Journal of Econometrics 227 (2022) 85–113

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Testing for episodic predictability in stock returns ☆

Matei Demetrescu [a], Iliyan Georgiev [b], Paulo M.M. Rodrigues [c], A.M. Robert Taylor [d,*]

[a] Institute for Statistics and Econometrics, Christian-Albrechts-University of Kiel, Germany
[b] Department of Economics, University of Bologna, Italy
[c] Banco de Portugal and NOVA School of Business and Economics, Portugal
[d] Essex Business School, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom

ARTICLE INFO

Article history:
Received 13 November 2018
Received in revised form 15 November 2019
Accepted 5 January 2020
Available online 15 February 2020

JEL classification:
C12
C22

Keywords:
Predictive regression
Rolling and recursive IV estimation
Persistence
Endogeneity
Conditional and unconditional heteroskedasticity

ABSTRACT

Standard tests based on predictive regressions estimated over the full available sample data have tended to find little evidence of predictability in stock returns. Recent approaches based on the analysis of subsamples of the data suggest in fact that predictability where it occurs might exist only within so-called "pockets of predictability" rather than across the entire sample. However, these methods are prone to the criticism that the subsample dates are endogenously determined such that the use of standard critical values appropriate for full sample tests will result in incorrectly sized tests leading to spurious findings of stock returns predictability. To avoid the problem of endogenously-determined sample splits, we propose new tests derived from sequences of predictability statistics systematically calculated over subsamples of the data. Specifically, we will base tests on the maximum of such statistics from sequences of forward and backward recursive, rolling, and double-recursive predictive subsample regressions. We develop our approach using the over-identified instrumental variable-based predictability test statistics of Breitung and Demetrescu (2015). This approach is based on partial-sum asymptotics and so, unlike many other popular approaches including, for example, those based on Bonferroni corrections, can be readily adapted to implementation over sequences of subsamples. We show that the limiting null distributions of our proposed test statistics depend in general on whether the putative predictor is strongly or weakly persistent and on any heteroskedasticity present (indeed on any time-variation present in the unconditional variance matrix of the innovations), the latter even if the subsample statistics are based on heteroskedasticity-robust standard errors. As a consequence, we develop fixed regressor wild bootstrap implementations of the tests which we demonstrate to be first-order asymptotically valid. Finite sample behaviour against a variety of temporarily predictable processes is considered. An empirical application to US stock returns illustrates the usefulness of the new predictability testing methods we propose.

☆ We are grateful to the guest editor, Michael McAleer, two anonymous referees, Serena Ng and Torben Anderson for their helpful and constructive comments on earlier versions of this paper. Demetrescu gratefully acknowledges the support of the German Research Foundation (DFG) through the project DE 1617/4-2. Rodrigues gratefully acknowledges financial support from the Portuguese Science Foundation (FCT) through project PTDC/EGE-ECO/28924/2017, and (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, LISBOA-01-0145-FEDER-022209), POR Lisboa (LISBOA-01-0145-FEDER-007722, LISBOA-01-0145-FEDER-022209) and POR Norte (LISBOA-01-0145-FEDER-022209). Taylor gratefully acknowledges financial support provided by the Economic and Social Research Council of the United Kingdom under research grant ES/R00496X/1.

* Corresponding author.
E-mail address: robert.taylor@essex.ac.uk (A.M.R. Taylor).

https://doi.org/10.1016/j.jeconom.2020.01.001

1. Introduction

A large body of empirical research has been undertaken investigating whether stock returns can be predicted. Therein, a wide range of financial and macroeconomic variables have been considered as putative predictors for returns, including valuation ratios such as the dividend-price ratio, dividend yield, earnings-price ratio, book-to-market ratio, various interest rates and interest rate spreads, and macroeconomic variables including inflation and industrial production.

Early empirical studies, including Fama (1981), Keim and Stambaugh (1986), Campbell (1987), Campbell and Shiller (1988a,b), Fama and French (1988, 1989) and Fama (1990), often found significant evidence of in-sample predictability of U.S. stock index returns, at least over relatively long horizons. It has since been argued, however, that these findings could be spurious. Nelson and Kim (1993) and Stambaugh (1999) show that strongly persistent predictors lead to biased coefficients in predictive regressions if the innovations driving the predictors are correlated with returns, as is argued to be the case for many of the variables used as predictors; e.g., the stock price is a component of both the return and the dividend yield. Goyal and Welch (2003) show that the persistence of dividend-based valuation ratios increased significantly over the typical sample periods used in empirical studies, and argue that, as a consequence, out-of-sample predictions using these variables are no better than from a no-change strategy. Predictability tests which are asymptotically valid when the predictor is strongly persistent and driven by innovations which are correlated with returns have been proposed in Cavanagh et al. (1995), Campbell and Yogo (2006), Kostakis et al. (2015), Breitung and Demetrescu (2015), Elliott et al. (2015) and Jansson and Moreira (2006), *inter alia*. When such robust techniques are used the statistical evidence of predictability is considerably weaker and often disappears completely; see, among others, Ang and Bekaert (2007), Boudoukh et al. (2007), Welch and Goyal (2008) and Breitung and Demetrescu (2015).

The foregoing approaches are based on a maintained assumption that the coefficients of the predictive regression model are constant over time. However, there are several reasons to suspect that if returns are predictable, then it is likely to be a time-varying phenomenon. The business cycle, time-varying risk aversion, rare disasters, structural breaks, speculative bubbles, investor's market sentiment, and regime changes in monetary policy have all be cited as possible reasons; see, e.g., Pesaran and Timmermann (2002). For example, significant changes in monetary policy and financial regulations could lead to shifts in the relationship between macroeconomic variables and the fundamental value of stocks, via the impact of these changes on economic growth and the growth rates of earnings and dividends. Timmermann (2008) argues that for most time periods returns are not predictable but that there are 'pockets in time' where evidence of local predictability is seen. In particular, if predictability exists as a result of market inefficiency rather than because of time-varying risk premia, then rational investors will attempt to exploit its presence to earn abnormal profits. Assuming a large-enough proportion of investors are rational, this behaviour will eventually cause the predictive power of the relevant predictor to be eliminated. If a variable begins to have predictive power for returns then a window of predictability might exist before investors learn about that relationship, but it will eventually disappear; see, in particular, Paye and Timmermann (2006), Timmermann (2008) and Farmer et al. (2018). It therefore seems reasonable to consider the possibility that the predictive relationship might change over time, so that over a long span of data one may observe some windows of time during which predictability occurs.

A growing body of empirical evidence is supportive of the view that the slope parameter in prediction models for returns varies over time. Henkel et al. (2011) find that return predictability in the stock market appears to be closely linked to economic recessions with dividend yield and term structure variables displaying predictive power only during recessions. Similarly, Gargano et al. (2019) find that commodity returns are predictable using macroeconomic information, but again only during recessions. Lettau and Ludvigsson (2001) find evidence of instability in the predictive ability of the dividend and earnings yield in the second half of the 1990s. Goyal and Welch (2003) and Ang and Bekaert (2007) find instability in prediction models for U.S. stock returns based on the dividend yield in the 1990s. Other studies which report evidence of time-varying behaviour in stock return predictability include Barberis (2000), Lettau and van Nieuwerburgh (2008), Welch and Goyal (2008), Pástor and Stambaugh (2009, 2012), Pettenuzzo and Timmermann (2011), Dangl and Halling (2012), Gonzalo and Pitarakis (2012), Rapach and Wohar (2006) and Giannetti (2007), *inter alia*. In the context of predicting the equity premium, Kolev and Karapandza (2017) find that, for a given set of predictors, alternative data splits often lead to strongly contradictory outcomes concerning return predictability. Paye and Timmermann (2006) undertake a comprehensive analysis of prediction model instability for international stock market indices using conventional Bai–Perron structural break tests and report statistically significant evidence of structural breaks for many of the countries considered, arguing that the "[e]mpirical evidence of predictability is not uniform over time and is concentrated in certain periods" (*op. cit.* p. 312). Paye and Timmermann (2006) also cite a number of applied studies which find significant evidence of in-sample (ex post) predictability in returns data but yet find very weak evidence of out-of-sample (ex ante) predictability, and argue that a possible explanation is structural instability in the predictive relations involved.

A limitation of many of the statistical techniques used in previous research on the instability of return prediction models is that they are not designed for use with highly persistent, endogenous predictors. Paye and Timmermann (2006) investigate the effects of persistence and endogeneity of the regressors on the Bai–Perron tests for structural breaks using Monte Carlo simulations. Their simulations reveal that size distortions, whereby parameter change is falsely signalled when none is present, can be substantial. They also show that some of the tests lack power in this context because of the large amount of noise typically present in predictive regression models. Moreover, because tests from predictive regression models based on the full sample of available data will have relatively low power to detect short windows of predictability,

a number of these studies applied such tests to separate subsamples of the data (data splits) with the timings of those subsamples either chosen by the practitioner or performed over a large set of possible subsamples of the data. In both cases the critical values which would apply to the test run on the full sample cannot validly be used. In the former case because the subsamples are endogenously determined. For the latter case the probability of spuriously signalling a predictive relationship when none is present will tend to one as the number of subsamples considered increases; see Inoue and Rossi (2005) for a detailed discussion of this problem in relation to the use of t-tests.

With these issues in mind, our goal is to provide size controlled tests designed to detect predictability regardless of whether it applies across the entire available sample or within pockets of predictability. Our proposed tests are based on predictability statistics obtained from sequences of predictive regressions computed over subsamples of the data. In particular we will consider forward and backward recursive sequences of predictive regressions as well as rolling and double-recursive sequences. For each of these sequences of statistics the proposed test will be given by the largest (in absolute value) outcome. Because the range of subsamples considered in each sequence is set without reference to the particular data set involved this avoids the problem of endogenously chosen sample splits. Moreover, by considering the maximum of the sequences we avoid the spurious detection issues discussed in Inoue and Rossi (2005). As we will demonstrate in the Monte Carlo experiments considered in the paper, each of these sequences has particular patterns of local predictability that it is well designed to detect. For example, the test based on the (forward) reverse recursive sequence of statistics is suited to detecting (beginning-of-sample) end-of-sample pockets of predictability. As such, the reverse recursive based tests could usefully be employed in an on-going monitoring exercise for the emergence of predictive regimes. Because both the forward and reverse recursive sequences contain the usual full sample predictability test, they also deliver tests which have power to detect predictability which holds over the whole sample. For a given window width, tests based on a rolling sequence of statistics are designed to pick up a window of predictability, of roughly the same length, within the data. The double-recursive sequence amounts to considering all possible window width rolling sequences, subject to a minimum width. These then are useful for picking up multiple predictive regimes of potentially different lengths within the data.

The approach we take has implications for the type of predictability statistics that can be used, as it will be necessary to characterise the joint behaviour of the full sequence of statistics in order to conduct inference. Some commonly used testing approaches such as those based on the Bonferroni inequality (e.g. Campbell and Yogo, 2006) or on bias corrections (e.g. Amihud and Hurvich, 2004) are difficult to implement for sequences of statistics; the same holds for technically more involved procedures such as those proposed by Jansson and Moreira (2006) or Elliott et al. (2015). Rather we will use tests based on instrumental variable [IV] estimation which benefit from the fact that closed-form expressions for the test statistics exist, which can be characterised using familiar partial-sum-based asymptotics. Specifically we will adapt the full sample methods from Kostakis et al. (2015) who propose the use of the so-called extended IV, or IVX, approach, and Breitung and Demetrescu (2015) who propose the combination of several instruments with complementary properties. While the marginal null limiting distribution (at least when based on heteroskedasticity-robust standard errors) of any one of the subsample statistics in the sequences considered does not depend on either the degree of persistence or endogeneity of the regressors in the predictive regression, or on any heteroskedasticity present in the shocks, we show this is not the case for the limiting null distribution of the *maximum statistics* from these sequences. We therefore propose fixed regressor wild bootstrap implementations of the maximum tests and demonstrate that these are first-order asymptotically valid under heteroskedasticity, irrespective of whether the regressors are strongly or weakly persistent.

The paper is organised as follows. Section 2 introduces the time-varying predictive regression model we consider together with the assumptions needed for our analysis. Section 3 reviews the standard full sample IV-based predictability tests, while Section 4 details the subsample implementations of these statistics across the various sequence types discussed above and discusses relevant instruments that can validly be used in the context of our proposed approach. Representations for the limiting distributions of these statistics under both the null and local alternatives are provided and are shown to depend on any heteroskedasticity present, regardless of whether the putative predictor follows a strongly persistent process (modelled as near-integrated) or a weakly persistent process (modelled as a stable autoregression). Moreover, the form of these limiting distributions depends on whether the predictor is near-integrated or weakly dependent, even under homoskedasticity. Section 5 discusses fixed regressor wild bootstrap implementations of our proposed tests and demonstrates the first-order asymptotic validity of these. Section 6 presents the results from a Monte Carlo analysis into the finite sample behaviour of the tests under both the null hypothesis of no predictability and alternatives where local predictability occurs within the data. An application to monthly U.S. stock returns data is presented in Section 7. Section 8 concludes. Accompanying on-line supplementary material provides detailed proofs of the technical results given in the paper along with extensions to allow for multiple predictors and a general deterministic component, additional material relating to our empirical application and additional simulation results.

The notation $\mathcal{D}^k$ will be used to denote the space of càdlàg real functions on $[0,1]^k$ equipped with the Skorokhod topology, and we abbreviate $\mathcal{D}^1$ to $\mathcal{D}$. The weak convergence of probability measures on both function spaces (in particular, on $\mathcal{D}^k$) and on $\mathbb{R}^k$ is denoted by $\Rightarrow$. We reserve the notation P, E etc. for probability, expectation etc. with respect to the distribution of the original data and use P^*, E^* etc. for probability, expectation etc. induced by the data and the wild bootstrap multipliers (denoted $\{R_t\}$) conditionally on the data. The notation $\overset{w}{\Rightarrow}_p$ stands for weak convergence in probability; specifically, $\zeta_T^* \overset{w}{\Rightarrow}_p \zeta$ holds for random elements ζ_T^* and ζ, not necessarily defined on the same probability space, if $\mathrm{E}^* f(\zeta_T^*) \to \mathrm{E} f(\zeta)$ in P-probability for all bounded continuous real functions f with matching domain. In the

special case $\zeta = 0 \in \mathbb{R}$, we recall that $\zeta_T \overset{w}{\Rightarrow}_p 0$ (equivalently, $\zeta_T = o_{p^*}(1)$ in P-probability) means that $\mathrm{P}^*(|\zeta_T| > \varepsilon) \to 0$ in P-probability for every $\varepsilon > 0$. Finally, $\zeta_T = O_{p^*}(1)$ in P-probability signifies that for every $\varepsilon > 0$ there exists a $K > 0$ such that $\mathrm{P}\left\{\mathrm{P}^*(|\zeta_T| > K) < \varepsilon\right\} > 1 - \varepsilon$ for all T. The o_p and O_p symbols retain their usual meaning.

2. The episodic predictive regression model

The basic predictive regression model for stock returns, y_t, allowing for time-variation in the slope coefficient on the predictor variable, is taken to be of the form

$$y_t = \beta_0 + \beta_{1,t} x_{t-1} + u_t, \qquad t = 1, \ldots, T \tag{2.1}$$

where x_t, $t = 0, \ldots, T$, is observed and satisfies the data generating process [DGP]

$$x_t = \mu_x + \xi_t, \qquad t = 0, \ldots, T \tag{2.2a}$$

$$\xi_t = \rho\, \xi_{t-1} + v_t, \qquad t = 1, \ldots, T \tag{2.2b}$$

with ξ_0 a mean zero $O_p(1)$ variate. The innovations u_t form a martingale difference [MD] sequence, while v_t is allowed to exhibit weak serial dependence. For expositional simplicity we have only allowed for a single predictive regressor, x_{t-1}, and an intercept in (2.1). Generalisations to the case where the predictive regression contains multiple predictors and/or a general deterministic component of the form considered in section 3.2 of Breitung and Demetrescu (2015) are detailed in section S.2 of the supplementary material.

The DGP in (2.1) generalises the constant parameter predictive regression model by allowing the slope coefficient on x_{t-1} to vary over time, thereby allowing for changes over time in the predictive content of the regressor x_{t-1}. The constant parameter predictive regression model obtains by setting a constant slope parameter such that $\beta_{1,t} = \beta_1$, for all $t = 1, \ldots, T$. Our interest will focus in this paper on testing the usual null hypothesis that $(y_t - \beta_0)$ is a MD sequence and, hence, that y_t is not predictable by x_{t-1}, which entails that $\beta_{1,t} = \beta_1 = 0$, for all $t = 1, \ldots, T$, in (2.1). In contrast to the extant literature which tests this null hypothesis against the alternative that y_t is predictable by x_{t-1} with a constant slope parameter holding across the whole sample, that is $\beta_1 \neq 0$, under the maintained hypothesis that $\beta_{1,t} = \beta_1$, for all $t = 1, \ldots, T$, we will test against alternatives such that $\beta_{1,t} \neq 0$ for some t but without imposing constancy on $\beta_{1,t}$. Some structure obviously needs to be placed on the class of alternative hypotheses we may consider and this will be formalised below.

As discussed in the Introduction it is important to allow for the possibility of high persistence in the predictor variable x_t and to allow the shocks driving the predictor, v_t in (2.2), to be correlated with the unpredictable component of stock returns, u_t in (2.1). As regards the latter, we will allow u_t and v_t to be contemporaneously correlated and heteroskedastic; exact conditions will be detailed in Assumption 3. For the former, we allow ρ in (2.2) to satisfy the following assumption.

Assumption 1. Exactly one of the two following conditions holds true:

1. **Weakly persistent predictors:** The autoregressive parameter ρ in (2.2) is fixed and bounded away from unity, $|\rho| < 1$.
2. **Strongly persistent predictors:** The autoregressive parameter ρ in (2.2) is local-to-unity with $\rho := 1 - \frac{c}{T}$ where c is a fixed non-negative constant.

Remark 1. Many predictors are strongly persistent, exhibiting sums of sample autoregressive coefficients which are close to unity. Near-integrated asymptotics has been found to provide better approximations for the behaviour of test statistics in such circumstances; see, *inter alia*, Elliott and Stock (1994). However, a large part of the literature works with models which take x_t to be generated from a stable autoregressive process; see, for example, Amihud and Hurvich (2004). Assumption 1 allows for either of these possibilities to hold on x_t. ◇

We will develop tests for the null hypothesis that y_t is not predictable by x_{t-1} in any subsample, which do not require the practitioner to know which of Assumption 1.1 or Assumption 1.2 holds in (2.2), nor indeed what the precise value of ρ is in either case. Moreover, we aim to develop tests which possess non-trivial asymptotic *local* power against DGPs where predictability is present. Predictive regressions for stock returns typically exhibit small R^2 and low signal-to-noise ratios (see, *inter alia*, Campbell, 2008, and Phillips, 2015) so departures from the null, should predictability be present, are small. We will therefore conduct our theoretical analysis of the large sample properties of the tests we discuss under local alternatives such that the slope parameter $\beta_{1,t}$ is local-to-zero for an asymptotically non-vanishing set of the sample observations. The localisation rate (or Pitman drift) will need to be such that $\beta_{1,t}$ is specified to lie in a neighbourhood of zero which shrinks with the sample size, T. The appropriate Pitman drift is dictated by which of Assumption 1.1 and Assumption 1.2 holds in (2.2). Where x_t is near-integrated the appropriate rate is T^{-1}, while for weakly dependent x_{t-1}, the rate is $T^{-1/2}$. The different localisation rates reflect the fact that near-integration implies a much stronger signal from the predictor x_{t-1}. Moreover, tests based on the maxima from sequences of subsample predictability test statistics can only deliver non-trivial asymptotic local power in cases where an asymptotically non-vanishing fraction of the data is such that $\beta_{1,t} \neq 0$ holds on the DGP. For example, if $\beta_{1,t} \neq 0$ at one time point only, then although this would formally violate

the null that y_t is not predictable by x_{t-1}, this data point would, as $T \to \infty$, however be dominated by the remaining $T-1$ data points where the null hypothesis holds. Formally, in our framework we specify $\beta_{1,t}$ to satisfy the following assumption.

Assumption 2. In the context of (2.1) and (2.2), let $\beta_{1,t} := n_T^{-1} b(t/T)$, where $b(\cdot)$ is a piecewise Lipschitz-continuous real function on $[0, 1]$, with $n_T = \sqrt{T}$ under Assumption 1.1, and $n_T = T$ under Assumption 1.2.

Using the framework of Assumption 2 we can then equivalently write our null hypothesis that $\beta_{1,t} = 0$, for all $t = 1, \ldots, T$, as

$$H_0: \text{ The function } b(\tau) \text{ is identically zero for all } \tau \in [0, 1]. \tag{2.3}$$

We can now also formally specify the alternative hypothesis as,

$$H_{1,b(\cdot)}: \text{ The function } b(\cdot) \text{ is non-zero over at least one non-empty open interval contained in } [0, 1]. \tag{2.4}$$

Remark 2. The alternative hypothesis specified by $H_{1,b(\cdot)}$ is very general but entails that at least one subset of the sample observations (this need not be a strict subset, so it could contain all of the sample observations) comprising contiguous observations exists for which $\beta_{1,t} \neq 0$, and where the size of this subset is proportional to the sample size T. Notice that under $H_{1,b(\cdot)}$ the integral of $|b(\cdot)|$ on $[0, 1]$ is non-zero and it is this property which qualifies $H_{1,b(\cdot)}$ as a genuine (local) alternative. Moreover, as we will establish later, the form that $b(\cdot)$ takes under H_1 determines the local power offsets obtained in the limiting distributions of the statistics we propose. Notice also that, under $H_{1,b(\cdot)}$, $b(\cdot)$ may be zero in certain parts of its domain and it may also change magnitude and/or sign over its domain; the former corresponds to data points where $\beta_{1,t} = 0$, while the latter reflects observations for which $\beta_{1,t}$ does not have a fixed magnitude and/or sign across the full sample. ◇

We conclude this section by detailing in Assumption 3 the conditions that we will place on the disturbances u_t and v_t in (2.1) and (2.2).

