupWald statistic for testing H_0^A our next goal is to illustrate the finite sample adequacy of our asymptotic approximation and empirically illustrate our theoretical findings. It will also be important to highlight the equivalence of the limiting results obtained in Proposition 1 to the Brownian Bridge type of limit documented in Andrews (1993) and for which Hansen (1997) obtained p-value approximations and Estrella (2003) exact p-values. Naturally, this allows us to

evaluate the size properties of our tests as well.

Our data generating process (DGP) under H_0^A is given by the following set of equations

$$\begin{aligned} y_t &= \alpha + \beta x_{t-1} + u_t \\ x_t &= \left(1 - \frac{c}{T}\right) x_{t-1} + v_t \\ v_t &= \rho v_{t-1} + e_t, \end{aligned} \tag{16}$$

with u_t and e_t both $NID(0,1)$ while the fitted model is given by (1) with q_t assumed to follow the AR(1) process $q_t = \phi q_{t-1} + u_{qt}$ with $u_{qt} = NID(0,1)$. Regarding the covariance structure of the random disturbances, letting $z_t = (u_t, e_t, u_{qt})'$ and $\mathbf{\Sigma}_z = E[z_t z_t']$, we use

$$\mathbf{\Sigma}_z = \begin{pmatrix} 1 & \sigma_{ue} & \sigma_{uu_q} \\ \sigma_{ue} & 1 & \sigma_{eu_q} \\ \sigma_{uu_q} & \sigma_{eu_q} & 1 \end{pmatrix}$$

which allows for a sufficiently general covariance structure while imposing unit variances. Note also that our chosen covariance matrix parameterisation allows the threshold variable to be contemporaneously correlated with the shocks to y_t. All our H_0^A based size experiments use $N = 5000$ replications and set $\{\alpha, \beta, \rho, \phi\} = \{0.01, 0.10, 0.40, 0.50\}$ throughout. Since our initial motivation is to explore the theoretically documented robustness of the limiting distribution of $SupWald^A$ to the presence or absence of endogeneity, we consider the two scenarios given by

$$\begin{aligned} DGP_1 &: \{\sigma_{ue}, \sigma_{uu_q}, \sigma_{eu_q}\} = \{-0.5, 0.3, 0.4\} \\ DGP_2 &: \{\sigma_{ue}, \sigma_{uu_q}, \sigma_{eu_q}\} = \{0.0, 0.0, 0.0\}. \end{aligned}$$

The implementation of all our Sup based tests assume 10% trimming at each end of the sample.

Table 1 below presents some key quantiles of the $SupWald^A$ distribution (see Proposition 1) simulated using moderately small sample sizes and compares them with their asymptotic counterparts. Results are displayed solely for the DGP_1 covariance structure since the corresponding figures for DGP_2 were almost identical.

Table 1: Critical Values of $SupWald^A$

	$DGP_1, T = 200$				$DGP_1, T = 400$				∞
	$c = 1$	$c = 5$	$c = 10$	$c = 20$	$c = 1$	$c = 5$	$c = 10$	$c = 20$	
2.5%	2.18	2.21	2.21	2.19	2.31	2.24	2.24	2.27	2.41
5.0%	2.53	2.52	2.57	2.50	2.65	2.63	2.62	2.63	2.75
10.0%	3.01	3.07	2.99	2.99	3.13	3.10	3.11	3.12	3.27
90.0%	10.20	10.46	10.48	10.39	10.28	10.23	10.20	10.30	10.46
95.0%	12.07	12.03	12.13	12.19	11.85	12.05	12.11	12.08	12.17
97.5%	13.82	13.76	13.85	13.84	13.74	13.57	13.91	13.64	13.71

Looking across the different values of c as well as the different quantiles we note an excellent adequacy of the T=200 and T=400 based finite sample distributions to the asymptotic counterpart tabulated in Andrews (1993) and Estrella (2003). This also confirms our results in Proposition 1 and provides empirical support for the fact that inferences are robust to the magnitude of c. Note that with T=200 the values of $(1 - c/T)$ corresponding to our choices of c in Table 1 are 0.995, 0.975, 0.950 and 0.800 respectively. Thus the quantiles of the simulated distribution appear to be highly robust to a wide range of persistence characteristics.

Naturally, the fact that our finite sample quantiles match closely their asymptotic counterparts even under T=200 is not sufficient to claim that the test has good size properties. For this purpose we have computed the empirical size of the $SupWald^A$ based test making use of the *pvsup* routine of Hansen (1997). The latter is designed to provide approximate p-values for test statistics whose limiting distribution is as in (4). Results are presented in Table 2 below which concentrates solely on the DGP_1 covariance structure.

Table 2: Size Properties of $SupWald^A$

	T=200			T=400			T=200, BOOT			T=400, BOOT		
	2.5%	5.0%	10%	2.5%	5.0%	10%	2.5%	5.0%	10.0%	2.5%	5.0%	10.0%
c=1	2.60	4.70	8.90	2.50	4.60	9.60	3.01	6.20	11.14	3.62	5.98	11.02
c=5	2.50	4.90	9.30	2.40	4.90	9.30	2.98	6.36	11.86	3.38	6.08	11.02
c=10	2.80	4.80	9.20	2.70	5.10	9.30	3.26	6.42	12.00	3.26	5.64	10.66
c=20	2.60	4.80	9.50	2.50	5.00	9.60	3.20	6.42	11.32	3.26	6.16	11.40

From the figures presented in the two left panels of Table 2 we again note the robustness of the empirical size estimates of $SupWald^A$ to the magnitude of the noncentrality parameter. Overall the size estimates match their nominal counterparts quite accurately even under a moderately small sample size. It is also interesting to compare the asymptotic approximation in (4) with that occuring when x_t is assumed to follow an AR(1) with $|\rho| < 1$ rather than the local to unit root specification we have adopted in this

paper. Naturally, under pure stationarity the results of Hansen (1996, 1999) apply and inferences can be conducted by simulating critical values from the asymptotic distribution that is the counterpart to (3) obtained under pure stationarity and following the approach outlined in the aforementioned papers. This latter approach is similar to an external bootstrap but should not be confused with the idea of obtaining critical values from a bootstrap distribution. The obvious question we are next interested in documenting is which approximation works better when x_t is a highly persistent process? For this purpose the two right hand panels of Table 2 above also present the corresponding empirical size estimates obtained using the asymptotic approximation and its external bootstrap style implementation developed in Hansen (1996, 1999) and justified by the multiplier central limit theorem (see Van der Vaart and Wellner (1996)). Although our comparison involves solely the size properties of the test and should therefore be interpreted cautiously the above figures suggest that the nuisance parameter free Brownian Bridge based asymptotic approximation does a good job in matching empirical with nominal sizes when ρ is close to the unit root frontier. Proceeding using Hansen (1996)'s approach on the other hand suggests a mild oversizeness of the procedure which does not taper off as T is allowed to increase.

