Inferring the Predictability Induced by a Persistent Regressor in a Predictive Threshold Model*

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

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

February 16, 2016

Abstract

We develop tests for detecting possibly episodic predictability induced by a persistent predictor. Our framework is that of a predictive regression model with threshold effects and our goal is to develop operational and easily implementable inferences when one does not wish to impose à priori restrictions on the parameters of the model other than the slopes corresponding to the persistent predictor. Differently put our tests for the null hypothesis of no predictability against threshold predictability remain valid without the need to know whether the remaining parameters of the model are characterised by threshold effects or not (e.g. shifting versus non-shifting intercepts). One interesting feature of our setting is that our test statistics remain unaffected by whether some nuisance parameters are identified or not. We subsequently apply our methodology to the predictability of aggregate stock returns with valuation ratios and document a robust countercyclicality in the ability of some valuation ratios to predict returns in addition to highlighting a strong sensitivity of predictability based results to the time period under consideration.

Keywords: Predictive Regressions, Threshold Effects, Predictability of Stock Returns.

*We wish to thank seminar participants at the University of Nottingham, Cambridge University, Durham University, LSE, Surrey University, Barcelona GSE, ESSEC Paris, Oxford University, the ESEM 2014 meetings in Toulouse, the 2014 IAAE conference at Queen Mary and the ENSAI for their comments. We are grateful to Dietmar Ferger, Tassos Magdalinos and Peter Phillips for some very helpful comments. We are also particularly indebted to Oscar Martinez for his review and suggestions. All errors are ours. The first author wishes to thank the Spanish MINECO under grant ECO2013-46395 for financial support and the second author thanks the ESRC for financial support under grant RES-000-22-3983. Address for correspondence: Jean-Yves Pitarakis, University of Southampton, School of Social Sciences, Department of Economics, SO17 1BJ, United-Kingdom.

1 Introduction

Predictive regressions are simple regression models in which a highly persistent variable is used as a predictor of a noisier time series. The econometric difficulties that arise due to the combination of a persistent regressor and possible endogeneity have generated an enormous literature aiming to improve inferences in such settings. Common examples include the predictability of stock returns with valuation ratios, the predictability of GDP growth with interest rates amongst numerous others (see for instance Valkanov (2003), Lewellen (2004), Campbell and Yogo (2006), Jansson and Moreira (2006), Rossi (2007), Bandi and Perron (2008), Ang and Bekaert (2008), Wei and Wright (2013) and more recently Kostakis, Magdalinos and Stamatogiannis (2015, KMS2015 thereafter)).

In a recent paper Gonzalo and Pitarakis (2012) have extended the linear predictive regression model into one that allows the strength of predictability to vary across economic episodes such as expansions and recessions. This was achieved through the inclusion of threshold effects which allowed the parameters of the model to switch across regimes driven by an external variable. Within this piecewise linear setting the authors developed a series of tests designed to detect the presence of threshold effects in *all* the parameters of the model by maintaining full linearity within the null hypotheses (i.e. restricting both intercepts and slopes to be stable throughout the sample). Differently put this earlier work was geared towards uncovering regimes within a predictive regression setting rather than determining the predictability of a particular predictor per se.

The goal of this paper is to develop a toolkit that will allow practitioners to test the null hypothesis of no predictability induced by a persistent regressor explicitly without restricting the remaining parameters of the model (e.g. intercepts may or may not exhibit threshold effects). Indeed, a researcher may wish to assess the presence of predictability induced solely by some predictor x_t while remaining agnostic about the presence or absence of regimes in the remaining parameters. Moreover, in applications involving return predictability with valuation ratios such as the dividend yield and a threshold variable proxying the business cycle, rejection of the null of no predictability on the basis of a null hypothesis that restricts all the parameters of the model as in Gonzalo and Pitarakis (2012) may in fact be driven by the state of the business cycle rather than the regime specific predictability induced by the dividend yield itself.

The type of inference we consider in this paper naturally raises important identification issues which we address by exploring the feasibility of conducting inferences on the relevant slope parameters that are immune to any knowledge about the behaviour of the intercepts and in particular to whether the latter are subject to regime shifts or not. Our null hypothesis of interest here allows for the possibility of having *nuisance* parameters that may or may not switch across regimes. This is fundamentally different from the setting considered in Gonzalo and Pitarakis (2012) where the intercepts were also restricted to be equal under the null hypothesis of no predictability and the only nuisance parameter was the unknown threshold parameter itself.

Our proposed inferences are based on a standard Wald type test statistic whose distribution we derive under the null hypothesis of no predictability induced by a highly persistent regressor. The limiting distribution of our test statistic evaluated at a particular location of the threshold parameter is then shown to be immune to whether the remaining parameters of the model shift or not. Since the limiting distribution in question depends on a series of nuisance parameters it is not directly usable for inferences unless one wishes to impose an exogeneity assumption on the predictor. Using an Instrumental Variable approach we subsequently introduce a modified Wald statistic whose new distribution is shown to be standard and free of nuisance parameters under a very general setting.

The plan of the paper is as follows. Section 2 presents our operating model and the underlying probabilistic assumptions. Section 3 develops the large sample inferences. Section 4 illustrates their properties and usefulness via a rich set of simulations. Section 5 applies our proposed methods to the predictability of aggregate US equity returns using a wide range of valuation ratios and threshold variables and Section 6 concludes.

2 The Model and Assumptions

We operate within the same setting as in Gonzalo and Pitarakis (2012). Our predictive regression model with threshold effects or *Predictive Threshold Regression* (PTR) is given by

$$y_{t+1} = (\alpha_1 + \beta_1 x_t) I(q_t \leq \gamma) + (\alpha_2 + \beta_2 x_t) I(q_t > \gamma) + u_{t+1} \tag{1}$$

where the highly persistent predictor x_t is modelled as the nearly integrated process

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

with $c > 0$ and $q_t = \mu_q + u_{qt}$ denoting the stationary threshold variable with distribution function $F(.)$. Before proceeding further it is useful to reformulate our specification in (1) in matrix form. In doing so we make use of the property $I(q_t \leq \gamma) \equiv I(F(q_t) \leq \lambda) \equiv I_{1t}$ and $I(q_t > \gamma) \equiv I(F(q_t) > \lambda) \equiv I_{2t}$ with $\lambda \equiv F(\gamma)$ so that in what follows the threshold parameter can be referred to as as either γ or λ interchangeably. We now rewrite (1) as

$$y = Q_\lambda \alpha + X_\lambda \beta + u \tag{3}$$

with $Q_\lambda = [I_1 \ I_2]$ and $X_\lambda = [x_1 \ x_2]$ stacking the elements $(I_{1t} \ I_{2t})$ and $(x_t I_{1t} \ x_t I_{2t})$ respectively and $\alpha = (\alpha_1 \ \alpha_2)'$, $\beta = (\beta_1 \ \beta_2)'$. Given the assumptions that will be imposed on q_t (e.g. strict stationarity and ergodicity) it is useful to note that $E[I_{1t}] = \lambda$ and $E[I_{2t}] = 1 - \lambda \ \forall t$ and throughout the paper it will be understood that $\lambda \in \Lambda = [\underline{\lambda}, \overline{\lambda}]$ with $0 < \underline{\lambda} < \lambda < \overline{\lambda} < 1$. Note that (1) is the same parameterisation as the one used in Gonzalo and Pitarakis (2012) but its key features are repeated here for self containedness considerations. When relevant we will also refer to the true value of the threshold parameter as either γ_0 or λ_0.

Our main goal is to focus on the sole predictive power of x_t without imposing any restrictions on the α's. Note for instance that a null hypothesis such as $\alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ may be rejected solely due to $\alpha_1 \neq \alpha_2$ while continuing to be compatible with an environment in which x_t has no predictive content. It is this aspect that we wish to address in the present paper whose goal is to develop inferences about the β's without imposing any constraints on the α's in the sense that they may or may not be regime dependent within the underlying DGP. More specifically we will be interested in exploring testing strategies for testing the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ while allowing the α's to be free in the background. This is an important departure from the framework in Gonzalo and Pitarakis (2012) where we considered Sup over λ type tests of various null hypotheses which were also restricting the intercepts themselves in addition to β_1 and β_2 (e.g. $\alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$). More importantly in this paper our inferences will be based on a Wald statistic evaluated at a particular estimator of the threshold parameter (as opposed to taking its supremum over λ) which ensures that its limiting distribution under $\beta_1 = \beta_2 = 0$ is unaffected by whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ and is nuisance parameter free.

We next outline our operating assumptions regarding the probabilistic properties of u_t, v_t, q_t and their joint interactions. Throughout this paper we let the random disturbance v_t be described by the linear process $v_t = \Psi(L)e_{vt}$ with the polynomial $\Psi(L) = \sum_{j=0}^{\infty} \Psi_j L^j$ having $\Psi(1) \neq 0$, $\Psi_0 = 1$ and absolutely summable coefficients. We also let $\zeta_t = (u_t, e_{vt})'$ and introduce the filtration $\mathcal{F}_t = \sigma(\zeta_s, u_{qs} | s \leq t)$.

ASSUMPTIONS *A1: $E[\zeta_t|\mathcal{F}_{t-1}] = 0$, $E[\zeta_t\zeta_t'|\mathcal{F}_{t-1}] = \widetilde{\Sigma} > 0$, $\sup_t E\zeta_{it}^4 < \infty$. A2: The sequence $\{u_{qt}\}$ is strictly stationary, ergodic, strong mixing with mixing numbers α_m such that $\sum_{m=1}^{\infty} \alpha^{\frac{1}{m}-\frac{1}{r}} < \infty$ for some $r > 2$. A3: The probability density function $f_q(.)$ of q_t is bounded away from zero and ∞ over each bounded set.*

Assumption A1 requires the error process driving (1) to be a martingale difference sequence with respect to $\mathcal{F}_t$ hence ruling out serial correlation in u_t (but not in v_t or q_t) while also imposing conditional homoskedasticity. Both v_t and q_t are allowed to be sufficiently general dependent processes. This setting mimics closely the standard framework used in the predictive regression literature (e.g. Campbell and Yogo (2006), Jansson and Moreira (2006)) and is in fact slightly more general since we do allow v_t to be serially correlated. At this stage it is also important to clarify our stance regarding the joint interactions of our variables. Our assumptions about the dependence structure of the random disturbances together with the finiteness of moments requirements imply that a Functional Central Limit Theorem holds for $w_t = (u_t, u_t I_{1t-1}, v_t)$. More formally $T^{-\frac{1}{2}} \sum_{t=1}^{[Tr]} w_t \Rightarrow (B_u(r), B_u(r,\lambda), B_v(r)' = BM(\Omega)$ with $\Omega = \sum_{k=-\infty}^{\infty} E[w_0 w_k']$. Our analysis will impose a particular structure on Ω which governs and restricts the joint interactions of u_t, v_t and q_t. More specifically we impose

$$\Omega \;=\; \begin{pmatrix} \sigma_u^2 & \lambda\sigma_u^2 & \sigma_{uv}\Psi(1) \\ \lambda\sigma_u^2 & \lambda\sigma_u^2 & \lambda\sigma_{uv}\Psi(1) \\ \sigma_{uv}\Psi(1) & \lambda\sigma_{uv}\Psi(1) & \sigma_e^2\Psi(1)^2 \end{pmatrix} \tag{4}$$

where $\sigma_u^2 = E[u_t^2]$, $\sigma_e^2 = E[e_{vt}^2]$ and since $E[u_t e_{v,t-j}] = 0$ we also write $\sigma_{uv} = E[u_t v_t] = E[u_t e_{vt}] = \sigma_{ue}$. The chosen structure of Ω is general enough to encompass the standard setting used in the linear predictive regression literature that typically imposes $\{u_t, v_t\}$ to be a martingale difference sequence and u_t and v_t solely contemporaneously correlated. Our assumptions allow us to operate within a similar environment while also permitting the shocks to the threshold variable to be contemporaneously correlated with u_t and/or v_t. As in Caner and Hansen (2001) and Pitarakis (2008), $B_u(r,\lambda)$ refers to a two-parameter Brownian Motion which is a zero mean Gaussian process with covariance kernel $(r_1 \wedge r_2)(\lambda_1 \wedge \lambda_2)\sigma_u^2$ so that we implicitly also operate under the requirement that $E[u_t^2|q_{t-1}, q_{t-2}, \ldots] = \sigma_u^2$ as well as $E[u_t v_t|q_{t-1}] = E[u_t v_t] \equiv \sigma_{uv}$ and $E[u_t v_{t-k}|q_{t-1}, q_{t-2}, \ldots] = 0 \;\forall k \geq 1$. Given our nearly integrated specification for x_t and A1-A3 above it is also clear (see Phillips (1988)) that $x_{[Tr]}/\sqrt{T} \Rightarrow J_c(r)$ with

$J_c(r) = B_v(r) + c\int_0^r e^{(r-s)c} B_v(s)ds$ denoting a scalar Ornstein-Uhlenbeck process. For later use we also define the demeaned versions of $J_c(r)$ and $B_u(r)$ as $J_c^*(r) = J_c(r) - \int J_c(r)$ and $B_u^*(r) = B_u(r) - \int B_u(r)$.