Assumption 3. Let

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & B(L) \end{pmatrix} \mathbf{H}(t/T) \begin{pmatrix} a_t \\ e_t \end{pmatrix}, \quad \text{with} \quad \begin{pmatrix} a_t \\ e_t \end{pmatrix} \sim \text{White Noise}\,(0, \mathbf{I}_2), \tag{2.5}$$

where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix and:

1. $\boldsymbol{\zeta}_t := (a_t, e_t)'$ is a uniformly L_4-bounded martingale difference sequence which is such that $\sup_t \mathrm{E}\left|\mathrm{E}\left(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t' - \mathbf{I}_2 | \boldsymbol{\zeta}_{t-m}, \boldsymbol{\zeta}_{t-m-1}, \ldots\right)\right| \to 0$ as $m \to \infty$;
2. $\mathbf{H}(\cdot) := \begin{pmatrix} h_{11}(\cdot) & h_{12}(\cdot) \\ h_{21}(\cdot) & h_{22}(\cdot) \end{pmatrix}$ is a matrix of piecewise Lipschitz-continuous bounded functions on $(-\infty, 1]$, which is of full rank at all but a finite number of points;
3. $B(L)$, where L denotes the usual lag operator, is an invertible lag polynomial with $b_0 = 1$ and 1-summable coefficients, $\sum_{j\geq 0} j\left|b_j\right| < \infty$, for which $\omega := \sum_{j\geq 0} b_j > 0$.

Remark 3. The structure in (2.5) imposes that the disturbances u_t are uncorrelated with the increments of x_t at all (positive) lags. Where $\boldsymbol{\zeta}_t$ is independent and identically distributed [IID], this structure would entail that x_{t-1} is weakly exogenous with respect to u_t, and we will continue (with an abuse of language) to use the same term as a shorthand to describe this structure irrespective of whether $\boldsymbol{\zeta}_t$ is IID or not. Assumption 3.3 allows the increments to the predictor x_{t-1} to be serially correlated. These dynamics are not restricted beyond a 1-summability regularity condition on the moving average representation, as is typical in this literature; see, for example, Breitung and Demetrescu (2015) and Kostakis et al. (2015). ◇

Remark 4. Assumption 3 allows for quite general forms of heteroskedasticity in $\left(\tilde{u}_t, \tilde{v}_t\right)' := \mathbf{H}(t/T)(a_t, e_t)'$ and hence in u_t and v_t. In particular, Assumption 3.1 imposes a MD structure on $\boldsymbol{\zeta}_t$ allowing for conditional heteroskedasticity which is natural for the empirical applications to financial data we have in mind. Assumption 3.1 also imposes finite fourth moments; while daily returns often display very fat tails (see, for example, Nicolau and Rodrigues, 2019) such that the assumption of finite fourth order moments might not be a suitable assumption for daily data, standard predictive regression models have tended to be run on lower frequency data (monthly, quarterly or even annual data) where infinite kurtosis does not appear to be a concern. Assumption 3.1 places uniformity conditions on the cross-product moments of the innovations which limits the degree of serial dependence allowed in the conditional variances; these conditions are satisfied, for example, by strictly stationary and ergodic MD sequences with finite variance. Assumption 3.2 allows for unconditional time heteroskedasticity in the innovations through the matrix $\mathbf{H}(\tau)$. Where $\mathbf{H}(\tau)$ is diagonal for all $\tau \in [0, 1]$ the innovations $\left(\tilde{u}_t, \tilde{v}_t\right)'$ can display time-varying variances but are contemporaneously uncorrelated, so in the case where $\boldsymbol{\zeta}_t$ is IID with independent components, this would entail that x_t is strictly exogenous with respect to u_t (again we will use

this terminology, with an abuse of language, whether $\boldsymbol{\zeta}_t$ is IID or not). Importantly, the off-diagonal elements of $\mathbf{H}(\tau)\mathbf{H}(\tau)'$ (i.e., the covariance matrix of $\left(\tilde{u}_t, \tilde{v}_t\right)'$) are not imposed to be zero, thus allowing for contemporaneous and time-varying correlation among the innovations. The structure placed on $\mathbf{H}(\tau)$ by Assumption 3.2 allows for a wide class of models for the behaviour of the variance matrix of the innovations including single or multiple (co-) variance shifts, variances which follow a broken trend, and smooth transition variance shifts. As discussed in Breitung and Demetrescu (2015, p. 360), such patterns are plausible with macro and financial data and it is therefore important to use tests which are robust to such behaviour to avoid the possibility of spurious rejection of the null because of non-constancy in the variance matrix rather than genuine predictability from x_{t-1}. ◇

Under Assumption 1.1, x_t is a particular case of a *locally* stationary process which admits a time-varying variance when the series v_t displays time-varying volatility. In fact, the variance of the putative predictor is given as $\mathrm{Var}\left(x_{\lfloor \tau T \rfloor}\right) \approx \bar{\sigma}^2_{\xi}(\rho)\left(h^2_{21}(\tau) + h^2_{22}(\tau)\right)$, where $\bar{\sigma}^2_{\xi}(\rho)$ denotes the sum of the squared coefficients of the lag polynomial $(1-\rho L)^{-1}B(L)$ (which is finite in the stable autoregression case). This form of heteroskedasticity impacts on the inferential procedures based on subsample sequences of statistics discussed in this paper. Furthermore, time-varying volatility where present in the regression errors, u_t, and in the instrumental variables used in constructing the statistics can also affect the behaviour the sequences of statistics.

Heteroskedasticity has analogous effects under Assumption 1.2 (near-integration), though the transmission mechanism is somewhat different. In particular, under Assumption 3 we have that $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor} \mathbf{H}(t/T)\boldsymbol{\zeta}_t \Rightarrow \int_0^{\tau} \mathbf{H}(s)\mathrm{d}\boldsymbol{W}(s) =: (U(\tau), V(\tau))'$ on $\mathcal{D}^2$, where $\boldsymbol{W}$ is a two-dimensional standard Wiener process (see e.g. the invariance principle given in Lemma 1 of Boswijk et al., 2016), such that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor}\begin{pmatrix} u_t \\ v_t \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 \\ 0 & \omega \end{pmatrix}\int_0^{\tau} \mathbf{H}(s)\mathrm{d}\boldsymbol{W}(s) =: \begin{pmatrix} U(\tau) \\ \omega V(\tau) \end{pmatrix} \tag{2.6}$$

on $\mathcal{D}^2$. The processes $U(\tau)$ and $\omega V(\tau)$ are individually time-transformed Brownian motions whose correlation may also vary over time; their covariance at time τ is given by $\omega\int_0^{\tau}\mathbf{H}(s)\mathbf{H}(s)'\mathrm{d}s$. Under Assumption 1.2 x_t satisfies the invariance principle, $\frac{1}{\sqrt{T}}x_{\lfloor \tau T \rfloor} \Rightarrow \omega J_{c,\mathrm{H}}(\tau)$, where $J_{c,\mathrm{H}}(\tau)$ is an Ornstein–Uhlenbeck-type process driven by $V(\tau)$, i.e., $J_{c,\mathrm{H}}(\tau) := \int_0^{\tau} e^{-c(\tau-s)}\mathrm{d}V(s)$. Notice that $J_{c,\mathrm{H}}(\tau)$ is a heteroskedastic process when $\mathbf{H}(\cdot)$ is not constant: the quadratic variation processes of $U(\tau)$ and $V(\tau)$, given by $[U](\tau) := \int_0^{\tau}\left(h^2_{11}(s)+h^2_{12}(s)\right)\mathrm{d}s$ and $[V](\tau) := \int_0^{\tau}\left(h^2_{21}(s)+h^2_{22}(s)\right)\mathrm{d}s$, respectively, are nonlinear in general, and their quadratic covariation process is given by $[UV](\tau) := \int_0^{\tau}\left(h_{11}(s)h_{21}(s)+h_{12}(s)h_{22}(s)\right)\mathrm{d}s$.

3. Full sample predictability tests

Consider the maintained hypothesis that the slope parameter $\beta_{1,t}$ in (2.1) is constant, such that $\beta_{1,t} = \beta_1$, for all $t = 1, \ldots, T$. This yields the standard constant parameter predictive regression

$$y_t = \beta_0 + \beta_1 x_{t-1} + u_t, \qquad t = 1, \ldots, T. \tag{3.1}$$

A number of procedures have been developed for testing $H_0 : \beta_1 = 0$ in (3.1) against the local alternative $H_c : \beta_1 = n_T^{-1}b_1$, with b_1 a non-zero constant. Of these the simplest is the standard (full sample) ordinary least squares [OLS] t-test for the significance of x_{t-1} in (3.1). While standard normal asymptotic theory applies to the t-statistic under Assumption 1.1 provided the errors are homoskedastic (although this can be weakened by using heteroskedasticity-robust standard errors), it does not under Assumption 1.2 where the limiting null distribution of the t-statistic is nonstandard and depends on the local-to-unity parameter c unless x_t is strictly exogenous with respect to u_t.

Tests robust to c have been developed in Elliott and Stock (1994), who propose a Bayesian mixture procedure, and Cavanagh et al. (1995) and Campbell and Yogo (2006) who develop tests based on conservative bounds, and Jansson and Moreira (2006), who conduct inference on the basis of conditionally sufficient statistics. However, these procedures are all developed for the case where x_t is near-integrated, i.e. such that Assumption 1.2 holds, and for the case of homoskedastic disturbances. Variable addition [VA] techniques (see Breitung and Demetrescu (2015, p. 359) for a literature review) can be used to develop predictability tests which can be validly used regardless of whether x_t is local-to-unity or stationary. However, these VA-based tests have only trivial asymptotic local power against the Pitman rate, T^{-1}, where x_t is near-integrated. Breitung and Demetrescu (2015) show that the finite sample power of the VA-based tests is indeed very low relative to the tests designed for the use with near-integrated x_t when the AR parameter ρ in (2.2) is close to unity. They also develop modifications of the VA approach but some loss of power still remains. Gorodnichenko et al. (2012) proposed tests based on quasi-differencing but like the original VA-based tests these only have power in $T^{-1/2}$ neighbourhoods of the null.

Breitung and Demetrescu (2015) also examine tests based on the instrumental variables [IV] approach. They show that these can be validly implemented in the presence of endogeneity and uncertain regressor persistence and heteroskedasticity of the form specified in Section 2. The basic idea underlying IV estimation of the predictive regression model is to use instruments such that the instrument has lower persistence than the regressor x_{t-1} (so-called type-I instruments), or is such that the instrument is strictly exogenous with respect to u_t (so-called type-II instruments). Formal conditions which must hold on these instruments are given in Breitung and Demetrescu (2015).

A range of possible type-I instruments is given in Breitung and Demetrescu (2015, p. 361). These comprise: (i) a short memory instrument whereby we generate $z_{t-1} = (1-\bar{\alpha}L)_+^{-1}\Delta x_{t-1} := \Delta x_{t-1} + \bar{\alpha}\Delta x_{t-2} + \cdots + \bar{\alpha}^{t-2}\Delta x_1$ with $|\bar{\alpha}| < 1$; (ii) a mildly integrated instrument, generated as $z_{t-1} = (1-\alpha_T L)_+^{-1}\Delta x_{t-1}$, for $\alpha_T := 1 - aT^{-\gamma}$ with $a > 0$, $0 < \gamma < 1$; (iii) a fractionally integrated instrument, generated as $z_{t-1} = (1-L)^{1-d^*}x_{t-1}\mathbb{I}(t > 0) := \Delta_+^{1-d^*}x_{t-1}$ for some $d^* \in (0, 1/2)$; (iv) a long differences instrument, generated as $z_{t-1} = x_{t-1} - x_{t-k_T}$ for $K_T := \min\{\lfloor KT^{\upsilon}\rfloor, t-1\}$ for some $0 < \upsilon < 1$ and positive constant K. The use of the mildly integrated instrument in (ii) is an example of the so-called IVX approach of Phillips and Magdalinos (2009). In each case the generated instrument is, by design, free of a stochastic trend and hence less persistent than a near-integrated process, regardless of whether x_{t-1} is near-integrated or stationary. Being filtered versions of x_{t-1}, these instruments are driven by the same innovations and it is therefore expected that they provide valid instruments for x_{t-1}; at the same time, the reduced persistence leads to standard inference. Breitung and Demetrescu (2015, p. 362) also discuss the following type-II instruments: (i) a generated random walk, $z_{t-1} = (1-L)_+^{-1}w_{t-1}$ where $w_t \sim IID(0, \sigma_w^2)$ with w_t independent of u_t and v_t; (ii) deterministic functions of time, such as $z_{t-1} = (t-1)$ or $z_{t-1} = \sin(\pi(t-1)/2T)$, and (iii) Cauchy instruments, $z_{t-1} = \text{sign}(x_{t-1})$. Each of these is exogenous with respect to u_t by construction. However, they do not exploit any specific information about x_t, other than where x_t is near-integrated in which case they will be correlated with x_t; see Phillips (1998).

Simulation evidence in Breitung and Demetrescu (2015) shows that tests based on type-II instruments are significantly more powerful than those based on type-I instruments when x_t is strongly persistent. However, these instruments will be weak, in the sense that they will be almost uncorrelated with the regressor, where x_t is stationary. In such cases, Breitung and Demetrescu (2015) show that the resulting IV test for $\beta_1 = 0$ in (3.1) will have only trivial power. In order to simultaneously exploit the settings which result in superior power properties for the IV approach based on type-I and type-II instruments, Breitung and Demetrescu (2015) recommend the use of a test which combines two instruments for x_{t-1}, one of each type, which we denote by $z_{I,t-1}$ and $z_{II,t-1}$, collected into the vector $\boldsymbol{z}_{t-1} := \left(z_{I,t-1}, z_{II,t-1}\right)'$ for $t = 1, \ldots, T$. The general form of the resulting full sample IV-combination test statistic of Breitung and Demetrescu (2015), implemented with Eicker–White standard errors to account for heteroskedasticity satisfying Assumption 3, is given by

$$t_{\beta_1} := \frac{\boldsymbol{A}_T'\mathbf{B}_T^{-1}\boldsymbol{C}_T}{\sqrt{\boldsymbol{A}_T'\mathbf{B}_T^{-1}\mathbf{D}_T\mathbf{B}_T^{-1}\boldsymbol{A}_T}} \tag{3.2}$$

where $\boldsymbol{A}_T := \sum_{t=1}^T \hat{x}_{t-1}\hat{\boldsymbol{z}}_{t-1}$, $\mathbf{B}_T := \sum_{t=1}^T \hat{\boldsymbol{z}}_{t-1}\hat{\boldsymbol{z}}_{t-1}'$, $\boldsymbol{C}_T := \sum_{t=1}^T \hat{\boldsymbol{z}}_{t-1}\hat{y}_t$ and $\mathbf{D}_T := \sum_{t=1}^T \hat{\boldsymbol{z}}_{t-1}\hat{\boldsymbol{z}}_{t-1}'\hat{u}_t^2$, with $\hat{y}_t$, $\hat{x}_{t-1}$ and $\hat{\boldsymbol{z}}_{t-1}$ denoting demeaned versions of y_t, x_{t-1} and $\boldsymbol{z}_{t-1}$, respectively, so that, for $\boldsymbol{w}_t$ generically denoting either y_t, x_{t-1} or $\boldsymbol{z}_{t-1}$, $\hat{\boldsymbol{w}}_t := \boldsymbol{w}_t - \frac{1}{T}\sum_{s=1}^T \boldsymbol{w}_s$, and where $\hat{u}_t$ denotes the regression residuals from estimating (3.1). For the reasons outlined in Remark 4 of Breitung and Demetrescu (2015), the IV-combination test must be run as two-sided and so we accordingly consider tests based on the square of t_{β_1}; that is $t_{\beta_1}^2$. The limiting null distribution of $t_{\beta_1}^2$ is χ_1^2 under either Assumption 1.1 or 1.2; see Breitung and Demetrescu (2015) for details.

A variety of choices for the residuals $\hat{u}_t$ used in constructing $\mathbf{D}_T$ is possible. A natural choice is the IV regression residuals so that $\hat{u}_t := y_t - \hat{\beta}_0^{iv} - \hat{\beta}_1^{iv}x_{t-1}$, where $\hat{\beta}_j^{iv}$ denotes the two-stage least squares [2SLS] estimator of β_j, $j = 0, 1$. However, both Breitung and Demetrescu (2015) and Kostakis et al. (2015) recommend the use of OLS residuals on the grounds that they represent the best linear projection of y_t on x_{t-1} regardless of the persistence of the putative predictor, and that their finite-sample behaviour appears to be more stable than that of IV residuals. Finally, one could also use residuals computed under the null; i.e., $\hat{u}_t := y_t - \frac{1}{T}\sum_{s=1}^T y_s$. Under the local alternatives considered in Assumption 2, these three possible choices can be shown to be asymptotically equivalent to one another in so far as the behaviour of (the suitably normalised) $\mathbf{D}_T$ is concerned.

As we will subsequently see, a special case of the large sample results which will be presented in Section 4 is that the full-sample test based on $t_{\beta_1}^2$ has non-trivial asymptotic local power against $H_{1,b(\cdot)}$ for both weakly and strongly persistent regressors. This property of the full sample IV-based test statistic obtains through the limiting behaviour of the sample cross-product moment $\boldsymbol{A}_T$. In particular, its two components are not of the same order of magnitude; therefore, upon normalisation, one of these terms will converge to zero and so all weight is placed on the other instrument. Which instrument gets full weight depends on the persistence of x_{t-1}. The type-II instrument is selected for strongly persistent predictors (i.e., those satisfying Assumption 1.2), while the type-I instrument is selected for weakly persistent predictors (i.e., those satisfying Assumption 1.1); see the proof of Lemma S.6 in the supplementary material for details. As a result, regardless of the degree of persistence of the regressor, the appropriate instrument is chosen in the limit.

However, as the simulation results in Section 6 demonstrate, the finite sample power of the full sample test can be quite low against such "pocket" alternatives. In the next section we therefore propose tests based on sequences of subsample implementations of the IV-combination test statistic. IV-based techniques are particularly useful to consider because the corresponding subsample-specific statistics may be expressed in terms of partial sums, whose behaviour may in turn be characterised in a tractable manner. This is not the case, for instance, with the test of Campbell and Yogo (2006) or those of Elliott and Müller (2006) and Elliott et al. (2015), where the analysis of the joint behaviour of subsample-specific statistics is considerably more involved.

4. Subsample IV-combination tests for predictability

Our aim is to develop predictability tests with good power to detect temporary periods of predictability irrespective of whether the putative predictor, x_{t-1} is stable or near-integrated, and which are robust to the presence of heteroskedasticity in the data. To that end, we will base our testing approach on the computation of the IV-combination predictability statistics outlined in the previous section in the context of (3.2) computed not over the full available sample but over various sequences of subsamples of the data. For each such sequence we consider, our proposed test will be based on the maximum (in absolute value) statistic within that sequence. By taking the maximum over these sequences, we therefore base our test on the particular subsample within the given sequence of subsamples where the predictability statistic gives the strongest signal of predictability.

4.1. Choice of instruments

Before laying out our subsample IV-combination testing approach, we first need to state some regularity conditions which must hold on the type-I and type-II instruments such that we can validly use a testing strategy based on sequences of subsample IV-combination predictability statistics. We will then discuss the choice of instruments to use in practice which satisfy these conditions.

Assumption 4 details the conditions which need to hold on the type-I instrument used.

Assumption 4.

Let $z_{I,t}$ obey the following conditions:

1. Under either Assumption 1.1 or Assumption 1.2: $\mathrm{E}\left(\zeta_t|\zeta_{t-1},\zeta_{t-2},\ldots,z_{I,t-1},z_{I,t-2},\ldots\right)=0$, there exists $\delta_I \geq 0$ such that $T^{-\delta_I}z_{I,t}$ is uniformly L_4 -bounded, $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{1+\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}\right|\xrightarrow{p}0$, and $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{1+\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}u_t^2\right|=O_p(1)$.
2. Under Assumption 1.1, and jointly on $\mathcal{D}$
   (a) $\frac{1}{T^{1+\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}\xi_{t-1}\Rightarrow K_{z_Ix}(\tau)$, where $K_{z_Ix}(\tau)$ is a Hölder-continuous stochastic process of some order $\alpha>0$ and nonzero w.p.1;
   (b) $\frac{1}{T^{1+2\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}^2\xrightarrow{p}K_{z_I^2}(\tau)$, where $K_{z_I^2}(\tau)$ is a deterministic Hölder-continuous function of some order $\alpha>0$ and strictly increasing;
   (c) $\frac{1}{T^{1/2+\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}u_t\Rightarrow G_I(\tau)$, where $G_I(\tau)$ is a continuous process with independent increments (and therefore, Gaussian), with $G_I(0)=0$ a.s., zero mean function, strictly increasing variance function $[G_I](\tau)$ and variance profile defined as $\eta_I(\tau):=\frac{[G_I](\tau)}{[G_I](1)}$;
   (d) $\frac{1}{T^{1+2\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}^2u_t^2\xrightarrow{p}[G_I](\tau)$.
3. Under Assumption 1.2,
   (a) $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{3/2+\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}\xi_{t-1}\right|\xrightarrow{p}0$;
   (b) $\frac{1}{T^{1+2\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}^2\xrightarrow{p}K_{z_I^2}(\tau)$ on $\mathcal{D}$, where $K_{z_I^2}(\tau)$ is a deterministic Hölder-continuous function of some order $\alpha>0$ and strictly increasing;
   (c) $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{1/2+\delta_I}}\sum_{t=1}^{\lfloor\tau T\rfloor}z_{I,t-1}u_t\right|=O_p(1)$;
   (d) $\frac{1}{T^{1+2\delta_I}}\sum_{t=1}^{T}z_{I,t-1}^2u_t^2=O_p(1)$.

