2020-13

Working paper. Economics

ISSN 2340-5031

OUT OF SAMPLE PREDICTABILITY IN PREDICTIVE REGRESSIONS WITH MANY PREDICTOR CANDIDATES

Jesús Gonzalo and Jean-Yves Pitarakis

uc3m | Universidad Carlos III de Madrid

Serie disponible en http://hdl.handle.net/10016/11

Web: http://economia.uc3m.es/

Correo electrónico: departamento.economia@eco.uc3m.es

Out of Sample Predictability in Predictive Regressions with Many Predictor Candidates

Jesús Gonzalo
Universidad Carlos III de Madrid
Department of Economics
jesus.gonzalo@uc3m.es

Jean-Yves Pitarakis
University of Southampton
Department of Economics
j.pitarakis@soton.ac.uk

October 18, 2020

Abstract

This paper is concerned with detecting the presence of out of sample predictability in linear predictive regressions with a potentially large set of candidate predictors. We propose a procedure based on out of sample MSE comparisons that is implemented in a pairwise manner using one predictor at a time and resulting in an aggregate test statistic that is standard normally distributed under the null hypothesis of no linear predictability. Predictors can be highly persistent, purely stationary or a combination of both. Upon rejection of the null hypothesis we subsequently introduce a predictor screening procedure designed to identify the most active predictors.

Keywords: Forecasting, Predictive Regressions, High Dimensional Predictability.

JEL: C12, C32, C52, C53.

1. Introduction

A vast body of research over the past decade has been concerned with developing estimation and forecasting techniques that can accommodate the availability of large datasets. An important objective driving this literature is to obtain accurate out of sample predictions of a response variable via suitable covariate screening and model selection techniques. These typically involve a vast amount of nested model permutations and predictor configurations that may often exceed the available sample size and are typically handled via shrinkage methods that can simultaneously achieve regularization and variable selection objectives in data rich environments. Popular examples include the LASSO, Ridge and their numerous variants (see Mullainathan and Spiess (2017) for an overview of these methods in the context of economic applications).

The detection of predictability within linear regression settings has also been the subject of extensive research in the econometrics literature. The broadly labelled topic of predictive regressions for instance has become an important field of research in its own right due to the specificities associated with economic data and the complications that these may cause for estimation and inference (e.g. persistent nature of many financial and economic predictors, endogeneity, low signal to noise ratios, imbalance in the persistence properties of predictand and predictors). Unlike the above mentioned statistical literature however, predictive regressions as explored in econometrics have been mainly concerned with in-sample significance testing in single predictor environments (see Gonzalo and Pitarakis (2019) and references therein for an overview of this literature).

Our objective in this paper is to consider this predictive regression environment as commonly explored in the econometrics literature and propose a method for detecting the potential presence of out-of-sample linear predictability when the latter is induced by one or more predictors from a potentially large pool of candidate predictors. These predictors could be purely stationary or highly persistent without affecting the validity of our proposed approach and without the need for the investigator to have knowledge of these properties. Our environment is that of a potentially large number of nested specifications that also include an intercept only model which we view as the benchmark model or the maintained

theory. More specifically, we focus our attention on testing this benchmark specification against the alternative hypothesis that at least one of the predictors under consideration is active in the sense of improving out of sample MSEs relative to the benchmark.

The approach introduced in this paper is able to accommodate a large number of predictors as it relies on multiple pairwise comparisons of the benchmark model with a larger model that includes solely one predictor at a time. These pairwise MSE comparisons are implemented via the repeated evaluation of a test statistic suitable for out of sample predictive accuracy comparisons in nested environments. The resulting series of test statistics are subsequently reassembled into a single aggregate statistic allowing us to test the null of no predictability against the alternative that at least one of the predictors is active.

Upon rejection of the benchmark model the important question as to which predictors are the most important drivers of predictability also arises. To address this question we subsequently introduce a covariate screening method that allows us to identify the key predictor that most improves the accuracy of forecasts of the response variable relative to the forecasts based on the benchmark model.

Our operating environment is particularly relevant to economic and financial applications where one is interested in the maintained hypothesis of no predictability whereby the response variable of interest is best described by a martingale difference process (e.g. excess stock returns, currency returns, consumption growth amongst many others). Suppose for instance that a researcher wishes to assess the presence of predictability in stock returns with the well known Goyal and Welch predictors consisting mostly of highly persistent series (e.g. valuation ratios, interest rates) combined with a variety of macroeconomic indicators. The methods developed in this paper are designed to explore such issues in a particularly simple and computationally feasible way. Recent applications across a variety of fields and that are relevant to our setting include Molodotsova and Papell (2015), Li Tsiakas and Wang (2015), Jacobsen, Jiang and Zhang (2019), Rapach, Strauss, Tu and Zhou (2019) amongst numerous others.

The plan of the paper is as follows. Section 2 introduces our modelling environment and the key test statistics that are used to implement our predictive accuracy comparisons. Section 3 develops the asymptotic theory under the benchmark specification followed by a comprehensive local power analysis. Section 4 is concerned with the detection of the most

relevant predictors upon rejection of the benchmark model. Section 5 demonstrates the finite sample properties of our methods through a comprehensive simulation based exercise and Section 6 concludes. All proofs are relegated to the appendix.

2. Models and Theory

Let $\{y_t\}$ denote a scalar random process. Given a sample of size $t = 1, \ldots, n$, we wish to assess the presence of linear one-step ahead predictability in y_t. If present, predictability is induced by at least one predictor from a finite pool of p predictors $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$. Predictability is understood to be present whenever an intercept only benchmark model is rejected in favour of a larger model on the basis of out-of-sample MSE based comparisons. Thus the generic framework within which we operate is given by the predictive regressions

$$y_{t+1} = \theta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_{t+1} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and u_t is a random disturbance term. For later use we also define $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\beta}')'$ and $\boldsymbol{w}_t = (1, \boldsymbol{x}_t')'$ so that (1) can equivalently be expressed as $y_{t+1} = \boldsymbol{\theta}'\boldsymbol{w}_t + u_{t+1}$. The predictors collected in $\boldsymbol{x}_t$ can accommodate candidates with different degrees of persistence as commonly encountered in economic applications (e.g. highly persistent versus less persistent). Nevertheless our approach does not rely on any knowledge of the persistence properties of the pool of predictors available to the investigator.

We view (1) as encompassing a family of nested linear predictive regressions including the benchmark specification given by

$$y_{t+1} = \theta_0 + u_{t+1}. \tag{2}$$

Given the above framework the main goal of this paper is to address the following questions. Suppose a researcher has access to a pool of predictors collected within $\boldsymbol{x}_t$. Is at least one of these predictors active relative to the benchmark model in (2)? In the affirmative, is it possible to identify which one of the p predictors has the strongest influence in the sense of improving forecast accuracy the most relative to the benchmark?

To formalise our environment we let $\hat{y}_{0,t+1|t}$ denote the one-step ahead forecasts of y_{t+1} obtained from the benchmark model in (2) and $\hat{y}_{j,t+1|t}$, $j = 1, \ldots, p$, the one-step ahead forecasts of y_{t+1} obtained from (1) using one predictor at a time from the available collection of p predictors and inclusive of a fitted intercept. The corresponding forecast errors are $\hat{e}_{0,t+1|t} = y_{t+1} - \hat{y}_{0,t+1|t}$ and $\hat{e}_{j,t+1|t} = y_{t+1} - \hat{y}_{j,t+1|t}$. Out of sample forecasts are constructed recursively with an expanding window approach. We estimate each predictive regression via recursive least-squares starting from an initial window of size $t = 1, \ldots, k_0$ and progressively expanding the estimation window up to $n - 1$. Throughout this paper k_0 is taken to be a given a fraction π_0 of the sample size and we write $k_0 = [n\pi_0]$ for some $\pi_0 \in (0, 1)$. Under the benchmark model we have $\hat{\theta}_{0t} = \sum_{s=1}^{t} y_s/t$ leading to the unconditional mean forecasts $\hat{y}_{0,t+1|t} = \hat{\theta}_{0t}$. Under the larger models estimated with an intercept and one predictor at a time we have $\hat{\theta}_{jt} = (\sum_{s=1}^{t} \widetilde{w}_{j,s-1}\widetilde{w}'_{j,s-1})^{-1} \sum_{s=1}^{t} \widetilde{w}_{j,s-1} y_s$ for $\widetilde{w}_{jt} = (1, w_{jt})$ and $w_{jt} \in \{x_{1t}, \ldots, x_{pt}\}$ with forecasts obtained as $\hat{y}_{j,t+1|t} = \hat{\theta}'_{jt}\widetilde{w}_{jt}$ for $t = k_0, \ldots, n - 1$.

At the end of this pseudo out-of-sample exercise we obtain the $p + 1$ sequences of forecast errors $\{\hat{e}_{0,t+1|t}\}_{t=k_0}^{n-1}$ and $\{\hat{e}_{j,t+1|t}\}_{t=k_0}^{n-1}$ which form the basis of our inferences. Throughout this paper the maintained null hypothesis is that the population MSEs of the benchmark model and the larger models are equal in the sense that $\boldsymbol{\beta} = 0$ in (1) or equivalently model (2) holds. The alternative of interest is that there is at least one active predictor w_{jt} in the sense that $E[\hat{e}^2_{0,t+1|t} - \hat{e}^2_{j,t+1|t}] > 0$ for at least one $j \in \{1, 2, \ldots, p\}$.

Addressing the two questions stated above raises three key challenges. The first one arises from the fact that we wish to conduct out of sample predictive accuracy comparisons in a nested setting (e.g. intercept only model versus single predictor specifications), rendering traditional sample MSE comparisons ineffective as under the null hypothesis of equal predictive accuracy all forecast errors under consideration will be asymptotically identical, leading to normalised sample MSE spreads identically equal to zero in the limit (and similarly for their variances). The second challenge is a dimensionality related complication as we wish our method to be computationally feasible to implement despite the availability of a potentially large pool of predictors. The third challenge has to do with the identification of active predictors upon rejection of the null hypothesis. Although numerous covariate screening procedures have been developed in the statistics literature (e.g. sure independence screening) the validity of most of these relies on assumptions that are not tenable in our time series

environment.