3 Large Sample Inference

Since within model (1) the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is compatible with either $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ in a first instance it will be important to establish the large sample properties of our threshold parameter estimator $\hat{\lambda}$ (or $\hat{\gamma}$) under the two alternative scenarios on the intercepts. As our focus is on inferences about β and mainly for notational convenience it will also be useful to reparameterise (3) in its canonical form. More specifically, letting $M_Q = I - Q_\lambda(Q'_\lambda Q_\lambda)^{-1}Q_\lambda$ we can equivalently express (3) as

$$y^* \;=\; X_\lambda^* \, \beta + u^* \tag{5}$$

with $y^* = M_Q y$, $X_\lambda^* = M_Q X_\lambda$ and $u^* = M_Q u$.

3.1 Threshold Parameter Estimation

The threshold parameter estimator we consider throughout this paper is based on the least squares principle and defined as

$$\widehat{\lambda} \;=\; \arg\min_{\lambda} S_T(\lambda) \tag{6}$$

with $S_T(\lambda)$ denoting the concentrated sum of squared errors function obtained from (3) or (5) under the restriction $\beta_1 = \beta_2 = 0$ i.e. $S_T(\lambda) = y' M_Q y$. Recall that throughout this paper we use $\hat{\lambda}$ and $\hat{\gamma} = \arg\min_\gamma S_T(\gamma)$ interchangeably. Naturally, the behaviour of $\hat{\lambda}$ is expected to depend on whether the underlying true model has $\alpha_1 \neq \alpha_2$ (i.e. identified threshold parameter) or $\alpha_1 = \alpha_2$ in which case λ vanishes from the true model. The following Proposition summarises the large sample behaviour of $\hat{\lambda}$ under the two scenarios.

Proposition 1. *Under Assumptions A1-A3, $H_0 : \beta_1 = \beta_2 = 0$ and as $T \to \infty$ we have (i) $T|\hat{\lambda} - \lambda_0| = O_p(1)$ when $\alpha_1 \neq \alpha_2$ and (ii) $\hat{\lambda} \xrightarrow{d} \lambda^*$ with $\lambda^* = \arg\max_{\lambda \in \Lambda}[B_u(\lambda) - \lambda B_u(1)]^2/\lambda(1-\lambda)$ when $\alpha_1 = \alpha_2$.*

When $\beta_1 = \beta_2 = 0$ is imposed on the fitted model and $\alpha_1 \neq \alpha_2$ we have a purely stationary mean shift specification and the result in part (i) of Proposition 1 is intuitive and illustrates the T-consistency of the least squares based threshold parameter estimator. This is in fact a well known result in the literature which we report for greater coherence with our subsequent analysis (see Hansen (2000) and Gonzalo and Pitarakis (2002)). The result in part (ii) of Proposition 1 is particularly interesting and highlights the fact that the threshold parameter estimator obtained from a model that is linear and contains no threshold effects converges in distribution to a random variable given by the maximum of a normalised squared Brownian Bridge process. Although the maximum of a Brownian Bridge is well known to be a uniformly distributed random variable an explicit expression or closed form density for λ^* is to our knowledge not available in the literature.

We next concentrate on the limiting distribution of a Wald type test statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ in (1).

3.2 Testing $H_0 : \beta_1 = \beta_2 = 0$

Using the canonical representation in (5) and for a given $\lambda \in (0, 1)$ we can write the standard OLS based Wald statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ as

$$W_T^{ols}(\lambda) = \hat{\beta}(\lambda)'(X_\lambda^{*\prime} X_\lambda^*)\hat{\beta}(\lambda)/\hat{\sigma}_u^2(\lambda) \tag{7}$$

with $\hat{\beta}(\lambda) = (X_\lambda^{*\prime} X_\lambda^*)^{-1} X_\lambda^{*\prime} y$ and $\hat{\sigma}_u^2(\lambda)$ referring to the residual variance estimated from the unrestricted specification. In what follows $W_T^{ols}(\hat{\lambda})$ will denote the Wald statistic evaluated at the estimated threshold parameter $\hat{\lambda}$ as defined in (6) and its limiting behaviour is summarised in the following Proposition.

Proposition 2 *Under the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, assumptions A1-A3 and as $T \to \infty$ we have*

$$W_T^{ols}(\hat{\lambda}) \Rightarrow \frac{\left[\int J_c^*(r) dB_u(r, 1)\right]^2}{\sigma_u^2 \int J_c^*(r)^2} + \chi^2(1) \tag{8}$$

regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$.

Proposition 2 highlights the usefulness of the Wald statistic for conducting inferences about the $\beta's$ without having to take a stand on whether the $\alpha's$ are regime dependent or not. The interesting point here is the fact that the limiting distribution of the Wald statistic evaluated at $\hat{\lambda}$ is the same regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ in the underlying model. One shortcoming of our expression in (8) is caused by the presence of the unknown noncentrality parameter c making it difficult to tabulate in practice. Due to the allowed correlation between B_u and B_v it is also the case that the first component in the right hand side of (8) will depend on σ_{uv}. There is however an instance under which the limiting distribution simplifies considerably as summarised in Proposition 3 below.

Proposition 3 *Under the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, assumptions A1-A3 together with the requirement that $\sigma_{uv} = 0$ in (4) and as $T \to \infty$ we have*

$$W_T^{ols}(\hat{\lambda}) \Rightarrow \chi^2(2) \tag{9}$$

regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$.

The above result highlights a unique scenario whereby the magnitude of the noncentrality parameter no longer enters the asymptotics of the Wald statistic despite a nearly integrated parameterisation in the DGP. See also Rossi (2005) for interesting similarities between our asymptotics in Proposition 2 and distributions arising within a related structural break framework.

In order to address the limitations of our result in (8) we next introduce an Instrumental Variable based Wald statistic designed in such a way that its limiting distribution remains a nuisance parameter free $\chi^2(2)$ random variable regardless of whether σ_{uv} is zero or not. This is achieved through an IV method developed in Phillips and Magdalinos (2009) in the context of the cointegration literature and which we adapt to our current context (see also Breitung and Demetrescu (2014)). The key idea is to instrument

x_t with a slightly less persistent version of itself using its own innovations (hence the IVX terminology). Letting $\phi_T = (1 - c_z/T^\delta)$ for some $c_z > 0$ (say $c_z = 1$ as discussed in Phillips and Magdalinos (2009) and KMS2015) and $\delta \in (0,1)$ the IVX variable is constructed as $\tilde{h}_t = \sum_{j=1}^{t} \phi_T^{t-j} \Delta x_j$. Within our present context and for $i = 1,2$ we instrument $x_t I_{it}$ in (1) with $\tilde{h}_t I_{it}$. Letting $\tilde{h}_i$ denote the vector stacking the $\tilde{h}_t I_{it}'s$ and $H_\lambda = [\tilde{h}_1 \; \tilde{h}_2]$ the IVX estimator of β in (5) can be formulated as

$$\hat{\beta}^{ivx}(\lambda) \;\; = \;\; (H_\lambda^{*\prime} X_\lambda^*)^{-1} H_\lambda^{*\prime} y^* \tag{10}$$

with $H_\lambda^* = M_Q H_\lambda$. Noting that the projection $P_Q = Q_\lambda (Q_\lambda' Q_\lambda)^{-1} Q_\lambda'$ is effectively analogous to applying a *regime specific demeaning* the above formulation of the IVX estimator also helps highlight its invariance to using either H_λ or H_λ^* as IVs since $H_\lambda^{*\prime} X_\lambda^* = H_\lambda' X_\lambda^*$ and $H_\lambda^{*\prime} y^* = H_\lambda' y^*$. The IV based Wald statistic for testing $\beta_1 = \beta_2 = 0$ in (1) (or (5)) can now be formulated as

$$W_T^{ivx}(\lambda) \;\; = \;\; \hat{\beta}^{ivx} \left[(H_\lambda^{*\prime} X_\lambda^*)^{-1} (H_\lambda^{*\prime} H_\lambda^*) (H_\lambda^{*\prime} X_\lambda^*)^{-1} \right]^{-1} \hat{\beta}^{ivx} / \hat{\sigma}_u^2(\lambda) \tag{11}$$

and its limiting distribution is summarised in Proposition 4 below.

Proposition 4 *Under the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, assumptions A1-A3 and as $T \to \infty$ we have $W_T^{ivx}(\hat{\lambda}) \Rightarrow \chi^2(2)$ regardless of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$.*

The above result provides a convenient test statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ since inferences can be based on a limiting distribution that does not depend on c or any endogeneity induced parameter (as opposed to our formulation in (8)) and are immune to whether the intercepts shift or not. The parameter δ used in the construction of the IVX variables controls the degree of persistence of the instruments and plays a key role in ensuring that the Wald based asymptotics are free of the influence of the noncentrality parameter c. It is also important to highlight the fact that although δ is a necessary user-input in the construction of $W_T^{ivx}(\hat{\lambda})$ it does not play any role in its limiting distribution which is nuisance parameter free and valid for all $\delta \in (0,1)$. This of course does not preclude the fact that particular choices of δ may have important finite sample effects and size/power tradeoffs when basing inferences on $W_T^{ivx}(\hat{\lambda})$, an issue we explore and address below.

As shown in KMS2015 and Phillips and Magdalinos (2009) and as it is also the case for our estimator in (10) the price to pay for the convenient mixed normal limit of $\hat{\beta}^{ivx}$ which in turn leads to the χ^2 approximation of the associated Wald statistic is a rate of convergence that is slightly lower than T and given by $O(T^{\frac{1+\delta}{2}})$, suggesting that a choice of δ that is close to 1 may be the most appropriate when constructing the IVX variables. This is an issue we document and explore comprehensively in the simulations that follow but before doing so we wish to discuss in greater detail the key factors that may influence the impact of δ on the finite sample size and power properties of $W_T^{ivx}(\hat{\lambda})$ such as the strength of the correlation between u_t and v_t and adapt the practical recommendations of KMS2015 to our predictive threshold context.

Our Monte-Carlo simulations below robustly demonstrate that for moderate degrees of correlation between u_t and v_t our IVX based statistic displays excellent size control regardless of the magnitude of δ and a power that increases with δ albeit stabilising for magnitudes in the vicinity of 0.9. This naturally suggests that a choice of δ in the range $[0.85, 0.95]$ should offer a good compromise between finite sample size and power with only minor finite sample implications whether one uses $\delta = 0.85$ or $\delta = 0.95$ or another magnitude of similar order. When the correlation between u_t and v_t is allowed to be close to 1 however as it may happen in numerous finance applications inferences based on $W_T^{ivx}(\hat{\lambda})$ are characterised by important size distortions that increase and deteriorate with δ. These finite sample properties we observe within our setting mirror *exactly* the properties of the IVX Wald statistic documented in the linear predictive regression setting of KMS2015 and prompted the authors to introduce an intuitive finite sample correction to the formulation of their IVX based Wald statistic which they show offers excellent size control even under strong degrees of endogenity combined with a power that grows as δ approaches 1. The proposed finite sample correction does not alter the first order asymptotic approximation of the IVX based Wald statistic hence allowing KMS2015 to argue that for practical purposes their proposed correction resolves the issue of choosing a suitable δ. Size is perfectly controlled regardless of the magnitude of δ while power increases monotonically with δ and mirroring our earlier point above stabilises for magnitudes in the vicinity of 0.85-0.95. This naturally leads us to adapt the finite sample correction of KMS2015 to our own specification with threshold effects. It is important to reiterate however that the proposed correction aplied to $W_T^{ivx}(\hat{\lambda})$ does not affect its first order limit theory which remains as in Proposition 4.