Remark 5. The conditions placed on $z_{I,t-1}$ by Assumption 4 can differ depending on whether Assumption 1.1 or Assumption 1.2 holds. This distinction is germane in cases where $z_{I,t-1}$ is constructed from x_{t-1}; see the examples listed in Section 3. In such cases δ_I may take different values for the same instrument, and, similarly, $K_{z_I^2}(\tau)$ may take different shapes under Assumption 1.1 and Assumption 1.2. We do not, however, make this explicit to ease notation. Assumption 4.1 complements the condition in Assumption 3.1 to ensure that the innovations u_t are uncorrelated with the instruments. Assumption 4.2 is new compared to Breitung and Demetrescu (2015) , and is required because we explicitly consider the behaviour of the IV-combination statistics under DGPs which can allow for either weak or strong persistence in the (putative) predictors; it requires the instruments to have stochastic properties similar to those of a stable autoregression driven by heteroskedastic innovations. Assumption 4.3 is the analogue of Assumption 3 of Breitung and Demetrescu (2015) but is considerably less restrictive: rather than the weak convergence of suitably normalised cross-product sample moments required there, we only require uniform boundedness in probability. Assumption 4 .3(b) regarding the partial sums of the squared instrument is new, but would appear fairly mild. Our conditions are weaker than those of Breitung and Demetrescu (2015) as we only consider the IV-combination statistic with two instruments, one of type-I and the other of type-II. ◇

Remark 6. Although the weak convergence in Assumption 4.2 is joint, we do not specify the dependence structure between the limiting processes because our asymptotic results will hold irrespective of this structure. We note, however, that the variance profile, $\eta_l(\cdot)$, which turns out to play an important role in our asymptotics under stability (Assumption 1.1) depends on *both* the choice of type-I instrument and the DGP (specifically, on $\mathbf{H}(\cdot)$ and the unconditional variance of u_t); see Lemma 1 for an example. Similarly, the limiting processes $K_{z_{lX}}(\cdot)$ and $K_{z_l^2}(\cdot)$ also depend on both the DGP and the choice of instrument; again, see Lemma 1 for an example. $\diamond$

Assumption 5 details the corresponding regularity conditions on the type-II instrument.

Assumption 5. The variable $z_{II,t}$ is deterministic and, for some function $Z(\tau)$, Hölder-continuous of order $\alpha > 1/2$, and some $\delta_{II} \geq 0$, satisfies $T^{-\delta_{II}} z_{II,\lfloor \tau T \rfloor} \to Z(\tau)$ in $\mathcal{D}$ where $Z(\cdot)$ is such that, for all $0 \leq \tau_1 < \tau_2 \leq 1$, $\int_{\tau_1}^{\tau_2} \tilde{Z}^2_{\tau_1,\tau_2}(s)\mathrm{d}s \neq 0$ with $\tilde{Z}_{\tau_1,\tau_2}(s) := Z(s) - \frac{1}{\tau_2-\tau_1}\int_{\tau_1}^{\tau_2} Z(s)\mathrm{d}s$.

Remark 7. Notice that the conditions stated in Assumption 5 do not involve the persistence of the regressor because the type-II instruments are exogenous. Assumption 5 essentially coincides with Assumption 4 of Breitung and Demetrescu (2015), up to minor differences. While Assumption 4 of Breitung and Demetrescu (2015) allows for stochastic $z_{II,t}$, it also requires the average cross-products of the instrument and the regression error to have a mixed Gaussian limiting distribution, such that it actually affords little additional flexibility in the choice of type-II instruments relative to Assumption 5. Indeed, under the above assumptions it holds, for example, that $\frac{1}{T^{1/2+\delta_{II}}}\sum_{t=1}^{\lfloor \tau T \rfloor} z_{II,t-1}u_t \Rightarrow \int_0^\tau Z(s)\mathrm{d}U(s)$ which is immediately seen to be a Gaussian process, given that Z is deterministic. However, the quadratic variation process of $\int_0^\tau Z(s)\mathrm{d}U(s)$, $\int_0^\tau Z^2(s)\left(h_{11}^2(s)+h_{12}^2(s)\right)\mathrm{d}s$, is in general nonlinear in τ and depends, analogously to the case of Assumption 1.1, on both the DGP and the choice of the instrument $z_{II,t-1}$. Finally, notice that $Z(\cdot)$ is not permitted to be constant for any of the subsamples over which the test statistics are computed, as this would entail perfect multicollinearity in those subsamples. $\diamond$

We also require further regularity conditions regarding the interaction of the type-I and type-II instruments used. These are now collected in Assumption 6.

Assumption 6. For instruments $z_{I,t}$ and $z_{II,t}$ satisfying the conditions of Assumptions 4 and 5, respectively, it is also required that: 1. $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{1+\delta_I+\delta_{II}}}\sum_{t=1}^{\lfloor \tau T \rfloor} z_{I,t-1}z_{II,t-1}\right| \xrightarrow{p} 0$; and 2. $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{1+\delta_I+\delta_{II}}}\sum_{t=1}^{\lfloor \tau T \rfloor} z_{I,t-1}z_{II,t-1}u_t^2\right| = O_p(1)$.

Remark 8. Breitung and Demetrescu (2015) do not impose such conditions explicitly as they are implied by the stricter set of assumptions under which they work. For instance, Assumption 6.1 would be implied by the weak convergence of the partial sums of $z_{I,t-1}$ in Assumption 3 of Breitung and Demetrescu, but we do not require such weak convergence here because Assumption 4.1 on the uniform boundedness of the partial sums of the type-I instrument, $\sup_{\tau\in[0,1]}\left|\frac{1}{T^{1+\delta_I}}\sum_{t=1}^{\lfloor \tau T \rfloor} z_{I,t-1}\right| \xrightarrow{p} 0$, suffices for our purposes (and is, for example, implied by Assumption 3 of Breitung and Demetrescu under near-integration). Indeed, Assumption 6.1 only differs through the weights $T^{-\delta_{II}}z_{II,t-1}$, which are deterministic; Assumption 6.2 can be seen as a randomly weighted version thereof, with weights $T^{-\delta_{II}}z_{II,t-1}u_t^2$. Notice that Assumption 6.1 entails that the (appropriately scaled) type-I and type-II instruments are mutually asymptotically orthogonal for all subsamples of the data, $t = \lfloor \tau_1 T \rfloor + 1, \ldots, \lfloor \tau_2 T \rfloor$, such that $0 \leq \tau_1 < \tau_2 \leq 1$. $\diamond$

In the context of their full-sample predictability tests, Breitung and Demetrescu (2015) consider the following choice for the type-II instrument, $z_{II,t}$,

$$z_{II,t-1} = \sin\left(\frac{k\pi(t-1)}{2T}\right) \tag{4.1}$$

where k is a positive integer chosen by the practitioner. Breitung and Demetrescu (2015) find that the best performing IV-combination test obtains for $k = 1$ in (4.1). For the type-I instrument we use the IVX approach which has become popular in predictive regressions; see, among others, Gonzalo and Pitarakis (2012), Phillips and Lee (2013) and Kostakis et al. (2015). This entails setting

$$z_{I,t-1} := \sum_{j=0}^{t-2} \varrho^j \Delta x_{t-1-j} \quad \text{with} \quad \varrho := 1 - \frac{a}{T^\gamma} \tag{4.2}$$

for some $a > 0$ and $\gamma \in (0, 1)$. In Lemma 1 we show that these two instruments satisfy the set of conditions required by Assumptions 4–6.[1]

[1] We will formally establish this result for only these two instruments which will subsequently be used in both our Monte Carlo study and empirical application. We conjecture, however, that the other examples of type-I and type-II instruments considered in Breitung and Demetrescu (2015, pp. 361–362) will also satisfy Assumption 4–6.

Lemma 1. *Let Assumptions 1 and 3 hold with $\boldsymbol{\zeta}_t$ strictly stationary and ergodic such that, for some $\vartheta > 0$, $\sup_{t\in\mathbb{Z}} \left|\mathrm{E}\left(\left(\tilde{v}_t^2 - \mathrm{E}\left(\tilde{v}_t^2\right)\right)\tilde{v}_{t-j}\tilde{v}_{t-k}\right)\right| \leq C\,(jk)^{-1/2-\vartheta/2}$. Then, Assumption 4–6 are satisfied by $\boldsymbol{z}_{t-1} := \left(z_{I,t-1}, z_{II,t-1}\right)'$ when $z_{II,t-1}$ and $z_{I,t-1}$ are as defined in (4.1), for any positive integer k, and (4.2), respectively. In particular, we have*

1. *Under Assumption 1.1, $\delta_I = 0$, $K_{z_I x}(\tau) = K_{z_I^2}(\tau) = \bar{\sigma}_\xi^2(\rho)[V](\tau)$ and $G_I(\tau)$ is a time-transformed Brownian motion given as $T^{-1/2}\sum_{t=1}^{\lfloor \tau T\rfloor}\xi_{t-1}u_t \Rightarrow G_I(\tau)$, where this weak convergence result holds jointly on $\mathcal{D}^3$ with the weak convergence given in (2.6), and*
2. *Under Assumption 1.2, $\delta_I = \gamma/2$ and $K_{z_I^2}(\tau) = \frac{\omega^2}{a}[V](\tau)$.*

Remark 9. The additional assumptions required to ensure the validity of the IVX instrument are relatively mild. Strict stationarity and ergodicity restrict the weak stationarity of $\boldsymbol{\zeta}_t$ required in Assumption 3 such that the asymptotic behaviour of sample averages can be accounted for, as required for example in Assumption 4.2. The additional condition on the rate of decay of $\mathrm{E}\left(\left(\tilde{v}_t^2 - \mathrm{E}\left(\tilde{v}_t^2\right)\right)\tilde{v}_{t-j}\tilde{v}_{t-k}\right)$ limits the amount of serial dependence allowed in the second order moments of the series. This rate is obviously satisfied when $\mathrm{E}\left(\left(\tilde{v}_t^2 - \mathrm{E}\left(\tilde{v}_t^2\right)\right)\tilde{v}_{t-j}\tilde{v}_{t-k}\right) = 0$, but is much weaker than that condition and, hence, still allows for asymmetric volatility clustering. ◇

Remark 10. Under Assumption 1.1, the processes $K_{z_I x}(\tau)$ and $K_{z_I^2}(\tau)$ are both proportional to the quadratic variation of $V(\tau)$, the limit process of the suitably normalised partial sums of ξ_t under stability. This demonstrates the usefulness of the IVX instrument in that, under stability, $z_{I,t-1}$ is approximately equal to the stochastic component ξ_{t-1} of the (putative) predictor, x_{t-1}, such that IVX effectively delivers the optimal instrument for x_{t-1} under Assumption 1.1. For a choice of type-I instrument other than IVX this is, in general, not true and one obtains different processes $K_{z_I x}(\tau)$ and $K_{z_I^2}(\tau)$ whose properties depend on the particular choice made; see also Corollary 2 of Breitung and Demetrescu (2015). Our large sample results will, however, be established under Assumptions 4 and 5 and, as such, will hold irrespective of the particular shape or properties of $K_{z_I x}(\tau)$ and $K_{z_I^2}(\tau)$. Furthermore, under Assumption 1.2 the IVX instrument will turn out to be dominated uniformly over all subsamples by the type-II instrument, such that the precise properties of $K_{z_I^2}$ will not be relevant under near-integration.[2] ◇

4.2. Subsample-based predictability tests

For type-I and type-II instruments satisfying Assumptions 4–6, we can proceed to develop subsample implementations of the IV-combination predictability test discussed in Section 3. To provide a unified notation for such subsample statistics it will prove useful to define the subsample-specific analogues $\boldsymbol{A}_T(\tau_1,\tau_2)$, $\mathbf{B}_T(\tau_1,\tau_2)$, $\boldsymbol{C}_T(\tau_1,\tau_2)$ and $\mathbf{D}_T(\tau_1,\tau_2)$ of the full-sample quantities $\boldsymbol{A}_T$, $\mathbf{B}_T$, $\boldsymbol{C}_T$ and $\mathbf{D}_T$, respectively, used to construct the standard IV-combination statistic, t_{β_1} of (3.2). These are defined analogously to their full-sample counterparts but for a sample consisting of observations $t = \lfloor \tau_1 T\rfloor + 1, \ldots, \lfloor \tau_2 T\rfloor$, so that, for example, $\boldsymbol{A}_T(\tau_1,\tau_2) := \sum_{t=\lfloor \tau_1 T\rfloor+1}^{\lfloor \tau_2 T\rfloor}\tilde{x}_{t-1}\tilde{\boldsymbol{z}}_{t-1}$ where $\tilde{y}_t$, $\tilde{x}_{t-1}$ and $\tilde{\boldsymbol{z}}_{t-1}$ are now subsample-specific demeaned versions of y_t, x_{t-1} and $\boldsymbol{z}_{t-1}$, respectively, so that, for $\boldsymbol{w}_t$ generically denoting either y_t, x_{t-1} or $\boldsymbol{z}_{t-1}$, $\tilde{\boldsymbol{w}}_t := \boldsymbol{w}_t - \frac{1}{\lfloor \tau_2 T\rfloor - \lfloor \tau_1 T\rfloor}\sum_{s=\lfloor \tau_1 T\rfloor+1}^{\lfloor \tau_2 T\rfloor}\boldsymbol{w}_s$. The full-sample quantity is recovered on setting $\tau_1 = 0$ and $\tau_2 = 1$. Precise definitions of these quantities are provided (in partial sum notation) in section S.3.1 of the supplementary material.

If it was known that a pocket of predictability might occur over the particular subsample $t = \lfloor \tau_1 T\rfloor + 1, \ldots, \lfloor \tau_2 T\rfloor$, then it would be logical to compute the subsample IV-combination statistic[3]

$$t_{\beta_1}(\tau_1,\tau_2) := \frac{\boldsymbol{A}_T'(\tau_1,\tau_2)\,\mathbf{B}_T^{-1}(\tau_1,\tau_2)\,\boldsymbol{C}_T(\tau_1,\tau_2)}{\sqrt{\boldsymbol{A}_T'(\tau_1,\tau_2)\,\mathbf{B}_T^{-1}(\tau_1,\tau_2)\,\mathbf{D}_T(\tau_1,\tau_2)\,\mathbf{B}_T^{-1}(\tau_1,\tau_2)\,\boldsymbol{A}_T(\tau_1,\tau_2)}} \tag{4.3}$$

and a test for predictability in this specific subsample could be obtained by comparing $(t_{\beta_1}(\tau_1,\tau_2))^2$ with the $\chi^2(1)$ distribution. Indeed, this would be nothing more than the approach of Breitung and Demetrescu (2015) applied to the particular subsample $t = \lfloor \tau_1 T\rfloor + 1, \ldots, \lfloor \tau_2 T\rfloor$. Such a test would be expected to have considerably more power to detect a regime of predictability over the subsample $t = \lfloor \tau_1 T\rfloor + 1, \ldots, \lfloor \tau_2 T\rfloor$ than would the full sample test based on t_{β_1} of (3.2) because the former would be calculated only for sample points where a predictive relationship holds.

In practice it is unlikely the practitioner will know which specific subsample(s) of the data might admit predictive regimes. While some previous applied studies in the literature have considered a variety of sample splits and also looked at the evolution of predictive regression statistics over a sequence of subsamples, these studies have tended to signal

[2] It should be noted, however, that $K_{z_I^2}(\tau)$ is also proportional to $[V](\tau)$ under near-integration, albeit with a different constant of proportionality; this is a consequence of the fact that $z_{I,t}$ is mildly integrated in this case.

[3] In the context of $\mathbf{D}_T(\tau_1,\tau_2) := \sum_{t=1}^{T}\tilde{\boldsymbol{z}}_{t-1}\tilde{\boldsymbol{z}}_{t-1}'\tilde{u}_t^2$, the residuals, $\tilde{u}_t^2$, are now the subsample analogues of the full sample residuals, $\hat{u}_t^2$, used in the construction of the full-sample statistic t_{β_1} in (3.2). The three possible choices discussed there can also be used here for the subsample $t = \lfloor \tau_1 T\rfloor + 1, \ldots, \lfloor \tau_2 T\rfloor$. As with the full sample statistic, these three are asymptotically equivalent in so far as the behaviour of (the suitably normalised) $\mathbf{D}_T(\tau_1,\tau_2)$ is concerned.

the presence of a predictive episode based on comparing each of these subsample statistics with the critical value that would apply when running a test for predictability on a single *known* subsample. As discussed in Section 1, this induces either multiple testing and/or endogenously determined breakdate problems and, hence, does not deliver size-controlled tests; see, *inter alia*, Inoue and Rossi (2005). In order to control for these issues, the critical value of the test needs to reflect the searching element involved. This can be done by basing one's test on certain functionals of the sequence of subsample predictability statistics considered. Given we are testing the null of no predictability against the alternative of predictability in at least one subsample of the data, an approach based on the maximum of the sequence of subsample predictability statistics considered would seem appropriate. The specific sequences of statistics that we take the maximum over must also be entirely agnostic of the data to avoid any endogenous selection bias; we could not, for example, validly choose to take the maximum statistic from the sequence of subsamples where previous studies had argued predictability holds.

There is an extensive literature on testing for fluctuations in the parameters of linear regression models; see, *inter alia*, Kuan and Hornik (1995). Common choices of agnostic sequences of statistics used include forward and reverse recursive sequences, rolling sequences, and double-recursive sequences. We will adopt these choices here and base our tests on the maximum statistic taken over each of these sequences of statistics. These can be formally defined as follows:

• The sequence of *forward recursive* statistics is given by $\{(t_{\beta_1}(0,\tau))^2\}_{\tau_L\le\tau\le 1}$, where the parameter $\tau_L \in (0,1)$ is chosen by the user. The forward recursive regression approach uses $\lfloor T\tau_L\rfloor$ start-up observations, where τ_L is the *warm-in* fraction, and then calculates the sequence of subsample predictive regression statistics $(t_{\beta_1}(0,\tau))^2$ for $t = 1,\dots,\lfloor\tau T\rfloor$, with τ travelling across the interval $[\tau_L, 1]$. The statistic formed as the maximum taken across this sequence is then,

$$\mathcal{T}^f := \max_{\tau_L\le\tau\le 1} (t_{\beta_1}(0,\tau))^2. \tag{4.4}$$

• The sequence of *backward recursive* statistics is given by $\{(t_{\beta_1}(\tau,1))^2\}_{0\le\tau\le\tau_U}$ with $\tau_U \in (0,1)$ again chosen by the user. In this case one calculates the sequence of subsample predictive regression statistics $(t_{\beta_1}(\tau,1))^2$ for $t = \lfloor\tau T\rfloor + 1,\dots,T$, with τ travelling across the interval $[0,\tau_U]$. The maximum statistic from this backward recursive sequence is then,

$$\mathcal{T}^b := \max_{0\le\tau\le\tau_U} (t_{\beta_1}(\tau,1))^2. \tag{4.5}$$

• The sequence of *rolling* statistics is given by $\{(t_{\beta_1}(\tau,\tau+\Delta\tau))^2\}_{0\le\tau\le 1-\Delta\tau}$ where the user-defined parameter $\Delta\tau \in (0,1)$. The rolling regression approach calculates the sequence of subsample statistics $(t_{\beta_1}(\tau,\tau+\Delta\tau))^2$ for $t = \lfloor\tau T\rfloor + 1,\dots,\lfloor\tau T\rfloor + \lfloor T\Delta\tau\rfloor$, where $\Delta\tau$ is the window fraction with $\lfloor T\Delta\tau\rfloor$ the window width, with τ travelling across the interval $[0, 1-\Delta\tau]$. The maximum statistic from this rolling sequence is then,

$$\mathcal{T}^r := \max_{0\le\tau\le 1-\Delta\tau} (t_{\beta_1}(\tau,\tau+\Delta\tau))^2. \tag{4.6}$$

• Finally, the *double-recursive* sequence of statistics is given by $\{(t_{\beta_1}(\tau_1,\tau_2))^2\}_{\substack{0\le\tau_1,\tau_2\le 1\\ \tau_2-\tau_1\ge\Delta\tau}}$, where $\Delta\tau \in (0,1)$ is again a user-defined parameter. The double-recursive approach calculates a double indexed sequence of subsample statistics $(t_{\beta_1}(\tau_1,\tau_2))^2$ for $t = \lfloor\tau_1 T\rfloor + 1,\dots,\lfloor\tau_2 T\rfloor$, for all subsamples such that $0 \le \tau_1 < \tau_2 \le 1$ and where $\tau_2 - \tau_1 \ge \Delta\tau$. Notice that this entails that the forward recursive sequence discussed above is calculated across all possible warm-in fractions such that $\tau_L \ge \Delta\tau$, which is why this sequence is referred to as double-recursive.[4] The maximum statistic from the double-recursive sequence is then,

$$\mathcal{T}^d := \max_{\substack{0\le\tau_1,\tau_2\le 1\\ \tau_2-\tau_1\ge\Delta\tau}} (t_{\beta_1}(\tau_1,\tau_2))^2. \tag{4.7}$$

Remark 11. The full sample IV-combination statistic $t^2_{\beta_1}$ of (3.2) is contained within the forward recursive sequence of statistics and obtains by setting $\tau = 1$, and similarly is contained within the backward recursive sequence for $\tau = 0$. It is also contained within the double-recursive sequence for $\tau_1 = 0$ and $\tau_2 = 1$. Notice also that if we set $\Delta\tau = 1$ in the context of the rolling sequence then this would collapse to the single full sample statistic, $t^2_{\beta_1}$. ◇

Tests based on the maximum from each of the foregoing sequences of subsample statistics have particular patterns of local predictability that they will be well designed to detect. Tests based on the forward recursive sequence of statistics are designed to detect pockets of predictability which start at or near the start of the full sample period available to the practitioner. The longer the duration of such an episode the more powerful these tests will be, other things being equal, because they are based on a sequence of increasing subsamples all starting from the first data point. By analogy, tests based on the reverse recursive sequence of subsample statistics are designed to detect end-of-sample pockets of predictability. As such, reverse recursive based tests could therefore usefully be employed in an on-going monitoring exercise for the emergence of predictive regimes. Because both the forward and reverse recursive sequences, and indeed the double-recursive sequence, contain the usual full sample predictability statistic, regardless of the choice

[4] Notice that this double sequence also obtains by calculating the rolling sequence discussed above for all possible rolling window widths between $\Delta\tau$ and 1 inclusive.

of the trimming parameters, they also deliver tests which have power to detect predictability which holds over the whole sample, although in this particular case they would not be expected to be as powerful as the standard full sample IV-combination test which is clearly designed for that specific alternative hypothesis.