The issue of predictive performance testing in nested environments has attracted considerable attention in the forecasting literature following the observation that Diebold-Mariano (DM) type constructions (Diebold and Mariano (1995), West (1996)) are not suitable since under the null hypothesis of equal predictive ability the pair of models being compared become identical in the limit. Consequently suitably normalised sample MSE spreads and their variance both converge to zero asymptotically resulting in statistics with ill-defined limits. In the context of predictive regressions this problem has been addressed through the use of alternative normalisations of sample MSEs, resulting in test statistics with well defined but non-standard limits requiring bootstrap based approaches (see McCracken (2007), West (2006), Clark and McCracken (2013) and references therein). More recently, alternative solutions involving modifications to DM type statistics that result in conventional standard normal asymptotics regardless of the nested nature of competing models have also been developed in Pitarakis (2020). These are similar in spirit to the way Vuong type model selection tests (Vuong (1989)) have been recently adapted to accommodate both nested and non-nested environments via sample splitting and related approaches (see Schennah (2017), Hsu and Shi (2017)). This is also the route we will continue to take in this paper by introducing a novel DM type test statistic for the initial pairwise model comparisons between the benchmark model in (2) and the p larger models containing one predictor at a time. The formulation of our test statistic relies on the same principles as the statistics introduced in Pitarakis (2020) but does not involve any discarding of sample information. The resulting DM type statistics associated with the p pairwise model comparisons are subsequently reassembled into an aggregate statistic designed to test whether at least one of the p predictors is active in predicting y_{t+1}.

The idea of considering one predictor at a time makes the practical implementation of our approach trivial regardless of the size of the pool of predictor candidates and is here justified by the fact that our null hypothesis is given by the benchmark model in (2). This is very much in the spirit of Ghysels et al. (2020) where the authors developed a procedure for testing the statistical significance of a large number of predictors through functionals (e.g. maximum) of multiple individual t-statistics obtained from models estimated with one regressor at a time and a benchmark model with none of the explanatory variables included.

An important advantage unique to our own setting however is the fact that each of the pairwise DM type statistics will have identical limits under our null hypothesis, making the exercise of constructing an aggregate statistic trivial. The average of these individual statistics will by construction also have the same limiting distribution. Differently put we do not need to be concerned with the behaviour of the covariances of the p individual test statistics and the nested nature of our setting is here used to our advantage.

Before proceeding further it is also useful to mention the recent but already extensive literature on screening for relevant predictors in high dimensional settings and which is related to our second concern of identifying dominant predictors upon rejection of the benchmark model in (2). In this context, a particularly popular approach has been based on ranking marginal correlations via marginal linear regressions (see Fan and Lv (2008), Tang, Wang and Barut (2017), McKeague and Qian (2015) amongst numerous others). In McKeague and Qian (2016) for instance the authors developed a test for the presence of at least one significant predictor via a maximum correlation type of approach between each predictor and predictand. Within our own context and upon rejection of the benchmark model we aim to identify at least one of the active predictors among the pool of p predictors using the above mentioned aggregate test statistic instead. An important aspect accommodated by our framework is the possibility that the pool of predictors contains dependent series with different persistence properties in addition to being possibly correlated (following a VAR process for instance) as it is the norm with economic data.

We now introduce and motivate the DM type test statistic which will be used to conduct pairwise predictive performance comparisons between the benchmark model and each of the p marginal predictive regression. Recalling that the key complication arising from the underlying nestedness of models is that in the limit $\hat{e}^2_{0,t+1|t}$ and $\hat{e}^2_{j,t+1|t}$ will be identical under the null hypothesis, we split the evaluation of the forecast errors associated with the benchmark model across two subsamples of size m_0 and $n-k_0-m_0$ respectively. This *double counted* sample MSE of the benchmark model can then be compared with twice the sample MSE of the larger model via the following base formulation

$$\mathcal{D}_n(m_0,j)=\frac{\sqrt{n-k_0}}{\hat{\omega}_n(m_0)}\left(\left(\frac{\sum_{t=k_0}^{k_0+m_0-1}\hat{e}^2_{0,t+1}}{m_0}+\frac{\sum_{t=k_0+m_0}^{n-1}\hat{e}^2_{0,t+1}}{n-k_0-m_0}\right)-2\frac{\sum_{t=k_0}^{n-1}\hat{e}^2_{j,t+1}}{n-k_0}\right). \quad (3)$$

Here we take m_0 to be a user-defined parameter and express it as a fraction μ_0 of the effective sample size $n - k_0$, writing $m_0 = [(n - k_0)\mu_0]$. An alternative but equivalent way of thinking about (3) is as splitting the sample MSE of the benchmark model $\sum_{t=k_0}^{n-1} \hat{e}_{0,t+1}^2/(n - k_0)$ into the two parts $\sum_{t=k_0}^{k_0+m_0-1} \hat{e}_{0,t+1}^2/m_0$ and $\sum_{t=k_0+m_0}^{n-1} \hat{e}_{0,t+1}^2/(n - k_0 - m_0)$ and taking the average of these two subsample means instead of the grand mean. Within our specific context, the motivation for proceeding this way is that it avoids the variance degeneracy problems associated with nested model comparisons as explained more formally below. Indeed, under the null hypothesis the numerator of (3) will have a non-degenerate positive limiting variance provided that $\mu_0 \in (0, 1) \setminus \{1/2\}$. This idea of using averages of subsample means instead of grand means has been used in a variety of other contexts such as the construction of more accurate confidence intervals as discussed in Decrouez and Hall (2014) for instance. Finally, the normaliser $\hat{\omega}_n(m_0)$ appearing in the denominator of (3) is understood to be a consistent estimator of the long run variance of the numerator and whose characteristics together with suitable choices of μ_0 we postpone to further below.

Given the sequences $\{\hat{e}_{0,t+1|t}\}_{t=k_0}^{n-1}$, $\{\hat{e}_{j,t+1|t}\}_{t=k_0}^{n-1}$ $(j = 1, \ldots, p)$ and suitable choices for μ_0 and $\hat{\omega}_n(m_0)$ the quantities in (3) can be trivially obtained for each possible predictor $j = 1, \ldots, p$ resulting in p such statistics which we aggregate into the following overall statistic

$$\overline{\mathcal{D}}_n(m_0) = \frac{1}{p} \sum_{j=1}^{p} \mathcal{D}_n(m_0, j). \tag{4}$$

Intuitively, the test statistic in (4) compares the sample MSEs associated with the benchmark model with the average of the p individual sample MSEs obtained from the p single predictor based specifications. A large positive magnitude of $\overline{\mathcal{D}}_n(m_0)$ is expected to indicate that at least one of the p predictors improves the predictability of y_{t+1} relative to the benchmark model.

In what follows our first objective is to establish the limiting behaviour of (4) under the null hypothesis that there are no active predictors in helping predict y_{t+1}. We show that under weak assumptions on the probabilistic properties of (1) the above aggregate test statistic is standard normally distributed. We subsequently assess its local power properties against departures from (2) that are relevant to practitioners. This in turn allows us to formalise suitable choices for μ_0 in the practical implementation of (4).

Upon rejection of the null hypothesis the interesting question as to which predictor is the key driver of predictability arises. Although our goal here is not to develop a new covariate screening method, our framework does allow us to identify a key predictor through the analysis of the $\mathcal{D}_n(m_0, j)$ components that make up the test statistic in (4). We focus our attention on the following estimator

$$\hat{j}_n \in \arg \max_{j=1,\ldots,p} \mathcal{D}_n(m_0, j) \tag{5}$$

which we expect to be informative about the most important contributor to predictability i.e. the predictor that leads to the greatest reduction in out of sample MSEs relative to the benchmark model. A limitation of $\hat{j}_n$ is of course the fact that it allows us to identify only a single predictor. Nevertheless, in numerous economic applications this information can be extremely valuable as it isolates the key player that causes the rejection of a maintained martingale difference hypothesis for instance.

3. Limiting Distributions and Local Power

3.1. Asymptotics of $\overline{\mathcal{D}}_n(\mu_0)$ under the benchmark model

Our objective here is to obtain the limiting distribution of $\overline{\mathcal{D}}_n(\mu_0)$ under the null hypothesis of no predictability when the benchmark model in (2) holds. Our operating assumptions are collected under Assumption 1 below and consist of a collection of high level assumptions that are general enough to accommodate most environments commonly encountered in economics and finance applications.

Assumptions 1. (i) The sequence $\eta_t = u_{t+1}^2 - E[u_{t+1}^2]$ satisfies the functional central limit theorem $\sum_{t=k_0}^{[k_0+(n-k_0)r]} \eta_t/\sqrt{n-k_0} \xrightarrow{d} \phi\ W(r)$ for $r \in [0,1]$ with $W(r)$ denoting a standard scalar Brownian Motion and where $\phi^2 = \sum_{s=-\infty}^{\infty} \gamma_\eta(s) > 0$ for $\gamma_\eta(s) = E[\eta_t \eta_{t+s}]$. (ii) There is a $\hat{\phi}_n^2$ such that $\hat{\phi}_n^2 \xrightarrow{p} \phi^2 \in (0, \infty)$. (iii) For given $r_a \in [0,1]$ and $r_b \in [0,1]$ with $r_a < r_b$ and under the null hypothesis, the forecast errors satisfy $\sum_{t=[(n-k_0)r_a]}^{[(n-k_0)r_b]} (\hat{e}_{\ell,t+1|t}^2 - u_{t+1}^2)/\sqrt{n-k_0} = o_p(1)$ $\forall \ell \in \{0, 1, 2, \ldots, p\}$. (iv) μ_0 satisfies $\mu_0 \in (0,1) \setminus \{1/2\}$.

Assumption 1(i) requires the sequence of demeaned squared errors driving (1)-(2) to satisfy a suitable functional central limit theorem. Such a result would hold for instance if the u_t's are a conditionally homoskedastic or conditionally heteroskedastic martingale difference sequence with mild existence of moments restrictions. Assumption 1(ii) requires that a consistent estimator of the long run variance associated with the η_t's to be available. Under conditional homoskedasticity a trivial choice would be $\hat{\phi}^2 = \sum_t (\hat{u}_{t+1}^2 - \hat{\sigma}_u^2)^2/(n-k_0)$ while under conditional heteroskedasticity one may use a Newey-West type formulation as in Deng and Perron (2008). Assumption 1(iii) can be viewed as a correct specification assumption in the sense that under the null hypothesis squared forecast errors are understood to behave like their true counterparts. Such a property holds within a broad range of contexts as established in Berenguer-Rico and Nielsen (2019), including settings with purely stationary or highly persistent predictors. Assumption 1(iv) imposes a minor restriction on $m_0 = [(n-k_0)\mu_0]$ used in the construction of $\overline{\mathcal{D}}_n(\mu_0)$ to ensure that it has a non-degenerate asymptotic variance. To gain further intuition on this latter point it is useful to explicitly evaluate the limiting variance, say $\omega^2(\mu_0)$, of the numerator of (3) under the null hypothesis. Replacing $\hat{e}_{0,t+1}^2$ and $\hat{e}_{j,t+1}^2$ with $\eta_{t+1} = u_{t+1}^2 - E[u_{t+1}^2]$ in (3), rearranging and taking expectations results in the population variance

$$\omega^2(\mu_0) = \frac{(1-2\mu_0)^2}{\mu_0(1-\mu_0)}\phi^2 \tag{6}$$

so that the availability of a consistent estimator for ϕ^2 also ensures that $\omega_n^2(\mu_0)$ can be estimated consistently provided that μ_0 satisfies Assumption 1(iv).