The limiting χ^2 result in (11) naturally originates in the mixed Gaussianity of $\hat{\beta}^{ivx}$ in turn driven by the normality of a suitably normalised version of $H_\lambda^{*\prime} u^* \equiv H_\lambda' u^* = H_\lambda' u - H_\lambda' Q_\lambda (Q_\lambda' Q_\lambda)^{-1} Q_\lambda' u$ in (10) with the second component $H_\lambda' Q_\lambda (Q_\lambda' Q_\lambda)^{-1} Q_\lambda' u$ arising due to the presence of fitted intercepts (recall that $Q_\lambda = [I_1\ I_2]$) and which vanishes asymptotically. The first order asymptotic behaviour of $H_\lambda' u^*$ is driven by the asymptotic normality of a normalised version of $H_\lambda' u$. Although the second component $H_\lambda' Q(Q'Q)^{-1} Q' u$ vanishes asymptotically its presence can cause significant finite sample distortions compared to a setting with no fitted intercepts, distortions that are further amplified when the degree of correlation between u_t and v_t is large. KMS2015's correction which we adapt here is motivated by the need to neutralise this finite sample impact induced by the fitted intercepts and in proportion to how strongly correlated u_t and v_t are. The finite sample corrected $W_T^{ivx}(\hat{\lambda})$ adapted to our present context can be formulated as

$$W_T^{ivxc}(\lambda) \quad = \quad \hat{\beta}^{ivx\prime}(\lambda) \left[(H_\lambda^{*\prime} X_\lambda^*)^{-1} G_\lambda (H_\lambda^{*\prime} X_\lambda^*)^{-1} \right]^{-1} \hat{\beta}^{ivx}(\lambda) \tag{12}$$

with

$$G_\lambda \quad = \quad \hat{\sigma}_u^2(\lambda) \left(H_\lambda^{*\prime} H_\lambda^* + \hat{\rho}_{uv}^2 H_\lambda' Q_\lambda (Q_\lambda' Q_\lambda)^{-1} Q_\lambda' H_\lambda \right) \tag{13}$$

and where $\hat{\rho}_{uv}^2 = \hat{\omega}_{uv}^2 / \hat{\sigma}_u^2 \hat{\omega}_v^2$. Here $\hat{\omega}_{uv}$ denotes an estimator of the long run covariance between u_t and v_t and $\hat{\omega}_v^2$ an estimator of the long run variance of the $v_t's$ (see (4)). Note for instance that the correction in (13) will have little impact for small magnitudes of ρ_{uv} while playing an important finite sample adjustment role when the correlation between u_t and v_t is large, effectively neutralising the finite sample distortions resulting from the fitted intercepts. It is also useful to point out that when suitable normalisations are applied to $W_T^{ivxc}(\hat{\lambda})$ defined above, the correction term adjacent to $\hat{\rho}_{uv}^2$ will vanish asymptotically. Both $\hat{\omega}_{uv}$ and $\hat{\omega}_v^2$ can be estimated in a straightforward manner using Newey-West type estimators. For this purpose we proceed as in KMS2015 introducing a bandwidth parameter K_T such that $K_T \to \infty$ and $K_T/\sqrt{T} \to 0$ as $T \to \infty$ and using

$$\begin{aligned}
\hat{\omega}_v^2 &= \frac{1}{T}\sum_{t=1}^{T}\hat{v}_t^2 + \frac{2}{T}\sum_{\ell=1}^{K_T}\left(1-\frac{\ell}{K_T+1}\right)\sum_{t=\ell+1}^{T}\hat{v}_t\hat{v}_{t-\ell} \\
\hat{\omega}_{uv} &= \frac{1}{T}\sum_{t=1}^{T}\hat{u}_t\hat{v}_t + \frac{1}{T}\sum_{\ell=1}^{K_T}\left(1-\frac{\ell}{K_T+1}\right)\sum_{t=\ell+1}^{T}\hat{v}_t\hat{u}_{t-\ell} \qquad (14)
\end{aligned}$$

in the construction of $W_T^{ivxc}(\hat{\lambda})$. Our next goal is to comprehensively evaluate the finite sample properties of our IVX based test statistics with a particular emphasis on documenting the role played by δ and how best to select its magnitude in applied work.

4 Finite Sample Evaluation

The goal of this section is twofold. First we wish to demonstrate the validity and finite sample accuracy of our theoretical results presented in Propositions 2-4 through simulations. Second we wish to use our simulations to comprehensively illustrate the potential influence of δ on the finite sample size/power tradeoffs of our test statistics with the aim of achieving clear and reliable practical recommendations for the implementation of our IVX based statistics (e.g. for a broad range of experiments we consider magnitudes of δ ranging between 0.50 and 0.98 with increments of size 0.02). Due to space considerations we only present key outcomes while relegating a broad range of additional and supportive simulations to an online appendix.

We initially concentrate on the size properties of our test statistics. Our chosen DGP is given by (1) with $\beta_1 = \beta_2 = 0$. For the parameterisation of the intercepts we consider two scenarios. Namely, $\{\alpha_1, \alpha_2\} = \{1, 1\}$ and $\{\alpha_1, \alpha_2\} = \{1, 3\}$. In the latter case we set $\gamma_0 = 0.25$ with the threshold variable taken to follow the AR(1) process $q_t = 0.5q_{t-1} + u_{qt}$ while we set $v_t = 0.5v_{t-1} + e_{vt}$ for the shocks associated with the nearly integrated variable x_t. Finally we take (u_t, e_{vt}, u_{qt}) to be a Gaussian vector with covariance given by $\Sigma = \{(1, \sigma_{uv}, \sigma_{uq}), (\sigma_{uv}, 1, \sigma_{eq}), (\sigma_{uq}, \sigma_{eq}, 1)\}$. We initially focus on a scenario characterised by $\sigma_{uv} = 0$ and subsequently consider the more general case that allows contemporaneous correlations across all random disturbances. In this context we are particularly interested in the potential role played by a very strong correlation between u_t and v_t and how this may in turn interact with alternative choices of δ. For these reasons we conduct all our simulations by considering $\sigma_{uv} \in \{-0.9, -0.6, -0.3, 0.0\}$ and setting $(\sigma_{uv}, \sigma_{uq}, \sigma_{eq}) = (\sigma_{uv}, 0.2, 0.2)$.

We are initially interested in illustrating our result in Proposition 3 stating that the limiting distribution of the OLS based Wald statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ in (1) is $\chi^2(2)$ *regardless* of whether $\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$ and regardless of the magnitude of the noncentrality parameter c appearing in the DGP. Table 1 below displays the simulated finite sample critical values of $W_T^{ols}(\hat{\lambda})$ together with those of the $\chi^2(2)$ under $c = 1$ and $c = 10$. Overall we note an excellent match of the simulated quantiles with their asymptotic counterparts. It is also clear that varying c has little impact on the quantiles as expected by our theoretical result. Perhaps more importantly we note the robustness of the estimated quantiles to the two scenarios about the $\alpha's$. Even under moderately small sample sizes such as $T = 200$

the cutoffs of the asymptotic distribution of $W_T^{ols}(\hat{\lambda})$ under $\alpha_1 = \alpha_2$ and $\alpha_1 \neq \alpha_2$ remain extremely close as again confirmed by our theory.

Table 1: **Simulated Quantiles of $W_T^{ols}(\hat{\lambda})$ versus $\chi^2(2)$ under** $\sigma_{uv} = 0$

	10.0%	5.0%	2.5%	10.0%	5.0%	2.5%
χ_2^2	4.610	5.990	7.380	4.610	5.990	7.380
	$\alpha_1 = \alpha_2$, $c = 1$			$\alpha_1 \neq \alpha_2$, $c = 1$		
T=200	4.508	5.795	7.167	4.521	5.880	7.354
T=400	4.708	6.089	7.433	4.779	6.341	8.159
T=1000	4.692	5.981	7.418	4.592	5.723	6.948
	$\alpha_1 = \alpha_2$, $c = 10$			$\alpha_1 \neq \alpha_2$, $c = 10$		
T=200	4.481	6.056	7.841	4.494	5.959	7.381
T=400	4.561	6.094	7.638	4.619	5.845	7.287
T=1000	4.668	6.027	7.439	4.400	6.027	7.228

We next, concentrate on our IVX based Wald statistics and evaluate their empirical size properties across alternative scenarios on the σ_{uv}'s and δ's. As a benchmark scenario Table 2 initially reports empirical sizes for the $\sigma_{uv} = 0$ case which as expected corroborate our quantile based results of Table 1 while also highlighting the adequacy of $W_T^{ivx}(\hat{\lambda})$ and $W_T^{ivxc}(\hat{\lambda})$ when neither would have been truly needed here due to exogeneity. It is also important to note that size is very accurately controlled regardless of the magnitude of δ including magnitudes in the vicinity of 1.

Table 2: **Empirical Size of $W_T^{ivx}(\hat{\lambda})$, $W_T^{ivxc}(\hat{\lambda})$ and $W_T^{ols}(\hat{\lambda})$ (5% Nominal),** $\sigma_{uv} = 0.0$

	δ							δ						
δ	0.70	0.74	0.78	0.82	0.86	0.90	0.94	0.70	0.74	0.78	0.82	0.86	0.90	0.94
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	4.35	4.55	4.60	4.60	4.60	4.80	4.80	4.40	4.50	4.85	4.50	4.85	5.00	5.00
T=400	4.70	4.40	4.55	4.70	4.90	5.05	5.00	5.15	5.10	5.25	5.40	5.70	6.05	6.40
T=1000	5.40	5.25	5.45	5.20	5.40	5.45	5.70	4.20	4.50	4.60	5.05	4.95	4.75	4.75
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	4.35	4.50	4.55	4.55	4.55	4.65	4.75	4.35	4.40	4.60	4.40	4.70	4.60	4.75
T=400	4.65	4.30	4.55	4.70	4.85	5.05	4.90	5.15	5.10	5.25	5.35	5.65	6.05	6.35
T=1000	5.40	5.25	5.45	5.10	5.40	5.35	5.65	4.15	4.45	4.50	5.00	4.95	4.70	4.60
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	4.65	4.65	4.65	4.65	4.65	4.65	4.65	4.75	4.75	4.75	4.75	4.75	4.75	4.75
T=400	5.35	5.35	5.35	5.35	5.35	5.35	5.35	6.25	6.25	6.25	6.25	6.25	6.25	6.25
T=1000	4.90	4.90	4.90	4.90	4.90	4.90	4.90	4.45	4.45	4.45	4.45	4.45	4.45	4.45
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.75	5.90	5.70	5.50	5.55	5.45	5.70	5.30	5.45	5.40	5.25	5.50	5.15	5.25
T=400	5.60	5.60	5.90	5.65	5.75	5.70	5.65	4.90	4.80	5.00	4.85	4.65	4.75	4.70
T=1000	5.15	5.40	5.35	5.50	5.30	5.35	5.10	4.20	4.45	4.45	4.40	4.35	4.65	4.60
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.75	5.85	5.65	5.50	5.55	5.45	5.60	5.30	5.45	5.40	5.25	5.40	5.15	5.25
T=400	5.60	5.60	5.90	5.65	5.70	5.60	5.60	4.90	4.80	5.00	4.85	4.65	4.75	4.70
T=1000	5.15	5.40	5.35	5.50	5.30	5.35	5.10	4.20	4.45	4.45	4.40	4.35	4.65	4.60
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.10	5.10	5.10	5.10	5.10	5.10	5.10	4.85	4.85	4.85	4.85	4.85	4.85	4.85
T=400	5.55	5.55	5.55	5.55	5.55	5.55	5.55	4.40	4.40	4.40	4.40	4.40	4.40	4.40
T=1000	5.15	5.15	5.15	5.15	5.15	5.15	5.15	5.10	5.10	5.10	5.10	5.10	5.10	5.10

Table 3 presents size estimates under a nonzero but weak correlation between u_t and v_t. $W_T^{ivx}(\hat{\lambda})$

continues to offer excellent size control across all scenarios on the intercepts and non-centrality parameter and perhaps more importantly magnitudes of δ. Under $\alpha_1 = \alpha_2, c = 1$ for instance the average empirical size across the seven different magnitudes of δ ranging between 0.70 and 0.94 was 4.64%. Given the weak degree of endogeneity considered here we also note very similar outcomes characterising the OLS based Wald statistic $W_T^{ols}(\hat{\lambda})$.

With $\sigma_{uv} = -0.6$ Table 4 focuses on a scenario with a stronger correlation between u_t and v_t. We can immediately note the important distortions characterising the OLS based Wald statistic $W_T^{ols}(\hat{\lambda})$ which is clearly not suitable under endogeneity as also suggested by our theoretical result in Proposition 2. Here $W_T^{ivx}(\hat{\lambda})$ is seen to offer considerable improvements over $W_T^{ols}(\hat{\lambda})$. The match of empirical sizes to their nominal counterparts is good to excellent for moderate magnitudes of δ and although finite sample distortions start kicking in as δ approaches 1, overall the distortions appear acceptable especially for larger sample sizes. Also noteworthy is the excellent match of empirical sizes of $W_T^{ivx}(\hat{\lambda})$ based inferences to their nominal counterparts for slightly larger magnitudes of the non centrality parameter c. Finally and equally importantly the corrected version of our Wald statistic $W_T^{ivxc}(\hat{\lambda})$ is seen to be characterised by excellent size properties across all magnitudes of δ including when the latter are very close to 1. Under $\delta = 0.94$ and $T = 400$ for instance we note an empirical size of 4.90% for a 5% nominal size.