Before proceeding further, it is also important to document $SupWald^A$'s ability to correctly detect the presence of threshold effects via a finite sample power analysis. Our goal here is not to develop a full theoretical and empirical power analysis of our test statistics which would bring us well beyond our scope but to instead give a snapshot of the ability of our test statistics to lead to a correct decision under a series of fixed departures from the null. All our power based DGPs use the same covariance structure as our size experiments and are based on the following configurations for $\{\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma\}$ in (1): DGP_1^A $\{-0.03, -0.03, 1.26, 1.20, 0\}$, DGP_2^A $\{-0.03, 0.15, 1.26, 1.20, 0\}$ and DGP_3^A $\{-0.03, 0.25, 1.26, 1.26, 0\}$ thus covering both intercept only, slope only and joint intercept and slope shifts. In Table 3 below the figures represent correct decision frequencies evaluated as the number of times the pvalue of the test statistic leads to a rejection of the null using a 2.5% nominal level.

Table 3: Power Properties of $SupWald^A$

	DGP_1^A	DGP_2^A	DGP_3^A	DGP_1^A	DGP_2^A	DGP_3^A	DGP_1^A	DGP_2^A	DGP_3^A
	$c = 1$			$c = 5$			$c = 10$		
$T = 200$	0.73	0.73	0.15	0.39	0.44	0.14	0.20	0.26	0.14
$T = 400$	0.98	0.98	0.37	0.92	0.93	0.37	0.78	0.82	0.37
$T = 1000$	1.00	1.00	0.88	1.00	1.00	0.89	1.00	1.00	0.86

We note from Table 3 that power converges towards one under all three parameter configurations albeit quite slowly when only intercepts are characterised by threshold effects. The test displays good finite sample power even under $T = 200$ when the slopes are allowed to shift as in DGP_1^A and DGP_2^A. It is also interesting to note the negative influence of an increasing c on finite sample power under the

DGPs with shifting slopes. As expected this effect vanishes asymptotically since even for $T \geq 400$ the frequencies across the different magnitudes of c become very similar.

4.2 Testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$

We next turn to the null hypothesis given by $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$. As documented in Proposition 2 we recall that the limiting distribution of the $SupWald^B$ statistic is no longer free of nuisance parameters and does not take a familiar form when we operate under the set of assumptions characterising Proposition 1. However, one instance under which the limiting distribution of the $SupWald^B$ statistic takes a simple form is when we impose the exogeneity assumption as when considering the covariance structure referred to as DGP_2 above. Under this scenario the relevant limiting distribution is given by (6) and can be easily tabulated through standard simulation based methods.

For this purpose, Table 4 below presents some empirical quantiles obtained using $T = 200$, $T = 400$ and $T = 800$ from the null DGP $y_t = 0.01 + u_t$. As can be inferred from (6) we note that the quantiles are unaffected by the chosen magnitude of c and appear sufficiently stable across the different sample sizes considered. Viewing the $T = 800$ based results as approximating the asymptotic distribution for instance the quantiles obtained under $T = 200$ and $T = 400$ match closely their asymptotic counterparts.

Table 4. Critical Values of $SupWald^B$ under Exogeneity

	2.5%	5%	10%	90%	95%	97.5%
	c=1					
$T = 200$	2.59	3.03	3.58	11.73	13.63	15.36
$T = 400$	2.67	3.06	3.67	11.80	13.69	15.41
$T = 800$	2.67	3.15	3.78	11.71	13.42	15.35
	c=5					
$T = 200$	2.56	3.02	3.64	11.63	13.69	15.46
$T = 400$	2.65	3.06	3.69	11.97	13.79	15.85
$T = 800$	2.71	3.15	3.73	11.55	13.42	15.14

We next turn to the more general scenario in which one wishes to test H_0^B within a specification that allows for endogeneity. Taking our null DGP as $y_t = 0.01 + u_t$ and the covariance structure referred to as DGP_1 it is clear from Proposition 2 that using the critical values from Table 4 will lead to misleading results. This is indeed confirmed empirically with size estimates for $SupWald^B$ lying about two percentage points above their nominal counterparts (see Table 5 below). Using our IVX based test statistic in (11)-(12) however ensures that the above critical values remain valid even under the presence of endogeneity. Results for this experiment are also presented in Table 5 below. Table 5 also aims to highlight the

influence of the choice of the δ parameter in the construction of the IVX variable (see (10)) on the size properties of the test.

Table 5: Size Properties of $SupWald^{B,ivx}$ and $SupWald^{B}$ under Endogeneity

	2.5%	5.0%	10.0%	2.5%	5.0%	10.0%	2.5%	5%	10%
$T = 200$		$c = 1$			$c = 5$			$c = 10$	
$\delta = 0.70$	2.80	5.12	10.26	2.48	5.02	10.40	2.62	5.00	10.34
$\delta = 0.80$	2.84	5.60	10.38	2.52	5.08	10.78	2.70	5.10	10.40
$\delta = 0.90$	3.04	5.48	10.68	2.70	5.20	10.86	2.76	5.32	10.56
$SupWald^{B}$	3.54	6.36	12.28	3.06	5.94	11.52	2.98	5.72	11.14
$T = 400$		$c = 1$			$c = 5$			$c = 10$	
$\delta = 0.70$	3.02	5.66	11.06	3.00	5.36	10.60	2.74	5.32	10.14
$\delta = 0.80$	3.14	5.92	11.46	3.14	5.36	10.94	2.82	5.44	10.32
$\delta = 0.90$	3.42	6.28	12.08	3.24	5.52	11.04	2.82	5.48	10.52
$SupWald^{B}$	4.28	7.30	13.20	3.46	6.22	11.46	3.08	5.66	11.08
$T = 1000$		$c = 1$			$c = 5$			$c = 10$	
$\delta = 0.70$	2.74	5.14	10.24	2.62	4.96	10.22	2.50	4.72	10.18
$\delta = 0.80$	2.96	5.68	10.74	2.64	5.40	10.12	2.66	4.74	10.62
$\delta = 0.90$	3.30	5.90	11.50	2.92	5.42	10.06	2.64	4.96	10.44
$SupWald^{B}$	4.00	6.52	13.18	3.22	5.72	10.74	2.74	5.16	10.74

Overall, we note an excellent match of the empirical sizes with their nominal counterparts. As δ increases towards one, it is possible to note a very slight deterioration in the size properties of $SupWald^{B,ivx}$ with empirical sizes mildly exceeding their nominal counterparts. Looking also at the power figures presented in Table 6 below it is clear that as $\delta \to 1$ there is a very mild size power tradeoff that kicks in. This is perhaps not surprising since as $\delta \to 1$ the instrumental variable starts behaving like the original nearly integrated regressor. Overall, choices of δ in the 0.7-0.8 region appear to lead to very sensible results within our chosen simulations with almost unnoticeable variations in the corresponding size estimates. Even under $\delta = 0.9$ and looking across all configurations we can reasonably argue that the resulting size properties are good to excellent. Finally, the rows labelled $SupWald^{B}$ clearly highlight the unsuitability of this uncorrected test statistic whose limiting distribution is as in (5).