For a given window width, tests based on a rolling sequence of statistics are designed to pick up a window of predictability, of (roughly) the same length, within the data. As discussed above, the double-recursive sequence amounts to considering all possible window width rolling sequences, subject to a minimum window width. These then are useful for picking up multiple predictive regimes, of potentially different lengths, within the data. However, because the double-recursive sequence considers such a large number of possible subsamples of the data a test based on the maximum from this sequence would necessarily be expected to be less powerful than the recursive or rolling-based tests in scenarios for which the latter are designed. This is because the more statistics one considers in a sequence over which the maximum is taken the stricter the critical value needs to be to maintain a correctly sized test. So, for example, in the case where a pocket of predictability existed in the middle of the sample data of length say m observations, a test based on the maximum from the rolling sequence using a window width of m observations would be expected to be more powerful than a test based on the maximum of the double-recursive sequence because the critical value for the latter would be considerably larger than the former. However, a power advantage over the double-recursive test would not necessarily be expected to hold for the corresponding rolling tests where the window width was either smaller than m or greater than m. In the former case this would be because the maximum subsample length available over which predictability held (m observations) could never be utilised because the window width is less than m, while in the latter case all subsamples in the sequence will contain a mix of data points where predictability holds and where it does not. It is of course very hard to analytically predict what the relative finite sample power properties of the recursive, rolling and double recursive based tests will be in cases like these and so we will investigate these further using Monte Carlo experimentation in Section 6.

Before establishing the asymptotic properties of the maximum subsample statistics, it is worth briefly commenting on estimation of the location of any predictive windows in cases where our proposed tests reject. Even for the simplest possible case where $H_{1,b(\cdot)}$ of (2.4) implies predictability over just a single subsample of the data, say $t = \lfloor \tau_1 T \rfloor + 1, \ldots, \lfloor \tau_2 T \rfloor$, with $\tau_1 < \tau_2$ and where either $\tau_1 > 0$ or $\tau_2 < 1$, consistent estimation of τ_1 and τ_2 is not possible because of the Pitman localisation to zero placed on $\beta_{1,t}$ in this interval by Assumption 2. In practice, however, if a given maximum statistic rejects then a sensible estimate of τ_1 and τ_2 would be given by the start and end points of the subsample corresponding to the maximum value from the sequence of statistics from which a rejection was obtained. If one was looking to date possibly multiple windows of predictability then one could reapply the procedures outlined above to the data set excluding those sample points for which a first stage rejection occurred, and do so repeatedly until no rejection was obtained.

4.3. Asymptotic distributions

In Proposition 1 we now provide representations for the asymptotic distributions of the maximum subsample statistics defined in Section 4.2 under the appropriate local alternative, $H_{1,b(\cdot)}$.

Proposition 1. *Consider the model in (2.1) and (2.2) and let Assumption 2–6 hold. Then under the local alternative $H_{1,b(\cdot)}$ of (2.4):*

(i) *Under Assumption 1.1, as $T \to \infty$, it holds that,*

$$\mathcal{T}^f \Rightarrow \sup_{\tau\in[\tau_L,1]} \frac{\left(G_l(\tau) + \int_0^\tau b(s)\mathrm{d}K_{z_l x}(s)\right)^2}{[G_l](1)\,\eta_l(\tau)}$$

$$\mathcal{T}^b \Rightarrow \sup_{\tau\in[0,\tau_U]} \frac{\left(G_l(1) - G_l(\tau) + \int_\tau^1 b(s)\mathrm{d}K_{z_l x}(s)\right)^2}{[G_l](1)\,(1-\eta_l(\tau))}$$

$$\mathcal{T}^d \Rightarrow \sup_{\substack{0\le\tau_1,\tau_2\le 1\\ \tau_2-\tau_1\ge\Delta\tau}} \frac{\left(G_l(\tau_2) - G_l(\tau_1) + \int_{\tau_1}^{\tau_2} b(s)\mathrm{d}K_{z_l x}(s)\right)^2}{[G_l](1)\,(\eta_l(\tau_2)-\eta_l(\tau_1))}$$

$$\mathcal{T}^r \Rightarrow \sup_{0\le\tau\le 1-\Delta\tau} \frac{\left(G_l(\tau+\Delta\tau) - G_l(\tau) + \int_\tau^{\tau+\Delta\tau} b(s)\mathrm{d}K_{z_l x}(s)\right)^2}{[G_l](1)\,(\eta_l(\tau+\Delta\tau)-\eta_l(\tau))}$$

where $G_l(\cdot)$, $[G_l](\cdot)$, $K_{z_l x}(\cdot)$ and $\eta_l(\cdot)$, are as defined in Assumption 4.2.

(ii) *Under Assumption 1.2, as $T \to \infty$, it holds that,*

$$\mathcal{T}^f \Rightarrow \sup_{\tau_L \le \tau \le 1} \frac{\left(\int_0^\tau \tilde{Z}_{0,\tau}(s)\mathrm{d}U(s) + \omega \int_0^\tau \tilde{Z}_{0,\tau}(s)b(s)J_{c,\mathrm{H}}(s)\mathrm{d}s\right)^2}{\int_0^\tau \tilde{Z}_{0,\tau}^2(s)\mathrm{d}[U](s)}$$

$$\mathcal{T}^b \Rightarrow \sup_{0 \le \tau \le \tau_U} \frac{\left(\int_\tau^1 \tilde{Z}_{\tau,1}(s)\mathrm{d}U(s) + \omega \int_\tau^1 \tilde{Z}_{\tau,1}(s)b(s)J_{c,\mathrm{H}}(s)\mathrm{d}s\right)^2}{\int_\tau^1 \tilde{Z}_{\tau,1}^2(s)\mathrm{d}[U](s)}$$

$$\mathcal{T}^d \Rightarrow \sup_{\substack{0 \le \tau_1, \tau_2 \le 1 \\ \tau_2 - \tau_1 \ge \Delta\tau}} \frac{\left(\int_{\tau_1}^{\tau_2} \tilde{Z}_{\tau_1,\tau_2}(s)\mathrm{d}U(s) + \omega \int_{\tau_1}^{\tau_2} \tilde{Z}_{\tau_1,\tau_2}(s)b(s)J_{c,\mathrm{H}}(s)\mathrm{d}s\right)^2}{\int_{\tau_1}^{\tau_2} \tilde{Z}_{\tau_1,\tau_2}^2(s)\mathrm{d}[U](s)}$$

$$\mathcal{T}^r \Rightarrow \sup_{0 \le \tau \le 1-\Delta\tau} \frac{\left(\int_\tau^{\tau+\Delta\tau} \tilde{Z}_{\tau,\tau+\Delta\tau}(s)\mathrm{d}U(s) + \omega \int_\tau^{\tau+\Delta\tau} \tilde{Z}_{\tau,\tau+\Delta\tau}(s)b(s)J_{c,\mathrm{H}}(s)\mathrm{d}s\right)^2}{\int_\tau^{\tau+\Delta\tau} \tilde{Z}_{\tau,\tau+\Delta\tau}^2(s)\mathrm{d}[U](s)}$$

where $J_{c,\mathrm{H}}(\cdot)$, $U(\cdot)$ and $[U](\cdot)$ are defined in Section 2, and $\tilde{Z}_{\cdot,\cdot}(\cdot)$ is defined in Assumption 5.

Remark 12. Expressions for the limiting null distributions of the statistics can be obtained by omitting those terms involving the function $b(\cdot)$ from the representations given in Proposition 1. In what follows we will denote the resulting limiting null distributions of the $\mathcal{T}^f$, $\mathcal{T}^b$, $\mathcal{T}^d$ and $\mathcal{T}^r$ statistics under Assumption 1.1 as $\mathcal{T}_\infty^{f,I}$, $\mathcal{T}_\infty^{b,I}$, $\mathcal{T}_\infty^{d,I}$ and $\mathcal{T}_\infty^{r,I}$, respectively, and under Assumption 1.2 as $\mathcal{T}_\infty^{f,II}$, $\mathcal{T}_\infty^{b,II}$, $\mathcal{T}_\infty^{d,II}$ and $\mathcal{T}_\infty^{r,II}$, respectively. For any given $s \in \{f, b, d, r\}$, the limiting distribution $\mathcal{T}_\infty^{s,I}$, appropriate for the case where the predictor is weakly persistent satisfying Assumption 1.1, and $\mathcal{T}_\infty^{s,II}$, the corresponding limiting null distribution where the predictor is strongly persistent satisfying Assumption 1.2, have different functional forms. For example, while $\mathcal{T}_\infty^{s,II}$, $s = f, b, d, r$, all depend on the choice of type-II instrument, $\mathcal{T}_\infty^{s,I}$, $s = f, b, d, r$, do not. The impact of non-constancy in $\mathbf{H}(\cdot)$ on these limiting null distributions also differs between the strongly and weakly persistent cases; see the discussion in Remarks 15 and 16. ◇

Remark 13. All of the statistics in the sequences are exact invariant to both μ_x and β_0 by virtue of being based on subsample demeaned variables. Moreover, the vector of instruments used is, by construction, invariant to μ_x, because $z_{I,t}$ is based on *differences* of x_t for the instruments mentioned in Section 3, and $z_{II,t}$ is a deterministic function of time chosen by the user *without* reference to μ_x. Consequently the limiting representations in Proposition 1 do not depend on either μ_x or β_0. ◇

Remark 14. Under Assumption 1.1, local power depends indirectly on the persistence of the putative predictor, as measured by ρ and $B(L)$ through $K_{z_I x}(\cdot)$; see Lemma 1 for the particular example of the IVX instrument. Under Assumption 1.2, while the mean-reversion parameter c does not affect the limiting null behaviour of the maximum statistics, the local power functions depend explicitly on c through $J_{c,\mathrm{H}}(\cdot)$. In each case, the rule-of-thumb that the stronger the mean reversion, the lower the local power, seems to hold; see the Monte Carlo results in Section 6. ◇

For weakly persistent regressors, a time transformation can shed further light on the influence of heteroskedasticity. Under Assumption 4.2(c), the process $W(\cdot) := G_I(\eta^{-1}(\cdot))/\sqrt{[G_I](1)}$ is continuous with stationary independent increments, $W(0) = 0$ a.s. and $\mathrm{Var}(W(\tau)) = \tau$, and therefore, $W(\cdot)$ is a standard Wiener process. It follows that $G_I(\cdot) = \sqrt{[G_I](1)}W(\eta_I(\cdot))$ is a time-transformed Wiener process. Consequently, taking the limiting functional associated with $\mathcal{T}^f$ as an example, we have that $\sup_{\tau\in[\tau_L,1]} \frac{\left(G_I(\tau)+\int_0^\tau b(s)\mathrm{d}K_{z_I x}(s)\right)^2}{[G_I](1)\,\eta_I(\tau)} \overset{d}{=} \sup_{\tau\in[\tau_L,1]} \frac{\left(W(\eta_I(\tau))+\int_0^\tau b(s)\mathrm{d}\frac{K_{z_I x}(s)}{\sqrt{[G_I](1)}}\right)^2}{\eta_I(\tau)}$ with similar distributional identities holding for the remaining statistics. As the maximum of a function is invariant to monotonic transformations of the argument, we may set $r = \eta_I(\tau)$ and therefore obtain the following alternative representations of the limiting results in part (i) of Proposition 1.

Corollary 1. *Let the conditions of Proposition 1 hold. Then under Assumption 1.1, as $T \to \infty$,*

$$\mathcal{T}^f \Rightarrow \sup_{r\in[\eta_I(\tau_L),1]} \frac{\left(W(r) + \int_0^{\eta_I^{-1}(r)} b(s)\mathrm{d}\bar{K}_{z_I x}(s)\right)^2}{r}$$

$$\mathcal{T}^b \Rightarrow \sup_{r\in[0,\eta_I(\tau_U)]} \frac{\left(W(1) - W(r) + \int_{\eta_I^{-1}(r)}^1 b(s)\mathrm{d}\bar{K}_{z_I x}(s)\right)^2}{1-r}$$

$$\mathcal{T}^d \Rightarrow \sup_{\substack{0\le r_1,r_2\le 1\\ \eta_I^{-1}(r_2)-\eta_I^{-1}(r_1)\ge\Delta\tau}} \frac{\left(W(r_2)-W(r_1)+\int_{\eta_I^{-1}(r_1)}^{\eta_I^{-1}(r_2)} b(s)\mathrm{d}\bar{K}_{z_Ix}(s)\right)^2}{r_2-r_1}$$

$$\mathcal{T}^r \Rightarrow \sup_{0\le r\le\eta_I(1-\Delta\tau)} \frac{\left(W\left(\eta_I\left(\eta_I^{-1}(r)+\Delta\tau\right)\right)-W(r)+\int_{\eta_I^{-1}(r)}^{\eta_I^{-1}(r)+\Delta\tau} b(s)\mathrm{d}\bar{K}_{z_Ix}(s)\right)^2}{\eta_I\left(\eta_I^{-1}(r)+\Delta\tau\right)-r}$$

where $W(\cdot) \coloneqq G_I(\eta^{-1}(\cdot))/\sqrt{[G_I](1)}$ *is a standard Wiener process and* $\bar{K}_{z_Ix}(\cdot) \coloneqq K_{z_Ix}(\cdot)/\sqrt{[G_I](1)}$. *Moreover, the limiting null distributions discussed in Remark 12 are such that,*

$$\mathcal{T}_\infty^{f,I} \overset{d}{=} \sup_{r\in[\eta_I(\tau_L),\ 1]} \frac{(W(r))^2}{r}, \qquad \mathcal{T}_\infty^{b,I} \overset{d}{=} \sup_{r\in[0,\ \eta_I(\tau_U)]} \frac{(W(1)-W(r))^2}{1-r}$$

$$\mathcal{T}_\infty^{r,I} \overset{d}{=} \sup_{0\le r\le\eta_I(1-\Delta_\tau)} \frac{\left(W\left(\eta_I\left(\eta_I^{-1}(r)+\Delta_\tau\right)\right)-W(r)\right)^2}{\eta_I\left(\eta_I^{-1}(r)+\Delta_\tau\right)-r}$$

$$\mathcal{T}_\infty^{d,I} \overset{d}{=} \sup_{\substack{0\le r_1,r_2\le 1\\ \eta_I^{-1}(r_2)-\eta_I^{-1}(r_1)\ge\Delta\tau}} \frac{(W(r_2)-W(r_1))^2}{r_2-r_1}.$$

Remark 15. The results in Proposition 1 and Corollary 1 highlight that both the limiting null distributions and the local power functions of all of the tests depend, in general, on any unconditional heteroskedasticity present through the resulting non-constancy of $\mathbf{H}(\cdot)$. This holds irrespective of the persistence of the regressor x_t; moreover, heteroskedasticity has differing effects on the limiting distributions depending on the degree of persistence of x_t. At least under the null this may seem surprising, as Eicker–White standard errors are designed to robustify any of the subsample statistics, $0 \le \tau_1 < \tau_2 \le 1$, to heteroskedasticity (conditional or unconditional). However, this asymptotic invariance only holds *marginally* for a given statistic in the sequence; indeed, it can be shown for each of the sequences of statistics, and regardless of which of Assumptions 1.1 and 1.2 holds, any given statistic in the sequence has a marginal χ_1^2 limiting null distribution. The representations in Corollary 1, for example, show that under Assumption 1.1 the suprema are taken over statistics computed for various intervals whose endpoints depend on the variance profile $\eta_I(\cdot)$ defined in Assumption 4 .2, which depends in turn on both the DGP and the choice of type-I instrument. Moreover, under Assumption 1.2, the same phenomenon explains part (ii) of Proposition 1, with the additional complication that one cannot represent the subsample statistics more tractably using a time transformation due to the presence of the *subsample-demeaned* process Z. Here, too, heteroskedasticity depends on the choice of instrument (now the type-II instrument) in addition to the DGP. Under local alternatives, heteroskedasticity additionally enters by means of $K_{z_Ix}(\cdot)$ and $J_{c,\mathrm{H}}(\cdot)$, under Assumptions 1.1 and 1.2, respectively. It is important to emphasise that the precise effect of non-constancy of $\mathbf{H}(\cdot)$ due to unconditional heteroskedasticity on the limiting distributions of our maximum statistics depends on which of Assumption 1.1 or 1.2 holds. ◇

Remark 16. More generally, the impact of the DGP on the large sample behaviour of the statistics depends on the choice of instrument and on the persistence of the (putative) predictor. Consider first the results under Assumption 1.1. Here the limiting null distributions, $\mathcal{T}_\infty^{s,I}$, $s = f, b, d, r$, all depend on $\eta_I(\cdot)$ which in turn depends on the unconditional variance of u_t. In the case where $\eta_I(s) = s$, these limiting null distributions simplify to the suprema of squared standardised Wiener processes taken over the range of the subsamples. However, constancy of $\mathbf{H}(\cdot)$ is not sufficient to ensure linearity of $\eta_I(\cdot)$, because heteroskedasticity can still enter via the instrument $z_{I,t-1}$. Under the local alternative, the key quantity controlling power is $K_{z_Ix}(\cdot)$ which can be deterministic under Assumption 1.1 (see, for example, Lemma 1 for the case of the IVX instrument), and (upon normalisation) characterises the strength of the instrument $z_{I,t-1}$. However, $K_{z_Ix}(\cdot)$ also characterises the signal; other things equal, if x_t has a large marginal variance relative to u_t, then local power will increase. In the case of Assumption 1.2, local power depends on the process $J_{c,\mathrm{H}}(\cdot)$ in a more intricate way, due to the fact that $J_{c,\mathrm{H}}$ and $\int \tilde{Z}\mathrm{d}U$ may be dependent. Clearly, local power is influenced by all three factors c, ω and $\mathbf{H}(\cdot)$. The effect of the elements of $\mathbf{H}(\cdot)$ is not easy to disentangle, as can be seen from the expressions given for the quadratic variation processes of $U(\cdot)$ and $V(\cdot)$ at the end of Section 2. ◇

In Corollary 2 we detail the limiting distributions of the full sample statistic $t_{\beta_1}^2$ of (3.2) under the local alternative, $H_{1,b(\cdot)}$ of (2.4).

Corollary 2. *Let the conditions of Proposition 1 hold. Then under* $H_{1,b(\cdot)}$, *as* $T \to \infty$ *: (i) Under Assumption 1. 1,* $t_{\beta_1}^2 \Rightarrow \left(W(1)+\int_0^1 b(s)\mathrm{d}\bar{K}_{z_Ix}(s)\right)^2$; *(ii) under Assumption 1.2,* $t_{\beta_1}^2 \Rightarrow \left(\int_0^1 \tilde{Z}^2(s)\mathrm{d}[U](s)\right)^{-1}\left(\int_0^1 \tilde{Z}(s)\mathrm{d}U(s)+\omega\int_0^1 \tilde{Z}(s)b(s)J_{c,\mathrm{H}}(s)\mathrm{d}s\right)^2$, *where* $\tilde{Z}(s) \coloneqq Z(s)-\int_0^1 Z(s)$.

Remark 17. From Corollary 2, under the null hypothesis H_0 of (2.3), $t^2_{\beta_1} \Rightarrow W(1)^2$ under Assumption 1.1, while $t^2_{\beta_1} \Rightarrow \left(\int_0^1 \tilde{Z}^2(s)\mathrm{d}[U](s)\right)^{-1}\left(\int_0^1 \tilde{Z}(s)\mathrm{d}U(s)\right)^2$ under Assumption 1.2 with $\int_0^1 \tilde{Z}(s)\mathrm{d}U(s) \sim N\left(0,\ \int_0^1 \tilde{Z}^2(s)\mathrm{d}[U](s)\right)$. Consequently, $t^2_{\beta_1}$ is seen to possess a standard χ^2_1 limiting null distribution regardless of whether x_t is stable or near-integrated. Moreover, the results in Corollary 2 show that the full sample IV-combination test exhibits non-trivial power against the class of local alternatives we consider in this paper; that is, it has power to detect predictive episodes. However, local power depends indirectly on heteroskedasticity which influences the stochastic properties of $K_{z_t x}(\cdot)$ and $J_{c,\mathrm{H}}(\cdot)$; see Remarks 15 and 16. ◇

Remark 18. Where $\beta_{1,t} = \beta_1 \neq 0$, for all $t = 1, \ldots, T$, the results in Corollary 2 specialise to the standard local power of the full sample IV-combination test based on $t^2_{\beta_1}$. For type-II instruments without demeaning, one recovers the result of Breitung and Demetrescu (2015, Theorem 2.2). ◇

To summarise, the limiting null distributions of the maximum subsample statistics all depend both on any heteroskedasticity present and on whether the putative predictor x_t is a near-integrated or weakly dependent process. This poses significant problems for conducting inference not encountered with tests based on the full sample statistic, $t^2_{\beta_1}$ of (3.2). We next demonstrate that these issues can be solved using fixed regressor wild bootstrap implementations of the subsample tests.

5. Bootstrap implementation

As the results in the previous section show, implementing tests based on the $\mathcal{T}^s$, $s = f, b, d, r$, statistics from Section 4.2 will require us to address the fact that their limiting null distributions depend on any unconditional heteroskedasticity present in u_t and v_t, and on whether the predictor x_{t-1} is weakly dependent or near-integrated. To account for the former we employ a wild bootstrap resampling scheme applied to the demeaned dependent variable $\hat{y}_t := y_t - \frac{1}{T}\sum_{t=1}^T y_t$, while for the latter we use the observed outcomes on $\mathbf{x} := [x_0, x_1, \ldots, x_T]'$ and $\mathbf{z} := [\boldsymbol{z}'_0, \boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_T]'$ as a fixed regressor and fixed instrument vector, respectively, when implementing the bootstrap procedure.