Our first proposition below establishes the limiting distribution of $\overline{\mathcal{D}}_n(\mu_0)$ when the maintained benchmark model holds.

Proposition 1. *Under the benchmark model in (2), assumptions 1(i)-(iv) and as* $n \to \infty$ *we have*

$$\overline{\mathcal{D}}_n(\mu_0) \xrightarrow{d} \mathcal{Z}. \tag{7}$$

with $\mathcal{Z}$ denoting a standard normally distributed random variable.

The result stated in Proposition 1 allows us to test for the presence of at least one active predictor for any finite number of such predictors. As it is customary in this literature (7) is implemented using one-sided (right tail) tests so that a rejection of the null provides support for the availability of at least one active predictor that helps generate more accurate forecasts than the benchmark model.

3.2. Asymptotic Power Properties of $\overline{\mathcal{D}}_n(\mu_0)$

We next explore the ability of $\overline{\mathcal{D}}_n(\mu_0)$ to detect predictability induced by one or more of the available p predictors. Two aspects we are particularly interested in exploring are the influence of the persistence properties of predictors on power and the role played by the ad-hoc choice of μ_0 in the construction of $\overline{\mathcal{D}}_n(\mu_0)$. We analyse local power within the following parameterisation

$$y_{t+1} = \boldsymbol{\beta}_n' \boldsymbol{x}_t + u_{t+1} \tag{8}$$

with $\boldsymbol{\beta}_n = n^{-\gamma}\boldsymbol{\beta}^*$ for $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)'$ and where we have abstacted from the inclusion of an intercept for notational simplicity and with no loss of generality. Note of course that within (8) all but one of the β_i^*'s may be zero and from a notational point of view $\boldsymbol{x}_t$ is understood to contain all of the p predictor candidates. We let

$$\mathcal{I}^* = \{1 \leq j \leq p \colon \beta_j^* \neq 0\} \tag{9}$$

denote the set of active predictors with $|\mathcal{I}^*| = q \geq 1$ referring to the size of the true model (i.e. the number of nonzero β_i^*'s in (8)).

In what follows we establish the local power properties of $\overline{\mathcal{D}}_n(\mu_0)$ across three scenarios. In a first instance we take all p components of $\boldsymbol{x}_t$ to be stationary and ergodic processes (scenario **A**) with $\mathcal{I}^*$ in (9) containing one or more of these $\boldsymbol{x}_t$'s. We then focus on the case where the $\boldsymbol{x}_t$'s are parameterised as persistent processes (scenario **B**). Finally our last scenario sets $\boldsymbol{x}_t = (\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t})'$ in (8) with $\boldsymbol{x}_{1,t}$ and $\boldsymbol{x}_{2,t}$ containing non-persistent and persistent

predictors respectively (scenario **C**). In this latter instance the specification in (8) takes the following form

$$y_{t+1} = \boldsymbol{\beta}'_{1n}\boldsymbol{x}_{1t} + \boldsymbol{\beta}'_{2n}\boldsymbol{x}_{2t} + u_{t+1} \tag{10}$$

with $\boldsymbol{x}_{1t} = (x_{1,t}, \ldots, x_{p_1,t})$ and $\boldsymbol{x}_{2t} = (x_{p_1+1,t}, \ldots, x_{p,t})$ so that the pool of p predictors is sub-divided into two types ranging from $1, \ldots, p_1$ and $p_1 + 1, \ldots, p$ respectively. The slope parameter vectors are in turn specified as $\boldsymbol{\beta}_{1n} = n^{-\gamma_1}\boldsymbol{\beta}_1^*$ for $\boldsymbol{\beta}_1^* = (\beta_{1,1}^*, \ldots, \beta_{1,p_1}^*)'$ and $\boldsymbol{\beta}_{2n} = n^{-\gamma_2}\boldsymbol{\beta}_2^*$ for $\boldsymbol{\beta}_2^* = (\beta_{2,p_1+1}^*, \ldots, \beta_{2,p}^*)'$. This mixed environments require us to also modify the formulation of the active set of predictors included in the DGP. For this purpose we let

$$\mathcal{I}_1^* = \{1 \leq j \leq p_1 \colon \beta_{1,j}^* \neq 0\} \tag{11}$$

$$\mathcal{I}_2^* = \{p_1 + 1 \leq j \leq p \colon \beta_{2,j}^* \neq 0\} \tag{12}$$

with $|\mathcal{I}_1^*| = q_1$ and $|\mathcal{I}_2^*| = q_2$. In this setting the specification in (10) is therefore understood to have q_1 active predictors satisfying scenario A and q_2 active predictors satisfying scenario B.

Assumption 2A below summarises our operating framework when all predictors are assumed to be purely stationary.

Assumption 2A. (i) Assumptions 1(i), 1(ii) and 1(iv) hold. (ii) The model in (8) holds with $\gamma = 1/4$. (iii) The p predictors satisfy $\sup_{\lambda \in [0,1]} |\sum_{t=1}^{[n\lambda]} x_{it}x_{jt}/n - \lambda E[x_{it}x_{jt}]| = o_p(1)$ in $\lambda \geq 0$ and $\sum_{t=1}^{[n\lambda]} x_{it}u_{t+1}/\sqrt{n} = O_p(1)$ for $i, j = 1, \ldots, p$.

Note that part (i) of Assumption 2A excludes 1(iii) as we are no longer operating under the null hypothesis. Part (ii) sets the rate at which we explore departures from the null. The remainder parts essentially require that a uniform law of large number applies to the predictors and that a suitable CLT holds ensuring the uniform boundedness of relevant sample moments.

Regarding the scenario involving solely persistent predictors we parameterise these as

mildly integrated processes via

$$x_{jt} = \left(1 - \frac{c_j}{n^\alpha}\right) x_{jt-1} + v_{jt} \quad j = 1, \ldots, p \tag{13}$$

where $c_j > 0$, $\alpha \in (0,1)$ and v_{jt} denotes a random disturbance term. The high level assumptions we impose under Assumption 2B explicitly accommodate dynamics such as (13) and follow directly from Phillips and Magdalinos (2009). We also let $\mathbf{\Sigma}_{vv}$ denote the $p \times p$ covariance of the $v'_{jt}s$ and refer to its diagonal components as $\sigma^2_{v_j}$ and its off-diagonal components as $\sigma_{v_i v_j}$ respectively.

Assumption 2B. (i) Assumptions 1(i), 1(ii) and 1(iv) hold. (ii) The model in (8) holds with $\gamma = (1+2\alpha)/4$ for $\alpha \in (0,1)$. (iii) The p predictors follow the process in (13) and satisfy $\sum_{t=1}^{[n\lambda]} x_{it}x_{jt}/n^{1+\alpha} \xrightarrow{p} \lambda\sigma_{v_i v_j}/(c_i + c_j)$, $\sum_{t=1}^{[n\lambda]} x_{jt}^2/n^{1+\alpha} \xrightarrow{p} \lambda\sigma^2_{v_j}/(2c_j)$ and $\sum_{t=1}^{[n\lambda]} x_{jt}u_{t+1}/n^{\frac{1+\alpha}{2}} = O_p(1)$ for $i, j = 1, \ldots, p$.

We note that in the context of our specification in (8), Assumption 2B(iii) is guaranteed to hold when the predictors follow the mildly integrated process in (13) as established in Lemmas 3.1-3.3 of Phillips and Magdalinos (2009). Our last assumption accommodates an environment that combines stationary and persistent predictors.

Assumption 2C. (i) Assumptions 1(i), 1(ii) and 1(iv) hold. (ii) The model in (10) holds with $\gamma_1 = 1/4$ and $\gamma_2 = (1+2\alpha)/4$ for $\alpha \in (0,1)$. (iii) The pool of p predictors consists of p_1 predictors satisfying Assumptions 2A(ii)-(iii) and $p_2 = p - p_1$ predictors satisfying Assumptions 2B(ii)-(iii).

Local Power under Stationarity (scenario A)

Proposition 2A: *Under Assumption 2A, $q := |\mathcal{I}^*|$ active predictors in (8) with associated*

slope parameters $\beta_i = n^{-1/4}\beta_i^*$ *for* $i \in \mathcal{I}^*$*, and as* $n \to \infty$ *we have*

$$\overline{\mathcal{D}}_n(\mu_0) \xrightarrow{d} \mathcal{Z} + \ g(\mu_0, \pi_0, \phi) \ \ \frac{1}{p}\sum_{j=1}^{p}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{E[x_{it}x_{jt}]}{\sqrt{E[x_{jt}^2]}}\right)^2 \tag{14}$$

where

$$g(\mu_0, \pi_0, \phi) = \frac{2\sqrt{1-\pi_0}\sqrt{\mu_0(1-\mu_0)}}{\sqrt{\phi^2}(1-2\mu_0)}. \tag{15}$$

The result in (14) establishes the consistency of our proposed test and its ability to detect departures from the constant mean model in (2) when predictors are taken to be stationary processes. An obvious implication of (14) is that under fixed alternatives $\overline{\mathcal{D}}_n(\mu_0) \to \infty$ and more specifically

$$\overline{\mathcal{D}}_n(\mu_0) \overset{H_1}{=} O_p(\sqrt{n}). \tag{16}$$

REMARK 1. The local power result in Proposition 2A has been obtained under local departures from the null that are of order $n^{-1/4}$ rather than the conventional square root rates one typically observes in stationary environments. This is not really due to the way the test statistic $\overline{D}_n(\mu_0)$ has been constructed. The main reason for operating under such a rate comes from the use of squared errors which result in the squaring of the relevant parameters in the DGP.