Next, we also considered the finite sample power properties of our $SupWald^{B,ivx}$ statistic through a series of fixed departures from the null based on the following configurations for $\{\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma\}$: DGP_1^B $\{0.01, 0.01, 0.05, 0.05, 0\}$, DGP_2^B $\{-0.03, 0.25, 0.05, 0.05, 0\}$ and DGP_3^B $\{0.01, 0.25, 0, 0, 0\}$. Results for this set of experiments are presented in Table 6 below.

Table 6: Power Properties of $SupWaldB^{ivx}$

	DGP_1^B			DGP_2^B			DGP_3^B		
$c=1,T$	200	400	1000	200	400	1000	200	400	1000
$\delta=0.70$	0.81	0.97	1.00	0.89	0.99	1.00	0.17	0.37	0.87
$\delta=0.80$	0.89	0.99	1.00	0.94	1.00	1.00	0.17	0.37	0.87
$c=5,T$	200	400	1000	200	400	1000	200	400	1000
$\delta=0.70$	0.71	1.00	1.00	0.85	1.00	1.00	0.16	0.36	0.87
$\delta=0.80$	0.79	1.00	1.00	0.89	1.00	1.00	0.16	0.36	0.87
$c=10,T$	200	400	1000	200	400	1000	200	400	1000
$\delta=0.70$	0.51	1.00	1.00	0.74	1.00	1.00	0.16	0.36	0.87
$\delta=0.80$	0.58	1.00	1.00	0.78	1.00	1.00	0.16	0.36	0.86

The above figures suggest that our modified $SupWald^{B,ivx}$ statistic has good power properties under moderately large sample sizes. We again note that violating the null restriction that affects the slopes leads to substantially better power properties than scenarios where solely the intercepts violate the equality constraint.

4.3 Testing $H_0^C : \beta_1 = \beta_2 = 0$

Our initial objective here is to document the accuracy of the $\chi^2(2)$ approximation for our main IVX based test statistic $W_T^{C,ivx}(\lambda)$ defined in (15) and designed to make our inferences robust to endogeneity and to the magnitude of c. When referring to the arguments of our Wald statistics in what follows we will make use of γ and $\lambda = F(\gamma)$ interchangeably. We consider two DGPs having $\alpha_1 = \alpha_2$ and $\alpha_1 \neq \alpha_2$ respectively. In the first case $W_T^{C,ivx}(\gamma)$ is evaluated at an ad-hoc choice of γ while in the second case we consider $W_T^{C,ivx}(\hat{\gamma})$. For our $\alpha_1 = \alpha_2$ based experiments we also present the corresponding results for $W_T^{C,ivx}(\gamma|\alpha_1 = \alpha_2)$. All our experiments below use $\delta = 0.7$ in the construction of the IVX variable and set $c_z = 1$ in (10).

Table 7: Quantiles of $W_T^{C,ivx}(\gamma)$ and $W_T^{C,ivx}(\hat{\gamma})$ under Endogeneity (T=400)

	2.5%	5.0%	10.0%	90.0%	95.0%	97.5%	
$\alpha_1 = \alpha_2$	$DGP : y_{t+1} = 0.01 + u_{t+1}$						
$W_T^{C,ivx}(\gamma = 0 \vert \alpha_1 = \alpha_2)$	0.048	0.100	0.210	4.657	6.052	7.308	$c = 1$
$W_T^{C,ivx}(\gamma = 0 \vert \alpha_1 = \alpha_2)$	0.049	0.100	0.213	4.616	6.018	7.471	$c = 5$
$W_T^{C,ivx}(\gamma = 0 \vert \alpha_1 = \alpha_2)$	0.050	0.109	0.211	4.650	6.034	7.359	$c = 10$
$W_T^{C,ivx}(\gamma = 0)$	0.056	0.112	0.223	4.768	6.181	7.433	$c = 1$
$W_T^{C,ivx}(\gamma = 0)$	0.047	0.091	0.200	4.680	5.992	7.291	$c = 5$
$W_T^{C,ivx}(\gamma = 0)$	0.050	0.103	0.209	4.599	6.066	7.369	$c = 10$
$\alpha_1 \neq \alpha_2$	$DGP : y_{t+1} = -0.03I_{1t}(0) + 0.25I_{2t}(0) + u_{t+1}$						
$W_T^{C,ivx}(\hat{\gamma})$	0.055	0.113	0.203	4.783	6.172	7.460	$c = 1$
$W_T^{C,ivx}(\hat{\gamma})$	0.055	0.106	0.217	4.706	6.132	7.619	$c = 5$
$W_T^{C,ivx}(\hat{\gamma})$	0.056	0.111	0.212	4.639	6.222	7.618	$c = 10$
$\chi^2(2)$	**0.051**	**0.103**	**0.211**	**4.605**	**5.991**	**7.378**	

Table 7 above highlights how good a job the IVX based transformation of our original Wald statistic is doing in matching the theoretical quantiles of the $\chi^2(2)$ distribution even under a moderately large sample size such as $T = 400$. Under the constant intercepts scenario we also note that $W_T^{C,ivx}(\gamma|\alpha_1 = \alpha_2)$ leads to quantiles marginally closer to those of the $\chi^2(2)$ when compared with $W_T^{C,ivx}(\gamma)$. This makes intuitive sense since when $\alpha_1 = \alpha_2$, $W_T^{C,ivx}(\gamma)$ implements the test within an unnecessarily overfitted model.

We next assess the finite sample properties of $W_T^{C,ivx}(\gamma)$ through a series of size based experiments that distinguish across the two scenarios of interest on the intercepts using the same two DGPs as in Table 7. Results are presented in Table 8 below. Note that as in Table 7 above all our experiments make use of a DGP with endogeneity. We make use of $W_T^{C,ivx}(\gamma)$ with an ad-hoc choice of γ for the DGP with $\alpha_1 = \alpha_2$ while we use $W_T^{C,ivx}(\hat{\gamma})$ for the DGP with $\alpha_1 \neq \alpha_2$.