Table 3: **Empirical Size of $W_T^{ivx}(\hat{\lambda})$, $W_T^{ivxc}(\hat{\lambda})$ and $W_T^{ols}(\hat{\lambda})$ (5% Nominal),** $\sigma_{uv} = -0.3$

	δ							δ						
	0.70	0.74	0.78	0.82	0.86	0.90	0.94	0.70	0.74	0.78	0.82	0.86	0.90	0.94
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	4.25	4.75	4.85	4.95	5.30	5.15	5.25	5.00	5.55	5.90	6.20	6.30	6.05	6.15
T=400	3.95	4.05	4.50	4.60	4.85	5.20	5.35	5.60	5.80	6.05	6.50	6.75	7.25	7.15
T=1000	5.80	6.20	5.85	6.10	6.15	6.50	6.45	6.30	6.35	6.65	6.35	6.30	6.40	6.50
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	4.15	4.45	4.50	4.80	4.70	4.60	4.65	4.75	5.25	5.30	5.55	5.25	5.30	5.00
T=400	3.85	3.70	4.05	4.20	4.45	4.70	4.90	5.40	5.20	5.50	5.75	5.90	6.00	5.95
T=1000	5.70	6.00	5.70	5.70	5.75	6.05	5.90	6.10	6.15	6.25	5.60	5.45	5.45	5.70
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	5.80	5.80	5.80	5.80	5.80	5.80	5.80	6.60	6.60	6.60	6.60	6.60	6.60	6.60
T=400	6.20	6.20	6.20	6.20	6.20	6.20	6.20	6.15	6.15	6.15	6.15	6.15	6.15	6.15
T=1000	6.85	6.85	6.85	6.85	6.85	6.85	6.85	6.65	6.65	6.65	6.65	6.65	6.65	6.65
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	6.10	5.95	5.85	6.10	6.15	6.15	6.15	5.40	5.20	5.25	5.40	5.60	5.55	5.65
T=400	5.25	5.20	5.30	5.10	5.15	5.20	5.35	5.00	5.30	5.35	5.70	5.60	5.75	5.70
T=1000	4.95	5.45	5.30	5.40	5.45	5.35	5.60	5.40	5.40	5.40	5.75	5.55	5.50	5.65
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.95	5.80	5.80	6.05	6.05	5.95	6.00	5.40	5.10	5.20	5.40	5.60	5.55	5.55
T=400	5.20	5.15	5.25	5.10	5.10	5.15	5.20	5.00	5.30	5.35	5.60	5.55	5.75	5.45
T=1000	4.95	5.30	5.25	5.35	5.40	5.30	5.45	5.40	5.30	5.40	5.70	5.50	5.40	5.65
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.40	5.40	5.40	5.40	5.40	5.40	5.40	5.60	5.60	5.60	5.60	5.60	5.60	5.60
T=400	5.25	5.25	5.25	5.25	5.25	5.25	5.25	4.90	4.90	4.90	4.90	4.90	4.90	4.90
T=1000	5.30	5.30	5.30	5.30	5.30	5.30	5.30	5.35	5.35	5.35	5.35	5.35	5.35	5.35

Table 4: **Empirical Size of $W_T^{ivx}(\hat{\lambda})$, $W_T^{ivxc}(\hat{\lambda})$ and $W_T^{ols}(\hat{\lambda})$ (5% Nominal), $\sigma_{uv} = -0.6$**

	δ							δ						
	0.70	0.74	0.78	0.82	0.86	0.90	0.94	0.70	0.74	0.78	0.82	0.86	0.90	0.94
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	5.95	6.15	6.55	6.80	7.70	8.35	8.65	6.75	7.25	7.70	8.05	8.40	8.80	9.15
T=400	5.75	6.25	6.45	6.80	6.90	7.15	7.55	6.00	6.45	6.75	6.95	7.30	7.20	7.15
T=1000	5.60	6.05	6.30	6.05	6.65	6.90	7.15	4.80	4.95	5.45	5.70	5.40	5.95	6.65
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	5.05	5.25	5.25	5.30	5.50	5.65	5.90	5.60	5.90	5.85	5.65	5.65	5.10	5.25
T=400	5.20	5.20	5.05	5.15	5.05	4.80	4.90	5.20	5.25	5.40	5.25	5.20	4.95	4.80
T=1000	4.95	5.00	4.95	4.75	4.75	4.45	4.60	4.05	4.05	4.10	4.00	3.80	3.90	4.20
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	10.25	10.25	10.25	10.25	10.25	10.25	10.25	10.35	10.35	10.35	10.35	10.35	10.35	10.35
T=400	10.75	10.75	10.75	10.75	10.75	10.75	10.75	9.85	9.85	9.85	9.85	9.85	9.85	9.85
T=1000	9.80	9.80	9.80	9.80	9.80	9.80	9.80	10.05	10.05	10.05	10.05	10.05	10.05	10.05
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	4.85	4.80	5.10	5.00	5.15	5.15	5.00	5.55	5.65	5.80	5.80	5.95	6.15	6.30
T=400	6.00	5.75	5.40	5.60	5.75	6.00	5.95	4.55	4.95	5.20	5.40	5.65	6.00	6.10
T=1000	5.20	5.60	5.50	5.30	5.20	5.30	5.10	4.85	5.20	5.15	5.00	5.00	4.95	5.30
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	4.45	4.55	4.75	4.85	4.85	4.70	4.55	5.50	5.50	5.55	5.40	5.65	5.75	5.95
T=400	5.85	5.55	5.30	5.30	5.35	5.45	5.35	4.30	4.85	5.00	5.25	5.40	5.70	5.65
T=1000	5.15	5.50	5.35	5.15	4.90	4.95	4.70	4.70	5.05	5.05	4.75	4.75	4.65	4.70
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.10	5.10	5.10	5.10	5.10	5.10	5.10	6.50	6.50	6.50	6.50	6.50	6.50	6.50
T=400	6.30	6.30	6.30	6.30	6.30	6.30	6.30	5.90	5.90	5.90	5.90	5.90	5.90	5.90
T=1000	5.20	5.20	5.20	5.20	5.20	5.20	5.20	6.00	6.00	6.00	6.00	6.00	6.00	6.00

Next, Table 5 treats the important case of $\sigma_{uv} = -0.9$ which brings the DGP closer to the type of endogeneity we encounter when dealing with financial returns and valuation ratios. Our results highlight the remarkable robustness and usefulness of $W_T^{ivxc}(\hat{\lambda})$ under more extreme endogeneity scenarios combined with choices of δ in the vicinity of unity. This modified IVX based Wald statistic is seen to offer excellent size properties with empirical sizes in the region of 4.5%-5.1% for a 5% nominal size. Also noteworthy is the robustness of this feature to alternative magnitudes of c and to whether the intercepts are allowed to shift or not.

Regarding the uncorrected IVX based Wald statistic $W_T^{ivx}(\hat{\lambda})$, although its size properties are adequate for magnitudes of δ around 0.70 it is clear that it will lead to too many spurious rejections of the null unless impractically large sample sizes become available. As discussed earlier and in analogy with KMS2015 the root cause of this phenomenon originates in the estimation of intercepts in the fitted specification. Important finite sample distortions appear to affect the term $H_\lambda' Q_\lambda (Q_\lambda' Q_\lambda)^{-1} Q_\lambda' u$ in $H_\lambda^{*\prime} u^*$ despite the fact that it vanishes asymptotically and is dominated by $H_\lambda' u$.

Table 5: **Empirical Size of $W_T^{ivx}(\hat{\lambda})$, $W_T^{ivxc}(\hat{\lambda})$ and $W_T^{ols}(\hat{\lambda})$ (5% Nominal),** $\sigma_{uv} = -0.9$

	δ							δ						
	0.70	0.74	0.78	0.82	0.86	0.90	0.94	0.70	0.74	0.78	0.82	0.86	0.90	0.94
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	7.85	8.45	9.10	9.75	10.60	11.00	11.60	7.85	8.45	9.25	9.55	10.15	10.60	11.30
T=400	7.85	8.25	9.15	9.90	10.40	11.15	11.95	6.95	7.35	7.60	7.85	9.30	10.35	10.65
T=1000	6.65	7.20	8.00	8.25	9.40	10.65	11.20	7.15	8.05	8.40	9.05	10.00	11.20	12.00
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	5.30	5.10	5.20	4.95	4.65	4.55	4.25	5.05	4.75	4.75	4.35	4.35	4.15	4.25
T=400	5.60	5.40	5.25	5.00	4.55	4.50	4.40	5.15	5.25	4.90	4.60	4.70	4.60	4.55
T=1000	5.10	4.90	4.75	4.80	4.40	4.50	4.30	5.20	5.25	5.05	4.80	4.45	4.40	4.20
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 1$							$\alpha_1 \neq \alpha_2$, $c = 1$						
T=200	15.05	15.05	15.05	15.05	15.05	15.05	15.05	15.15	15.15	15.15	15.15	15.15	15.15	15.15
T=400	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.85	14.85	14.85	14.85	14.85	14.85	14.85
T=1000	15.20	15.20	15.20	15.20	15.20	15.20	15.20	15.50	15.50	15.50	15.50	15.50	15.50	15.50
$W_T^{ivx}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	6.05	5.95	5.85	5.90	6.15	6.45	6.75	5.25	5.50	5.50	5.70	5.80	5.75	6.05
T=400	5.75	5.85	6.05	6.00	6.15	6.20	6.20	6.05	5.80	5.95	6.25	6.40	6.45	6.40
T=1000	6.20	6.35	6.65	6.50	6.65	6.95	6.85	5.70	5.95	6.05	6.10	6.50	6.75	6.80
$W_T^{ivxc}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	5.40	5.20	5.10	5.15	5.05	5.10	5.35	4.85	4.85	4.85	4.95	4.95	4.80	5.00
T=400	5.55	5.40	5.30	5.10	4.95	4.85	4.85	5.45	5.55	5.45	5.45	5.35	5.30	5.10
T=1000	6.00	6.05	6.05	5.95	5.85	5.75	5.70	5.50	5.70	5.50	5.75	5.85	6.05	5.90
$W_T^{ols}(\hat{\lambda})$	$\alpha_1 = \alpha_2$, $c = 10$							$\alpha_1 \neq \alpha_2$, $c = 10$						
T=200	7.25	7.25	7.25	7.25	7.25	7.25	7.25	6.40	6.40	6.40	6.40	6.40	6.40	6.40
T=400	6.85	6.85	6.85	6.85	6.85	6.85	6.85	7.55	7.55	7.55	7.55	7.55	7.55	7.55
T=1000	7.55	7.55	7.55	7.55	7.55	7.55	7.55	7.45	7.45	7.45	7.45	7.45	7.45	7.45

In summary our size experiments have demonstrated the suitability and usefulness of an IVX type of approach for conducting inferences about episodic predictability in small to moderate samples. Our IVX based Wald statistic $W_T^{ivx}(\hat{\lambda})$ provides excellent size control under weak to moderate correlations between u_t and v_t regardless of the magnitude of δ including values in the vicinity of 1. For strong to extreme degrees of correlations between u_t and v_t however the same statistic can lead to serious size distortions for magnitudes of δ in excess of 0.70 and its corrected version $W_T^{ivxc}(\hat{\lambda})$ should be preferred given the excellent size control of the latter for any magnitude of the pair $\{\sigma_{uv}, \delta\}$.

We next focus on the ability of $W_T^{ivxc}(\hat{\lambda})$ to detect fixed departures from the null hypothesis across a broad range of scenarios and parameterisations. Results are presented in Table 6 below. More importantly we here address the issue of the impact of δ on finite sample size power trade-offs more thoroughly by considering fine increments of δ ranging between 0.40 and 0.94 under both the null and various alternatives, with the outcomes compiled within Figures 1-2 below.

Focusing on the results presented in Table 6 first we note a clear progression of empirical power towards 100% as the sample size increases, with $W_T^{ivxc}(\hat{\lambda})$ achieving power close to 100% under T=1000 and across all intercept, noncentrality parameter scenarios and any magnitude of δ. Concentrating on the case $\sigma_{uv} = -0.9$ we can also observe that empirical power is steadily increasing with δ with a spread in empirical power of about 10% between $\delta = 0.70$ and $\delta = 0.82$ albeit with a clear stabilisation for magnitudes of δ in the vicinity of the 0.85-0.95 range. Under $\{\alpha_1 = \alpha_2, c = 1, T = 200\}$ for instance an empirical power of 83.8% when $\delta = 0.82$ can be compared with 86.4% when $\delta = 0.90$ and 86.6% when $\delta = 0.94$, a pattern that carries through across most parameterisations.