To gain further intuition on the formulation of the second component in the right hand side of (14) it is useful to specialise the result to a single active predictor scenario. Suppose that there is a single active predictor, say x_{at}, with associated slope parameter $\beta_{an} = n^{-1/4}\beta_a^*$. It now follows directly from (14) that

$$\overline{\mathcal{D}}_n(\mu_0) \xrightarrow{d} \mathcal{Z} + g(\mu_0, \pi_0, \phi)\ (\beta_a^*)^2\ E[x_{at}^2]\ \frac{1}{p}\sum_{j=1}^{p}\rho_{a,j}^2. \tag{17}$$

It is here interesting to note the role played by the correlation between the single predictor x_{at} driving the DGP in (8) and the remaining components of the predictor pool (i.e. the

irrelevant candidates). The higher this correlation is the stronger we expect power to be. This clearly conforms with intuition since the models are estimated with one predictor at a time. A particular fitted specification containing a predictor other than x_{at} and therefore misspecified will nevertheless continue to dominate the intercept only model in an MSE sense provided that this *wrong* predictor contains relevant information about x_{at}. Note also that this does not mean that in an environment where all predictors in the pool are uncorrelated with x_{at} power will vanish as we have $\rho^2_{a,a} = 1$ by construction, implying that the second component in the right hand side of (17) will always be strictly positive under our assumptions. Note however that in such instances where all candidate predictors are uncorrelated with x_{at} the size of the predictor pool p will have a detrimental impact on power.

Another important feature that can be inferred from (17) is the favourable impact that the variance of x_{at} has on power. The more persistent x_{at} is, the better the power is expected to be. This hints at the fact that the presence of persistent predictors in the pool will improve the detection ability of our test. Perhaps more interestingly we can also note that the role of persistence may manifest itself not only via $E[x^2_{at}]$ but also via $\rho^2_{a,j}$ due to the well known spurious correlation phenomenon characterising persistent processes. These issues are explored in the next proposition that focuses on the local power properties of $\overline{\mathcal{D}}_n(\mu_0)$ under persistence.

Local Power under Persistence (scenario B)

Proposition 2B: *Under Assumption 2B,* $q := |\mathcal{I}^*|$ *active predictors in (8) with associated slope parameters* $\beta_i = n^{-(1+2\alpha)/4}\beta_i^*$ *for* $i \in \mathcal{I}^*$ *and as* $n \to \infty$ *we have*

$$\overline{\mathcal{D}}_n(\mu_0) \xrightarrow{d} \mathcal{Z} + g(\mu_0, \pi_0, \phi)\ \frac{1}{p}\sum_{j=1}^{p}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{\sigma_{v_iv_j}}{\sqrt{\sigma^2_{v_j}}}\sqrt{\frac{2c_j}{(c_i+c_j)^2}}\right)^2. \tag{18}$$

The result in (18) highlights the beneficial impact that predictor persistence will have on the detection ability of $\overline{\mathcal{D}}_n(\mu_0)$. This can also be observed by focusing on fixed alternatives

under which we can immediately infer from (18) that

$$\overline{\mathcal{D}}_n(\mu_0) \stackrel{H_1}{=} O_p(n^{\frac{1+2\alpha}{2}}). \tag{19}$$

It is also interesting to observe from (18) that if we were to restrict all predictors to have the same non-centrality parameter, say $c_i = c \ \forall i = 1, \ldots, p$ the expression reduces to

$$\overline{\mathcal{D}}_n(\mu_0) \stackrel{d}{\to} \mathcal{Z} + g(\mu_0, \pi_0, \phi) \ \frac{1}{p}\frac{1}{\sqrt{2c}} \sum_{j=1}^{p} \left(\sum_{i \in \mathcal{I}^*} \beta_i^* \frac{\sigma_{v_i v_j}}{\sqrt{\sigma_{v_j}^2}} \right)^2 \tag{20}$$

which also suggests that all other things being equal power is expected to improve for smaller magnitudes of this non-centrality parameter.

REMARK 2. It is useful to note that we have opted to capture persistence via a *mildly integrated* setting with the α parameter in (13) excluding the boundary case of a local to unit-root specification (see Phillips and Magdalinos (2007, 2009)). As our result in (18) captures a reasonably broad notion of persistence we omit the details associated with more extreme forms of persistence that can be modelled via local to unit-root processes that impose $\alpha = 1$. Nevertheless it is useful to point out that in such settings the divergence of our test statistic would be faster with (19) replaced by $O_p(n^{3/2})$ instead.

Local Power under Mixed Predictors (scenario C)

The last scenario we consider consists of an environment that blends purely stationary and persistent predictors. The pool of predictors now consists of p_1 purely stationary and p_2 persistent predictors with $p_1 + p_2 = p$ and we let $\mathcal{J}_1$ and $\mathcal{J}_2$ denote the sets associated with the stationary and persistent predictors respectively so that $|\mathcal{J}_1| = p_1$ and $|\mathcal{J}_2| = p - p_1$.

Proposition 2C: *Under Assumption 2C,* $q_1 := |\mathcal{I}_1^*|$ *and* $q_2 := |\mathcal{I}_2^*|$ *active predictors in (10) with associated slope parameters* $\beta_{1,i} = n^{-1/4}\beta_{1,i}^*$ *for* $i \in \mathcal{I}_1^*$ *and* $\beta_{2,i} = n^{-(1+2\alpha)/4}\beta_{2,i}^*$ *for* $i \in \mathcal{I}_2^*$

we have as $n \to \infty$

$$\overline{\mathcal{D}}_n(\mu_0) \xrightarrow{d} \mathcal{Z} + g(\mu_0, \pi_0, \phi)\ \frac{1}{p}\left(\sum_{j\in\mathcal{J}_1}\left(\sum_{i\in\mathcal{I}_1^*}\beta_i^*\frac{E[x_{it}x_{jt}]}{\sqrt{E[x_{jt}^2]}}\right)^2 + \sum_{j\in\mathcal{J}_2}\left(\sum_{i\in\mathcal{I}_2^*}\beta_i^*\frac{\sigma_{v_iv_j}}{\sqrt{\sigma_{v_j}^2}}\sqrt{\frac{2c_j}{(c_i+c_j)^2}}\right)^2\right) \quad (21)$$

REMARK 3. The expressions in (14), (18) and (21) provide useful insights on suitable choices about μ_0 when constructing our main test statistic. We note for instance that μ_0 affects local power via $g(\mu_0, \pi_0, \phi)$ as defined in (15) suggesting that a choice for μ_0 in the vicinity of 0.5 may result in the most favourable power outcomes, all other things being equal.

4. Detecting active Predictors

Upon rejection of the benchmark model it becomes interesting to explore ways of identifying the predictors driving these departures from the null hypothesis. In this context we distinguish between two settings and obtain the corresponding limiting behaviour of $\hat{j}_n \in \arg\max_{j=1,\ldots,p}\mathcal{D}_n(m_0, j)$ which selects the predictor that results in the greatest MSE spread relative to the benchmark model.

In a first instance we evaluate the large sample behaviour of $\hat{j}_n$ when the DGP contains a single active predictor (i.e. $q = |\mathcal{I}^*| = 1$ in (9)) that can be either stationary or persistent. We subsequently extend our analysis to environments with multiple predictors (i.e. $q > 1$ in $\mathcal{I}^*$) that are again assumed to be of the same type in their persistence properties (i.e. all stationary or all persistent). Finally we consider the case of mixed predictors as in (11)-(12) with the joint presence of stationary and persistent active predictors numbering q_1 and q_2 respectively. The large sample behaviour of $\hat{j}_n$ is summarised in the following Proposition.

Proposition 3. *(i) Under Assumptions 2A or 2B and as* $n \to \infty$ *we have* $\hat{j}_n \xrightarrow{p} j_0 \in \mathcal{I}^*$ *for* $q \geq 1$. *(ii) Under Assumptions 2C and as* $n \to \infty$ *we have* $\hat{j}_n \xrightarrow{p} j_0 \in \mathcal{I}_1^* \cup \mathcal{I}_2^*$.

When the DGP consists solely of a single predictor (stationary or persistent), part (i) of Proposition 3 implies that $\hat{j}_n$ will be consistent for that true predictor asymptotically. When

there are multiple predictors of the same type the same result implies that $\hat{j}_n$ remains consistent for one of the $q > 1$ active predictors i.e. $\hat{j}_n$ is consistent for one of the true components in $\mathcal{I}^*$. Part (ii) of Proposition 3 relates to a scenario with mixed active predictors and states that in such a mixed setting $\hat{j}_n$ will continue to point to one of the true predictors which may come from any of the two sets.

Using the results provided in the proof of Proposition 3 it is useful to illustrate the mixed predictor scenario via a simple example of a predictive regression with two active predictors, say $y_{t+1} = \beta_{an} x_{at} + \beta_{bn} x_{bt} + u_{t+1}$ with $x_{at} \in \mathcal{I}_1^*$, $x_{bt} \in \mathcal{I}_2^*$ and as before $\beta_{an} = \beta_a^*/n^{1/4}$ and $\beta_{bn} = \beta_b^*/n^{(1+2\alpha)/4}$. Proposition 3(ii) clearly applies and implies that $\hat{j}_n$ will asymptotically point to either x_{at} or x_{bt}. From (53) and (55) in the appendix, we have that $\hat{j}_n$ will asymptotically point to x_{bt} (the persistent predictor) if

$$(\beta_b^*)^2 > \frac{E[x_{1t}^2]}{(\sigma_{v_b}^2/2c_b)}(\beta_a^*)^2 \tag{22}$$

and to x_{at} otherwise. It is now interesting to observe from (22) that $\hat{j}_n$ is expected to pick x_{bt} when the squared slope associated with this predictor exceeds the scaled slope of x_{at} with the scaling factor given by the ratio of the variances of the two predictors. As the ratio of these variances is likely to be small due the higher persistence of x_{bt} the procedure is also more likely to identify the persistent predictor unless the slope associated with x_{at} is particularly large relative to that of x_{bt}.

5. Experimental Properties

This section aims to document the empirical properties of our proposed test as well as the ability of $\hat{j}_n$ to detect relevant predictors in finite samples. In a first instance we focus on the size and power properties of $\overline{\mathcal{D}}_n(\mu_0)$ which we then follow with experiments documenting the correct decision frequencies associated with $\hat{j}_n$.

Before proceeding with our experimental design however it is important to revisit the practical implementation of $\overline{\mathcal{D}}_n(\mu_0)$ which requires imposing a suitable choice for μ_0 and a suitable estimator for $\omega^2(\mu_0)$ as defined in (6). We conduct our size experiments across a

variety of choices for μ_0 covering the $(0,1)\setminus\{1/2\}$ interval with all our results showing strong robustness to alternative magnitudes of μ_0 when it comes to size related outcomes. However, specific choices of μ_0 do play an important role when it comes to the power properties of $\overline{\mathcal{D}}_n(\mu_0)$ as also highlighted in our formal local power analysis. Our simulation based results confirm that magnitudes of μ_0 in the vicinity of 0.5 lead to the best power outcomes (e.g. $\mu_0 = 0.4$) with good size control.