Table 8: Size Properties of $W_T^{C,ivx}(\gamma)$ and $W_T^{C,ivx}(\hat{\gamma})$ under Endogeneity

	$\alpha_1 = \alpha_2$				$\alpha_1 \neq \alpha_2$		
$W_T^{C,ivx}(\gamma = 1)$	2.5%	5.0%	10.0%	$W_T^{C,ivx}(\hat{\gamma})$	2.5%	5.0%	10.0%
$T = 200$	3.10	6.20	10.90	$T = 200$	2.60	5.10	10.50
$T = 400$	2.80	5.80	11.00	$T = 400$	2.90	5.10	10.20
$T = 1000$	2.50	5.10	10.40	$T = 1000$	2.80	5.30	10.60

Under $\alpha_1 = \alpha_2$ our test statistic is evaluated at the ad-hoc choice of $\gamma = 1$ and despite a mild oversizeness under T=200 we note a good overall match of empirical and nominal sizes. Note that $W_T^{C,ivx}(\gamma = 1)$ is evaluated on the fully unrestricted model (1) despite our knowledge of the DGP having

$\alpha_1 = \alpha_2$ (see our discussion following Proposition 4). Results across alternative magnitudes of γ were very similar and therefore omitted. Similar properties are also observed when $\alpha_1 \neq \alpha_2$ with the test statistic evaluated at $\hat{\gamma}$.

5 Regime Specific Predictability of Returns with Valuation Ratios

One of the most frequently explored specification in the financial economics literature has aimed to uncover the predictive power of valuation ratios such as Dividend Yields for future stock returns via significance tests implemented on simple linear regressions linking r_{t+1} to DY_t. The econometric complications that arise due to the presence of a persistent regressor together with endogeneity issues have generated a vast methodological literature aiming to improve inferences in such models commonly referred to as predictive regressions (e.g. Valkanov (2003), Lewellen (2004), Campbell and Yogo (2006), Jansson and Moreira (2006), Ang and Bekaert (2007) among numerous others).

Given the multitude of studies conducted over a variety of sample periods, methodologies, data definitions and frequencies it is difficult to extract a clear consensus on predictability. From the recent analysis of Campbell and Yogo (2006) there appears to be statistical support for some very mild DY based predictability with the latter having substantially declined in strength post 1995 (see also Lettau and Van Nieuwerburgh (2008)). Using monthly data over the 1946-2000 period Lewellen (2004) documented a rather stronger DY based predictability using a different methodology that was mainly concerned with small sample bias correction. See also Cochrane (2008) for a more general overview of this literature.

Our goal here is to reconsider this potential presence of predictability through our regime based methodology focusing on the DY predictor. More specifically, using growth in Industrial Production (IP) as our threshold variable proxying for aggregate macro conditions our aim is to assess whether the data support the presence of regime dependent predictability induced by good versus bad economic times. Theoretical arguments justifying the possible existence of episodic instability in predictability have been alluded to in the theoretical setting of Menzly, Santos and Veronesi (2004) and more recently Henkel, Martin and Nardari (2009) explored the issue empirically using Bayesian methods within a Markov-Switching setup. We will show that our approach leads to a novel view and interpretation of the predictability phenomenon and that its conclusions are robust across alternative sample periods. Moreover our findings may provide an explanation for the lack of robustness to the sample period documented in existing linearity based work. An alternative strand of the recent predictive regression literature or more generally the forecasting literature has also explored the issue of predictive instability through the allowance of time variation via structural breaks and the use of recursive estimation techniques. A general message that has come out from this research is the omnipresence of model instability and the important influence of time variation on forecasts (see Rapach and Wohar (2006), Rossi (2005, 2006), Timmermann (2008) amongst others). Our own research is also motivated by similar concerns but focuses on explicitly

identifying predictability episodes induced by a particular variable such as a business cycle proxy.

Our analysis will be based on the same CRSP data set as the one considered in the vast majority of predictability studies (value weighted returns for NYSE, AMEX and NASDAQ). Throughout all our specifications the dividend yield is defined as the aggregate dividends paid over the last 12 months divided by the market capitalisation and is logged throughout (LDY therefater). For robustness considerations we will distinguish between returns that include dividends and returns that exclude dividends. Finally, using the 90-day T-Bills all our inferences will also distinguish between raw returns and their excess counterparts. Following Lewellen (2004) we will restrict our sample to the post-war period. We will concentrate solely on monthly data since the regime specific nature of our models would make yearly or even quarterly data based inferences less reliable due to the potentially very small size of the sample. We will subsequently explore the robustness of our results to alternative sample periods.

Looking first at the stochastic properties of the dividend yield predictor over the 1950M1-2007M12 period it is clear that the series is highly persistent as judged by a first order sample autocorrelation coefficient of 0.991. A unit root test implemented on the same series unequivocally fails to reject the unit root null. The IP growth series is stationary as expected displaying some very mild first order serial correlation and clearly conforming to our assumptions about q_t in (1)-(2). Before proceeding with the detection of regime specific predictability we start by assessing return predictability within a linear specification as it has been done in the existing literature. Results across both raw and excess returns are presented in Table 9 below with VWRETD denoting the returns inclusive of dividends and VWRETX denoting the returns ex-dividends. The columns named as p and p_{HAC} refer to the standard and HAC based pvalues.

Table 9. Linear Predictability $r_{t+1} = \alpha_{DY} + \beta_{DY} LDY_t + u_{t+1}$

VWRETD	$\hat{\beta}_{DY}$	p_{HAC}	p	R^2	**VWRETX**	$\hat{\beta}_{DY}$	p_{HAC}	p	R^2
$1950-2007$	0.010	0.011	0.008	0.9%	$1950-2007$	0.008	0.054	0.046	0.4%
$1960-2007$	0.010	0.056	0.037	0.6%	$1960-2007$	0.008	0.142	0.110	0.3%
$1970-2007$	0.009	0.069	0.056	0.6%	$1970-2007$	0.007	0.170	0.148	0.2%
$1980-2007$	0.011	0.059	0.042	0.9%	$1980-2007$	0.009	0.131	0.103	0.5%
$1990-2007$	0.014	0.153	0.105	0.8%	$1990-2007$	0.001	0.207	0.152	0.5%
Excess					**Excess**				
$1950-2007$	0.009	0.025	0.019	0.7%	$1950-2007$	0.007	0.102	0.087	0.3%
$1960-2007$	0.007	0.210	0.169	0.2%	$1960-2007$	0.004	0.417	0.372	0.0%
$1970-2007$	0.006	0.269	0.240	0.1%	$1970-2007$	0.004	0.665	0.479	0.0%
$1980-2007$	0.007	0.253	0.208	0.2%	$1980-2007$	0.005	0.439	0.392	0.0%
$1990-2007$	0.013	0.198	0.138	0.6%	$1990-2007$	0.011	0.263	0.196	0.0%

The coefficient estimates of Table 9 refer to the OLS estimates of β_{DY} in the regression $r_{t+1} = \alpha + \beta_{DY} LDY_t + u_{t+1}$. Focusing first on the VWRETD series our results conform with the consensus that predictability has been vanishing from the late 80s onwards (see for instance Campbell and Yogo (2006)). The remaining pvalues suggest some mild predictability especially when considering the entire 1950-2007 sample range. Interestingly as we switch from raw to excess returns the picture changes considerably with most pvalues strongly pointing towards the absence of any predictability. Given these pvalue magnitudes it is difficult to conceive that any methodological improvements may reverse the big picture. Also worth pointing out is the fact that a conventional test for heteroskedasticity implemented on the above specifications failed to reject the null of no heteroskedasticity. This is particularly reassuring since one of our assumptions leading to our theoretical results in Propositions 1 and 2 ruled out the presence of heteroskedasticity.