Table 6: **Empirical Power of $W_T^{ivxc}(\hat{\lambda})$ (5% Nominal Size)**

				δ							δ			
$\beta_1 = 0$	0.70	0.74	0.78	0.82	0.86	0.90	0.94	0.70	0.74	0.78	0.82	0.86	0.90	0.94
$\beta_2 = 0.025$				$\alpha_1 = \alpha_2$, $c = 1$, $\sigma_{uv} = -0.6$							$\alpha_1 \neq \alpha_2$, $c = 1$, $\sigma_{uv} = -0.6$			
T=200	24.3	26.9	29.5	31.6	33.5	34.6	35.3	31.4	34.6	37.1	39.5	41.4	42.9	43.7
T=400	66.2	70.7	75.4	78.7	81.6	83.2	84.9	72.2	77.5	81.0	83.6	85.7	87.3	88.6
T=1000	96.5	97.5	98.3	99.2	99.6	100.0	100.0	96.8	97.5	98.4	99.0	99.5	99.8	99.9
$\beta_2 = 0.05$														
T=200	66.7	69.6	73.4	76.7	79.7	80.7	82.0	72.8	76.6	79.4	81.1	83.3	84.4	85.5
T=400	95.9	97.1	98.0	98.8	99.3	99.5	99.7	95.5	96.5	97.7	98.6	99.3	99.5	99.7
T=1000	98.9	99.2	99.5	99.7	100.0	100.0	100.0	98.7	99.3	99.7	100.0	100.0	100.0	100.0
$\beta_2 = 0.025$				$\alpha_1 = \alpha_2$, $c = 10$, $\sigma_{uv} = -0.6$							$\alpha_1 \neq \alpha_2$, $c = 10$, $\sigma_{uv} = -0.6$			
T=200	11.6	11.9	12.1	12.6	13.3	14.0	14.1	15.8	16.5	17.1	17.9	18.3	19.1	19.4
T=400	33.3	35.6	38.2	40.5	41.9	43.3	44.2	49.4	52.6	55.3	57.4	59.2	60.8	61.8
T=1000	98.3	98.9	99.1	99.5	99.7	99.7	99.7	99.8	99.8	99.9	99.9	100.0	100.0	100.0
$\beta_2 = 0.05$														
T=200	32.6	34.8	36.4	38.1	39.5	41.0	41.6	51.6	53.9	55.9	57.2	58.4	59.8	60.4
T=400	89.1	90.9	92.5	94.1	94.5	94.8	95.0	96.8	97.5	98.2	98.5	98.7	98.8	98.8
T=1000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$\beta_2 = 0.025$				$\alpha_1 = \alpha_2$, $c = 1$, $\sigma_{uv} = -0.9$							$\alpha_1 \neq \alpha_2$, $c = 1$, $\sigma_{uv} = -0.9$			
T=200	24.3	26.7	28.9	30.2	31.3	32.2	32.6	32.7	36.1	38.2	40.0	41.5	43.2	43.8
T=400	72.0	77.3	80.9	83.8	85.9	87.2	88.0	80.1	84.0	86.9	89.2	90.7	92.3	92.7
T=1000	97.5	97.8	98.7	99.2	99.6	99.8	99.9	97.6	98.5	99.2	99.5	99.8	99.9	100.0
$\beta_2 = 0.05$														
T=200	73.0	78.0	81.2	83.8	85.0	86.4	86.6	78.4	82.8	85.2	87.3	88.9	89.6	89.8
T=400	96.9	97.6	98.4	98.8	99.2	99.5	99.7	96.3	97.7	98.2	98.9	99.1	99.4	99.6
T=1000	99.2	99.5	99.7	99.8	99.9	100.0	100.0	98.8	99.1	99.5	99.8	100.0	100.0	100.0
$\beta_2 = 0.025$				$\alpha_1 = \alpha_2$, $c = 10$, $\sigma_{uv} = -0.9$							$\alpha_1 \neq \alpha_2$, $c = 10$, $\sigma_{uv} = -0.9$			
T=200	12.2	13.0	13.8	14.4	14.6	15.1	15.7	15.9	16.9	17.5	18.1	18.7	18.9	18.8
T=400	36.1	38.6	40.7	42.9	44.8	46.2	47.5	52.1	56.5	59.2	61.9	63.8	65.8	67.3
T=1000	99.8	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$\beta_2 = 0.05$														
T=200	35.7	38.7	40.5	41.8	43.1	44.4	45.5	51.7	54.6	57.3	60.0	61.8	62.9	63.9
T=400	94.3	95.7	96.6	97.4	98.0	98.3	98.4	99.2	99.5	99.7	99.7	99.9	99.9	100.0
T=1000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

The analysis presented in Figures 1-2 below is also highly informative when it comes to assessing the influence of δ and for providing practical guidelines on its choice. We note that the $W_T^{ivxc}(\hat{\lambda})$ statistic displays excellent size control as judged by the horizontal line across 5% while its power is seen to increase with δ, typically stabilising for magnitudes in the vicinity of or greater than 0.85. This suggests that selecting a δ that is close to 0.9 should provide reliable finite sample inferences with only marginal differences if it is slightly above or below 0.9. This is also supported by our application below where our test statistics are seen to have very similar pvalues for any magnitude of δ between 0.80 and 0.95.

Figure 1: Size and Power of $W_T^{ivxc}(\hat{\lambda})$ across δ (5% Nominal Size)

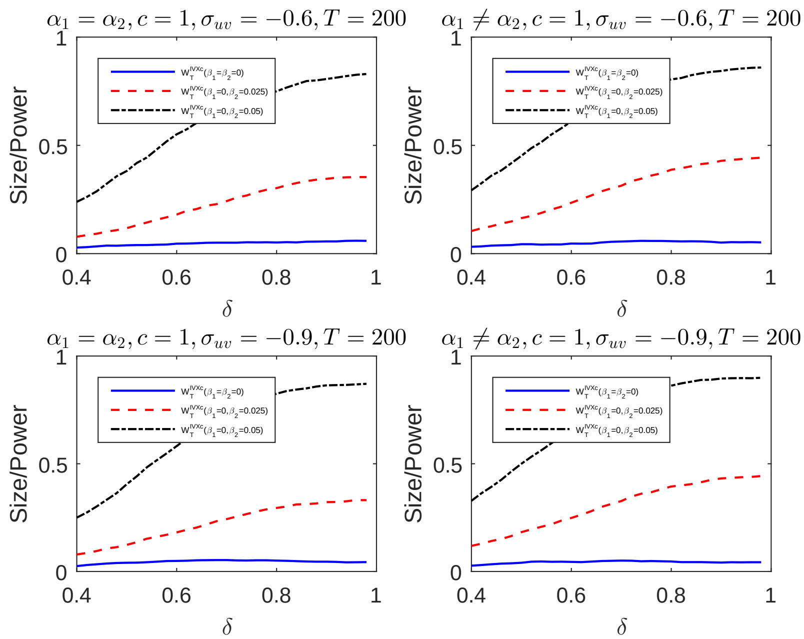
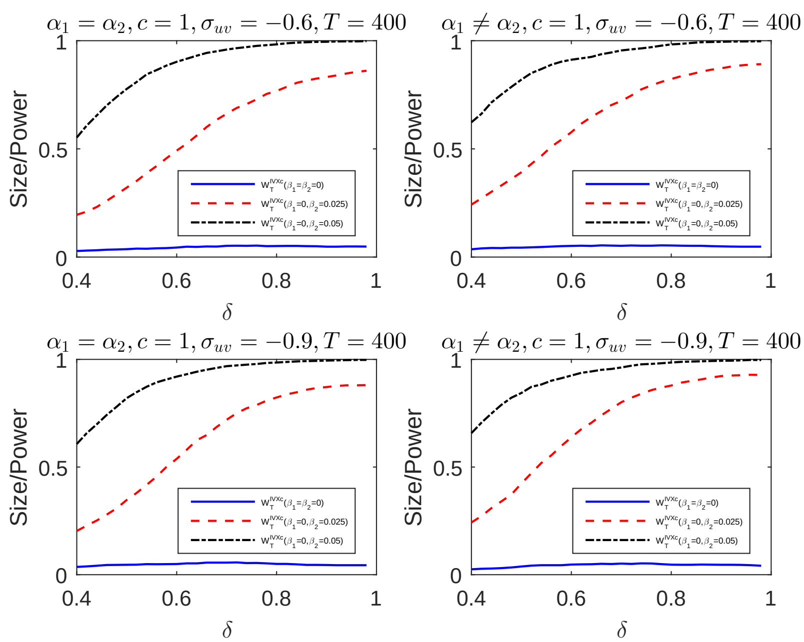


Figure 2: Size and Power of $W_T^{ivxc}(\hat{\lambda})$ across δ (5% Nominal Size)



5 Valuation Ratio Based Return Predictability

Due to its ability to let the data determine the presence or absence of regime specific behaviour in predictive regressions, our threshold setting is particularly suited for exploring the presence of time varying return predictability when time variation is driven by economic episodes such as recessions and expansions rather than calendar time per se. The new inference theory developed in this paper is an important complement to the two test statistics proposed in Gonzalo and Pitarakis (2012) allowing us to distinguish between regime specific predictability truly induced by a particular predictor such as the dividend yield and regime specific behaviour that may arise solely due to the variable used for generating the regimes (e.g. average returns varying across business cycle regimes).

Despite a large literature geared towards testing for the linear predictability of stock returns with valuation ratios such as the dividend yield it is only recently that empirical work has recognised the possibility that predictability may be kicking in occasionally depending on the state of the economy. In Gonzalo and Pitarakis (2012) for instance, using aggregate US data over the 1950-2007 period we established a strong countercyclical property to dividend yield based predictability of stock returns with an R^2 as high as 17% in the weak or negative growth regime, dropping to 0% during expansions (see also Henkel, Martin and Nardari (2011) who reached similar conclusions using a different statistical framework). More recently Gargano (2013) also reached similar conclusions using the dividend to Price ratio as a predictor while also proposing a theoretical framework that embeds this recessionary period based predictability of stock returns within a consumption based asset pricing model. Earlier research that highlighted the importance of a changing environment on predictability include Pesaran and Timmermann (1995), Paye and Timmermann (2006) amongst numerous others.

We here consider the question of episodic predictability of aggregate US market returns using four fundamental valuation ratios given by the dividend yield (DY), the book-to-market ratio (BM), the dividend to Price ratio (DP) and the earnings yield (EP). Although they serve a different purpose we also contrast inferences based on our $W_T^{ivxc}(\hat{\lambda})$ statistic developed here with inferences based on the $SupB^{ivx}$ statistic developed in Gonzalo and Pitarakis (2012) and which was designed to test $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$. As the latter null hypothesis could be rejected due to unequal intercepts we are here able to infer whether episodic predictability is directly induced by the valuation ratios under consideration.

The potential influence of economic conditions on predictability is captured by the threshold variable q_t for which we consider three alternative choices proxying business cycle conditions. In addition to the monthly growth rate in Industrial Production (IPGR) we considered in Gonzalo and Pitarakis (2012) (data item INDPRO retrieved from the Fred database) we also implement our analysis using a selection of composite indicators of real economic activity commonly tracked by policy makers. Namely the 3 month moving average of the Chicago Fed's National Activity index (CFNAIMA, 1967:05-2013:12) and the Aruoba-Diebold-Scotti business conditions index (ADS, 1960:04-2013:12) with the shortcoming that these two series are available from the 60s onwards whereas IPGR can cover the full sample period for which returns and valuation ratios are available. As the ADS index (see Aruoba, Diebold and Scotti (2009)) is designed to track the economy in *real time* it is constructed as a daily index which we transformed into a monthly series by selecting its end of the month values. Given our operating assumptions we verified the stationarity properties of the above three threshold variables through a standard ADF test which led to strong rejections of the unit root null for all cases.

Compared to our analysis in Gonzalo and Pitarakis (2012) where we had focused solely on DY over

1950-2007 we also extend our sample to cover the 1927-2013 period using the recently extended Goyal and Welch data set (see Goyal and Welch (2014) and Welch and Goyal (2008)). The specific return series we are considering is the recently revised excess returns series referred to as $Mkt - RF$ in Kenneth French's data library with Mkt referring to the value weighted returns of all CRSP firms listed on the NYSE, AMEX or NASDAQ and RF the one month T-Bill return.

Table 7 below presents our empirical results across various values of δ for the $W_T^{ivxc}(\hat{\lambda})$ and $SupB^{ivx}$ statistics. Note that in analogy to the correction we applied to our IVX based Wald statistic we also implemented the same correction to the $SupB^{ivx}$ statistic of Gonzalo and Pitarakis (2012) and referred to as $SupB^{ivxc}$ hereafter (see Remark A1 in the appendix). Although not reported here inferences based on $SupB^{ivx}$ led to outcomes identical to those based on $SupB^{ivxc}$ across all magnitudes of δ. Outcomes of the $SupA$ statistic designed to test the null of linearity $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ rather than predictability per se are also included for reference. The underlying theory for this test was developed in Gonzalo and Pitarakis (2012).