A second issue related to the the practical implementation of $\overline{\mathcal{D}}_n(\mu_0)$ has to do with a suitable choice for an estimator of $\omega^2(\mu_0)$. As we wish to explicitly isolate the properties of our *one predictor at a time* method in what follows we abstract from additional layers of complications and contaminations related to the quality of the estimator of $\omega^2(\mu_0)$ used to evaluate $\overline{\mathcal{D}}_n(\mu_0)$. For this reason we conduct our experiments in the simplest possible conditionally homoskedastic setting and use

$$\hat{\omega}^2(\mu_0) = \frac{(1-2\mu_0)^2}{\mu_0(1-\mu_0)} \frac{\sum_{t=k_0}^{n-1}(\hat{u}_{t+1}^2 - \hat{\sigma}_u^2)^2}{n-k_0} \tag{23}$$

as an estimator of $\omega^2(\mu_0)$. The estimator in (23) does leave the choice of $\hat{\sigma}_u^2$ open in the sense that we could use the $\hat{u}_t$'s from either the benchmark model or the larger model. Although this is asymptotically immaterial when it comes to the large sample properties of $\overline{\mathcal{D}}_n(\mu_0)$ we expect that the use of residuals obtained from the larger model will result in better finite sample power properties as it is commonly the case for test statistics such as CUSUM and CUSUMSQ. This is the option we also adopt in what follows.

Finally, in the context of our size and power experiments we also consider a modification to $\overline{\mathcal{D}}_n(\mu_0)$ designed to enhance its power properties without affecting its null limiting distribution. Our proposed approach is in the spirit of Fan, Liao and Yao (2015) and involves augmenting our test statistic with a quantity that converges to 0 under the null while diverging under the alternative of at least one active predictor. Consider for instance the quantity $d_{nj} = \sum_{t=k_0}^{n-1}(\hat{e}_{0,t+1|t} - \hat{e}_{j,t+1|t})^2/(n-k_0)$. Within our nested context and under the benchmark model we clearly have $d_{nj} = O_p(n^{-1/2})\ \forall j = 1,\ldots,p$ while under the alternative whereby the true model contains at least one active predictor $\widetilde{d}_{nj} \equiv \sqrt{n-k_0}\ d_{nj}/\hat{\omega}(\mu_0) = O_p(1)\ \forall j = 1,\ldots,p$ with the associated limiting random variable being strictly positive. This prompts us to

propose the following augmentation to $\overline{\mathcal{D}}_n(m_0)$

$$\overline{\mathcal{D}}_n^d(m_0) = \frac{1}{p}\sum_{j=1}^{p}(\mathcal{D}_n(m_0, j) + \tilde{d}_{nj}) \tag{24}$$

which we expect to have a favourable impact on power while being size neutral asymptotically and whose properties we explore via the simulations below.

5.1. Finite Sample Size and Power Properties of $\overline{\mathcal{D}}_n(m_0)$ and $\overline{\mathcal{D}}_n^d(m_0)$

Empirical Size

For our size experiments the DGP is given by the benchmark specification in (2) where we set $\theta_0 = 0$ with no loss of generality. The pool of p predictors is taken to follow the VAR(1) process $\boldsymbol{x}_t = \boldsymbol{\Phi}\boldsymbol{x}_{t-1} + \boldsymbol{v}_t$ with $\boldsymbol{v}_t \sim N(0, \boldsymbol{\Sigma}_{vv})$ which we parameterise in ways that can distinguish between weakly or uncorrelated features and more strongly correlated features. In what follows we also let $\boldsymbol{\Omega}$ denote the covariance matrix of $(u_t, \boldsymbol{v}_t)'$ and write $\boldsymbol{\Omega} = \{\sigma_u^2, \boldsymbol{\sigma}_{uv}'\}, \{\boldsymbol{\sigma}_{uv}, \boldsymbol{\Sigma}_{vv}\}\}$. Our experiments involving either purely stationary or purely persistent predictors are conducted across the following three alternative configurations of $\boldsymbol{\Omega}$

(i) $\boldsymbol{\Omega}_0$: $\sigma_u^2 = 1, \boldsymbol{\sigma}_{uv} = \mathbf{0}_{p\times 1}, \boldsymbol{\Sigma}_{vv} = \boldsymbol{I}_p$,

(ii) $\boldsymbol{\Omega}_1$: $\sigma_u^2 = 1, \boldsymbol{\sigma}_{uv} = \mathbf{0}_{p\times 1}, \boldsymbol{\Sigma}_{vv} = [0.5^{|i-j|}]_{1\leq j\leq p}$,

(iii) $\boldsymbol{\Omega}_2$: $\sigma_u^2 = 1, \boldsymbol{\sigma}_{uv} = (-0.5)^{|1-j|}, \boldsymbol{\Sigma}_{vv} = [0.5^{|i-j|}]_{1\leq j\leq p}$

and two alternative scenarios on the persistence properties of the predictors, with

(a) $\boldsymbol{\Phi} = 0.50\ \boldsymbol{I}_p$

(b) $\boldsymbol{\Phi} = 0.95\ \boldsymbol{I}_p$.

Accordingly we label these size related DGPs as DGP-S(a-i)-(a-iii) and DGP-S(b-i)-(b-iii).

For the scenario corresponding to a mixed pool of predictors we set $\boldsymbol{\Phi} = diag(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)$ with $\boldsymbol{\Phi}_1 = 0.50\ \boldsymbol{I}_{p_1}$ and $\boldsymbol{\Phi}_2 = 0.95\ \boldsymbol{I}_{p_2}$ for $p_2 = p - p_1$. The corresponding covariance matrix, say $\boldsymbol{\Omega}_3$, is here formulated as $\boldsymbol{\Omega}_3 = \{\{1, \mathbf{0}', \mathbf{0}'\}, \{\mathbf{0}_{p_1\times 1}, \boldsymbol{\Sigma}_{v_1v_1}, \mathbf{0}_{p_1\times p_2}\}, \{\mathbf{0}_{p_2\times p_1}, \mathbf{0}_{p_2\times p_1}, \boldsymbol{\Sigma}_{v_2v_2}\}\}$

with $\mathbf{\Sigma}_{v_1v_1} = [0.5^{|i-j|}]$ for $i, j = 1, \ldots, p_1$ and $\mathbf{\Sigma}_{v_2v_2} = [0.5^{|i'-j'|}]$ for $i', j' = 1, \ldots, p_2$. We label this latter DGP as DGP-S(c).

Empirical size outcomes are obtained for $p \in \{10, 20\}$, $\mu_0 \in \{0.3, 0.4\}$ and samples of size $n = 500$ with $\pi_0 = 0.25$ used as the starting point for generating recursive forecasts i.e. $n - k_0 = 375$. Results across the two test statistics and all of the above parameterisations are collected in Table 1 below.

We can first highlight the fact that the *one predictor at a time* approach based on $\mathcal{D}_n(\mu_0)$ and $\mathcal{D}_n^d(\mu_0)$ appears to be robust to the size of the predictor pool with almost identical size estimates obtained across $p = 10$ and $p = 20$. Outcomes are also clearly robust to alternative predictor dynamics and their persistence characteristics as expected from our result in Proposition 1. We can note for instance very little variability in size outcomes across the seven DGPs that span various persistence characteristics.

From the size estimates based on $\overline{\mathcal{D}}_n(\mu_0)$ we note that the statistic appears to be undersized under $\mu_0 = 0.4$ while maintaining good size control under $\mu_0 = 0.3$. More importantly, its augmented version $\overline{\mathcal{D}}_n^d(\mu_0)$ appears to offer good size control across all configurations and chosen magnitudes of μ_0. Our earlier local power analysis established that setting μ_0 in the vicinity of 0.5 should result in more favourable power outcomes relative to other magnitudes in $(0, 1)$. The size estimates of Table 1 suggest potential size-power trade-offs (albeit moderate) when one proceeds this way and uses $\mathcal{D}_n(\mu_0 = 0.4)$ for instance. Focusing on $\mathcal{D}_n^d(\mu_0)$ on the other hand, this augmented test statistic appears to offer excellent size control even for $\mu_0 = 0.4$ and such a choice comes across as a good compromise between size and power. Our power experiments below provide extensive support for this intuition.

Empirical Power

We consider predictive regressions with either one or two active predictors parameterised as

$$y_{t+1} = \beta_{an} x_{a,t} + \beta_{bn} x_{b,t} + u_{t+1} \tag{25}$$

Table 1: Empirical Size

DGP-S	$a(i)$	$a(ii)$	$a(iii)$	$b(i)$	$b(ii)$	$b(iii)$	c
				$p = 10$			
$\overline{\mathcal{D}}_n(\mu_0 = 0.3)$	0.090	0.092	0.091	0.086	0.086	0.086	0.089
$\overline{\mathcal{D}}^d_n(\mu_0 = 0.3)$	0.113	0.111	0.111	0.111	0.112	0.111	0.112
$\overline{\mathcal{D}}_n(\mu_0 = 0.4)$	0.062	0.061	0.062	0.056	0.056	0.056	0.057
$\overline{\mathcal{D}}^d_n(\mu_0 = 0.4)$	0.105	0.105	0.105	0.106	0.109	0.108	0.105
				$p = 20$			
$\overline{\mathcal{D}}_n(\mu_0 = 0.3)$	0.090	0.091	0.091	0.086	0.086	0.086	0.088
$\overline{\mathcal{D}}^d_n(\mu_0 = 0.3)$	0.109	0.108	0.109	0.110	0.110	0.109	0.106
$\overline{\mathcal{D}}_n(\mu_0 = 0.4)$	0.071	0.071	0.071	0.067	0.066	0.064	0.067
$\overline{\mathcal{D}}^d_n(\mu_0 = 0.4)$	0.107	0.111	0.109	0.113	0.116	0.115	0.115

with $\beta_{an} = \beta^*_a/n^{0.25}$, $\beta_{bn} = \beta^*_b/n^{0.675}$ and the following three scenarios

(i) $\mathbf{\Phi} = 0.50\ \boldsymbol{I}_p$, $\beta^*_a \in \{1, 2, 3, 4, 5\}$, $\beta^*_b = 0$ and $\mathbf{\Omega} = \mathbf{\Omega}_1$, $x_{at} \in \{x_{1t}, \ldots, x_{pt}\}$, $x_{at} = x_{1t}$

(ii) $\mathbf{\Phi} = 0.95\ \boldsymbol{I}_p$, $\beta^*_a = 0$, $\beta^*_b \in \{5, 6, 7, 8, 9\}$ and $\mathbf{\Omega} = \mathbf{\Omega}_1$, $x_{bt} \in \{x_{1t}, \ldots, x_{pt}\}$, $x_{bt} = x_{1t}$

(iii) $\mathbf{\Phi}_1 = 0.50\ \boldsymbol{I}_{p_1}$, $\mathbf{\Phi}_2 = 0.95\ \boldsymbol{I}_{p-p_1}$, $x_{at} \in \{x_{1,t}, \ldots, x_{p_1 t}\}$, $x_{bt} \in \{x_{p_1+1,t}, \ldots, x_{p,t}\}$, $(\beta^*_a, \beta^*_b) \in (\{1, 2, 3, 4, 5\}, \{5, 6, 7, 8, 9\})$ and $\mathbf{\Omega} = \mathbf{\Omega}_3$, $x_{at} = x_{1t}$, $x_{bt} = x_{p_1+1,t}$,

which we label as DGP-P(i)-(iii). It is also useful to point out that our chosen slope parameterisations translate into $\beta_{an} \in \{0.211, 0.423, 0.634, 0.846, 1.057\}$ for the stationary predictors and $\beta_{bn} \in \{0.075, 0.090, 0.106, 0.121, 0.136\}$ for the persistent predictors, highlighting the point that power is evaluated as the DGP moves further away from the null.