Next, focusing on the returns that exclude dividend payments it is again the case that with pvalues as high as 0.665 the null of no predictability cannot be rejected. Results appear to also be robust across different starting periods except perhaps under the full 1950-2007 range under which we note a mild rejection of the null. It is also important to note that all results were robust across HAC versus non-HAC standard errors. This latter point is particularly important since our assumptions surrounding (1)-(2) rule out serial correlation and heteroskedasticity in u_t.

Overall the above linearity based results corroborate the view that predictability is at best mildly present and its strength appears to have declined. Perphaps more importantly Table 9 also suggests that one should be particularly cautious and worry about robustness considerations when assessing DY induced predictability of returns since findings may be extremely sensitive to data definitions, frequency and chosen sample period. At this stage it is also important to reiterate that our analysis in Table 9 is mainly meant to provide a comparison benchmark for our subsequent regime based inferences rather than reverse findings from the existing literature. This is also the reason why we do not explore outcomes based on alternative methodologies as developed in the recent econometric literature.

The fact that numerous studies documented a decline in predictability characterising the 90s could also be due to the fact that predictability kicks in during particular economic episodes. Table 10 below presents the results of our tests of the hypotheses $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$, $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $H_0^C : \beta_1 = \beta_2 = 0$ as applied to the VWRETD series ($*$ indicates rejection at 2.5%). Since results for the return series that exclude dividends as well as their excess counterparts were both qualitatively and quantitatively similar in what follows we concentrate solely on the VWRETD series.

Table 10. Regime Specific Predictability

	$SupWald^{A}$	$SupWald^{B,ivx}$			$W_T^{C,ivx}(\gamma=0)$	$W_T^{C,ivx}(\hat{\gamma})$
		$\delta=0.7$	$\delta=0.8$	$\delta=0.9$	$\delta=0.7$	$\delta=0.7$
$1950-2007$	20.75 (0.001)	26.75^{*}	28.87^{*}	30.21^{*}	10.57 (0.005)	6.77 (0.034)
$1960-2007$	18.98 (0.002)	23.24^{*}	23.40^{*}	23.46^{*}	8.16 (0.017)	4.09 (0.129)
$1970-2007$	17.73 (0.004)	21.64^{*}	21.82^{*}	21.77^{*}	7.62 (0.022)	6.17 (0.046)
$1980-2007$	24.52 (0.000)	27.73^{*}	28.60^{*}	28.96^{*}	10.84 (0.004)	8.00 (0.018)
$1990-2007$	28.87 (0.000)	29.52^{*}	30.18^{*}	31.10^{*}	7.89 (0.019)	20.05 (0.000)

The evidence presented in Table 10 comfortably points towards the presence of regime specific predictability since both H_0^A and H_0^B are strongly rejected. We also note that inferences based on $SupWald^{B,ivx}$ appear robust to alternative choices of δ in the construction of the IVX variable. Regarding the null given by $H_0^C : \beta_1 = \beta_2 = 0$ it is again strongly rejected under the assumption of equality of the intercepts with $W_T^C(\gamma)$ evaluated at the ad-hoc choice of $\gamma = 0$. If we operate under $\alpha_1 \neq \alpha_2$, results appear to be slightly less clearcut albeit mostly pointing towards rejection of the null (e.g. pvalue of 0.034). It is also interesting to note that unlike the linear case inferences appear to be robust to the starting period. One should be cautious however when interpreting inferences such as the ones based on the 1990-2007 period due to sample size limitations which are further exacerbated when fitting a threshold specification.

Recalling that the R^2's characterising the various linear specifications were clustered around values close to zero (see Table 9) it is also useful to highlight the remarkable jump in goodness of fit in our proposed threshold model presented in (17) below. Our results strongly point towards the presence of very strong predictability during *bad times* when the growth in IP (variable ΔLIP_t) is negative while no or very weak predictability during expansionary periods or normal times. More specifically, over the 1950-2007 period we have

$$\hat{r}_{t+1} = \begin{cases} 0.1606_{(0.0357)} + 0.0441_{(0.0107)} LDY_t & \Delta LIP_t \leq -0.0036,\ R_1^2 = 17.47\%,\ N_1 = 131 \\ 0.0135_{(0.0161)} + 0.0010_{(0.0045)} LDY_t & \Delta LIP_t > -0.0036,\ R_2^2 = 0.00\%,\ N_2 = 564 \end{cases} \tag{17}$$

with a joint R^2 of 3.88%. Estimated standard erros are in parentheses. Besides being interesting in its own right this result may also help explain the conflicting results obtained in the recent literature where the samples considered included or excluded data on the late 90s and 00s, a period with few recessions. Even with the reduction in the sample size it is quite remarkable that the goodness of fit can jump from a magnitude close to zero to about 17% in one subset. Overall our results strongly support DY based predictability in US returns but occurring solely during *bad times*. Note for instance that more than half of the periods during which $\Delta LIP_t \leq -0.0036$ coincide with the NBER recessions. The strength of this predictability is very strong and unlikely to be sensitive to the methodology or our assumptions. Interestingly and through a different methodology, our findings about the presence of strong return predictability during bad times also corroborate the findings in Henkel, Martin and Nardari (2009). Using Bayesian inference techniques on a Markov Switching VAR setup in which they

consider multiple predictors in addition to the Dividend Yield the authors document a substantial jump in predictive strength of variables such as DY, short term rates, term structure etc during recessions.

6 Conclusions

The goal of this paper was to develop inference methods useful for detecting the presence of regime specific predictability in predictive regressions. We obtained the limiting distributions of a series of Wald statistics designed to test the null of linearity versus threshold type nonlinearity, the joint null of linearity and no predictability and the null of no predictability induced by x. One important feature of the limiting distribution that arises in the first case is the fact that it does not depend on any unknown nuisance parameters thus making it straightforward to use. This is an unusual occurence in this literature where under a purely stationary framework (as opposed to a nearly integrated one) it is well known that limiting distributions typically depend on unknown population moments of the underlying models.

Our empirical application also leads to the interesting result that US return series are clearly predictable using valuation ratios such as DY but this predictability kicks in solely during bad times and would therefore be masked in studies that operate within linear specifications.

Finally, it is worth mentioning some important extensions to the present work. A useful extension we are currently considering involves introducing long horizon variables into (1)-(2). This would offer an interesting parallel to the linear predictive regression literature which has often distinguished long versus short horizon predictability. Other important extensions include extending (1)-(2) to allow for more than two regimes following some of the methods developed in Gonzalo and Pitarakis (2002) while further statistical properties (e.g. confidence intervals) of objects such as the estimated threshold parameter may be explored using the subsampling methodology of Gonzalo and Wolf (2005).