We now outline our fixed regressor wild bootstrap approach in Algorithm 1. We will then demonstrate the asymptotic validity of this approach in Proposition 2.

Algorithm 1.

Step 1 Construct the wild bootstrap innovations $y^*_t := \hat{y}_t R_t$, where $\hat{y}_t := y_t - \frac{1}{T}\sum_{t=1}^T y_t$ are the demeaned sample observations on y_t, and R_t, $t = 1, \ldots, T$, is an *IID* $N(0, 1)$ sequence independent of the data.[5]

Step 2 Using the bootstrap sample data $\left(y^*_t, x_{t-1}, \boldsymbol{z}'_{t-1}\right)'$, in place of the original sample data $\left(y_t, x_{t-1}, \boldsymbol{z}'_{t-1}\right)'$, construct the bootstrap analogues of the statistics $\mathcal{T}^s$, $s = f, b, d, r$, from Section 4.2. Denote these bootstrap statistics as $\mathcal{T}^{s*}$, $s = f, b, d, r$.

Step 3 Define the bootstrap p-values as $P^{s,*}_T := 1 - G^{s,*}_T(\mathcal{T}^s)$, $s = f, b, d, r$, with $G^{s,*}_T(\cdot)$ denoting the conditional (on the original data) cumulative distribution function (cdf) of $\mathcal{T}^{s*}$, $s = f, b, d, r$. In practice, the $G^{s,*}_T(\cdot)$, $s = f, b, d, r$, will be unknown, but can be simulated in the usual way by repeating Steps 1 and 2 a large number, say B, times to obtain empirical analogues of $G^{s,*}_T(\cdot)$, $s = f, b, d, r$. The $\{R_t\}_{t=1}^T$ variables used in Step 1 must also be independent across the B bootstrap replications.

Step 4 The wild bootstrap test of the null hypothesis H_0 of (2.3) at level α based on $\mathcal{T}^s$ rejects if $P^{s,*}_T \leq \alpha$, $s = f, b, d, r$.

Remark 19. The bootstrap statistics $\mathcal{T}^{s*}$, $s = f, b, d, r$, are calculated treating both x_{t-1} and the vector of instruments, $\boldsymbol{z}_{t-1}$, as fixed; i.e., they are calculated using the *same* observed x_{t-1} and $\boldsymbol{z}_{t-1}$ as were used in the construction of $\mathcal{T}^s$, $s = f, b, d, r$. This aspect is crucial for delivering bootstrap tests that are asymptotically valid regardless of whether x_t satisfies Assumption 1.1 or 1.2 and without knowledge of which of these holds. In particular, the same instrument (either type-I or type-II, depending on the true regressor persistence) gets full asymptotic weight in both the original 2SLS and the bootstrap t-ratios (see section S.3 of the supplementary material). ◇

Remark 20. The wild bootstrap generating y^*_t in Step 1 of Algorithm 1 replicates the pattern of unconditional heteroskedasticity present in the original innovations, as conditionally on $\hat{y}_t$, y^*_t is independent over time with zero mean and variance $\hat{y}^2_t$. Moreover, any heteroskedasticity present in x_{t-1} and $\boldsymbol{z}_{t-1}$ is replicated through the fixed regressor/instrument aspect of the bootstrap statistics. In particular, as u_t is a MD sequence, it is anticipated that the bootstrap will replicate the variance properties of either $z_{I,t-1}u_t$ or $z_{II,t-1}u_t$, depending on the degree of persistence exhibited by x_t. Having fixed the regressor and the instruments when bootstrapping, the analogous terms in the bootstrap test statistics are given by $z_{I,t-1}\hat{y}_t R_t$ and $z_{II,t-1}\hat{y}_t R_t$ with variances $z^2_{I,t-1}\hat{y}^2_t$ and $z^2_{II,t-1}\hat{y}^2_t$, respectively. Using the result detailed in the last sentence of Remark 19, it is then seen that the correct variance profile is replicated in the limit. For full details see the proof of Proposition 2. ◇

[5] The Gaussianity assumption on R_t is standard in the literature and simplifies the proof of Proposition 2. This can, however, be generalised such that R_t is any IID sequence with $\mathrm{E}(R_t) = 0$, $\mathrm{E}(R^2_t) = 1$ and $\mathrm{E}(R^4_t) < \infty$.

Remark 21. Step 1 of Algorithm 1 is based on residuals obtained under the null hypothesis. It is straightforward to show that the large sample properties of corresponding bootstrap tests based on either the OLS or IVX residuals from estimating the predictive regression over the full sample are unaltered from those given here. Moreover, albeit more computationally intensive, one could also use the analogous subsample implementations of any of these three full sample residuals. ◇

In Proposition 2 we now demonstrate the large sample validity of the fixed regressor wild bootstrap implementation of the tests from Section 4.2. In particular, we show that our proposed bootstrap in Algorithm 1 correctly replicates the first order asymptotic null distributions of the statistics given in Remark 12 under both the null hypothesis and local alternatives.

Proposition 2. *Let the conditions of Proposition 1 hold. Then, as $T \to \infty$, under either the null hypothesis H_0 of (2.3) or the local alternative $H_{1,b(\cdot)}$ of (2.4): (i) under Assumption 1.1, as $T \to \infty$, it holds that $\mathcal{T}^{f*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{f,I}$, $\mathcal{T}^{b*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{b,I}$, $\mathcal{T}^{r*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{r,I}$, and $\mathcal{T}^{d*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{d,I}$; (ii) under Assumption 1.2 as $T \to \infty$, it holds that, $\mathcal{T}^{f*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{f,II}$, $\mathcal{T}^{b*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{b,II}$, $\mathcal{T}^{r*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{r,II}$, and $\mathcal{T}^{d*} \overset{w}{\Rightarrow}_p \mathcal{T}_\infty^{d,II}$.*

A consequence of Proposition 2 is that we obtain asymptotically correctly sized tests when using bootstrap critical values obtained using Algorithm 1. We now formalise this result in Corollary 3.

Corollary 3. *As $T \to \infty$, under H_0, $\mathrm{P}^*(\mathcal{T}^{s*} \leq \mathcal{T}^s) \Rightarrow Unif[0,1]$, for each of $s = f, b, d, r$, where P^* denotes probability conditional on the original sample $\left(y_t, x_{t-1}, \boldsymbol{z}_{t-1}'\right)'$, $t = 1, \ldots, T$.*

Remark 22. Corollary 3 establishes the asymptotic validity of our proposed bootstrap tests. This result holds without knowledge of whether the (putative) predictor x_t satisfies Assumption 1.1 or 1.2, and holds regardless of any heteroskedasticity present in u_t and v_t satisfying Assumption 3. ◇

Remark 23. Proposition 2 shows that each of the bootstrap statistics $\mathcal{T}^{s*}$, $s = f, b, d, r$, attains the same first order limiting distribution under both the null hypothesis and local alternatives as that attained under the null hypothesis by the corresponding original (non-bootstrap) statistic $\mathcal{T}^s$, $s = f, b, d, r$. An immediate consequence of this is that each of the wild bootstrap tests proposed in Algorithm 1 will admit the same asymptotic local power function as the (infeasible) size-adjusted test based on the corresponding original statistic $\mathcal{T}^s$, $s = f, b, d, r$. ◇

Remark 24. The full sample IV-combination statistic, $t_{\beta_1}^2$, of Breitung and Demetrescu uses Eicker–White standard errors to correct the limiting null distribution of the statistic for non-constancy in $\mathbf{H}(\cdot)$ due to unconditional heteroskedasticity in the innovations. Because it is still necessary to implement the subsample maximum tests using a wild bootstrap it would be feasible to replace the Eicker–White standard errors used in the computation of the subsample $\left(t_{\beta_1}(\tau_1, \tau_2)\right)^2$ statistic in (4.3) and its bootstrap equivalent, computed in Step 2 of Algorithm 1, with conventional standard errors. While this would alter the limiting representations given for the maximum statistics in Proposition 1 and Corollary 1, it can be shown that the resulting wild bootstrap tests would still be asymptotically valid with an analogous result to that in Corollary 3 holding. In this case the wild bootstrap tests would attain the same asymptotic local power functions as (infeasible) size-corrected implementations of the (non-bootstrap) maximum tests based on conventional standard errors. These asymptotic local power functions will not in general coincide with those obtained for the statistics based on Eicker–White standard errors, but they would where $\mathbf{H}(\cdot)$ is constant. ◇

Remark 25. The bootstrap validity results given in this section also apply to a fixed regressor wild bootstrap implementation of the full sample IV-combination test based on $t_{\beta_1}^2$. In particular, this will satisfy a result of the form given in Corollary 3 and will have the same asymptotic local power function as the test based on $t_{\beta_1}^2$ using χ_1^2 critical values, discussed in Section 4.3. As with the discussion for the subsample maximum statistics in Remark 24, one could replace Eicker–White standard errors with conventional standard errors without losing asymptotic validity. ◇

6. Numerical results

We use Monte Carlo simulation methods to investigate the finite sample performance of the bootstrap implementations of the subsample-based predictability tests $\mathcal{T}^f$, $\mathcal{T}^b$, $\mathcal{T}^r$ and $\mathcal{T}^d$ proposed in Section 4 for testing the null hypothesis of no predictability in (2.3); i.e., $H_0 : \beta_{1,t} = 0$, for all $t = 1, \ldots, T$, against the alternative $H_{1,b(\cdot)}$ of (2.4) that predictability holds across some subset of the sample data. Data are generated from (2.1)–(2.2). In Section 6.1 we explore the empirical size properties of these tests comparing with the corresponding full sample IV-combination test of Breitung and Demetrescu (2015), $t_{\beta_1}^2$ of (3.2). In Section 6.2 we compare the finite sample local power properties of these tests against a variety of DGPs displaying temporary predictability.

Following the discussion in Section 4.1, we base both the full sample t_{β_1} IV-combination statistic in (3.2) and the corresponding subsample $t_{\beta_1}(\tau_1, \tau_2)$ statistics in (4.3) on the instrument vector $\boldsymbol{z}_{t-1} := \left(z_{I,t-1}, z_{II,t-1}\right)'$ with the type-II instrument, $z_{II,t-1}$, defined as in (4.1) with $k = 1$, and the type-I instrument, $z_{I,t-1}$, given by the IVX choice of Kostakis et al.

(2015) defined as in (4.2), with $a = 1$ and $\gamma = 0.95$.[6] Excepting the IVX instrument, $z_{l,t-1}$, all variables and instruments entering the estimated predictive regressions are demeaned, as in the main text. As discussed in Kostakis et al. (2015, p. 1514) the IVX instrument, $z_{l,t-1}$, does not need to be demeaned as the slope estimator in the predictive regression is invariant to whether $z_{l,t-1}$ is demeaned or not. In order to correct for the finite sample effects of estimating the intercept term in (2.1), which are most pronounced for highly persistent regressors which are strongly correlated with the predictive model's innovations, Kostakis et al. (2015, p. 1516) recommend the use of a finite-sample correction factor. We also found that this correction factor led to significant improvements in the finite sample properties of our proposed tests and hence it is implemented in all of the numerical and empirical results we report.

All simulations are preformed in MATLAB, versions R2018a and R2018b, using the Mersenne Twister random number generator. All results pertain to the nominal 5% level; qualitatively similar results were obtained for other conventional significance levels. All of the subsample tests are computed according to Algorithm 1 using 399 bootstrap replications; the bootstrap tests are denoted $\mathcal{T}^{s*}$, $s = f, b, d, r$. Following Banerjee et al. (1992), we set $\tau_L = 1/4$ and $\tau_U = 3/4$ in the context of the forward and backward recursive statistics, respectively, and $\Delta\tau = 1/3$ for the rolling and double recursive statistics. The empirical size simulations were based on 5000 Monte Carlo replications and the local power simulations on 1000 replications, with the exception of the double recursive tests where 1000 replications were used for size and 500 for power because of the much higher computing time required. For the full sample $t^2_{\beta_1}$ test, results for versions based on the asymptotic χ^2_1 critical value and on a fixed regressor wild bootstrap are reported, the latter using 399 bootstrap replications. For all of the bootstrap tests, two versions are reported. The first is based on statistics using Eicker–White standard errors while for the second, following the discussion in Remark 24, conventional standard errors are used. These two variants are distinguished apart by the additional "*NW*" nomenclature in the subscripts of the latter. Following the discussion in Section 3 and footnote 4.2, all of the reported statistics use residuals, $\hat{u}_t$, computed under the null hypothesis.

6.1. Empirical size

We first investigate the finite sample size properties of our proposed tests. To that end, we consider the simulation DGP given by (2.1)–(2.2) with $\beta_{1,t} = \beta_1 = 0$ for all $t = 1, \ldots, T$. Results are reported for $T = 250$ and $T = 500$. In generating the simulation data we set the intercepts β_0 and μ_x in (2.1) and (2.2), respectively, to zero with no loss of generality. The autoregressive process characterising the dynamics of the putative predictor, x_t, in (2.2) was initialised at $\xi_0 = 0$. Results are reported for a range of values of the autoregressive parameter ρ in (2.2) that cover both stationary and persistent predictors; in particular, for $\rho := 1 - c/T$ we consider $c \in \{0, 2.5, 5, 10, 20, 0.5T\}$. Notice that $c = 0.5T$ corresponds to $\rho = 0.5$, such that the autoregressive parameter is fixed and stable.

In our simulation DGP the innovation vector $(u_t, v_t)'$ is drawn from an i.i.d. bivariate Gaussian distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_t := \begin{bmatrix} \sigma^2_{ut} & \phi\sigma_{ut}\sigma_{vt} \\ \phi\sigma_{ut}\sigma_{vt} & \sigma^2_{vt} \end{bmatrix}$. Notice, therefore, that ϕ corresponds to the correlation between the innovations u_t and v_t. Results are reported in Table 1 for the case where $\phi = 0$, and in Table 2 for the case where $\phi = -0.90$.[7] We report results for the case where the innovations are homoskedastic, $\sigma^2_{ut} = \sigma^2_{vt} = 1$ (labelled DGP1 in Tables 1 and 2), and for the case where there is a contemporaneous one-time break of equal magnitude in the variances of u_t and v_t. Following the simulation designs considered in Georgiev et al. (2019), two such heteroskedastic cases are considered: (i) an upward change in variance (labelled DGP2 in Tables 1 and 2) such that $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + 4\mathbb{I}(t > \lfloor 0.5T \rfloor)$, and (ii) a downward change (labelled DGP3 in Tables 1 and 2) where $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + \frac{1}{4}\mathbb{I}(t > \lfloor 0.5T \rfloor)$, where in each case $\mathbb{I}(\cdot)$ denotes the indicator function, taking the value one when its argument is true and zero otherwise. DGP2 and DGP3 allow us to examine the impact of unconditional heteroscedasticity, both in isolation and in its interaction with ϕ, on the finite sample size of the tests. In each of DGP2 and DGP3 a fourfold change in variance is seen which is likely to be of rather larger magnitude than we might expect to see in practice, but serves to illustrate how the tests behave in the presence of a large change in unconditional volatility. We also considered further DGPs allowing for stationary GARCH(1,1) with different degrees of persistence coupled with either Gaussian or t-distributed innovations, thereby allowing for unconditionally heteroskedastic and fat-tailed innovations. These results were qualitatively similar to those reported here for DGP1 and can be found in the supplementary material.

Consider first the results pertaining to the homoskedastic DGP1. A comparison of the results in Table 1 for $\phi = 0$ and Table 2 for $\phi = -0.90$ shows that, in the homoskedastic case at least, the correlation parameter ϕ has relatively little impact on the size properties of the tests. For the full sample tests there is relatively little difference between the tests based on the asymptotic χ^2_1 critical value and the fixed regressor wild bootstrap. Similarly, as might be expected, there is little to choose between the versions of the full sample tests with Eicker–White standard errors and those with conventional standard errors. For the subsample tests, there is a general trend towards undersizing in the Eicker–White versions in cases where the putative predictor, x_{t-1}, displays persistence at or close to a unit root process. This is most

[6] We also considered tests based on using the fractionally integrated instrument suggested on page 363 of Breitung and Demetrescu (2015) for $z_{l,t-1}$. We do not report these results here as the IVX choice performed better in our results, but they can be obtained from the authors on request.

[7] In predictive regression models for the equity premium employing valuation ratios as predictors (e.g. the dividend-price ratio, earnings-price ratio), as we shall do in the empirical application in Section 7, the relevant innovation terms are strongly negatively correlated, hence our choice of $\phi = -0.90$.

Table 1
Empirical Rejection Frequencies under the Null Hypothesis, H_0. Nominal 5% significance level. DGP1-DGP3 with $\phi = 0$.

c	$t^{2*}_{\beta_1}$	$t^{2*}_{\beta_1,NW}$	$t^{2}_{\beta_1}$	$t^{2}_{\beta_1,NW}$	$\mathcal{T}^{f*}$	$\mathcal{T}^{b*}$	$\mathcal{T}^{f*}_{NW}$	$\mathcal{T}^{b*}_{NW}$	$\mathcal{T}^{r*}$	$\mathcal{T}^{r*}_{NW}$	$\mathcal{T}^{d*}$	$\mathcal{T}^{d*}_{NW}$
DGP1: $T = 250$, $\phi = 0$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1$												
0.0	0.045	0.046	0.046	0.045	0.037	0.036	0.049	0.051	0.017	0.057	0.015	0.067
2.5	0.047	0.047	0.049	0.045	0.032	0.032	0.048	0.049	0.013	0.053	0.008	0.059
5.0	0.043	0.044	0.043	0.041	0.031	0.032	0.047	0.044	0.009	0.050	0.012	0.062
10.0	0.042	0.043	0.040	0.039	0.033	0.039	0.046	0.048	0.014	0.050	0.015	0.057
20.0	0.049	0.048	0.046	0.044	0.039	0.039	0.044	0.048	0.025	0.050	0.028	0.056
0.5T	0.049	0.047	0.046	0.047	0.058	0.053	0.052	0.047	0.061	0.045	0.070	0.044
DGP2: $T = 250$, $\phi = 0$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + 4\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.047	0.048	0.032	0.059	0.040	0.044	0.057	0.050	0.024	0.052	0.020	0.048
2.5	0.044	0.047	0.036	0.066	0.033	0.033	0.053	0.048	0.018	0.055	0.009	0.059
5.0	0.043	0.043	0.035	0.068	0.036	0.033	0.056	0.046	0.015	0.054	0.008	0.061
10.0	0.046	0.045	0.037	0.078	0.037	0.038	0.057	0.048	0.018	0.054	0.020	0.055
20.0	0.048	0.047	0.041	0.084	0.044	0.041	0.058	0.049	0.027	0.051	0.028	0.055
0.5T	0.052	0.051	0.047	0.090	0.062	0.058	0.051	0.050	0.066	0.050	0.070	0.057
DGP3: $T = 250$, $\phi = 0$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + \frac{1}{4}\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.045	0.047	0.076	0.061	0.036	0.043	0.047	0.057	0.029	0.059	0.037	0.072
2.5	0.045	0.047	0.058	0.072	0.031	0.033	0.046	0.061	0.022	0.058	0.017	0.063
5.0	0.044	0.045	0.045	0.067	0.029	0.036	0.043	0.056	0.016	0.060	0.012	0.062
10.0	0.044	0.045	0.042	0.069	0.033	0.038	0.040	0.051	0.018	0.057	0.020	0.059
20.0	0.045	0.043	0.040	0.071	0.041	0.044	0.046	0.049	0.027	0.053	0.031	0.057
0.5T	0.047	0.045	0.045	0.084	0.056	0.062	0.049	0.053	0.069	0.049	0.063	0.038
DGP1: $T = 500$, $\phi = 0$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1$												
0.0	0.046	0.046	0.043	0.045	0.043	0.039	0.049	0.051	0.017	0.052	0.011	0.050
2.5	0.048	0.049	0.047	0.047	0.042	0.039	0.052	0.052	0.016	0.054	0.014	0.048
5.0	0.047	0.048	0.047	0.046	0.044	0.045	0.053	0.056	0.015	0.054	0.013	0.050
10.0	0.050	0.051	0.049	0.048	0.047	0.042	0.056	0.049	0.027	0.055	0.027	0.056
20.0	0.053	0.052	0.053	0.053	0.051	0.043	0.052	0.052	0.041	0.051	0.043	0.062
0.5T	0.047	0.047	0.044	0.045	0.053	0.053	0.048	0.050	0.065	0.054	0.072	0.058
DGP2: $T = 500$, $\phi = 0$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + 4\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.052	0.053	0.036	0.061	0.050	0.039	0.063	0.055	0.025	0.061	0.033	0.062
2.5	0.051	0.052	0.036	0.065	0.046	0.040	0.062	0.053	0.023	0.060	0.022	0.060
5.0	0.050	0.050	0.037	0.073	0.046	0.044	0.056	0.052	0.026	0.061	0.026	0.055
10.0	0.050	0.050	0.041	0.083	0.050	0.045	0.055	0.054	0.033	0.058	0.035	0.052
20.0	0.050	0.051	0.045	0.088	0.053	0.050	0.051	0.056	0.046	0.055	0.041	0.047
0.5T	0.052	0.050	0.045	0.092	0.066	0.054	0.053	0.050	0.063	0.056	0.069	0.050
DGP3: $T = 500$, $\phi = 0$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + \frac{1}{4}\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.050	0.051	0.085	0.073	0.045	0.042	0.056	0.062	0.029	0.066	0.033	0.062
2.5	0.052	0.053	0.065	0.077	0.043	0.041	0.056	0.060	0.028	0.063	0.020	0.083
5.0	0.053	0.051	0.050	0.075	0.042	0.053	0.051	0.060	0.025	0.065	0.025	0.078
10.0	0.052	0.053	0.049	0.076	0.044	0.049	0.054	0.058	0.030	0.062	0.026	0.075
20.0	0.052	0.051	0.048	0.081	0.047	0.056	0.053	0.058	0.041	0.057	0.041	0.064
0.5T	0.051	0.051	0.048	0.092	0.056	0.061	0.050	0.053	0.066	0.056	0.055	0.050

Notes: A superscript $*$ denotes tests run using the fixed regressor wild bootstrap outlined in Algorithm 1; $t^2_{\beta_1}$ and $t^2_{\beta_1,NW}$ denote the full sample IV-combination predictability tests of Breitung and Demetrescu (2015) based on the 5% asymptotic critical value from the χ^2_1 distribution and computed with Eicker–White [EW] and conventional standard errors, respectively, and $t^{2*}_{\beta_1}$ and $t^{2*}_{\beta_1,NW}$ their bootstrap analogues; $\mathcal{T}^{f*}$, $\mathcal{T}^{b*}$ and $\mathcal{T}^{f*}_{NW}$, $\mathcal{T}^{b*}_{NW}$, denote the maximum forward and backward recursive tests computed with EW and conventional standard errors, respectively; $\mathcal{T}^{r*}$ and $\mathcal{T}^{r*}_{NW}$ denote the maximum rolling tests computed with EW and conventional standard errors, respectively; $\mathcal{T}^{d*}$ and $\mathcal{T}^{d*}_{ols}$ denote the maximum double recursive tests computed with EW and conventional standard errors, respectively.

pronounced in the rolling and double recursive statistics. However, this undersizing is not seen with the versions of the subsample tests based on conventional standard errors. It is well known that Eicker–White standard errors can be heavily downward biased in small samples leading to incorrectly sized tests; see, for example, MacKinnon and White (1985).