Here DGP-P(i) corresponds to a setting with a single active stationary predictor that belongs to the pool of p stationary predictors. DGP-P(ii) considers the case of a single active predictor that is persistent and belongs to the pool of p persistent predictors. Finally, DGP-P(iii) considers the case of two active predictors. The first one belongs to the pool of p_1 stationary predictors and the second one belongs to the pool of $p_2 = p - p_1$ persistent predictors. Note also that for notational convenience and no loss of generality we have taken the active predictor to be x_{1t} in cases (i) and (ii) and the two active predictors to be $(x_{1t}, x_{p_1+1,t})$ in case (iii). Results for these experiments are presented in Table 2 below.

Table 2: Empirical Power

	p=10					p=20				
	DGP-P(i)					DGP-P(i)				
$\beta_{a,n}$	0.211	0.423	0.634	0.846	1.057	0.211	0.423	0.634	0.846	1.057
$\overline{\mathcal{D}}_n(\mu_0 = 0.3)$	0.147	0.348	0.730	0.953	0.997	0.108	0.190	0.362	0.611	0.848
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.3)$	0.250	0.717	0.989	1.000	1.000	0.161	0.407	0.747	0.972	0.998
$\overline{\mathcal{D}}_n(\mu_0 = 0.4)$	0.189	0.682	0.987	1.000	1.000	0.113	0.331	0.714	0.965	1.000
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$	**0.447**	**0.980**	**1.000**	**1.000**	**1.000**	**0.250**	**0.760**	**0.989**	**1.000**	**1.000**
	DGP-P(ii)					DGP-P(ii)				
$\beta_{b,n}$	0.075	0.090	0.106	0.121	0.136	0.075	0.090	0.106	0.121	0.136
$\overline{\mathcal{D}}_n(\mu_0 = 0.3)$	0.142	0.162	0.220	0.252	0.303	0.118	0.134	0.138	0.171	0.217
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.3)$	0.264	0.342	0.445	0.539	0.608	0.190	0.255	0.272	0.341	0.410
$\overline{\mathcal{D}}_n(\mu_0 = 0.4)$	0.188	0.251	0.331	0.426	0.520	0.115	0.165	0.168	0.233	0.312
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$	0.472	0.635	0.742	0.835	0.901	0.316	0.440	0.516	0.613	0.719
$\beta_{b,n}$	0.136	0.151	0.166	0.181	0.196	0.136	0.151	0.166	0.181	0.196
$\overline{\mathcal{D}}_n(\mu_0 = 0.3)$	0.303	0.354	0.414	0.477	0.493	0.217	0.246	0.248	0.290	0.344
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.3)$	0.608	0.704	0.765	0.821	0.871	0.410	0.488	0.522	0.562	0.657
$\overline{\mathcal{D}}_n(\mu_0 = 0.4)$	0.520	0.618	0.679	0.754	0.805	0.312	0.381	0.401	0.462	0.541
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$	**0.901**	**0.947**	**0.974**	**0.988**	**0.992**	**0.719**	**0.804**	**0.848**	**0.900**	**0.932**
	DGP-P(iii)					DGP-P(iii)				
$\beta_{a,n}$	0.211	0.423	0.634	0.846	1.057	0.211	0.423	0.634	0.846	1.057
$\beta_{b,n}$	0.075	0.090	0.106	0.121	0.136	0.075	0.090	0.106	0.121	0.136
$\overline{\mathcal{D}}_n(\mu_0 = 0.3)$	0.200	0.423	0.729	0.931	0.994	0.130	0.242	0.374	0.576	0.780
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.3)$	0.400	0.801	0.990	1.000	1.000	0.240	0.510	0.779	0.957	0.996
$\overline{\mathcal{D}}_n(\mu_0 = 0.4)$	0.330	0.767	0.985	1.000	1.000	0.168	0.422	0.734	0.947	0.995
$\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$	**0.730**	**0.993**	**1.000**	**1.000**	**1.000**	**0.434**	**0.878**	**0.995**	**1.000**	**1.000**

The top panel of Table 2 presents power outcomes in a stationary context (case (i)). As expected from our theoretical results power can be seen to increase towards 100% as β_{an} increases and we move further away from the null hypothesis. The DGP here contains a single active predictor and our *one at a time* approach has been implemented using predictor pools of size 10 and 20. Overall $\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$ based inferences are seen to provide excellent power outcomes even under a sizeable number of predictors and very moderate signal to noise ratios. Under $\beta_{an} = 0.423$ for instance we note that $\mathcal{D}_n(\mu_0 = 0.4)$ and $\mathcal{D}_n^d(\mu_0 = 0.4)$ lead to power estimates of 68.2% and 98.0% respectively.

The middle panel of Table 2 considers the case of persistent predictors while the DGP continues to be characterised by a single active predictor. For this scenario we provide a more detailed range of β_{bn} magnitudes which help highlight the favourable impact that persistence has on power. For comparable departures from the null we can indeed observe substantially better power outcomes. Under stationarity and $\beta_{an} = 0.423$ for instance $\overline{\mathcal{D}}_n(\mu_0 = 0.4)$ led to an empirical power of 98% while in the persistent setting such power magnitudes are achieved for substantially smaller departures from the null. The bottom panel of Table 2 considers the case of two active predictors with the first one being stationary and the second one persistent. We continue to note the excellent performance of both $\overline{\mathcal{D}}_n(\mu_0 = 0.4)$ and $\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$ across most parameterisations. Both test statistics do appear to lose power as p is allowed to increase however. This feature is not surprising as it follows directly from our theoretical analyses in Propositions 2A-2C. Nevertheless it is useful to point out that the drop in power of $\overline{\mathcal{D}}_n^d(\mu_0 = 0.4)$ is considerably more moderate (e.g. dropping from 99.3% to 87.8% under $(\beta_{an}, \beta_{bn}) = (0.423, 0.090)$ and as we double the pool of predictors).

5.2. Finite Sample Properties of $\hat{j}$

In this last set of experiments our goal is to illustrate the result in Proposition 3 by documenting the behaviour of $\hat{j}_n$ as n is allowed to grow. For this purpose we focus on DGP-P(iii) that contains two mixed predictors and evaluate the behaviour of $\hat{j}_n$ for a given slope configuration (β_{an}, β_{bn}) as n increases. More specifically, we proceed with $(\beta_{an}, \beta_{bn}) \in \{(0.423, 0.090), (0.423, 0.196)\}$ and track $\hat{j}_n$ across $n \in \{150, 250, 500\}$. We operate under $p = 10$ with the active stationary predictor being given by $x_{at} \equiv x_{1t}$ and the

persistent active predictor by $x_{bt} \equiv x_{6t}$. The pool of p=10 predictors consists of 5 stationary and 5 persistent components so that according to Proposition 3(ii) we expect to observe $\hat{j}_n \xrightarrow{p} j_0 \in \{1, 6\}$.

Under $(\beta_{an}, \beta_{bn}) = (0.423, 0.090)$ we note that the slope associated with the stationary predictor is substantially larger than that of the persistent predictor. Based on our analysis in (22) we expect that under this parameterisation $\hat{j}_n$ is likely to point to 1 more often than to 6. For $(\beta_{an}, \beta_{bn}) = (0.423, 0.196)$ the slope associated with the stationary predictor continues to dominate that of $x_{bt} = x_{6t}$. However as x_{6t} is more persistent it is now possible that $\hat{j}_n$ may point to 6 more often. These conjectures are confirmed by the empirical outcomes presented in Table 3 below.

Under $(\beta_{an}, \beta_{bn}) = (0.423, 0.090)$, $\hat{j}_n$ points to x_{1t} close to 100% of the times. We can also observe that the correct selection frequencies increase with n as expected from our result in Proposition 3. Within this parameterisation the large magnitude of the slope of x_{1t} makes it a dominant predictor that gets uniquely picked up by our proposed estimator.

Under $(\beta_{an}, \beta_{bn}) = (0.423, 0.196)$ the right panel of Table 3 highlights the greater ambiguity in outcomes due to the complex interplay between the magnitudes of the slopes and the greater persistence associated with x_{6t}. We note that as n increases the estimator increasingly concentrates on the persistent predictor x_{6t}.

Table 3: Selection frequencies based on $\hat{j}_n$

$p = 10$	$\hat{j}_n = 1$	$\hat{j}_n = 6$	$\hat{j}_n \notin \{1, 6\}$	$\hat{j}_n = 1$	$\hat{j}_n = 6$	$\hat{j}_n \notin \{1, 6\}$
	$(\beta_{an}, \beta_{bn}) = (0.423, 0.090)$			$(\beta_{an}, \beta_{bn}) = (0.423, 0.196)$		
$n = 150$						
$\mathcal{D}_n(\mu_0 = 0.4)$	0.910	0.064	0.026	0.400	0.586	0.014
$\mathcal{D}_n^d(\mu_0 = 0.4)$	0.919	0.059	0.022	0.380	0.604	0.016
$n = 250$						
$\mathcal{D}_n(\mu_0 = 0.4)$	0.961	0.036	0.003	0.290	0.708	0.002
$\mathcal{D}_n^d(\mu_0 = 0.4)$	0.964	0.034	0.002	0.280	0.720	0.000
$n = 500$						
$\mathcal{D}_n(\mu_0 = 0.4)$	0.984	0.016	0.000	0.157	0.843	0.000
$\mathcal{D}_n^d(\mu_0 = 0.4)$	0.989	0.011	0.000	0.150	0.850	0.000

6. Conclusions

This paper proposed a method of detecting the presence of out-of-sample predictability in the context of linear predictive regressions linking a response variable to one or more lagged predictors. The main novelty here is the ability of our method to accommodate a large number of predictors in a simple and computationally feasible way, including scenarios where the number of candidate predictors exceeds the sample size. An additional feature that is also particularly relevant to economic applications is the robustness of our proposed approach to the dynamic properties of predictors which can be noisy, persistent or a mixture of both.