Table 7: **Episodic Predictability of Stock Returns with Valuation Ratios**

	$W_T^{ivxc}(\hat{\lambda})$				$SupB^{ivxc}$				$SupA$
δ	0.70	0.78	0.86	0.94	0.70	0.78	0.86	0.94	
				IPGR (1927-2013)					
DY	6.69 [0.04]	6.51 [0.04]	5.59 [0.06]	4.61 [0.10]	33.26***	33.66***	33.28***	32.64***	27.54 [0.00]
BM	6.20 [0.05]	6.27 [0.04]	6.23 [0.04]	6.24 [0.04]	41.17***	41.7***	41.83***	41.81***	34.72 [0.00]
DP	4.53 [0.10]	4.68 [0.10]	4.13 [0.13]	3.42 [0.18]	22.32***	22.93***	22.9***	22.57***	19.19 [0.00]
EP	3.86 [0.15]	4.55 [0.10]	4.88 [0.09]	4.90 [0.09]	15.92***	16.61***	16.99***	17.13***	12.22 [0.05]
				IPGR (1940-2013)					
DY	4.32 [0.12]	5.61 [0.06]	6.02 [0.05]	5.80 [0.06]	24.11***	24.93***	25.03***	24.69***	20.02 [0.00]
BM	0.94 [0.63]	1.48 [0.48]	1.89 [0.39]	2.09 [0.35]	12.59 *	13.02 *	13.27 *	13.36 *	11.46 [0.07]
DP	3.00 [0.22]	4.20 [0.12]	4.66 [0.10]	4.55 [0.10]	22.29***	23.09***	23.25***	23.01***	19.37 [0.00]
EP	1.56 [0.46]	2.39 [0.30]	3.05 [0.22]	3.44 [0.18]	4.13	5.02	5.69	6.05	2.51 [0.98]
				IPGR (1950-2013)					
DY	2.41 [0.30]	2.76 [0.25]	2.63 [0.27]	2.17 [0.34]	23.52***	24.01***	24.09***	23.83***	21.53 [0.00]
BM	2.24 [0.33]	1.62 [0.44]	1.34 [0.51]	1.22 [0.54]	12.25 **	12.39 **	12.56 **	12.66 **	12.10 [0.05]
DP	1.63 [0.44]	2.05 [0.36]	2.03 [0.36]	1.70 [0.43]	21.54 **	22.04***	22.19***	22.04***	20.23 [0.00]
EP	0.67 [0.72]	0.84 [0.66]	1.29 [0.53]	1.61 [0.45]	3.53	3.81	4.30	4.75	3.39 [0.89]
				IPGR (1960-2013)					
DY	3.11 [0.21]	3.48 [0.18]	3.91 [0.14]	4.28 [0.12]	21.72***	21.62***	21.50***	21.43***	19.60 [0.00]
BM	0.37 [0.83]	0.10 [0.95]	0.21 [0.90]	0.44 [0.80]	10.92	10.92	10.93	10.94	10.88 [0.08]
DP	1.74 [0.42]	2.08 [0.35]	2.48 [0.29]	2.80 [0.25]	19.56***	19.64***	19.61***	19.56***	18.23 [0.00]
EP	3.22 [0.20]	1.92 [0.38]	1.31 [0.52]	1.12 [0.57]	3.18	3.20	3.27	3.36	2.65 [0.97]
				ADS (1960-2013)					
DY	4.68 [0.10]	4.98 [0.08]	5.26 [0.07]	5.45 [0.07]	17.19***	17.17***	17.05***	16.91***	14.99 [0.01]
BM	0.48 [0.79]	0.60 [0.74]	0.94 [0.63]	1.28 [0.53]	11.03	11.04	11.05	11.07	10.98 [0.08]
DP	2.96 [0.23]	3.27 [0.20]	3.60 [0.17]	3.84 [0.15]	14.73 **	14.85 **	14.82 **	14.74 **	13.38 [0.03]
EP	1.20 [0.55]	1.02 [0.60]	1.06 [0.59]	1.19 [0.55]	7.97 **	7.99 **	8.06 **	8.15 **	7.44 [0.30]
				CFNAIMA3 (1967-2013)					
DY	7.72 [0.02]	6.87 [0.03]	6.16 [0.05]	5.67 [0.06]	14.52 **	14.50 **	14.45 **	14.42 **	13.12 [0.03]
BM	5.74 [0.06]	5.08 [0.08]	4.47 [0.11]	4.06 [0.13]	12.51 *	12.51 *	12.52 *	12.52 *	12.50 [0.04]
DP	6.90 [0.03]	6.16 [0.05]	5.54 [0.06]	5.11 [0.08]	12.82 *	12.90 *	12.88 *	12.85 *	11.87 [0.06]
EP	4.62 [0.10]	4.54 [0.10]	4.30 [0.12]	4.09 [0.13]	9.52	9.51	9.56	9.63	9.07 [0.17]

Focusing first on the DY series with threshold effects driven by the full history of the growth rate in industrial production (IPGR 1927-2013) we note that on the basis of the $W_T^{ivxc}(\hat{\lambda})$ statistic and all magnitudes of δ the null of no episodic predictability induced by DY is rejected with a pvalue of 0.06 under $\delta = 0.86$ and a pvalue of 0.10 under $\delta = 0.94$. This further corroborates and strengthens our findings in Gonzalo and Pitarakis (2012) where we had documented the countercyclical predictability of DY over the 1950-2007 period on the basis of the $SupB^{ivx}$ statistic. Our new test statistic leads to rejections of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ as does $SupB^{ivxc}$ which tests $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$, suggesting that predictability over the full sample is truly driven by the DY predictor rather than unequal intercepts arising from our business cycle proxy.

Looking at the IPGR based subperiods we note that $SupB^{ivxc}$ based inferences continue to consistently reject across all scenarios while $W_T^{ivxc}(\hat{\lambda})$ based inferences attribute a more ambiguous role to the dividend yield as predictor when restricting the sample to the post 50s period. This suggests that over this particular subperiod, $SupA$ and $SupB^{ivxc}$ may in fact be rejecting their respective null hypotheses $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ mainly due to unequal intercepts i.e. the regime specific nature of return predictability may in fact be driven by our business cycle proxy rather than the DY predictor playing a distinct role across expansions versus recessions. This is in line with a recent branch of the predictability literature which argues that DY based predictability has declined due to greater dividend smoothing. Operating within a purely linear setting, KMS2015 documented a very weak return predictability using the dividend yield over the full sample and no evidence of predictability in the post 50s period. In our current context it is also important to point out that as we switch from the post 50s to the post 60s sample the $W_T^{ivxc}(\hat{\lambda})$ appears to revert and corroborate more strongly the earlier inferences based on the full IPGR sample.

Our use of alternative drivers of episodic predictability beyond IPGR is here helpful for exploring further the post-war period and assessing the robustness of our IPGR based results. Using both the ADS and CFNAIMA series as threshold variables we note strong rejections of the null hypothesis on the basis of our $W_T^{ivxc}(\hat{\lambda})$ statistic across all magnitudes of δ. Combined with our clear-cut results based on IPGR (1927-2013) we view our results as providing strong empirical evidence in support of countercyclical predictability of stock returns using DY. This finding also highlights the crucial importance that needs to be given to the time varying nature of predictability when evaluating the predictive power of any variable for future stock returns. It is also interesting to point out that our use of CFNAIMA3 as a threshold variable led to an estimate of the threshold parameter given by $\hat{\gamma} = -0.662$ which corresponds very precisely to the Chicago Fed guidelines of interpreting a CFNAIMA3 below -0.7 as signalling an increased likelihood that a recession has begun. Similarly, we obtained $\hat{\gamma} = -0.012$ for the cutoff associated with IPGR (1927-2013) effectively splitting the sample into periods of positive and negative Industrial Production growth. The ADS index led to $\hat{\gamma} = -0.99$, a negative magnitude also interpreted as signalling deteriorating economic conditions.

Our BM based inferences lead to more ambiguous outcomes and display greater sensitivity to both the choice of the threshold variable and periods of analysis. It is clear however that with the exception of the full historical sample period under IPGR there is very little support for any robust predictive power. An outcome that is also consistent with what has been documented in the linear predictive regression literature.

For the DP series and regardless of the sample period considered we note a consistent and strong rejection of the null hypotheses on the basis of the $SupB^{ivxc}$ statistic, indicating strong regime specific effects in the behaviour of stock returns. However in this instance and unlike the DY series our $W_T^{ivxc}(\hat{\lambda})$ test statistic mostly fails to reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. This suggests that the $SupB^{ivxc}$ based rejections were most likely driven by unequal intercepts and highlights the importance of our new inferences. Finally, regarding the predictive power of the earnings yield (EP) our results point to very little evidence of regime specific predictability. With the exception of the full sample period under IPGR, inferences based on both $W_T^{ivxc}(\hat{\lambda})$ and $SupB^{ivxc}$ are typically unable to reject their respective null hypotheses at reasonable significance levels.

6 Conclusions

We developed a toolkit for assessing the predictability induced by a single persistent predictor in an environment that allows predictability to kick in during particular economic episodes and affect all or only some parameters of the model. Our threshold based framework and testing methodology can be used to explore the possibility that the predictive power of highly persistent predictors such as interest rates, valuation ratios and numerous other economic and financial variables may be varying across time in an economically meaningful way with alternating periods of strong versus weak or no predictability. More importantly the core contribution of this paper was to provide a setting that allows us to distinguish predictability induced by a specific predictor from predictability that may be solely driven by economic episodes (e.g. stock returns differing across recessions and expansions). Our empirical results have highlighted the misleading or at best incomplete conclusions one may reach if such regime specific effects are ignored when assessing predictability.

Although our operating assumptions were closely aligned to the those commonly considered in the linear predictive regression literature and allowed for a rich interaction between the random disturbances driving our predictive threshold specification it is important to recognise the limitations of our conditional homoskedasticity restriction imposed on u_t. In the context of our application, a standard LM test for ARCH effects (up to order 12) in the residuals of our predictive threshold specifications under CFNAIMA3 and ADS did not reject the null hypothesis of no such effects at reasonable significance levels and similarly for IPGR within the post 50s sample but strong ARCH effects were supported by the data when considering the full sample period under IPGR (i.e. IPGR (1927-2013)).

In KMS2015 (Theorem 1) the authors showed that allowing for GARCH(p,q) errors within their *linear* predictive regression setting had no influence on the asymptotics of their IVX based Wald statistic. The key driver of this important and unusual result was the near integratedness of the predictor with the robustness to GARCH of the Wald statistic shown to fail under purely stationary predictors. Allowing for GARCH type effects in our setting can be particularly challenging when it comes to establishing the limiting properties of objects such as $\sum u_t I(q_{t-d} \leq \gamma)$ and $\sum u_t x_{t-1} I(q_{t-d} \leq \gamma)$ under very general dependence structures linking $u_t's$ and $q_t's$ while also allowing for ARCH type dependence in the $u_t's$ but it is an obvious extension we will consider in follow up work. Our on-line appendix provides a broad range of size simulations under GARCH effects and suggests very little impact on inferences based on $W_T^{ivxc}(\hat{\lambda})$, supporting the conjecture that KMS2015s result may also hold within our setting.