We next turn to the results for the two unconditionally heteroskedastic DGPs, DGP2 and DGP3. Consider first the full sample tests. As expected, the full sample test based on conventional standard errors and the asymptotic χ^2_1 critical value, $t^2_{\beta_1,NW}$, is unreliable in the presence of heteroskedasticity. These size distortions are considerably worse for $\phi = -0.90$ than for $\phi = 0$ when $c = 0$; for the other values of c considered the differences between $\phi = 0$ and $\phi = -0.90$ are much smaller. The size distortions observed with $t^2_{\beta_1,NW}$ are significantly ameliorated by the use of Eicker–White standard errors ($t^2_{\beta_1}$) in all but the case of DGP2 with $c = 0$ where no apparent improvements are seen. The bootstrap implementations of the full sample tests do a much better job at controlling finite sample size, regardless of whether Eicker–White or conventional standard errors are used, although some over-sizing is still seen for $\phi = -0.90$ when $c = 0$. There appears

Table 2
Empirical Rejection Frequencies under the Null Hypothesis, H_0. Nominal 5% significance level. DGP1-DGP3 with $\phi = -0.90$.

c	$t^{2*}_{\beta_1}$	$t^{2*}_{\beta_1,NW}$	$t^{2}_{\beta_1}$	$t^{2}_{\beta_1,NW}$	$\mathcal{T}^{f*}$	$\mathcal{T}^{b*}$	$\mathcal{T}^{f*}_{NW}$	$\mathcal{T}^{b*}_{NW}$	$\mathcal{T}^{r*}$	$\mathcal{T}^{r*}_{NW}$	$\mathcal{T}^{d*}$	$\mathcal{T}^{d*}_{NW}$
DGP1: $T = 250$, $\phi = -0.90$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1$												
0.0	0.069	0.073	0.069	0.074	0.037	0.035	0.047	0.055	0.011	0.053	0.018	0.055
2.5	0.055	0.056	0.055	0.057	0.030	0.040	0.037	0.058	0.011	0.052	0.020	0.062
5.0	0.053	0.052	0.051	0.050	0.034	0.045	0.039	0.055	0.012	0.051	0.021	0.064
10.0	0.057	0.055	0.056	0.058	0.038	0.048	0.044	0.059	0.017	0.055	0.029	0.065
20.0	0.060	0.060	0.057	0.056	0.046	0.051	0.050	0.062	0.029	0.061	0.034	0.067
$0.5T$	0.055	0.053	0.051	0.052	0.059	0.067	0.057	0.060	0.073	0.058	0.067	0.055
DGP2: $T = 250$, $\phi = -0.90$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + 4\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.057	0.057	0.044	0.071	0.032	0.031	0.026	0.046	0.009	0.030	0.018	0.021
2.5	0.052	0.053	0.044	0.076	0.035	0.032	0.027	0.051	0.011	0.028	0.012	0.031
5.0	0.054	0.053	0.046	0.076	0.037	0.037	0.033	0.054	0.013	0.032	0.014	0.038
10.0	0.054	0.051	0.046	0.077	0.042	0.043	0.037	0.055	0.018	0.039	0.019	0.051
20.0	0.054	0.052	0.049	0.080	0.047	0.046	0.042	0.054	0.032	0.045	0.036	0.055
$0.5T$	0.058	0.055	0.053	0.097	0.067	0.059	0.054	0.057	0.063	0.056	0.073	0.058
DGP3: $T = 250$, $\phi = -0.90$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + \frac{1}{4}\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.072	0.078	0.125	0.118	0.029	0.036	0.049	0.076	0.007	0.072	0.009	0.084
2.5	0.046	0.048	0.059	0.068	0.022	0.046	0.038	0.071	0.008	0.071	0.005	0.077
5.0	0.049	0.048	0.055	0.068	0.027	0.049	0.040	0.064	0.010	0.068	0.006	0.075
10.0	0.058	0.051	0.052	0.073	0.033	0.049	0.041	0.056	0.018	0.061	0.013	0.062
20.0	0.053	0.053	0.049	0.077	0.040	0.052	0.050	0.050	0.028	0.055	0.036	0.061
$0.5T$	0.053	0.052	0.047	0.091	0.058	0.067	0.056	0.056	0.064	0.056	0.070	0.061
DGP1: $T = 500$, $\phi = -0.90$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1$												
0.0	0.076	0.079	0.077	0.078	0.041	0.054	0.045	0.072	0.015	0.055	0.023	0.060
2.5	0.059	0.060	0.056	0.058	0.034	0.060	0.037	0.075	0.020	0.053	0.016	0.045
5.0	0.061	0.060	0.058	0.061	0.038	0.064	0.041	0.075	0.023	0.057	0.023	0.056
10.0	0.062	0.063	0.061	0.063	0.045	0.065	0.049	0.076	0.034	0.059	0.035	0.053
20.0	0.063	0.062	0.060	0.061	0.050	0.065	0.054	0.072	0.047	0.063	0.046	0.064
$0.5T$	0.050	0.049	0.049	0.050	0.060	0.058	0.053	0.053	0.070	0.059	0.067	0.058
DGP2: $T = 500$, $\phi = -0.90$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + 4\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.072	0.067	0.055	0.082	0.041	0.047	0.029	0.066	0.022	0.037	0.021	0.033
2.5	0.060	0.057	0.052	0.079	0.041	0.049	0.030	0.063	0.021	0.036	0.030	0.039
5.0	0.061	0.060	0.052	0.085	0.045	0.057	0.033	0.068	0.024	0.040	0.039	0.043
10.0	0.058	0.058	0.053	0.088	0.049	0.055	0.040	0.064	0.034	0.043	0.049	0.045
20.0	0.061	0.058	0.052	0.095	0.056	0.060	0.044	0.066	0.044	0.045	0.056	0.051
$0.5T$	0.048	0.048	0.045	0.084	0.063	0.052	0.051	0.050	0.067	0.057	0.073	0.054
DGP3: $T = 500$, $\phi = -0.90$ and $\sigma^2_{ut} = \sigma^2_{vt} = 1\mathbb{I}(t \leq \lfloor 0.5T \rfloor) + \frac{1}{4}\mathbb{I}(t > \lfloor 0.5T \rfloor)$												
0.0	0.082	0.089	0.133	0.129	0.043	0.054	0.060	0.092	0.019	0.092	0.017	0.100
2.5	0.053	0.053	0.066	0.074	0.029	0.071	0.047	0.090	0.022	0.093	0.023	0.081
5.0	0.054	0.054	0.056	0.076	0.033	0.068	0.047	0.077	0.027	0.083	0.026	0.083
10.0	0.060	0.058	0.057	0.086	0.040	0.069	0.048	0.070	0.033	0.076	0.033	0.078
20.0	0.061	0.058	0.055	0.092	0.047	0.063	0.054	0.063	0.044	0.070	0.049	0.074
$0.5T$	0.053	0.052	0.051	0.096	0.057	0.068	0.053	0.058	0.069	0.054	0.071	0.068

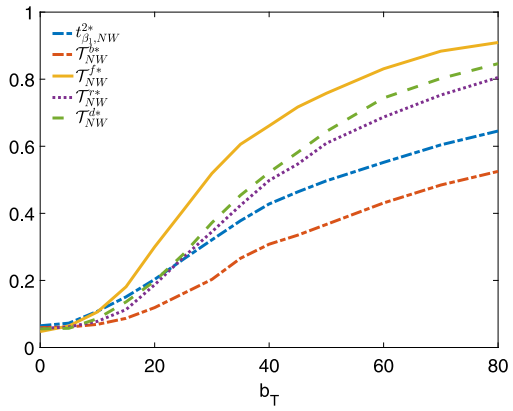
Notes: See notes to Table 1

to be no need to use Eicker–White standard errors with the fixed regressor bootstrap implementation of the full sample test.

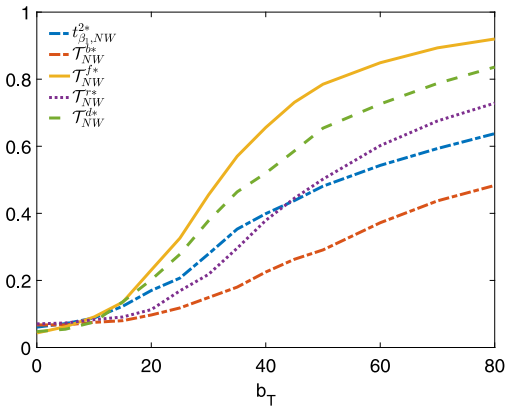
Consider next the subsample predictability tests. Undersizing, in many cases substantial, is again seen with the subsample bootstrap tests based on Eicker–White standard errors. As with the full sample tests, these effects tend to be larger, other things equal, for $\phi = -0.90$ *vis-à-vis* $\phi = 0$. As with the results for DGP1, the subsample bootstrap tests based on conventional standard errors are much less prone to this undersizing phenomenon, albeit some undersizing is seen in the case of DGP2 with $\phi = -0.90$ for small values of c, most notably for the rolling and double recursive tests. Moreover, under DGP3 with $\phi = -0.90$ some oversizing is seen in the persistent x_{t-1} cases for the backward recursive, rolling and double recursive tests. For $\phi = 0$ all of the subsample bootstrap tests implemented with conventional standard errors appear to display good finite sample size control.

6.2. Finite sample local power

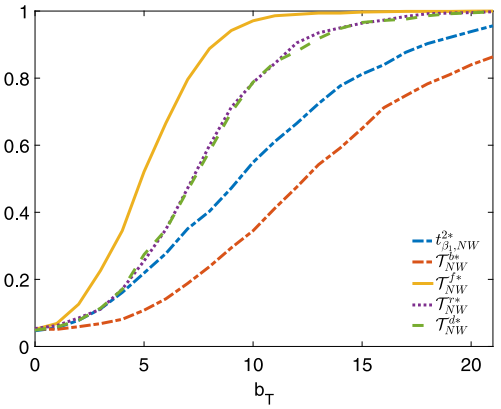
We now turn to an investigation into the relative finite sample local power properties of the tests. We again generate simulation data from DGP (2.1)–(2.2) but now for a variety of local alternatives satisfying $H_{1,b(\cdot)}$ of (2.4). To keep the set of results to a manageable level we report results only for $\phi = -0.90$, for the homoskedastic case, $\sigma^2_{ut} = \sigma^2_{vt} = 1$, for a sample



(a) $c = 0, \phi = -0.90$



(b) $c = 10, \phi = -0.90$



(c) $c = 0.5T, \phi = -0.90$

Fig. 1. Finite sample local power: Case 1, $T = 250$.

of size $T = 250$ and for three values of the persistence parameter, c, associated with x_t; specifically, $c = \{0, 10, 0.5T\}$. In all of our experiments the slope parameter β_{1t} in (2.1) is set to be local-to-zero. As specified by Assumption 2, for $c = 0$ and $c = 10$, where x_t is strongly persistent, we parameterise the slope parameter in (2.1) as $\beta_{1t} = b_{1t}/T$, and here we consider the following values of the Pitman drift parameter, $b_{1t} \in \{0, 5, \ldots, 80\}$. For the case of a weakly dependent predictor, $c = 0.5T$, we parameterise the slope parameter as $\beta_{1t} = b_{1t}/\sqrt{T}$, and here we consider the Pitman drift values $b_{1t} \in \{0, 1, \ldots, 21\}$.

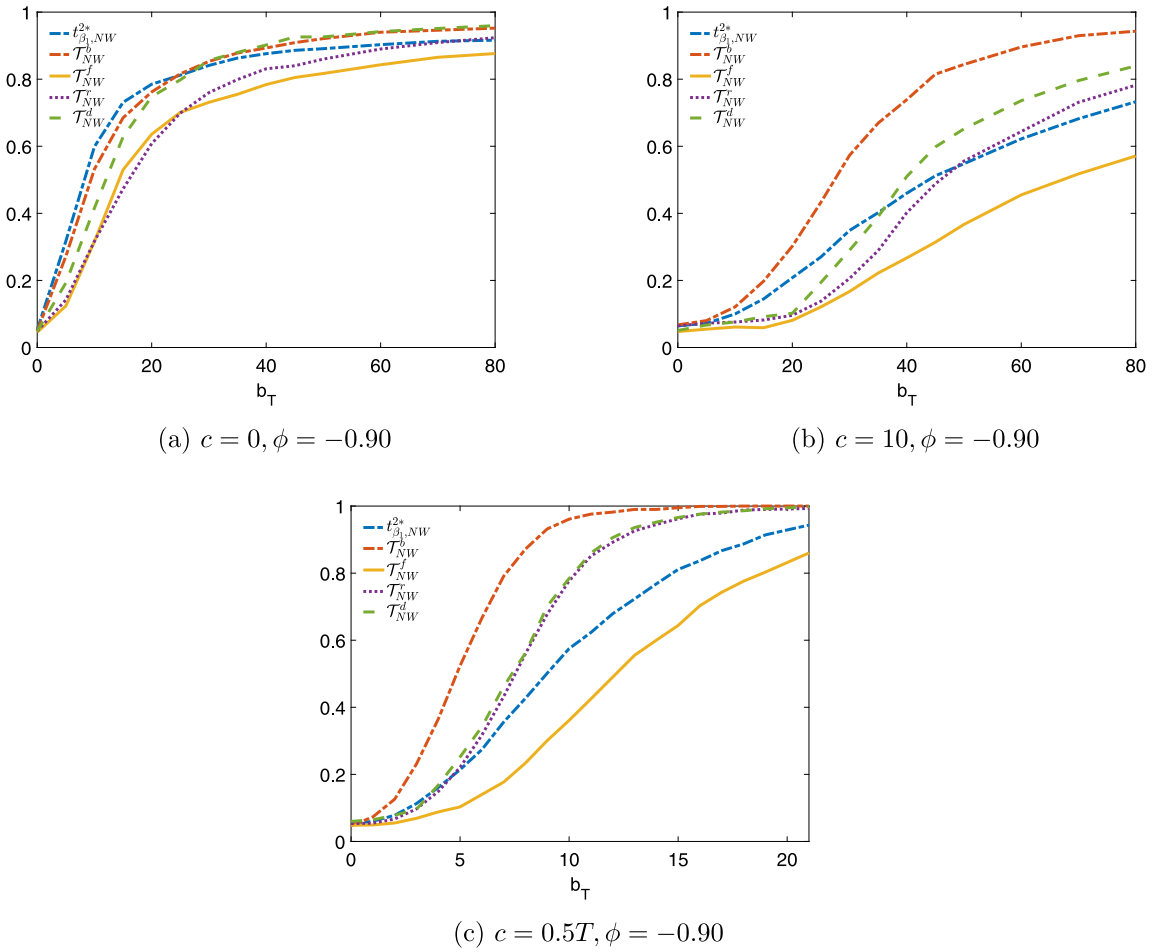
We report results for three distinct experimental cases, where episodes of predictability occur once in the sample either at the beginning, the end or within the sample. To that end, we consider the following three simulation DGPs, with the range of non-zero values of b_{1t} as outlined above,

$$\text{Case 1:} \begin{cases} b_{1t} > 0 & \text{for} \quad t = 1, \ldots, \lfloor T/5 \rfloor \\ b_{1t} = 0 & \text{for} \quad t = \lfloor T/5 \rfloor + 1, \ldots, T \end{cases} \qquad \text{Case 2:} \begin{cases} b_{1t} = 0 & \text{for} \quad t = 1, \ldots, \lfloor 4T/5 \rfloor \\ b_{1t} > 0 & \text{for} \quad t = \lfloor 4T/5 \rfloor + 1, \ldots, T \end{cases}$$

$$\text{Case 3:} \begin{cases} b_{1t} = 0 & \text{for} \quad t = 1, \ldots, \lfloor T/5 \rfloor \\ b_{1t} > 0 & \text{for} \quad t = \lfloor T/5 \rfloor + 1, \ldots, \lfloor 3T/5 \rfloor \\ b_{1t} = 0 & \text{for} \quad t = \lfloor 3T/5 \rfloor + 1, \ldots, T \end{cases}$$

All other aspects of the simulation design are as described previously.

Figs. 1–3 graph the simulated finite sample local power curves for each of Cases 1–3, respectively. Each figure contains power curves for the fixed regressor wild bootstrap implementations of the full sample $t^{2*}_{\beta_1,NW}$ test along with the subsample-based predictability tests $\mathcal{T}^{f*}_{NW}$, $\mathcal{T}^{b*}_{NW}$, $\mathcal{T}^{r*}_{NW}$ and $\mathcal{T}^{d*}_{NW}$. To aid presentation of the graphs, we have chosen only to report the versions of the bootstrap tests implemented with conventional standard errors. Results with Eicker–White



(a) $c = 0, \phi = -0.90$

(b) $c = 10, \phi = -0.90$

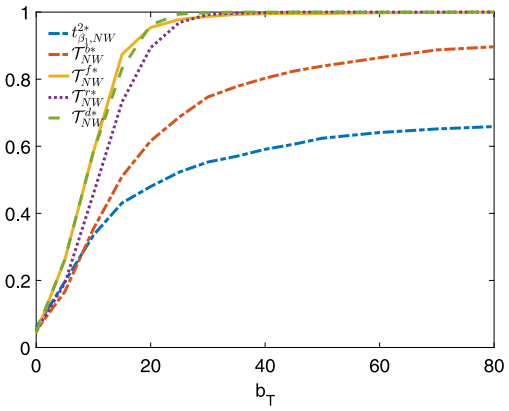
(c) $c = 0.5T, \phi = -0.90$

Fig. 2. Finite sample local power: Case 2, $T = 250$.

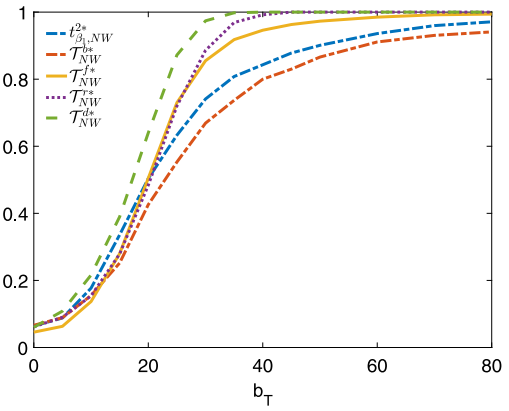
standard errors are available on request. In general the latter were less powerful (often considerably so) than the reported tests based on conventional standard errors.

Consider first the results pertaining to Case 1 in Fig. 1. Recall from Section 4.2 that the temporary predictability DGP in Case 1, with a pocket of predictability at the start of the sample, is one where we expect the forward recursive $\mathcal{T}^{f*}_{NW}$ test to perform best. Fig. 1 bears out this prediction. Regardless of the value of c, $\mathcal{T}^{f*}_{NW}$ is significantly more powerful than the other tests considered. The double recursive test, $\mathcal{T}^{d*}_{NW}$, also displays significant power gains over the full sample $t^{2\star}_{\beta_1,NW}$ test, for all of the values of c considered. The rolling test, $\mathcal{T}^{r*}_{NW}$, displays a similar power profile to $\mathcal{T}^{d*}_{NW}$ for $c = 0$ and $c = 0.5T$, but is significantly less powerful than $\mathcal{T}^{d*}_{NW}$ for $c = 10$. The least powerful tests among those considered are the backward recursive test, as expected, and the full sample $t^{2\star}_{\beta_1,NW}$ test. To illustrate, the empirical power of $t^{2\star}_{\beta_1,NW}$ at $b_T = 50$ is approximately 50% for both $c = 0$ and $c = 10$ while for $\mathcal{T}^{f*}_{NW}$ it is around 75%. For $c = 0.5T$ and $b_T = 10$ the power of $t^{2}_{\beta_1,NW}$ is about 55% while that of $\mathcal{T}^{f*}_{NW}$ is in excess of 95%. In the latter example both the rolling ($\mathcal{T}^{r*}_{NW}$) and double recursive ($\mathcal{T}^{d*}_{NW}$) tests have power of approximately 80%.