In numerous applications one is often interested in whether a particular variable is best described as a mean independent process as opposed to being predictable with some predictor belonging to a large information set. The researcher may not wish to take a stance on a particular predictor while also being constrained by dimensionality problems. The test we introduced in this paper is precisely designed to accommodate such environments by providing a way of testing the null of no predictability versus the alternative of at at least one active predictor drawn from a potentially large pool of candidate predictors. Although our primary focus is not about uncovering a true model or achieving full model selection consistency our framework does also allow us to consistently detect the *key player* or one of the *key players* which can be valuable information in itself.

Various limitations of our setting can become the subject of interesting further research and extensions. We conjecture for instance that it may be possible to consider an iterated approach to uncovering key predictors by proceeding sequentially, one at a time, using our estimator $\hat{j}_n$. Although conceptually straightforward such an extension would raise important technical challenges similar to well known post model selection related inference problems. A second potentially interesting extension would be to explore an alternative asymptotic framework for the limiting behaviour of our $\overline{D}_n(\mu_0)$ statistic whereby μ_0 would be made sample size dependent in a way that allows it to approach the variance degeneracy boundary of $\mu_0 = 1/2$ at some suitable rate.

REFERENCES

Berenguer-Rico, V. and B. Nielsen (2020). Cumulated Sum of Squares Statistics for Nonlinear and Nonstationary Regressions, *Econometric Theory 36(1)*, 1-47.

Chen, Y., and K. P. Tang (2013). What does the yield curve tell us about exchange rate predictability?, *Review of Economics and Statistics 95 (1)*, 185-205.

Clark, T.E., and M. W. McCracken (2013). Advances in Forecast Evaluation, in *Handbook of Economic Forecasting*, Chapter 20, 1107-1201, Elsevier.

Decrouez, G., and P. Hall (2014). Split sample methods for constructing confidence intervals for binomial and Poisson parameters, *J. R. Statist. Soc. B 76(5)*, 949-975.

Deng, A., and P. Perron (2008). The Limit Distribution of the Cusum of Squares Test under General Mixing Conditions, *Econometric Theory 24 (3)*, 809-822.

Diebold, F.X. and R. Mariano (1995). Comparing Predictive Accuracy, *Journal of Business and Economic Statistics 13 (3)*, 253-265.

Fan, J., and J. Liv (2008). Sure independence screening for ultrah dimensional feature space, *JRSS Series B, 70 (5)*, 849-911.

Fan, J., Liao, Y., and J. Yao (2015). Power Enhancement in High Dimensional Cross-Sectional Tests, *Econometrica 83 (4)*, 1497-1541.

Ghysels, E., Hill, J. B., and K. Moteigi (2020). Testing a large set of zero restrictions in regression models, with an application to mixed frequency Granger causality, *Journal of Econometrics 218 (2)*, 633-654.

Gonzalo, J., and J. Pitarakis (2019). Predictive Regressions, *Oxford Research Encyclopedia: Economics and Finance*, November 2019.

Hsu, Y., and X. Shi (2017). Model selection tests for conditional moment restriction models, *The Econometrics Journal 20 (1)*, 52–85.

Jacobsen, B., Jiang, F., and H. Zhang (2019). Ensemble Machine Learning and Stock Return Predictability, *SSRN 3310289*.

Li, J., Tsiakas, I. , and W. Wang (2015). Predicting exchange rates out of sample: can economic fundamentals beat the random walk?, *Journal of Financial Econometrics 13 (2)*, 293-341.

McCracken, M. (2007). Asymptotics for out of sample tests of Granger causality, *Journal*

of Econometrics 140 (2), 719-752.

McKeague, I. W., and M. Qian (2015). An adaptive resampling test for detecting the presence of significant predictors, *Journal of the American Statistical Association 110 (512)*, 1442-1433.

Molodotsova, T., and D. H. Papell (2009). Out-of-sample exchange rate predictability with Taylor rule fundamentals, *Journal of International Economics 77 (2)*, 167-180.

Mullainathan, S., and J. Spiess (2017). Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives 31 (2)*, 87-106.

Phillips, P. C. B., and A. Magdalinos (2007). Limit theory for moderate deviations from a unit root, *Journal of Econometrics 136 (1)*, 115-130.

Phillips, P. C. B., and A. Magdalinos (2009). Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors, *Econometric Theory 25 (2)*, 482-526.

Pitarakis, J. (2020). A Novel Approach to Predictive Accuracy Testing in Nested Environments, arXiv: 2008.08387v1.

Rapach, D. E., Strauss, J. K., Tu, J., and G. Zhou (2019). Industry return predictability: A machine learning approach, *Journal of Financial Data Science 1 (3)*, 9-28.

Schennah, S. M., and D. Wilhelm (2017). A Simple Parametric Model Selection Test, *Journal of the American Statistical Association 112 (520)*, 1663-1674.

Shi, X. (2015). A Nondegenerate Vuong Test, *Quantitative Economics 6 (1)*, 85–121.

Tang, Y., Wang H. J., and E. Barut (2017). Testing for the presence of significant covariates through conditional marginal regression, *Biometrika, 105 (1)*, 57-71.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses, *Econometrica 57 (2)*, 307–333.

West, K. (1996). Asymptotic Inference about Predictive Ability, *Econometrica 64 (5)*, 1067-1084.

West, K. (2006). Forecast Evaluation, in *Handbook of Economic Forecasting*, Volume 1, Graham Elliott, Clive W.J. Granger and Allan Timmermann eds.

APPENDIX

PROOF OF PROPOSITION 1. Under the null hypothesis Assumption 1(iii) implies that for given μ_0 we can write (4) as

$$\overline{\mathcal{D}}_n(m_0) = \frac{\sqrt{n-k_0}}{\hat{\omega}_n(m_0)} \left(\left(\frac{\sum_{t=k_0}^{k_0+m_0-1}(u_{t+1}^2 - \sigma_u^2)}{m_0} + \frac{\sum_{t=k_0+m_0}^{n-1}(u_{t+1}^2 - \sigma_u^2)}{n-k_0-m_0} \right) - 2\frac{\sum_{t=k_0}^{n-1}(u_{t+1}^2 - \sigma_u^2)}{n-k_0} \right) + o_p(1). \tag{26}$$

It now follows directly from Assumption 1(i) and the continuous mapping theorem that

$$\overline{\mathcal{D}}_n(m_0) \xrightarrow{d} \frac{1}{\omega(\mu_0)} \left(\phi \left(\frac{W(\mu_0)}{\mu_0} + \frac{W(1) - W(\mu_0)}{1-\mu_0} \right) - 2\phi W(1) \right). \tag{27}$$

As we operate under a given μ_0 it is now straightforward to observe that the variance of the expression between brackets in (27) is given by $\phi^2(1-2\mu_0)^2/(\mu_0(1-\mu_0))$. As $\hat{\omega}(\mu_0) \xrightarrow{p} \omega(\mu_0) = \phi(1-2\mu_0)/\sqrt{\mu_0(1-\mu_0)}$ it follows from Slutsky's theorem that $\overline{\mathcal{D}}_n(\mu_0) \xrightarrow{d} N(0,1)$ as required.

Before proceeding with the proofs of Propositions 2A, 2B and 2C we introduce a series of intermediate results and further notation that will be used throughout. As we operate under the hypothesis of at least one active predictor the true specifications under our three scenarios A, B and C are understood to be given by

$$y_{t+1} = \sum_{i \in \mathcal{I}^*} (\beta_i^*/n^{1/4})\, x_{it} + u_{t+1} \tag{28}$$

$$y_{t+1} = \sum_{i \in \mathcal{I}^*} (\beta_i^*/n^{(1+2\alpha)/4})\, x_{it} + u_{t+1} \tag{29}$$

and

$$y_{t+1} = \sum_{i \in \mathcal{I}_1^*} (\beta_i^*/n^{1/4})\, x_{it} + \sum_{i \in \mathcal{I}_2^*} (\beta_i^*/n^{(1+2\alpha)/4})\, x_{it} + u_{t+1} \tag{30}$$

respectively. We also recall that the fitted specification involving one predictor at a time is here given by

$$y_{t+1} = \beta_j x_{jt} + u_{t+1} \qquad j = 1, \ldots, p \tag{31}$$

so that using (28) and (29) we can write the recursively estimated slope parameters as

$$\hat{\beta}_{jt} = \frac{\sum_{i \in \mathcal{I}^*} \beta_i^* (\sum_{s=1}^t x_{is} x_{js})}{n^\gamma \sum_{s=1}^t x_{js}^2} + \frac{\sum_{s=1}^t x_{js} u_{s+1}}{\sum_{s=1}^t x_{js}^2} \tag{32}$$

for $\gamma = 1/4$ under scenario A and $\gamma = (1 + 2\alpha)/4$ under scenario B. For the mixed predictor scenario C and using (30) we have instead

$$\hat{\beta}_{jt} = \frac{\sum_{i \in \mathcal{I}_1^*} \beta_i^* (\sum_{s=1}^t x_{is} x_{js})}{n^{\gamma_1} \sum_{s=1}^t x_{js}^2} + \frac{\sum_{i \in \mathcal{I}_2^*} \beta_i^* (\sum_{s=1}^t x_{is} x_{js})}{n^{\gamma_2} \sum_{s=1}^t x_{js}^2} + \frac{\sum_{s=1}^t x_{js} u_{s+1}}{\sum_{s=1}^t x_{js}^2}. \tag{33}$$

The specifications in (28)-(30) are the DGPs under the local alternatives of interest and the $\hat{\beta}_{jt}$'s in (32)-(33) are the slope parameters estimated via recursive least squares when fitting (31). As for notational convenience we have abstracted from the inclusion of an intercept in the above specifications it is naturally understood that the forecasts under the null model will be taken as $\hat{y}_{0,t+1|t} = 0$ instead of $\sum_{j=1}^t y_j/t$. This has no bearing on any of the asymptotic results presented in Propositions 2A-2C. We can now write the forecast errors as

$$\begin{aligned} \hat{e}_{0,t+1|t} &= y_{t+1} - 0 \\ \hat{e}_{j,t+1|t} &= y_{t+1} - \hat{\beta}_{jt} x_{jt} \end{aligned} \tag{34}$$

with y_{t+1} given by either (28), (29) or (30).