APPENDIX

PROOF OF PROPOSITION 1: Since under $H_0 : \beta_1 = \beta_2 = 0$ the threshold model is given by $y_t = \alpha_1 I_{1t-1} + \alpha_2 I_{2t-1} + u_t$, all assumptions of Gonzalo and Pitarakis (2002) are satisfied implying the statement in (i). The result in Part (ii) follows by first noting that the minimiser of $S_T(\lambda)$ is numerically identical to the maximiser of the Wald statistic $W_T(\lambda)$ for testing $H_0 : \alpha_1 = \alpha_2$ in the above restricted specification. This Wald statistic is given by

$$W_T(\lambda) \quad = \quad \left(\frac{\sum u_t I_{1t-1}}{\sum I_{1t-1}} - \frac{\sum u_t I_{2t-1}}{\sum I_{2t-1}}\right)^2 \frac{\sum I_{1t-1} \sum I_{2t-1}}{T\hat{\sigma}_u^2(\lambda)} \tag{15}$$

with $\hat{\sigma}_u^2(\lambda)$ denoting the residual variance obtained from the above mean shift specification. Under $H_0 : \alpha_1 = \alpha_2$ and A1-A3 a suitable Law of Large Numbers (see White (2000, p.58)) ensures that $\hat{\sigma}_u^2(\lambda) \stackrel{p}{\to} \sigma_u^2$. From Caner and Hansen (2001) we have $\sum_{t=1}^T u_t I_{1t-1}/\sqrt{T} \Rightarrow B_u(\lambda)$. The strict stationarity and ergodicity of the $I'_{it}s$ further ensures that $\sum I_{1t-1}/T \stackrel{p}{\to} \lambda$ and $\sum I_{2t-1}/T \stackrel{p}{\to} (1-\lambda)$. It now follows from the Continuous Mapping Theorem that

$$W_T(\lambda) \quad \Rightarrow \quad \frac{[B_u(\lambda) - \lambda B_u(1)]^2}{\sigma_u^2 \lambda(1-\lambda)}. \tag{16}$$

The desired result then follows from the continuity of the argmax functional and the fact that the limit process has a unique maximum in Λ with probability 1 (see Theorem 2.7 in Kim and Pollard (1990)). ■

Before proceeding with the limiting properties of $W_T^{ols}(\hat{\lambda})$ we briefly set out the notation associated with each of its components under our DGP in (1) also applying suitable normalisations. Defining

$$\begin{aligned} g_{it}(\lambda) &\equiv \frac{\sum I_{it-1}}{T} \frac{\sum y_t x_{t-1} I_{it-1}}{T} - \frac{\sum y_t I_{it-1}}{\sqrt{T}} \frac{\sum x_t I_{it-1}}{T\sqrt{T}} \\ \Delta_{it}(\lambda) &\equiv \frac{\sum x_{t-1}^2 I_{it-1}}{T^2} \frac{\sum I_{it-1}}{T} - \left(\frac{\sum x_{t-1} I_{it-1}}{T\sqrt{T}}\right)^2 \end{aligned} \tag{17}$$

for $i = 1, 2$, standard algebra leads to

$$\frac{X_\lambda^{*\prime} y^*}{T} \quad = \quad \begin{pmatrix} \dfrac{g_{1t}(\lambda)}{\sum I_{1t-1}/T} \\ \\ \dfrac{g_{2t}(\lambda)}{\sum I_{2t-1}/T} \end{pmatrix} \tag{18}$$

and

$$\left(\frac{X_\lambda^{*\prime} X_\lambda^*}{T^2}\right)^{-1} \quad = \quad \begin{pmatrix} \dfrac{\sum I_{1t-1}/T}{\Delta_{1t}(\lambda)} & 0 \\ 0 & \dfrac{\sum I_{2t-1}/T}{\Delta_{2t}(\lambda)} \end{pmatrix}. \tag{19}$$

Given our null hypothesis of interest it is also useful to specialise (18) across the two scenarios on the intercepts, namely $y_t = \alpha + u_t$ when $\alpha_1 = \alpha_2$ and $y_t = \alpha_1 I_{1t-1}^0 + \alpha_2 I_{2t-1}^0 + u_t$ when $\alpha_1 \neq \alpha_2$. In this latter case I_{it-1}^0 refers to the indicator function evaluated at the true threshold parameter λ^0. We write

$$\left[\frac{X_\lambda^{*\prime} y^*}{T}\right]_{\alpha_1=\alpha_2} \quad = \quad \begin{pmatrix} \dfrac{g_{1t}(\lambda)|_{\alpha_1=\alpha_2}}{\sum I_{1t-1}/T} \\ \\ \dfrac{g_{2t}(\lambda)|_{\alpha_1=\alpha_2}}{\sum I_{2t-1}/T} \end{pmatrix} \tag{20}$$

and

$$\left[\frac{X_\lambda^{*\prime} y^*}{T}\right]_{\alpha_1\neq\alpha_2} \quad = \quad \begin{pmatrix} \dfrac{g_{1t}(\lambda)|_{\alpha_1\neq\alpha_2}}{\sum I_{1t-1}/T} \\ \\ \dfrac{g_{2t}(\lambda)|_{\alpha_1\neq\alpha_2}}{\sum I_{2t-1}/T} \end{pmatrix} \tag{21}$$

with

$$g_{it}(\lambda)|_{\alpha_1=\alpha_2} = \frac{\sum I_{it-1}}{T}\frac{\sum u_t x_{t-1} I_{it-1}}{T} - \frac{\sum u_t I_{it-1}}{\sqrt{T}}\frac{\sum x_{t-1} I_{it-1}}{T\sqrt{T}} \tag{22}$$

$$\begin{aligned} g_{it}(\lambda)|_{\alpha_1\neq\alpha_2} &= \frac{\sum I_{it-1}}{T}\left(\alpha_1\frac{\sum x_{t-1}I_{it-1}I^0_{1t-1}}{T} + \alpha_2\frac{\sum x_{t-1}I_{it-1}I^0_{2t-1}}{T} + \frac{\sum u_t x_{t-1}I_{it-1}}{T}\right) \\ &- \frac{\sum x_{t-1}I_{it-1}}{T}\left(\alpha_1\frac{\sum I_{it-1}I^0_{1t-1}}{T} + \alpha_2\frac{\sum I_{it-1}I^0_{2t-1}}{T} + \frac{\sum u_t I_{it-1}}{T}\right) \end{aligned} \tag{23}$$

Before proceeding with the proof of Proposition 2 we introduce the following auxiliary Lemma that is used for establishing the asymptotic properties of the sample moments in (23).

LEMMA A1. Under Assumptions A1-A3, $T|\hat{\lambda}-\lambda_0| = O_p(1)$ and letting $U_t \equiv F(q_t)$, as $T\to\infty$ we have

$$\frac{1}{\sqrt{T}}\sum I(U_{t-1}\leq\hat{\lambda})I(U_{t-1}\leq\lambda_0) - \frac{1}{\sqrt{T}}\sum I(U_{t-1}\leq\lambda_0) \quad \xrightarrow{p} \quad 0 \tag{24}$$

PROOF of LEMMA A1: We need to establish that for every $\varepsilon>0$ and $\delta>0$

$$\lim_{T\to\infty} P\left[\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[I\left(q_t<\widehat{\lambda}\right) - I\left(q_t<\lambda\right)\right]I\left(q_t<\lambda\right)\right| > \varepsilon\right] < \delta.$$

Given that

$$\begin{aligned} \left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[I\left(q_t<\widehat{\lambda}\right) - I\left(q_t<\lambda\right)\right]I\left(q_t<\lambda\right)\right| &\leq \left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[I\left(\lambda-\left|\widehat{\lambda}-\lambda\right|<q_t<\lambda+\left|\widehat{\lambda}-\lambda\right|\right)\right]\right| \\ &\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right) \end{aligned}$$

with $A_t(\lambda,d) = I(\lambda-|d|<q_t<\lambda+|d|)$, it will be enough to prove that

$$\lim_{T\to\infty} P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right) > \varepsilon\right] < \delta$$

for every $\varepsilon>0$ y $\delta>0$. Since $\hat{\lambda}$ is such that $T|\hat{\lambda}-\lambda_0| = O_p(1)$, therefore for every $\delta>0$, $\exists\Delta_\delta<\infty$ and an integer $T_\delta\geq 1$ such that

$$P\left[\left|\widehat{\lambda}-\lambda\right| > \frac{\Delta_\delta}{T}\right] < \delta \qquad \text{for } \forall T>T_\delta,$$

and also

$$\begin{aligned} P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right] &= P\left[\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right\}\cap\left\{\left|\widehat{\lambda}-\lambda\right|\leq\frac{\Delta_\delta}{T}\right\}\right] + \\ &+ P\left[\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right\}\cap\left\{\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right\}\right] \\ &\leq P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right] + P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right]. \end{aligned}$$

Using Markov's inequality

$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right] \leq \frac{\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1}{\varepsilon} \leq \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\|A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1}{\varepsilon}$$

and under our assumption on the boundedness of the pdf of q_t away from 0 and ∞ over each bounded set

$$\left\|A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1 = \left\|I\left(\lambda-\frac{\Delta_\delta}{T}<q_t<\lambda+\frac{\Delta_\delta}{T}\right)\right\|_1 \leq M\frac{\Delta_\delta}{T}$$

therefore

$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]\le\frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\|A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)\right\|_1}{\varepsilon}$$
$$\le\frac{\sqrt{T}M\frac{\Delta_\delta}{T}}{\varepsilon}\le\frac{M\Delta_\delta}{\varepsilon\sqrt{T}}.$$

Putting together these results we have that for every $\varepsilon>0$ and $\delta>0$ $\exists T_{\varepsilon\delta}<\infty$ such that for every $T>T_{\varepsilon\delta}$

$$P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right]<\delta$$
$$P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]<\delta$$

and then

$$\lim_{T\to\infty}P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\widehat{\lambda}-\lambda\right)>\varepsilon\right]\le\lim_{T\to\infty}P\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}A_t\left(\lambda,\frac{\Delta_\delta}{T}\right)>\varepsilon\right]+\lim_{T\to\infty}P\left[\left|\widehat{\lambda}-\lambda\right|>\frac{\Delta_\delta}{T}\right]<2\delta.$$

leading to the desired result. ■

PROOF OF PROPOSITION 2. We initially consider the case $\alpha_1\neq\alpha_2$. Given the T-consistency of $\hat{\lambda}$ for λ_0, $T|\hat{\lambda}-\lambda_0|=O_p(1)$, and our result in Lemma A1 we have

$$g_{it}(\hat{\lambda})|_{\alpha_1\neq\alpha_2}=\frac{\sum I^0_{it-1}}{T}\frac{\sum x_{t-1}u_tI^0_{it-1}}{T}-\frac{\sum u_tI^0_{it-1}}{\sqrt{T}}\frac{\sum x_{t-1}I^0_{it-1}}{T\sqrt{T}}+o_p(1),\tag{25}$$
$$\Delta_{it}(\hat{\lambda})=\frac{\sum I^0_{it-1}}{T}\frac{\sum x^2_{t-1}I^0_{it-1}}{T^2}-\left(\frac{\sum x_{t-1}I^0_{it-1}}{T\sqrt{T}}\right)^2+o_p(1).\tag{26}$$

Using Lemma 1 in Gonzalo and Pitarakis (2012), Theorem 1 in Caner and Hansen (2001) together with the continuous mapping theorem we have

$$\begin{aligned}
g_{1t}(\hat{\lambda})|_{\alpha_1\neq\alpha_2}&\Rightarrow\lambda_0\left(\int J_c(r)dB_u(r,\lambda_0)-B_u(\lambda_0)\int J_c(r)\right),\\
g_{2t}(\hat{\lambda})|_{\alpha_1\neq\alpha_2}&\Rightarrow(1-\lambda_0)\left(\int J_c(r)(dB_u(r)-dB_u(r,\lambda_0))-(B_u(1)-B_u(\lambda_0))\int J_c(r)\right),\\
\Delta_{1t}(\hat{\lambda})&\Rightarrow\lambda_0^2\int J^*_c(r)^2,\\
\Delta_{2t}(\hat{\lambda})&\Rightarrow(1-\lambda_0)^2\int J^*_c(r)^2.
\end{aligned}\tag{27}$$

Next, using (27) in (20)-(21) and rearranging gives

$$\frac{X^{*\prime}_{\hat{\lambda}}X^*_{\hat{\lambda}}}{T^2}\Rightarrow\int J^*_c(r)^2\begin{pmatrix}\lambda_0&0\\0&(1-\lambda_0)\end{pmatrix}\tag{28}$$

and

$$\frac{X^{*\prime}_{\hat{\lambda}}y^*}{T}\Rightarrow\begin{pmatrix}\int J_c(r)dB_u(r,\lambda_0)-B_u(\lambda_0)\int J_c(r)\\ \int J_c(r)(dB_u(r)-dB_u(r,\lambda_0))-(B_u(1)-B_u(\lambda_0))\int J_c(r)\end{pmatrix}.\tag{29}$$