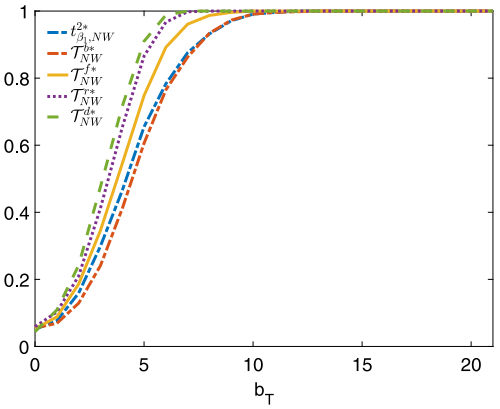
Consider next the results for Case 2, given in Fig. 2, where the pocket of predictability now occurs at the end of the sample. When x_t is weakly persistent the simulation DGP is approximately time-reversible and, as such, we would anticipate that all but the forward and backward recursive tests, whose relative behaviour would be expected to switch around, will behave similarly to how they behaved in Case 1 for the weakly dependent case. This is clearly seen to be the case in Fig. 2(c), with the backward recursive test now clearly the most powerful, the forward recursive test the least powerful, and the other tests all displaying almost identical power properties in Figs. 1(c) and 2(c). These patterns are also seen, albeit not as clearly, in a comparison of Figs. 1(b) and 2(b); the main difference being that the most of the tests (although not the double recursive test) tend to be slightly more powerful for $c = 10$ *vis-à-vis* $c = 0.5T$. The pattern of a general increase in power of the tests as c decreases for an end-of-sample pocket of predictability is very clearly continued in Fig. 2(a) for the case where $c = 0$ and x_t follows a pure unit root. Here, comparing with Fig. 1(a), we see that



(a) $c = 0, \phi = -0.90$



(b) $c = 10, \phi = -0.90$



(c) $c = 0.5T, \phi = -0.90$

Fig. 3. Finite sample local power: Case 3, $T = 250$.

all of the tests display considerably higher local power against an end-of-sample pocket of predictability than against a pocket of predictability at the start of the sample, and, comparing with Figs. 2(b) and 2(c), that the power of the tests is considerably higher than for $c = 10$ and $c = 0.5T$. A possible explanation for this improvement in power is the shape of the non-centrality term, $\int_{\tau_1}^{\tau_2} \tilde{Z}_{\tau_1,\tau_2}(s) b(s) J_{c,\mathrm{H}}(s) \mathrm{d}s$, entering the limiting distributions of the statistics under local alternatives in the case where x_t is strongly persistent. Clearly, end of sample predictability will be boosted from the larger magnitude of $J_{c,\mathrm{H}}(\tau)$ when τ is close to 1, and this will be most evident when $c = 0$. Interestingly, the full sample $t^{2\star}_{\beta_1,NW}$ test displays competitive power in Fig. 2(a) although it should be recalled from Table 2 that $t^{2\star}_{\beta_1,NW}$ is significantly over-sized in this case while the subsample tests are not.

Finally, the results in Fig. 3 pertain to Case 3, where the simulation DGP admits a window of predictability of size $\lfloor 2T/5 \rfloor$ within the sample. Here the double recursive test, $\mathcal{T}^{d*}_{NW}$, displays superior power to the other tests considered for both $c = 10$ and $c = 0.5T$ (Figs. 3(b) and 3(c) respectively), and is jointly most powerful along with the forward recursive $\mathcal{T}^{f*}_{NW}$ test for $c = 0$ (Fig. 3(a)). Notice also that for a given value of c, $\mathcal{T}^{d*}_{NW}$ displays considerably higher power under Case 3 than it does under both Cases 1 and 2. This is expected given that a larger window of predictive data is now present in the sample which the double recursive procedure is best able to exploit. Indeed, most of the tests considered display improved power performance compared to Figs. 1 and 2. This is particularly evident for the rolling test, $\mathcal{T}^{r*}_{NW}$, and again is to be expected given that a greater number of the subsample predictability statistics in the rolling sequence will contain data from a predictive period relative to the DGPs in Cases 1 and 2. For Case 3, the $\mathcal{T}^{b*}_{NW}$ test (as expected, given that the window of predictability begins early in the sample) and the full sample $t^{2}_{\beta_1,NW}$ test display the lowest power among the tests considered.

Table 3
Application to updated (Welch and Goyal, 2008) data: bivariate regressions — (1950:01–2017:12)

	$t^2_{\beta_1}$	$t^2_{\beta_1,NW}$	$\mathcal{T}^f$	$\mathcal{T}^f_{NW}$	$\mathcal{T}^b$	$\mathcal{T}^b_{NW}$	$\mathcal{T}^r$	$\mathcal{T}^r_{NW}$	$\mathcal{T}^d$	$\mathcal{T}^d_{NW}$	LM_x	$supF_x$
dp_{t-1}	0.472	0.457	10.088	**11.480**	5.608	6.885	6.882	8.280	10.284	12.519	**2.229**	**131.915**
	(0.481)	(0.492)	(0.121)	**(0.006)**	(0.340)	(0.182)	(0.729)	(0.287)	(0.998)	(0.194)	**(0.000)**	**(0.000)**
dy_{t-1}	0.581	0.568	**15.565**	**12.616**	6.241	7.849	11.143	9.318	10.252	**11.891**	**0.295**	**11.178**
	(0.414)	(0.408)	**(0.041)**	**(0.005)**	(0.252)	(0.133)	(0.507)	(0.182)	(0.072)	**(0.029)**	**(0.028)**	**(0.038)**
e/p_{t-1}	0.335	0.451	8.459	**9.189**	2.163	4.744	8.583	9.419	8.583	11.742	0.209	**38.229**
	(0.510)	(0.513)	(0.316)	**(0.043)**	(0.617)	(0.398)	(0.360)	(0.152)	(0.990)	(0.231)	(0.116)	**(0.000)**
de_{t-1}	0.291	0.490	**12.553**	**17.087**	0.291	0.490	13.399	**20.112**	15.145	**21.400**	0.192	5.031
	(0.594)	(0.575)	**(0.029)**	**(0.023)**	(0.852)	(0.851)	(0.058)	**(0.017)**	(0.099)	**(0.010)**	(0.210)	(0.355)
$rvol_{t-1}$	1.809	2.288	3.765	4.432	2.657	3.200	4.230	6.187	4.624	6.692	0.124	6.614
	(0.096)	(0.114)	(0.308)	(0.313)	(0.182)	(0.167)	(0.694)	(0.698)	(0.455)	(0.278)	(0.525)	(0.192)
bm_{t-1}	0.037	0.042	7.150	7.125	5.959	7.321	7.612	8.299	7.612	8.299	**0.342**	7.555
	(0.841)	(0.844)	(0.222)	(0.229)	(0.287)	(0.154)	(0.764)	(0.322)	(0.989)	(0.544)	**(0.012)**	(0.148)
$ntis_{t-1}$	0.041	0.059	5.634	5.235	1.648	2.600	9.383	10.543	9.679	10.874	0.375	8.102
	(0.830)	(0.821)	(0.180)	(0.312)	(0.623)	(0.546)	(0.074)	(0.059)	(0.101)	(0.114)	(0.070)	(0.180)
tbl_{t-1}	**7.001**	**9.408**	**11.763**	**15.989**	7.001	**9.408**	8.203	**11.618**	11.764	**16.588**	0.131	12.348
	(0.004)	**(0.004)**	**(0.003)**	**(0.001)**	(0.096)	**(0.040)**	(0.293)	**(0.035)**	(0.267)	**(0.029)**	(0.174)	(0.132)
lty_{t-1}	**4.178**	**5.524**	**11.036**	**13.410**	6.529	6.547	8.224	10.070	11.135	**14.308**	0.103	7.985
	(0.029)	**(0.026)**	**(0.013)**	**(0.009)**	(0.140)	(0.152)	(0.446)	(0.115)	(0.487)	**(0.044)**	(0.303)	(0.182)
ltr_{t-1}	**4.172**	**5.724**	**8.479**	10.313	4.438	6.021	8.145	9.702	8.902	10.709	0.163	6.407
	(0.045)	**(0.044)**	**(0.034)**	(0.055)	(0.136)	(0.167)	(0.082)	(0.115)	(0.061)	(0.094)	(0.341)	(0.325)
tms_{t-1}	2.726	3.075	**15.839**	**17.534**	4.769	5.133	12.142	**12.611**	**15.839**	**17.534**	0.350	**10.877**
	(0.084)	(0.084)	**(0.001)**	**(0.001)**	(0.145)	(0.150)	(0.061)	**(0.008)**	**(0.042)**	**(0.003)**	(0.060)	**(0.026)**
dfy_{t-1}	0.011	0.020	2.777	3.079	0.996	2.872	7.431	**19.169**	8.805	**19.473**	0.063	9.773
	(0.908)	(0.909)	(0.617)	(0.587)	(0.695)	(0.546)	(0.246)	**(0.044)**	(0.216)	**(0.046)**	(0.788)	(0.106)
dfr_{t-1}	0.768	1.535	4.487	5.475	2.407	5.398	11.218	7.685	11.218	13.141	0.156	6.134
	(0.477)	(0.434)	(0.475)	(0.416)	(0.336)	(0.346)	(0.228)	(0.521)	(0.295)	(0.336)	(0.537)	(0.377)
$infl_{t-1}$	3.042	**4.890**	9.083	**16.834**	3.999	6.138	11.705	11.516	12.195	**16.839**	0.326	**11.517**
	(0.070)	**(0.048)**	(0.276)	**(0.003)**	(0.269)	(0.155)	(0.477)	(0.060)	(0.722)	**(0.026)**	(0.102)	**(0.032)**

Notes: Numbers in parentheses are bootstrap p-values. Bold entries are those which are statistically significant at the 5% level (or stricter).

7. Empirical application

The data set used consists of monthly observations on the equity premium for the S&P Composite index calculated using CRSP's month-end values together with 14 different putative predictors, generically denoted x_t, and is taken from the updated monthly data set on Amit Goyal's website (`www.hec.unil.ch/agoyal/`) which is an extended version of the data set used by Welch and Goyal (2008). The data cover the period 1950:01–2017:12 ($T = 817$). We define the equity premium as in Goyal and Welch (2003) as the log return on the value-weighted CRSP stock market index minus the log return on the risk-free Treasury bill: $y_t = ep_t = \log(1 + R_{m,t}) - \log(1 + R_{f,t})$ where $R_{m,t}$ is the CRSP return and $R_{f,t}$ is the Treasury bill return. The variables are in log form (as in Goyal and Welch, 2003) and each of the predictors is lagged one period. A full list of the predictors together with graphs of the excess returns and the predictors can be found in the supplementary material.

Table 3 reports the outcome of the conventional IV-combination test from bivariate predictive regression models applied to the full sample of data. We report versions of the statistic using Eicker–White ($t^2_{\beta_1}$) and conventional ($t^2_{\beta_1,NW}$) standard errors. All of the IV-based test statistics computed in the empirical analysis follow the same specification as was used in the Monte Carlo experiments; that is, they are based on a combination of the IVX instrument, $z_{I,t-1}$ (as defined in (4.2), with $a = 1$ and $\gamma = 0.95$), and the sine instrument, $z_{II,t-1}$ (as defined in (4.1) with $k = 1$), with all of the observed variables and $z_{II,t-1}$ (but not $z_{I,t-1}$) entering the estimated predictive regressions demeaned, and with the finite-sample correction factor of Kostakis et al. (2015, p. 1516) implemented. Fixed regressor wild bootstrap p-values computed according to Algorithm 1 with 999 bootstrap replications are reported in parentheses. For most of the putative predictors considered, the results in Table 3 yield no statistically significant evidence of predictability. Exceptions are seen for the treasury bill rate (tbl_{t-1}), the long term government bond yield and rate of return series (lty_{t-1} and ltr_{t-1}, respectively), and inflation ($infl_{t-1}$) all of which are significant at the 5% level. Rejections of the null of no predictability are also seen at the 10% level for the term spread (tms_{t-1}) and the equity premium volatility ($rvol_{t-1}$) series.

To provide an insight into how stable the full sample predictive regressions are, Table 3 also reports the tests proposed in Georgiev et al. (2018) for the stability of the slope coefficient in the bivariate predictive regression of the equity premium on each (lagged) predictor. These tests are denoted LM_x and $supF_x$. The former is designed to test for the stability of the slope coefficient against a smoothly evolving slope change model and the latter against a one-time change in the slope. Bootstrap p-values calculated as outlined in Georgiev et al. (2018) for 999 bootstrap replications are reported in

parentheses. Significant rejections at the 5% level by at least one of these tests are observed for the predictive regressions involving the dividend price ratio (dp_{t-1}), dividend yield (dy_{t-1}), earnings price ratio (e/p_{t-1}), book to market ratio (bm_{t-1}), term spread (tms_{t-1}) and $infl_{t-1}$. The rejections seen for dp_{t-1} and e/p_{t-1} are particularly strong. A rejection at the 10% level is also seen for the net equity expansion ratio ($ntis_{t-1}$) predictor. Interestingly, for three of the four series (tbl_{t-1}, lty_{t-1} and ltr_{t-1}) for which the full sample IV-combination tests are significant at the 5% level these stability tests provide no evidence of structural instability in the slope coefficient.

To provide some additional insight into any time-varying behaviour present in the slope coefficients, Figs. 4 and 5 plot forward recursive and rolling IV (using the same choice of instruments as detailed above for the full sample IV-combination tests) slope estimates from the predictive regression of y_t on x_{t-1} and associated approximate 95% marginal confidence bands.[8] The warm-in fraction for the recursive sequence, τ_L, and the rolling window fraction, $\Delta\tau$, were both set at 1/4. In each case the horizontal axis dates correspond to the end of a given subsample. Commensurate with the results of the stability tests of Georgiev et al. (2018), these graphs highlight considerable time variation in the sequences of subsample slope estimates. A general pattern evident in Fig. 4 is a decline over time in the absolute value of the estimated slope coefficient with the recursive slope estimates generally tending to move closer to zero over time. This pattern can also be seen, albeit less clearly, in the rolling estimates in Fig. 5. This suggests that for some of these variables, any predictive ability they might have for the equity premium weakens over time. As a further heuristic device, rather than a formal statistical test, many of the graphs show some periods where the 95% marginal confidence intervals do not include zero, which is at least suggestive that pockets of predictability may be present in the data. Most of these episodes occur nearer the start of the data, such as, for example, with dy_{t-1}, but some are much longer lived as with, for example, the sequences of recursive estimates for tbl_{t-1}, ltr_{t-1}, tms_{t-1} and $infl_{t-1}$; recall that for tbl_{t-1}, ltr_{t-1} and $infl_{t-1}$, the full sample IV-combination tests gave significant rejections at the 5% level.

To pursue these findings further using statistically rigorous size-controlled methods, we next apply our proposed subsample-based predictability statistics. We report versions of the statistics using Eicker–White ($\mathcal{T}^f$, $\mathcal{T}^b$, $\mathcal{T}^r$ and $\mathcal{T}^d$) and conventional ($\mathcal{T}^f_{NW}$, $\mathcal{T}^b_{NW}$, $\mathcal{T}^r_{NW}$ and $\mathcal{T}^d_{NW}$) standard errors. Fixed regressor wild bootstrap p-values computed according to Algorithm 1 with 999 bootstrap replications are again reported in parentheses. In the computation of the forward and backward recursive statistics we set $\tau_L = 1/4$ and $\tau_U = 3/4$, respectively, while we set $\Delta\tau = 1/4$ for the rolling and double recursive statistics. The instruments used are as described above for the full sample statistics. Focusing on the forward recursive tests we see significant rejections at the 5% level (or stricter) of the null hypothesis of no predictability for each of dp_{t-1}, dy_{t-1}, e/p_{t-1}, de_{t-1}, tbl_{t-1}, lty_{t-1}, ltr_{t-1}, tms_{t-1} and $infl_{t-1}$; indeed, in many cases these rejections are also significant at the 1% level. While these rejections tally with those delivered by the full sample test for tbl_{t-1}, lty_{t-1}, ltr_{t-1} and $infl_{t-1}$, for the other series, all of which (other than de_{t-1}) fail the structural stability tests of Georgiev et al. (2018), these are series for which the full sample tests delivered no significant evidence of predictability. With the exception of dp_{t-1} and e/p_{t-1}, those series for which $\mathcal{T}^f_{NW}$ delivers a rejection at the 5% significance level also show rejections at the 5% level for at least one of the other subsample maximum tests reported. Additional evidence of temporary predictability at the 5% level (or stricter) is provided for dfy_{t-1} by both $\mathcal{T}^r_{NW}$ and $\mathcal{T}^d_{NW}$ (notice that for this series the $supF_x$ test is in fact very close to giving a rejection at the 10% level). A significant rejection at the 10% level is also provided for $ntis_{t-1}$ by the $\mathcal{T}^r_{NW}$ test.

To gain further insight, Fig. 6 graphs the forward recursive sequences of $t_{\beta_1}(\tau_1, \tau_2)$ subsample statistics for each case where a rejection at the 5% level is observed for the corresponding maximum test.[9] Also reported on these graphs are the 5% and 10% bootstrap critical values for the null distribution of the maximum statistic in the sequence, together with the 5% and 10% critical values from the χ^2_1 distribution (the marginal critical values which apply for any given subsample).

Consider first the graph in part (a) of Fig. 6 for the dividend price ratio, dp_{t-1}. Looking at the time path of the forward recursive subsample statistic we can see that for much of the first half of the sequence (up until roughly the early 1980s) the statistic exceeds the χ^2_1 5% critical value, suggesting that running the IV-combination test on any subsample of the data selected up until this point would have delivered a significant rejection at the (marginal) 5% level. After this sample endpoint no significant evidence of predictability would have been found. We can also see that a large number of exceedances of the 10% bootstrap critical value for the maximum are seen in the early part of the data, with exceedances of the 5% bootstrap critical value also seen, most notably in the mid 1970s. These results are suggestive that a pocket of predictability for returns existed for the predictor dp_{t-1} in the 1970s with peak predictability seen in the middle of that decade, and that since the 1980s onwards predictability appears to have evaporated. For the dividend yield, dy_{t-1}, a pocket of predictability appears to be present again from the early 1970s but lasting much longer, and with apparently stronger magnitude, displaying many more contiguous exceedances of the bootstrap critical values for the maximum than were seen for dp_{t-1}; indeed, here predictability appears to run until the early to mid 1990s. From the mid to late 1990s onwards the evidence for predictability disappears. Evidence for both the earnings price ratio, e/p_{t-1}, in part (c) and the dividend pay out ratio, de_{t-1}, in part (d) is less strong than for the previous two series (reflected in the considerably

[8] Denoting the IV slope estimate as $\hat{\beta}_1$, the confidence bands were computed as $\hat{\beta}_1 \pm 1.96 se(\hat{\beta}_1)$, where $se(\hat{\beta}_1)$ are the associated IV Eicker–White standard errors. These confidence bands should, however, be treated with caution as they are not joint 95% confidence bands for the entire sequence of slope estimates, but rather represent the marginal 95% confidence band at each point in the sequences of estimated slope coefficients.

[9] Where both the maximum tests based on Eicker–White and conventional standard errors reject we report the version with the smallest p-value; cf. Table 3. Corresponding graphs for the rolling sequences are available on request.

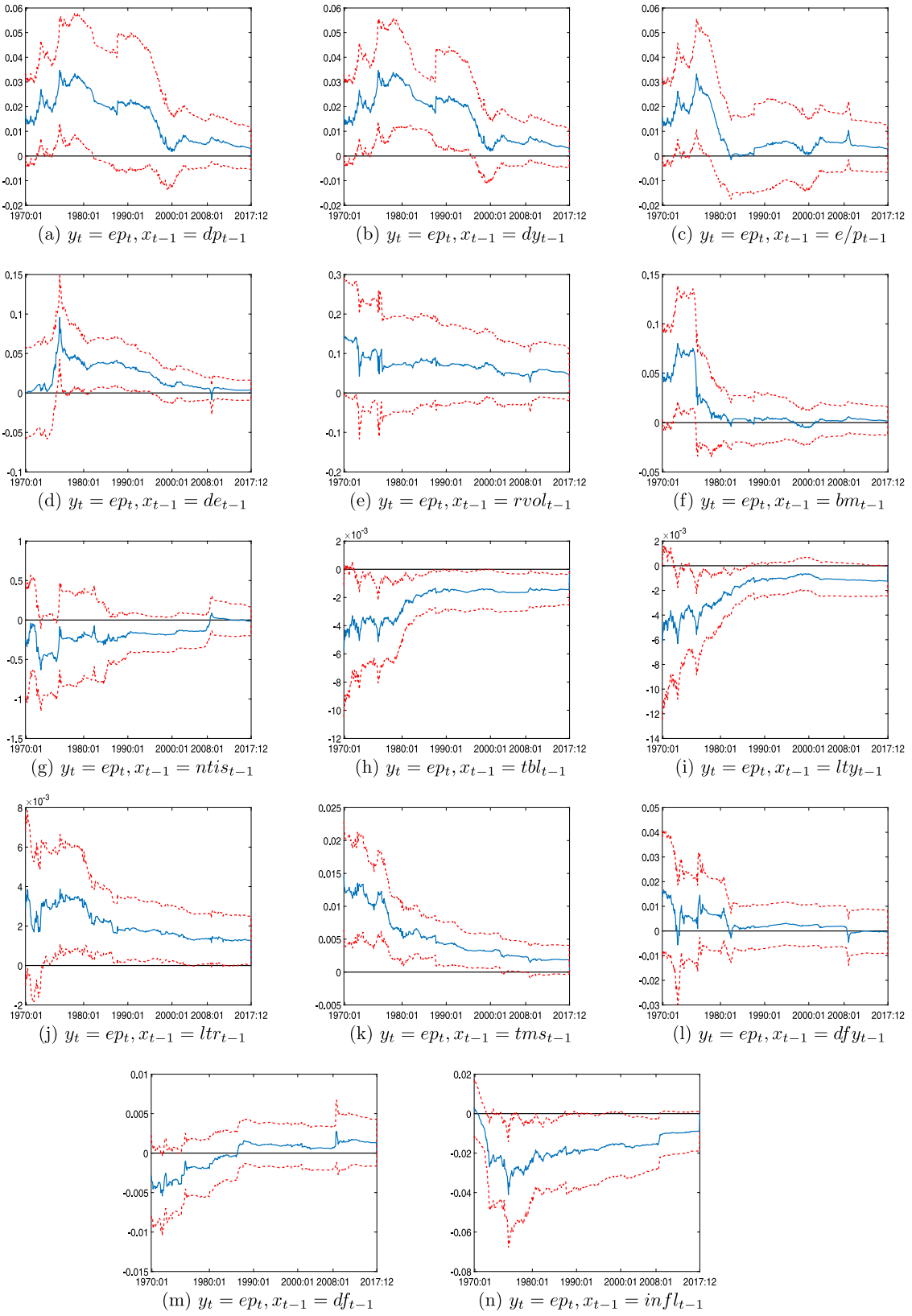
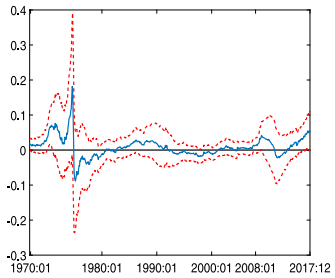
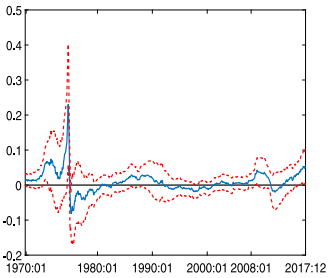


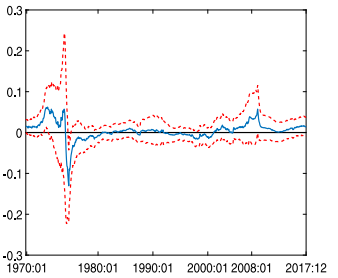
Fig. 4. Forward recursive slope estimates (solid line) and 95% confidence bands (dotted lines). Sample period 1950:01–2017:12. The y_t and x_t variable labels are as defined in the main text.