A key assumption under which we have operated ruled out heteroskedasticity and serial correlation in u_t. As our empirical application has documented however our results can continue to be extremely useful despite this limitation. This restriction is in fact the norm rather than the exception in any work that introduced nonlinearities parametrically or nonparametrically in models that contain persistent variables. Albeit challenging, we expect future work to also be directed towards tackling these issues.

APPENDIX

LEMMA 1: Under assumptions A1-A2 and as $T \to \infty$ we have (a) $\frac{\sum I_{1t}}{T} \xrightarrow{p} \lambda$, (b) $\frac{\sum x_t}{T^{\frac{3}{2}}} \Rightarrow \int_0^1 K_c(r)dr$, (c) $\frac{\sum x_t^2}{T^2} \Rightarrow \int_0^1 K_c^2(r)dr$, (d) $\frac{\sum x_{t-1}v_t}{T} \Rightarrow \int_0^1 K_c(r)dB_v(r)+\lambda_{vv}$. (e) $\frac{\sum x_{t-1}u_t}{T} \Rightarrow \int_0^1 K_c(r)dB_u(r,1)$, (f) $\frac{\sum x_t^2 I_{1t}}{T^2} \Rightarrow \lambda \int_0^1 K_c^2(r)dr$, (g) $\frac{\sum x_t I_{1t}}{T^{\frac{3}{2}}} \Rightarrow \lambda \int_0^1 K_c(r)dr$, (h) $\frac{\sum_{t=1}^{[Tr]} u_t I_{1t-1}}{\sqrt{T}} \Rightarrow B_u(r,\lambda)$, (i) $\frac{\sum x_{t-1}u_t I_{1t-1}}{T} \Rightarrow \int_0^1 K_c(r)dB_u(r,\lambda)$

PROOF OF LEMMA 1: (a) By assumptions A1-A2, I_{1t} is strong mixing with the same mixing numbers as q_t. The result then follows from a suitable law of large numbers (see White (2001, Sections 3.3-3.4)). (b)-(e) Under our assumptions A1-A2, the results follow directly from Lemma 3.1 in Phillips (1988). (f) Letting $X_{T,t} = x_t/\sqrt{T}$ and $X_T(r) = x_{[Tr]}/\sqrt{T}$ we can rewrite (f) as

$$\frac{1}{T}\sum X_{T,t}^2 I_{1t} = \lambda \frac{1}{T}\sum X_{T,t}^2 + \frac{1}{T}\sum X_{T,t}^2 (I_{1t}-\lambda). \tag{18}$$

Under A1-A2 and requiring $E|e_t|^p < \infty$ for some $p \geq 4$ we can make use of the strong approximation result $\sup_{r\in[0,1]} |X_T(r) - K_c(r)| = o_p(T^{-a})$ with $a = (p-2)/2p$ (see Lemma A.3 in Phillips (1998) and Phillips and Magdalinos (2007)) to obtain

$$\frac{1}{T}\sum X_{T,t}^2 = \int_0^1 K_c^2(r)dr + o_p(T^{-a}). \tag{19}$$

Indeed,

$$\begin{aligned}
\left|\int_0^1 X_T(r)^2 dr - \int_0^1 K_c(r)^2 dr\right| &\leq \int_0^1 \left|X_T(r)^2 - K_c(r)^2\right| dr \\
&= \int_0^1 |X_T(r) - K_c(r)|\,|X_T(r) + K_c(r)|\, dr \\
&\leq \sup_r |X_T(r) - K_c(r)| \left(\sup_r |X_T(r)| + \sup_r |K_c(r)|\right) \\
&= o_p(T^{-a}).
\end{aligned} \tag{20}$$

The above then leads to

$$\frac{1}{T}\sum X_{T,t}^2 I_{1t} - \lambda \int_0^1 K_c(r)^2 dr = \frac{1}{T}\sum X_{T,t}^2 (I_{1t}-\lambda) + o_p(T^{-a}) \tag{21}$$

holding uniformly $\forall \lambda \in \Lambda$. Finally, given that $\sup_{r\in[0,1]} |X_T(r)| = O_p(1)$ together with the fact that the result in (a) also holds uniformly over λ (see Lemma 1 in Hansen (1996)) we have $\sup_\lambda |\frac{1}{T}\sum X_{T,t}^2 I_{1t} - \lambda \int_0^1 K_c(r)^2 dr| = o_p(1)$ implying the required result. (g) Follows identical lines to the proof of (f). (h)-(i) Since our assumptions satify their Assumption 2 the result in (h) is Theorem 1 of Caner and Hansen (2001) while our result in (i) follows along the same lines as Theorem 2 of Caner and Hansen (2001).

PROOF OF PROPOSITION 1: It is initially convenient to reformulate $W_T^A(\lambda)$ under H_0^A as

$$\begin{aligned}
W_T^A(\lambda) = {} & [u'X_1 - u'X(X'X)^{-1}X_1'X_1][X_1'X_1 - X_1'X_1(X'X)^{-1}X_1'X_1]^{-1} \\
& [X_1'u - (X_1'X_1)(X'X)^{-1}X'u]/\hat{\sigma}_u^2.
\end{aligned} \tag{22}$$

With $D_T = diag(\sqrt{T}, T)$ we can write

$$D_T^{-1} X_1{}' X_1 D_T^{-1} \quad = \quad \begin{pmatrix} \frac{\sum I_{1t}}{T} & \frac{\sum x_t I_{1t}}{T^{\frac{3}{2}}} \\ \frac{\sum x_t I_{1t}}{T^{\frac{3}{2}}} & \frac{\sum x_t^2 I_{1t}}{T^2} \end{pmatrix} \tag{23}$$

and using Lemma 1 we have the following weak convergence results

$$D_T^{-1} X_1{}' X_1 D_T^{-1} \quad \Rightarrow \quad \begin{pmatrix} \lambda & \lambda \int_0^1 K_c(r) dr \\ \lambda \int_0^1 K_c(r) dr & \lambda \int_0^1 K_c^2(r) dr \end{pmatrix} \equiv \lambda \int_0^1 \overline{K}_c(r) \overline{K}_c(r)' \tag{24}$$

and

$$D_T^{-1} X' X D_T^{-1} \quad \Rightarrow \quad \int_0^1 \overline{K}_c(r) \overline{K}_c(r)' \tag{25}$$

where $\overline{K}_c(r) = (1, K_c(r))$. It now follows from the continuous mapping theorem that

$$[D_T^{-1} X_1{}' X_1 D_T^{-1} - D_T^{-1} X_1{}' X_1 (X'X)^{-1} X_1{}' X_1 D_T^{-1}]^{-1} \quad \Rightarrow \quad \frac{1}{\lambda(1-\lambda)} \left(\int_0^1 \overline{K}_c(r) \overline{K}_c(r)' \right)^{-1}. \tag{26}$$