Combining (28)-(29) into (7) and using $\hat{\sigma}^2(\hat{\lambda})\xrightarrow{p}\sigma^2_u$ leads to

$$\begin{aligned}
W^{ols}_T(\hat{\lambda})&\Rightarrow\frac{[\int J_cdB_u(r,\lambda_0)-B_u(\lambda_0)\int J_c(r)]^2}{\sigma^2_u\lambda_0\int J^*_c(r)^2}+\\
&\quad\frac{[\int J_c(dB_u(r)-dB_u(r,\lambda_0))-(B_u(1)-B_u(\lambda_0))\int J_c(r)]^2}{\sigma^2_u(1-\lambda_0)\int J^*_c(r)^2}\\
&\equiv\frac{[\int J^*_c(r)dG_u(r,\lambda_0)]^2}{\sigma^2_u\lambda_0(1-\lambda_0)\int J^*_c(r)^2}+\frac{[\int J^*_c(r)dB_u(r)]^2}{\sigma^2_u\int J^*_c(r)^2}\\
&\equiv\frac{[B_u(\lambda_0)-\lambda_0B_u(1)]^2}{\sigma^2_u\lambda_0(1-\lambda_0)}+\frac{[\int J^*_c(r)dB_u(r)]^2}{\sigma^2_u\int J^*_c(r)^2}
\end{aligned}\tag{30}$$

with $G_u(r,\lambda_0) = B_u(r,\lambda_0) - \lambda_0 B_u(r,1)$ denoting a Kiefer Process with covariance function $\sigma_u^2(r_1 \wedge r_2)\lambda_0(1-\lambda_0)$. The result in Proposition 2 then follows by noting that $J_c(r)$ and $G_u(r,\lambda_0)$ are uncorrelated and hence independent due to their Gaussianity so that $\int J_c^*(r)dG_u(r,\lambda) \equiv N(0, \sigma_u^2\lambda_0(1-\lambda_0)\int J_c^*(r)^2)$ conditionally on the realisation of $J_c(r)$. Thus normalising by $\sigma_u^2\lambda_0(1-\lambda_0)\int J_c^*(r)^2$ gives the $\chi^2(1)$ limit which is also the unconditional distribution since not dependent on the realisation of $J_c(r)$. The case $\alpha_1 = \alpha_2$ can be treated in a similar fashion with λ_0 replaced by the random variable λ^* in (30) as in Theorem 5 of Caner and Hansen (2001) with the nuance that our random maximiser λ^* does not depend on any nuisance parameters. The main result for this case then follows by noting that the first component in the right hand side of (30) evaluated at λ^* is a $\chi^2(1)$ random variable. This latter point is a consequence of the independence of λ^* and $[\int J_c^*(r)d\widetilde{G}_u(r,\lambda)]^2/\int J_c^*(r)^2$ or equivalently of $\int d\widetilde{G}_u(r,\lambda)$ (which λ^* is the maximiser of) and $[\int J_c^*(r)d\widetilde{G}_u(r,\lambda)]^2/\int J_c^*(r)^2$ where we let $\widetilde{G}_u(r,\lambda) \equiv G_u(r,\lambda)/\sqrt{\lambda(1-\lambda)}$. Indeed, letting $\widetilde{BB}_u(\lambda) \equiv [B_u(\lambda) - \lambda B_u(1)]^2/\lambda(1-\lambda)$ then given that $P[\widetilde{BB}_u(\lambda^*) \leq x|\lambda^* = \underline{\lambda}] = CHISQ(x)$ for any given $\underline{\lambda}$, independence here implies that the unconditional distribution of $\widetilde{BB}_u(\lambda^*)$ must also be $\chi^2(1)$. To note the independence of $\int d\widetilde{G}_u(r,\lambda) \equiv \zeta$ say, and $[\int J_c^*(r)d\widetilde{G}_u(r,\lambda)]/\sqrt{\int J_c^*(r)^2} \equiv M$ which is $N(0,1)$ as shown above, it is useful to point out that M is of the form $\mu'\zeta/\sqrt{\mu'\mu}$ and the two quantities have joint characteristic function $\psi(\zeta, M) = E[e^{it\zeta + isM}] = E[E[e^{it\zeta+isM}|\mu]]$. It is now straightforward to note that $\psi(\zeta, M) = E[e^{it\zeta}]E[e^{isM}]$ as $\widetilde{G}_u(r,\lambda)$ is independent of $B_v(r)$ and hence of $J_c^*(r)$ (see Gonzalo and Pitarakis (2012, p. 232) and Gonzalo and Pitarakis (2012, Supplementary Appendix Section 2.2).

PROOF OF PROPOSITION 3. The result follows directly from the independence of $B_u(r,\lambda)$ and $B_v(r)$ under $\sigma_{uv} = 0$ also implying the independence of $J_c^*(r)$ and $B_u(r,\lambda)$ and from which mixed normality follows. Noting also the independence of the two components in the right hand side of (30) established in Gonzalo and Pitarakis (2012). ■

Before proceeding with the proof of Proposition 4 it will be convenient to reformulate the components of (11) in an explicit and suitably normalised form. Defining

$$\begin{aligned}
m_{it}(\lambda) &\equiv \frac{\sum I_{it-1}}{T}\left[\left(\frac{\sum I_{it-1}}{T}\frac{\sum \tilde{h}_{t-1}^2 I_{it-1}}{T^{1+\delta}}\right) - \frac{1}{T^{1-\delta}}\left(\frac{\sum \tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\delta}}\right)^2\right] \\
\pi_{it}(\lambda) &\equiv \left(\frac{\sum I_{it-1}}{T}\frac{\sum \tilde{h}_{t-1}x_{t-1}I_{it-1}}{T^{1+\delta}} - \frac{\sum \tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\delta}}\frac{\sum x_{t-1}I_{it-1}}{T\sqrt{T}}\right)^2 \\
n_{it}(\lambda) &\equiv \frac{\sum I_{it-1}}{T}\frac{\sum y_t\tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\frac{\delta}{2}}} - \frac{1}{T^{\frac{1}{2}-\frac{\delta}{2}}}\left(\frac{\sum \tilde{h}_{t-1}I_{it-1}}{T^{\frac{1}{2}+\delta}}\frac{\sum y_t I_{it-1}}{\sqrt{T}}\right)
\end{aligned} \tag{31}$$

for $i = 1,2$ we can write

$$\frac{1}{T^{1+\delta}}H_\lambda^{*\prime}X_\lambda^*(H_\lambda^{*\prime}H_\lambda^*)^{-1}H_\lambda^{*\prime}X_\lambda^* = \begin{pmatrix} \frac{\pi_{1t}(\lambda)}{m_{1t}(\lambda)} & 0 \\ 0 & \frac{\pi_{2t}(\lambda)}{m_{2t}(\lambda)} \end{pmatrix} \tag{32}$$

and

$$T^{\frac{1+\delta}{2}}\hat{\beta}_{ivx}(\lambda) = \begin{pmatrix} \frac{n_{1t}(\lambda)}{\sqrt{\pi_{1t}(\lambda)}} \\ \frac{n_{2t}(\lambda)}{\sqrt{\pi_{2t}(\lambda)}} \end{pmatrix} \tag{33}$$

PROOF OF PROPOSITION 4. We concentrate on the case $\alpha_1 \neq \alpha_2$ with the underlying T-consistency of $\hat{\lambda}$ for λ_0. We also recall that $\tilde{h}_t = \sum_{j=1}^t \phi_T^{t-j}\Delta x_j$ and let $h_t = \sum_{j=1}^t \phi_T^{t-j}v_j$. It now follows directly from (31) and Lemma 3.1 in Phillips and Magdalinos (2009) that

$$\begin{aligned}
m_{it}(\hat{\lambda}) &= \left(\frac{\sum I_{it-1}^0}{T}\right)^2\frac{\sum h_{t-1}^2 I_{it-1}^0}{T^{1+\delta}} + o_p(1) \\
\pi_{it}(\hat{\lambda}) &= \left(\frac{\sum h_{t-1}I_{it-1}^0}{T^{\frac{1}{2}+\delta}}\frac{\sum x_{t-1}I_{it-1}^0}{T\sqrt{T}} - \frac{\sum I_{it-1}^0}{T}\frac{\sum h_{t-1}x_{t-1}I_{it-1}^0}{T^{1+\delta}}\right)^2 + o_p(1).
\end{aligned} \tag{34}$$

Under our assumptions A1-A3 the following deduce directly from Phillips and Magdalinos (2009, eq. (14))

$$\begin{aligned}
m_{1t}(\hat{\lambda}) &\Rightarrow \lambda_0^3\,\frac{\omega_v^2}{2} \\
m_{2t}(\hat{\lambda}) &\Rightarrow (1-\lambda_0)^3\,\frac{\omega_v^2}{2}
\end{aligned} \tag{35}$$

since $\sum h_{t-1}^2(I_{1t-1}^0 - \lambda_0)/T^{1+\delta} \xrightarrow{p} 0$. It also follows that

$$\begin{aligned} \pi_{1t}(\hat{\lambda}) &\Rightarrow \lambda_0^4 \left[\omega_v^2 + \int J_c^*(r)dJ_c(r)\right]^2 \\ \pi_{2t}(\hat{\lambda}) &\Rightarrow (1-\lambda_0)^4 \left[\omega_v^2 + \int J_c^*(r)dJ_c(r)\right]^2 \end{aligned} \tag{36}$$

so that

$$\frac{1}{T^{1+\delta}} H_{\hat{\lambda}}^{*\prime} X_{\hat{\lambda}}^* (H_{\hat{\lambda}}^{*\prime} H_{\hat{\lambda}}^*)^{-1} H_{\hat{\lambda}}^{*\prime} X_{\hat{\lambda}}^* \Rightarrow \frac{[\omega_v^2 + \int J_c^*(r)dJ_c(r)]^2}{\omega_v^2/2} \begin{pmatrix} \lambda_0 & 0 \\ 0 & 1-\lambda_0 \end{pmatrix} \tag{37}$$

Next, we also have

$$n_{it}(\hat{\lambda}) = \frac{\sum I_{it-1}^0}{T} \frac{\sum u_t h_{t-1} I_{it-1}^0}{T^{\frac{1}{2}+\frac{\delta}{2}}} + o_p(1) \tag{38}$$

and Lemma 3.2 in Phillips and Magdalinos (2009) together with (35) ensure the following holds

$$\begin{aligned} \frac{1}{T^{\frac{1}{2}+\frac{\delta}{2}}} \sum h_{t-1} u_t I_{1t-1}^0 &\Rightarrow N(0, \lambda_0 \sigma_u^2 \frac{\omega_v^2}{2}) \\ \frac{1}{T^{\frac{1}{2}+\frac{\delta}{2}}} \sum h_{t-1} u_t I_{2t-1}^0 &\Rightarrow N(0, (1-\lambda_0) \sigma_u^2 \frac{\omega_v^2}{2}) \end{aligned} \tag{39}$$

which when rearranged with (37) and using the continuous mapping theorem within $W_T^{ivx}(\hat{\lambda})$ leads to the desired result. The case $\alpha_1 = \alpha_2$ can be treated in a similar fashion with λ_0 replaced by the random variable λ^* as formulated in Proposition 1. ■

REMARK A1. The $SupB^{ivx}$ statistic developed in Gonzalo and Pitarakis (2012) was formulated as $SupB^{ivx} \equiv \sup_\lambda W_T^A(\lambda) + W_T^{ivx}(\beta = 0)$ with $W_T^A(\lambda)$ referring to the Wald statistic for testing $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ in (1) and $W_T^{ivx}(\beta = 0)$ was the simple IVX based Wald statistic for testing $H_0 : \beta = 0$ in $y_t = \alpha + \beta x_{t-1} + u_t$ i.e. exactly analogous to the Wald statistic developed in KMS2015. The finite sample corrected version of $SupB^{ivx}$ considered in our application above simply replaces $W_T^{ivx}(\beta = 0)$ with its formulation in KMS2015 (pp. 1514-1515, equations (19)-(21)).

REFERENCES

Ang, A. and G. Bekaert (2008), "Stock Return Predictability: Is it There?", *Review of Financial Studies*, 20,651-707.

Aruoba, S.B., Diebold, F.X. and Scotti, C. (2009), "Real-Time Measurement of Business Conditions," *Journal of Business and Economic Statistics*, 27, 417-27.

Bandi, F. and B. Perron (2008), "Long Run Risk Return Trade-Offs," *Journal of Econometrics*, 143, 349-374.

Breitung, J. and M. Demetrescu (2014), "Instrumental Variable and Variable Addition Based Inference in Predictive Regressions," *Journal of Econometrics*, forthcoming.

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

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

Gargano, A., (2013), "Return Predictability in Recessions: An Asset Pricing Perspective," Unpublished Manuscript.

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

Gonzalo, J., and Pitarakis, J. (2012), "Regime Specific Predictability in Predictive Regressions," *Journal of Business and Economic Statistics*, 30, 229-241.

Goyal, A., and Welch, I. (2014), "Update to *A Comprehensive Look at the Empirical Performance of Equity Premium Prediction*," Unpublished Manuscript.

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

Henkel, S. J., Martin, J. S., and Nardari, F. (2011), "Time-Varying Short-Horizon Predictability," *Journal of Financial Economics*, 99, 560-580.

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

Kim, J., and Pollard,D. (1990), "Cube Root Asymptotics," *Annals of Statistics*, 18, 191-219.

Kostakis, A., Magdalinos, A., and Stamatogiannis, M. (2015), "Robust Econometric Inference for Stock Return Predictability," *Review of Financial Studies,* 28, 1506-1553.

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

Paye, B. S. and Timmermann, A. (2006), "Instability of Return Prediction Models," *Journal of Empirical Finance*, 13, 274-315.

Pesaran, M. H., and Timmermann, A. (1995), "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance,* 50, 1201-2228.

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

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*, 76, 1207-1217.

Rossi, B., (2005), "Optimal Tests for Nested Model Selection with Underlying Parameter Stability," *Econometric Theory*, 21, 962-990.

Rossi, B., (2007), "Expectations Hypotheses Tests at Long Horizons," *Econometrics Journal*, 10, 1-26.

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

Wei, M., and Wright, J. (2013), "Reverse Regressions and Long Horizon Forecasting," *Journal of Applied Econometrics,* 28, 353-371.

Welch, I., and Goyal, A. (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21, 1455-1508.

White, H. (2000), *Asymptotic Theory for Econometricians.* Second Revised Edition. Academic Press.