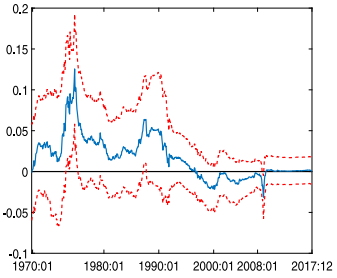
(a) $y_t = ep_t, x_{t-1} = dp_{t-1}$



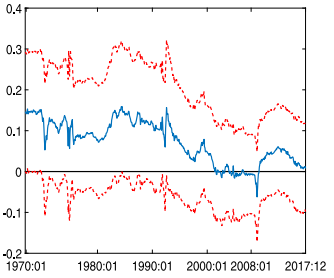
(b) $y_t = ep_t, x_{t-1} = dy_{t-1}$



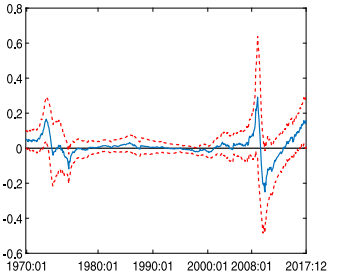
(c) $y_t = ep_t, x_{t-1} = e/p_{t-1}$



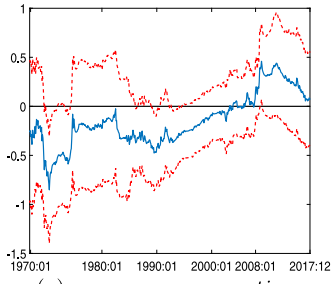
(d) $y_t = ep_t, x_{t-1} = de_{t-1}$



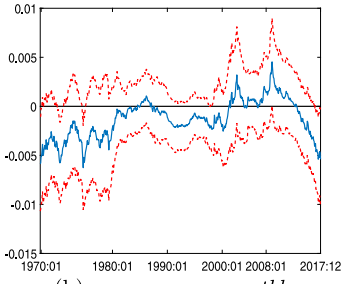
(e) $y_t = ep_t, x_{t-1} = rvol_{t-1}$



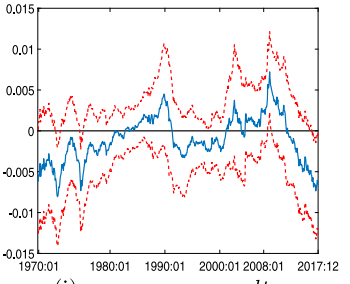
(f) $y_t = ep_t, x_{t-1} = bm_{t-1}$



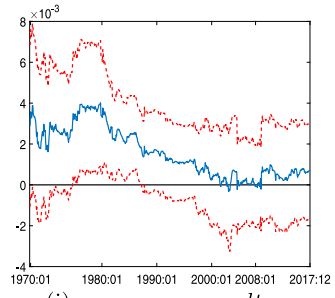
(g) $y_t = ep_t, x_{t-1} = ntis_{t-1}$



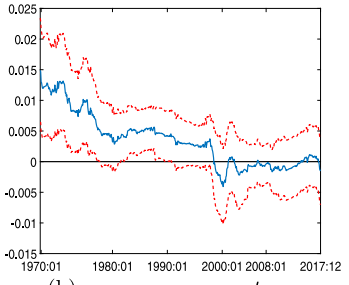
(h) $y_t = ep_t, x_{t-1} = tbl_{t-1}$



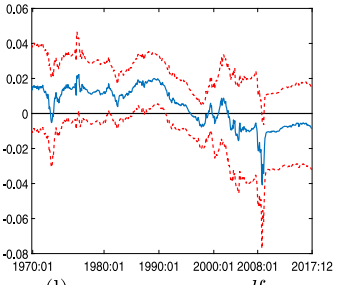
(i) $y_t = ep_t, x_{t-1} = lty_{t-1}$



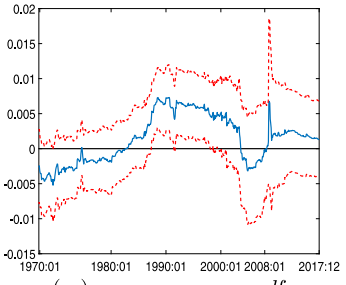
(j) $y_t = ep_t, x_{t-1} = ltr_{t-1}$



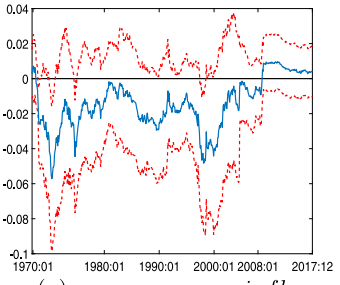
(k) $y_t = ep_t, x_{t-1} = tms_{t-1}$



(l) $y_t = ep_t, x_{t-1} = df y_{t-1}$



(m) $y_t = ep_t, x_{t-1} = df_{t-1}$



(n) $y_t = ep_t, x_{t-1} = infl_{t-1}$

Fig. 5. Rolling slope estimates (solid line) and 95% confidence bands (dotted lines). Sample period 1950:01–2017:12. The y_t and x_t variable labels are as defined in the main text.

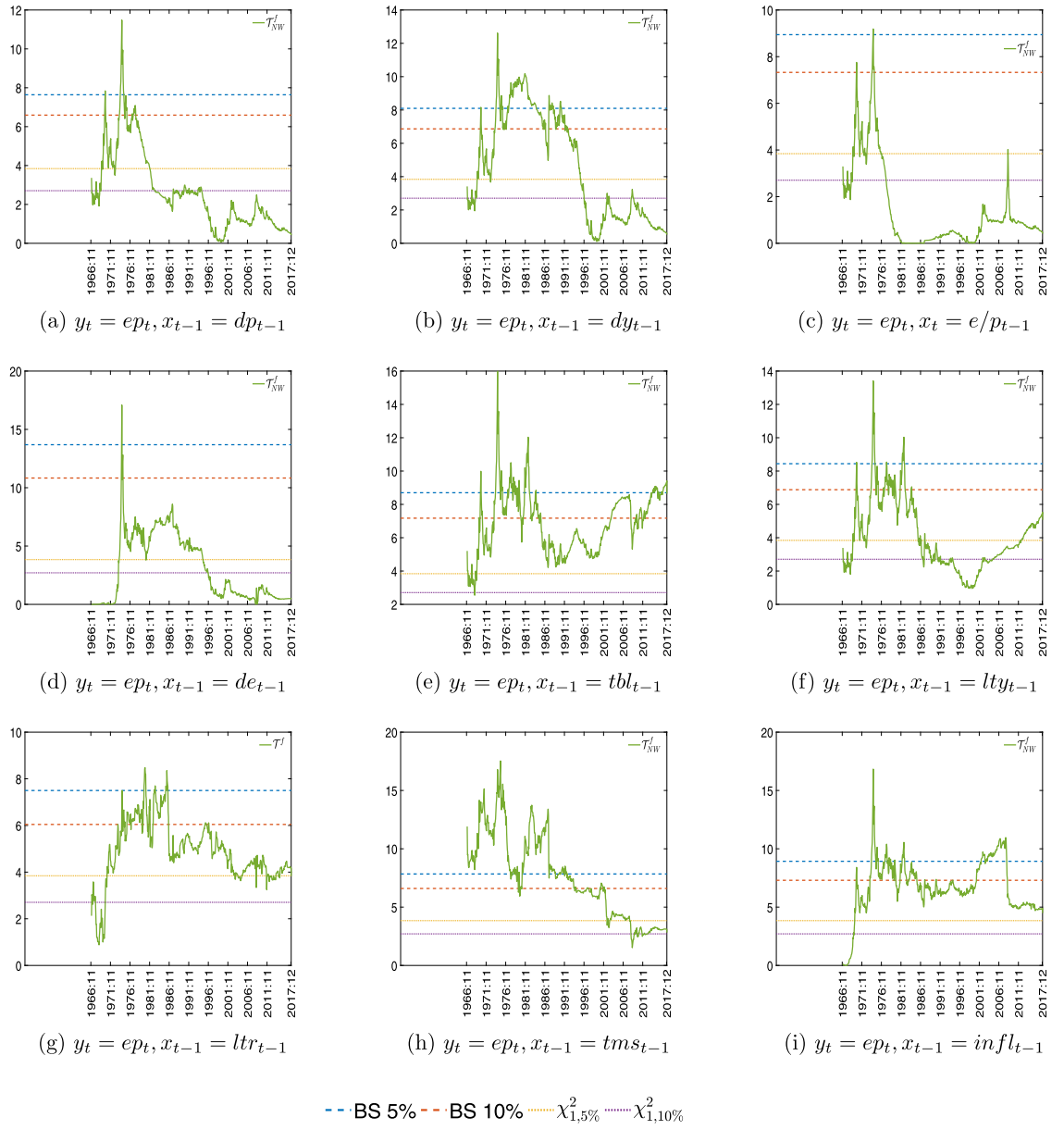


Fig. 6. Plots of forward recursive subsample statistics with marginal and bootstrap 10% and 5% critical values.

larger *p*-values for the maximum statistics for those series in Table 3), but again the period of predictability appears to be concentrated in mid 1970s. For both the treasury bill rate, tbl_{t-1}, in part (e) and the long term bond yield, lty_{t-1}, in part (f) there appears to be evidence of predictability across a window from the early 1970s until the mid 1980s, albeit the strength of predictability appears to waver somewhat over this period, particularly so for lty_{t-1}. For both of these series, there is also evidence that predictability is re-emerging from around the period of the recent financial crisis onwards, most notably so for tbl_{t-1} where a number of exceedances of the bootstrap critical values occur. In the case of tbl_{t-1} running the IV-combination test on almost any subsample of the data would yield a rejection at the 5% using the marginal χ^2_1 critical value. This observation is also true for the long term rate, ltr_{t-1}, in part (g) and for inflation, $infl_{t-1}$, in part (i). Recall that these are the three series for which the full sample IV-combination tests gave significant rejections at the 5% level. Finally for the tms_{t-1} series in part (h) predictability appears evident and consistently strong up until the mid 1990s after which the magnitude of predictability starts to tail off and then falls markedly around the time of the financial crisis onwards. In contrast, the full sample tests reveal no significant evidence (at the 5% level) of predictability from tms_{t-1}.

These examples highlight the advantage of considering the recursive sequence of statistics and their evolution through time rather than just full sample IV-combination tests, with much stronger evidence for predictability earlier in the sample than later for a number of the predictors considered.

8. Conclusions

Recent research has suggested that should stock returns be predictable, then this is likely to be a temporary phenomenon. Our motivation has been to develop tests with good power to detect such episodes. To avoid the problem of endogenously-determined sample splits, our proposed tests are derived from sequences of predictability statistics calculated over systematic subsamples of the data. The tests are based on the maxima of the instrumental variable-based predictability statistics of Breitung and Demetrescu (2015) taken across sequences of forward and backward recursive, rolling, and double-recursive predictive regressions. The limiting distributions of these statistics were shown to depend both on any heteroskedasticity present and on whether the putative predictor follows a near-integrated or weakly dependent process. To account for these dependencies, fixed regressor wild bootstrap implementations of the tests were proposed and shown to be first-order asymptotically valid. Monte Carlo simulation demonstrated that the tests display decent finite sample size control, and can be considerably more powerful in detecting temporary predictability than full sample tests. An empirical application to a well-known US monthly stock returns data set highlighted the ability of the new tests to detect predictability within the data where full sample tests could not.

We conclude with two suggestions for further research. First, we have focussed on tests based on subsample implementations of the IV-combination statistics of Breitung and Demetrescu (2015) which use two instruments per predictor. It should be possible to apply the same approach to subsample implementations of statistics which use only one instrument, such as the statistics considered in section 2.2 of Breitung and Demetrescu (2015) or the IVX statistic of Kostakis et al. (2015). Second, our proposed tests are based on an approach which assumes a linear predictive regression model with a constant slope parameter within the given subsample window and then bases a test on the fluctuations seen in the sequence of such statistics over a range of subsamples. As such, this approach is ambivalent about the true form of any time-variation present in the slope parameter and so would be expected to have reasonable power against a wide range of patterns of time-variation in the slope parameter, including those generated by threshold or other non-linear DGPs. LM-type tests could be developed based on an assumed non-linear model for the time-variation in the slope and would be expected to be more powerful than the tests developed here where this assumed model coincided with, or was at least a close approximation to, the true (unknown) DGP, but would likely have much lower power if the true DGP was not well approximated by the model.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.01.001.

References

Amihud, Y., Hurvich, C.M., 2004. Predictive regressions: A reduced-bias estimation method. J. Financ. Quant. Anal. 39, 813–841.
Ang, A., Bekaert, G., 2007. Stock return predictability: Is it there? Rev. Financ. Stud. 20, 651–707.
Banerjee, A., Lumsdaine, R.L., Stock, J.H., 1992. Recursive and sequential tests of the unit-root and trend-break hypotheses: theory and international evidence. J. Bus. Econom. Statist. 10, 271–287.
Barberis, N., 2000. Investing for the long run when returns are predictable. J. Finance 55, 225–264.
Boswijk, H.P., Cavaliere, G., Rahbek, A., Taylor, A.M.R., 2016. Inference on co-integration parameters in heteroskedastic vector autoregressions. J. Econometrics 192, 64–85.
Boudoukh, J.R., Michaely, M., Richardson, P., Roberts, M.R., 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. J. Finance 62, 877–915.
Breitung, J., Demetrescu, M., 2015. Instrumental variable and variable addition based inference in predictive regressions. J. Econometrics 187, 358–375.
Campbell, J.Y., 1987. Stock returns and the term structure. J. Financ. Econom. 18, 373–400.
Campbell, J.Y., 2008. Viewpoint: Estimating the equity premium. Canad. J. Econ. 41, 1–21.
Campbell, J.Y., Shiller, R.J., 1988a. Stock prices, earnings, and expected dividends. J. Finance 43, 661–676.
Campbell, J., Shiller, R.J., 1988b. The dividend-price ratio and expectations of future dividends and discount factors. Rev. Financ. Stud. 1, 195–228.
Campbell, J.Y., Yogo, M., 2006. Efficient tests of stock return predictability. J. Financ. Econom. 81, 27–60.
Cavanagh, C.L., Elliott, G., Stock, J.H., 1995. Inference in models with nearly integrated regressors. Econometric Theory 11, 1131–1147.
Dangl, T., Halling, M., 2012. Predictive regressions with time-varying coefficients. J. Financ. Econom. 106, 157–181.
Elliott, G., Müller, U.K., 2006. Efficient tests for general persistent time variation in regression coefficients. Rev. Econom. Stud. 73, 907–940.
Elliott, G., Müller, U.K., Watson, M.W., 2015. Nearly optimal tests when a nuisance parameter is present under the null hypothesis. Econometrica 83, 771–811.
Elliott, G., Stock, J.H., 1994. Inference in time series regression when the order of integration of a regressor is unknown. Econometric Theory 10, 672–700.
Fama, E.F., 1981. Stock returns, real activity, inflation, and money. Amer. Econ. Rev. 71, 545–565.
Fama, E.F., 1990. Stock returns, expected returns, and real activity. J. Finance 45, 1089–1108.
Fama, E.F., French, K.R., 1988. Dividend yields and expected stock returns. J. Financ. Econom. 22, 3–24.
Fama, E.F., French, K.R., 1989. Business conditions and expected returns on stocks and bonds. J. Financ. Econom. 25, 23–49.
Farmer, L., Schmidt, L., Timmermann, A.G., 2018. Pockets of predictability. CEPR Discussion Paper, DP12885.
Gargano, A., Pettenuzzo, D., Timmermann, A., 2019. Bond return predictability: economic value and links to the macroeconomy. Manage. Sci. 65, 508–540, https://doi.org/10.1287/mnsc.2017.2829.

Georgiev, I., Harvey, D.I., Leybourne, S.J., Taylor, A.M.R., 2018. Testing for parameter instability in predictive regression models. J. Econometrics 204, 101–118.
Georgiev, I., Harvey, D.I., Leybourne, S.J., Taylor, A.M.R., 2019. A bootstrap stationarity test for predictive regression invalidity. J. Bus. Econom. Statist. 37 (3), 528–541. http://dx.doi.org/10.1080/07350015.2017.1385467.
Giannetti, A., 2007. The short term predictive ability of earnings-price ratios: The recent evidence (1994-2003). Q. Rev. Econ. Finance 47, 26–39.
Gonzalo, J., Pitarakis, J.-Y., 2012. Regime-specific predictability in predictive regressions. J. Bus. Econom. Statist. 30, 229–241.
Gorodnichenko, Y., Mikusheva, A., Ng, S., 2012. Estimators for persistent and possibly non-stationary data with classical properties. Econometric Theory 28, 1003–1036.
Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Manage. Sci. 49, 639–654.
Henkel, S.J., Martin, J.S., Nardari, F., 2011. Time-varying short-horizon predictability. J. Financ. Econom. 99, 560–580.
Inoue, A., Rossi, B., 2005. Recursive predictability tests for real-time data. J. Bus. Econom. Statist. 23, 336–345.
Jansson, M., Moreira, M.J., 2006. Optimal inference in regression models with nearly integrated regressors. Econometrica 74, 681–714.
Keim, D.B., Stambaugh, R.F., 1986. Predicting returns in the stock and bond markets. J. Financ. Econom. 17, 357–390.
Kolev, G.I., Karapandza, R., 2017. Out-of-sample equity premium predictability and sample-split invariance. J. Bank. Financ. 34, 188–201.
Kostakis, A., Magdalinos, T., Stamatogiannis, M.P., 2015. Robust econometric inference for stock return predictability. Rev. Financ. Stud. 28, 1506–1553.
Kuan, C.-M., Hornik, K., 1995. The generalized fluctuation test: A unifying view. Econometric Rev. 14, 135–161.
Lettau, M., Ludvigsson, S., 2001. Consumption, aggregate wealth, and expected stock returns. J. Finance 56, 815–850.
Lettau, M., van Nieuwerburgh, S., 2008. Reconciling the return predictability evidence. Rev. Financ. Stud. 21, 1607–1652.
MacKinnon, J.G., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J. Econometrics 29, 305–325.
Nelson, C.R., Kim, M.J., 1993. Predictable stock returns: The role of small sample bias. J. Finance 48, 641–661.
Nicolau, J., Rodrigues, P.M.M., 2019. A new regression-based tail index estimator. Rev. Econ. Stat. 101 (4), 667–680.
Pástor, L., Stambaugh, R.F., 2009. Predictive systems: Living with imperfect predictors. J. Finance 64, 1583–1628.
Pástor, L., Stambaugh, R.F., 2012. Are stocks really less volatile in the long run? J. Finance 67, 431–478.
Paye, B.S., Timmermann, A., 2006. Instability of return prediction models. J. Empir. Financ. 13, 274–315.
Pesaran, M.H., Timmermann, A., 2002. Market timing and return prediction under model instability. J. Empir. Financ. 9, 495–510.
Pettenuzzo, D., Timmermann, A., 2011. Predictability of stock returns and asset allocation under structural breaks. J. Econometrics 164, 60–78.
Phillips, P.C.B., 1998. New tools for understanding spurious regressions. Econometrica 66, 1299–1325.
Phillips, P.C.B., 2015. Pitfalls and possibilities in predictive regression. J. Financ. Econom. 13, 521–555, 1990.
Phillips, P.C.B., Lee, J.H., 2013. Predictive regression under various degrees of persistence and robust long-horizon regression. J. Econometrics 177, 250–264.
Phillips, P.C.B., Magdalinos, T., 2009. Econometric Inference in the Vicinity of Unity. Working Paper. Singapore Management University.
Rapach, D.E., Wohar, M., 2006. Structural breaks and predictive regression models of aggregate U.S. stock returns. J. Financ. Econom. 4, 238–274.
Stambaugh, R.F., 1999. Predictive regressions. J. Financ. Econom. 54, 375–421.
Timmermann, A., 2008. Elusive return predictability. Int. J. Forecast. 24, 1–18.
Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21, 1455–1508.