We next focus on the limiting behaviour of $D_T^{-1} X' u$ and $D_T^{-1} X_1' u$. Looking at each component separately, setting $\sigma_u^2 = 1$ for simplicity and no loss of generality and using Lemma 1, we have

$$D_T^{-1} X_1{}' u \quad = \quad \begin{pmatrix} \frac{\sum I_{1t} u_{t+1}}{\sqrt{T}} \\ \frac{\sum x_t I_{1t} u_t}{T} \end{pmatrix} \Rightarrow \begin{pmatrix} B_u(r, \lambda) \\ \int_0^1 K_c(r) dB_u(r, \lambda) \end{pmatrix} \tag{27}$$

and

$$D_T^{-1} X' u \quad = \quad \begin{pmatrix} \frac{\sum u_{t+1}}{\sqrt{T}} \\ \frac{\sum x_t u_t}{T} \end{pmatrix} \Rightarrow \begin{pmatrix} B_u(r, 1) \\ \int_0^1 K_c(r) dB_u(r, 1) \end{pmatrix}. \tag{28}$$

The above now allows us to formulate the limiting behaviour of $D_T^{-1} X_1{}' u - \lambda D_T^{-1} X' u$ as

$$D_T^{-1} X_1{}' u - \lambda D_T^{-1} X' u \quad \Rightarrow \quad \int_0^1 \overline{K}_c(r) dG_u(r, \lambda) \tag{29}$$

where $G_u(r, \lambda) = B_u(r, \lambda) - \lambda B_u(r, 1)$. The result in (3) follows straightforwardly through the use of the continuous mapping theorem and standard algebra.

PROOF OF PROPOSITION 2: We rewrite our most general unrestricted specification in (1) as $y = \alpha_1 I_1 + \beta_1 x_1 + \alpha_2 I_2 + \beta_2 x_2 + u$. Within this notation lower case $x_i's$ stack $x_t I_{it}$ while the $I_i's$ stack I_{it} for $i = 1, 2$. We also recall that $X_i = (I_i \;\; x_i)$ for $i = 1, 2$. It is now convenient to reformulate (1) as $y = \alpha + \beta x + X_2 \eta + u$ where $\alpha = \alpha_1$, $\beta = \beta_1$ and $\eta = (\gamma, \delta)'$ with $\gamma = \alpha_2 - \alpha_1$ and $\delta = \beta_2 - \beta_1$ so that within this alternative parameterization $H_0^A : \eta = 0$ and $H_0^B : \eta = 0, \beta = 0$. Next, consider a most general (MG) model containing $(\mathbb{1} \;\; x \;\; X_2) = (X \;\; X_2)$, a partially restricted (PR) version containing $X = (\mathbb{1} \;\; x)$ and a fully restricted (FR) version containing just the vector of ones

$\mathbb{1}$. From standard projection algebra the sum of squared errors corresponding to each specification are $SSE_{MG} = y'M_{X,X_2}y$, $SSE_{PR} = y'M_Xy$ and $SSE_{FR} = y'M_{\mathbb{1}}y$ where $M_X = I - X(X'X)^{-1}X'$ and $M_{X,X_2} = M_X - M_XX_2(X_2'M_XX_2)^{-1}X_2'M_X$. It now trivially follows that we can write the Wald statistics corresponding to each hypothesis as $W_T^A(\lambda) = [y'M_Xy - y'M_{X,X_2}y]/\hat{\sigma}_u^2$ (PR against MG), $W_T^B(\lambda) = [y'M_{\mathbb{1}}y - y'M_{X,X_2}y]/\hat{\sigma}_u^2$ (FR against MG) and $W_T(\beta = 0) = [y'M_{\mathbb{1}}y - y'M_Xy]/\hat{\sigma}_{lin}^2$ (FR against PR). It can now immediately be observed that $W_T^B(\lambda) = W_T^A(\lambda) + (\hat{\sigma}_{lin}^2/\hat{\sigma}_u^2)W_T(\beta = 0)$. Under the null hypothesis $(\hat{\sigma}_{lin}^2/\hat{\sigma}_u^2) \xrightarrow{p} 1$ and therefore in large samples $W_T^B(\lambda) \approx W_T(\beta = 0) + W_T^A(\lambda)$ and $\sup_\lambda W_T^B(\lambda) \approx W_T(\beta = 0) + \sup_\lambda W_T^A(\lambda)$ as required. To obtain the limiting distribution in (5) it now suffices to use the results presented in Lemma 1 together with the CMT along lines identical to those in the proof of Proposition 1.

PROOF OF PROPOSITION 3: Follows directly from (11)-(12), Theorem 3.8 in Phillips and Magdalinos (2009), Lemma 1 and the use of the Continuous Mapping Theorem. Note that Theorem 3.8 in Phillips and Magdalinos (2009) has been obtained within a model with no fitted intercept however Kostakis, Magdalinos and Stamatogiannis (2010) and Magdalinos (2010) also established its validity in the more general setting that includes a constant term.

PROOF OF PROPOSITION 4: (i)-(ii) Letting W stack the elements of $[I_{1t}\ I_{2t}]$ and $M_W = I - W(W'W)^{-1}W'$ we can write (1) as $y^* = \beta_1x_1^* + \beta_2x_2^* + u^*$ with $x_i^* = M_Wx_i$ and where x_i stacks x_tI_{it}. We also let $\hat{\sigma}_*^2$ denote the corresponding residual variance from this canonical form. Under $H_0^C : \beta_1 = \beta_2 = 0$ it now immediately follows that $W_T^C(\lambda) = \sum_{i=1}^2 u'M_Wx_i(x_i'M_Wx_i)^{-1}x_i'M_Wu/\hat{\sigma}_*^2$. Focusing on $u'M_Wx_i$ we have $u'M_Wx_i/T = (\sum u_tx_tI_{it}/T) - (\sum u_tI_{1t}/\sqrt{T}\sum x_tI_{it}/T^{3/2})/(\sum I_{it}/T)$ so that using the CMT and our intermediate results in Proposition 1 we have $u'M_Wx_1/T \Rightarrow \int K_c^*(r)dB_u(r,\lambda)$ and $u'M_Wx_2/T \Rightarrow \int K_c^*(r)(dB_u(r,1) - dB_u(r,\lambda))$. Similarly $x_1'M_Wx_1/T^2 \Rightarrow \lambda\int K_c^*(r)^2$ and $x_2'M_Wx_2/T^2 \Rightarrow (1-\lambda)\int K_c^*(r)^2$. Combining and rearranging with the use of the CMT leads to

$$W_T^C(\lambda) \ \Rightarrow\ \frac{[\int K_c^*(r)dB_u(r,1)]^2}{\sigma_u^2\int K_c^*(r)^2} + \frac{[\int K_c^*(r)dG_u(r,\lambda)]^2}{\sigma_u^2\lambda(1-\lambda)\int K_c^*(r)^2} \equiv J_1 + J_2(\lambda). \tag{30}$$

Operating under the assumption that λ is a known constant, the second component in the rhs of (30) is a $\chi^2(1)$ random variable due to the independence of $K_c^*(r)$ and $G_u(r,\lambda)$. The result in (30) holds for any $\lambda \in (0,1)$ when $\alpha_1 = \alpha_2$ and for either λ_0 or $\hat{\lambda}$ when $\alpha_1 \neq \alpha_2$. Note that the T-consistency of $\hat{\lambda}$ estimated from the null model follows directly from Hansen (2000) or Gonzalo and Pitarakis (2002). (iii) The $\chi^2(2)$ outcome follows from three not necessarily related properties: (a) $J_2(\lambda)$ is $\chi^2(1)$ (see our discussion around (4)), (b) the well known fact that J_1 is $\chi^2(1)$ under exogeneity and (c) the independence of J_1 and $J_2(\lambda)$ which is discussed in our supplementary appendix.

REFERENCES

Andrews, D. W. K., (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, Vol. 61, pp. 821-856.

Ang, A. and Bekaert, G. (2007), "Stock Return Predictability: Is it There?," *Review of Financial Studies*, Vol. 20, pp. 651-707.

Bandi, F. and Perron, B. (2008), "Long-Run Risk-Return Trade-Offs," *Journal of Econometrics*, Vol. 143, pp. 349-74.

Campbell, J. Y. and Yogo, M. (2006), "Efficient tests of stock return predictability," *Journal of Financial Economics,* Vol. 81, pp. 27-60.

Caner, M. and Hansen, B. E. (2001), "Threshold Autoregression with a Unit Root," *Econometrica,* Vol. 69, pp. 1555-1596.

Cavanagh, C. L., G. Elliott, and Stock, J. H. (1995), "Inference in Models with Nearly Integrated Regressors," *Econometric Theory*, Vol. 11, pp. 1131-1147.

Chan, N. (1988), "The Parameter Inference for Nearly Nonstationary Time Series The Parameter Inference for Nearly Nonstationary Time Series," *Journal of the American Statistical Association*, Vol. 83, pp. 857-862.

Cochrane, J. H. (2008), "The dog that did not bark: a defense of return predictability," *Review of Financial Studies,* Vol. 21, pp. 1533-1575.

Elliott, G. (1998), "On the Robustness of Cointegration Methods when Regressors Almost Have Unit Roots," *Econometrica*, Vol. 66, pp. 149-158.

Estrella, A. (2003), "Critical Values and P Values of Bessel Process Distributions: Computation and Application to Structural Break Tests," *Econometric Theory*, Vol. 19, 1128-1143.

Gonzalo, J. and Pitarakis, J. (2002), "Estimation and Model Selection Based Inference in Single and Multiple Threshold Models," *Journal of Econometrics*, Vol. 110, pp. 319-352.

Gonzalo, J. and Wolf, M. (2005), "Subsampling Inference in Threshold Autoregressive Models," *Journal of Econometrics*, Vol. 127, pp. 201-224.

Hansen, B. E. (1996), "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, Vol. 64, pp. 413-430.

Hansen, B. E. (1997), "Approximate asymptotic p-values for structural change tests," *Journal of Business and Economic Statistics*, Vol. 15, pp. 60-67.

Hansen, B. E. (1999), "Testing for Linearity," *Journal of Economic Surveys*, Vol. 13, pp. 551-576.

Hansen, B. E. (2000), "Sample splitting and threshold estimation," *Econometrica*, Vol. 68, pp. 575-603.

Henkel, S. J., Martin, J. S. and Nardari, F. (2009), "Time-Varying Short-Horizon Return Predictability," AFA 2008 New Orleans Meetings Paper. Available at SSRN: http://ssrn.com/abstract=1101944.

Jansson, M. and Moreira, M. J. (2006), "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica*, Vol. 74, pp. 681-714.

Kostakis, A., Magdalinos, A. and Stamatogiannis, M. (2010), "Robust Econometric Inference for Stock Return Predictability," Unpublished Manuscript, University of Nottingham, UK.

Lettau, M. and Van Nieuwerburgh, S. (2008), "Reconciling the return predictability evidence," *Review of Financial Studies,* Vol. 21, pp. 1607-1652.

Lewellen, J. (2004), "Predicting returns with financial rations," *Journal of Financial Economics*, Vol. 74, pp. 209-235.

Magdalinos, A. (2010), Personal Communication.

Menzly, L., T. Santos and Veronesi, P. (2004), "Understanding predictability," *Journal of Political Economy,* Vol. 112, pp.

Park, J.Y. and Phillips, P. C. B. (1988), "Statistical Inference in Regressions With Integrated Processes: Part 1," *Econometric Theory*, Vol. 4, pp. 468-497.

Petruccelli, J.D. (1992), "On the approximation of time series by threshold autoregressive models," *Sankhya, Series B*, Vol. 54, pp. 54-61.

Phillips, P. C. B. (1988), "Regression Theory for Near-Integrated Time Series," *Econometrica*, Vol. 56, pp. 1021-1043.

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

Phillips, P. C. B. (1998), "New Tools for Understanding Spurious Regressions," *Econometrica*, Vol. 66, pp. 1299-1325.

Phillips, P. C. B. and Magdalinos, A. (2007), "Limit theory for moderate deviations from a unit root under weak dependence," *Journal of Econometrics*, Vol. 136, pp. 115-130.

Phillips, P. C. B. and Magdalinos, A. (2009), "Econometric Inference in the Vicinity of Unity," *Singapore Management University, CoFie Working Paper No. 7.*

Pitarakis, J. (2008), "Threshold Autoregressions with a unit root: Comment," *Econometrica*, Vol. 76, pp. 1207-1217.

Rapach, D. E. and Wohar, M. E. (2006), "Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns," *Journal of Financial Econometrics*, Vol. 4, pp. 238-274.

Rossi, B. (2005), "Optimal Tests for Nested Model Selection with underlying Parameter Instability," *Econometric Theory*, Vol. 21, pp. 962-990.

Rossi, B. (2006), "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability," *Macroeconomic Dynamics*, Vol. 10, pp. 20-38.

Saikkonen, P. (1991), "Asymptotically Efficient Estimation of Cointegrating Regressions," *Econometric Theory,* Vol. 7, pp. 1-21.

Saikkonen, P. (1992), "Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation," *Econometric Theory,* Vol. 8, pp. 1-27.

Timmermann, A. (2008), "Elusive Return Predictability," *International Journal of Forecasting*, Vol. 24, pp. 1-18.

Valkanov, R. (2003), "Long-horizon regressions: theoretical results and applications," *Journal of Financial Economics,* Vol. 68, pp. 201-232.

Van der Vaart, J. and Wellner, J. (1996), Weak Convergence and Empirical Processes: With Applications to Statistics. Springer-Verlag, New-York.

White, H. (2001), Asymptotic Theory for Econometricians. Academic Press.