LEMMA A1. Under Assumption 2A, $\hat{\beta}_{jt}$ as in (32) and $\forall j \in \{1, \ldots, p\}$ we have as $n \to \infty$

(i) $\sup_{r \in [\pi_0, 1]} \left| n^{1/4} \hat{\beta}_{j,[nr]} - \frac{1}{E[x_{jt}^2]} \sum_{i \in \mathcal{I}^*} \beta_i^* E[x_{it} x_{jt}] \right| = o_p(1)$

(ii) $\sup_{k_0 \leq t \leq n} \left| \frac{\sum_{\ell=k_0}^t \hat{\beta}_{j\ell} x_{j\ell} u_{\ell+1}}{\sqrt{n - k_0}} \right| = o_p(1)$

(iii) $\sup_{k_0\leq t\leq n}\left|\frac{\sum_{\ell=k_0}^{t}\hat{\beta}_{j\ell}^2 x_{j\ell}^2}{\sqrt{n-k_0}}-\frac{\sqrt{1-\pi_0}}{E[x_{jt}^2]}\left(\sum_{i\in\mathcal{I}^*}\beta_i^* E[x_{it}x_{jt}]\right)^2\right|=o_p(1)$

(iv) $\sup_{k_0\leq t\leq n}\left|\frac{\beta_i^*}{n^{1/4}}\frac{\sum_{\ell=k_0}^{t}\hat{\beta}_{j\ell}x_{i\ell}x_{j\ell}}{\sqrt{n-k_0}}-\sqrt{1-\pi_0}\;\beta_i^*\frac{E[x_{it}x_{jt}]}{E[x_{jt}^2]}(\sum_{i\in\mathcal{I}^*}\beta_i^* E[x_{it}x_{jt}])\right|=o_p(1)$

PROOF OF LEMMA A1. (i) From (32) we have

$$n^{1/4}\hat{\beta}_{jt}=\frac{\sum_{i\in\mathcal{I}^*}\beta_i^*(\sum_{s=1}^{t}x_{is}x_{js})}{\sum_{s=1}^{t}x_{js}^2}+n^{1/4}\frac{\sum_{s=1}^{t}x_{js}u_{s+1}}{\sum_{s=1}^{t}x_{js}^2} \tag{35}$$

and

$$n^{1/4}\sup_t|\hat{\beta}_{jt}|\leq\sup_t\left|\frac{\sum_{i\in\mathcal{I}^*}\beta_i^*(\sum_{s=1}^{t}x_{is}x_{js})}{\sum_{s=1}^{t}x_{js}^2}\right|+n^{1/4}\sup_t\left|\frac{\sum_{s=1}^{t}x_{js}u_{s+1}}{\sum_{s=1}^{t}x_{js}^2}\right|. \tag{36}$$

We can now note that

$$n^{1/4}\sup_t\left|\frac{\sum_{s=1}^{t}x_{js}u_{s+1}}{\sum_{s=1}^{t}x_{js}^2}\right|\leq\sup_t\left|\frac{t}{\sum_{s=1}^{t}x_{js}^2}\right|\frac{n^{1/4}}{t}\sup_t\left|\sum_{s=1}^{t}x_{js}u_{s+1}\right|\xrightarrow{p}0 \tag{37}$$

which follows directly from Assumption 2A(iii). This latter assumption now also leads to

$$\sup_t\left|\frac{\sum_{i\in\mathcal{I}^*}\beta_i^*(\sum_{s=1}^{t}x_{is}x_{js})}{\sum_{s=1}^{t}x_{js}^2}-\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{E[x_{it}x_{jt}]}{E[x_{jt}^2]}\right|=o_p(1) \tag{38}$$

as required. (ii) We write

$$\sup_{k_0\leq t\leq n}\left|\frac{\sum_{\ell=k_0}^{t}\hat{\beta}_{j\ell}x_{j\ell}u_{\ell+1}}{\sqrt{n-k_0}}\right|=\frac{1}{\sqrt{1-\pi_0}}\frac{1}{n^{1/4}}\sup_{r\in[\pi_0,1]}\left|\frac{\sum_{l=k_0}^{[nr]}(n^{1/4}\hat{\beta}_{jt})x_{j\ell}u_{\ell+1}}{\sqrt{n}}\right|+o_p(1). \tag{39}$$

The result in part (i) combined with Assumption 2A(iii) allows us to appeal to Theorem 3.3 in Hansen (1993) from which the statement in (ii) follows. For part (iii) it is sufficient to focus on

$$\frac{1}{\sqrt{n-k_0}}\sum_{\ell=k_0}^{n-1}\hat{\beta}_{j\ell}^2x_{j\ell}^2=\frac{1}{\sqrt{1-\pi_0}}\frac{1}{n}\sum_{\ell=k_0}^{n}(\sqrt{n}\hat{\beta}_{j\ell}^2)x_{j\ell}^2+o(1) \tag{40}$$

for which part (i) combined with Assumptions 2A(iii) ensures that

$$\frac{1}{\sqrt{n-k_0}}\sum_{\ell=k_0}^{n-1}\hat{\beta}_{j\ell}^2 x_{j\ell}^2 \xrightarrow{p} \sqrt{1-\pi_0}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{E[x_{it}x_{jt}]}{\sqrt{E[x_{jt}^2]}}\right)^2. \tag{41}$$

Part (iv) follows identical lines to part (iii) and its details are therefore omitted.

PROOF OF PROPOSITION 2A. Using y_{t+1} as in (28) in $\hat{e}_{0,t+1|t}=y_{t+1}^2$ from (34) we have

$$\begin{aligned}
\frac{\sum_{t=k_0}^{k_0+m_0-1}\hat{e}_{0,t+1|t}^2}{\sqrt{n-k_0}} &= \frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{\sum_{t=k_0}^{k_0+m_0-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2}{\sqrt{n}\sqrt{n-k_0}} \\
&\quad + \frac{2}{n^{1/4}}\sum_{i\in\mathcal{I}^*}\beta_i^*\left(\frac{\sum_{t=k_0}^{k_0+m_0-1}x_{it}u_{t+1}}{\sqrt{n-k_0}}\right) \\
&= \frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{\sum_{t=k_0}^{k_0+m_0-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2}{\sqrt{n}\sqrt{n-k_0}} + o_p(1) \\
&= \frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \mu_0\sqrt{1-\pi_0}\; E\left[\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it}\right]^2 + o_p(1)
\end{aligned} \tag{42}$$

where we made repeated use of Assumption 2A(iii). Proceeding as above it also follows that

$$\frac{\sum_{t=k_0+m_0}^{n-1}\hat{e}_{0,t+1|t}^2}{\sqrt{n-k_0}} = \frac{\sum_{t=k_0+m_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + (1-\mu_0)\sqrt{1-\pi_0}\; E\left[\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it}\right]^2 + o_p(1). \tag{43}$$

Next, we focus on $\hat{e}_{j,t+1|t}^2$ given by (34) with y_{t+1} as in (28). We have

$$\begin{aligned}
\frac{\sum_{t=k_0}^{n-1}\hat{e}_{j,t+1|t}^2}{\sqrt{n-k_0}} &= \frac{\sum_{t=k_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{1}{\sqrt{n(n-k_0)}}\sum_{t=k_0}^{n-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2 \\
&\quad + \frac{2}{n^{1/4}\sqrt{n-k_0}}\sum_{i\in\mathcal{I}^*}\beta_i^*(\sum_{t=k_0}^{n-1}x_{it}u_{t+1}) + \frac{1}{\sqrt{n-k_0}}\sum_{t=k_0}^{n-1}\hat{\beta}_{jt}^2 x_{jt}^2 \\
&\quad - \frac{2}{n^{1/4}\sqrt{n-k_0}}\sum_{i\in\mathcal{I}^*}\beta_i^*(\sum_{t=k_0}^{n-1}\hat{\beta}_{jt}x_{jt}x_{it}) \\
&\quad - \frac{2}{\sqrt{n-k_0}}\sum_{t=k_0}^{n-1}\hat{\beta}_{jt}x_{jt}u_{t+1}.
\end{aligned} \tag{44}$$

Appealing to Assumption 2A(iii) and using Lemma A1(ii)-(iii) in (44) also allows us to write

$$\frac{\sum_{t=k_0}^{n-1}\hat{e}_{j,t+1|t}^2}{\sqrt{n-k_0}} = \frac{\sum_{t=k_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + \sqrt{1-\pi_0}\; E\left[\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it}\right]^2 - \frac{\sqrt{1-\pi_0}}{E[x_{jt}^2]}\left(\sum_{i\in\mathcal{I}^*}\beta_i^* E[x_{it}x_{jt}]\right)^2 + o_p(1). \tag{45}$$

Using (42)-(45) in $\mathcal{D}_n(m_0, j)$ now gives

$$\mathcal{D}_n(m_0,j) = \frac{1}{\omega(\mu_0)}\left(\frac{n-k_0}{m_0}\frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{n-k_0}{n-k_0-m_0}\frac{\sum_{t=k_0+m_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}}\right) - 2\frac{\sum_{t=k_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + 2\sqrt{1-\pi_0}\frac{1}{\omega(\mu_0)}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{E[x_{it}x_{jt}]}{\sqrt{E[x_{jt}^2]}}\right)^2 + o_p(1) \tag{46}$$

leading to the desired result.

LEMMA B1. Under Assumption 2B, $\hat{\beta}_{jt}$ as in (32) and $\forall j \in \{1,\ldots,p\}$ we have as $n\to\infty$

(i) $\sup_{r\in[\pi_0,1]}\left|n^{(1+2\alpha)/4}\hat{\beta}_{j,[nr]} - \sum_{i\in\mathcal{I}^*}\beta_i^*\,\frac{\sigma_{v_iv_j}}{\sigma_{v_j}^2}\left(\frac{2c_j}{c_i+c_j}\right)\right| = o_p(1)$

(ii) $\sup_{k_0\le t\le n}\left|\frac{\sum_{\ell=k_0}^{t}\hat{\beta}_{j\ell}x_{j\ell}u_{\ell+1}}{\sqrt{n-k_0}}\right| = o_p(1)$

(iii) $\sup_{k_0\le t\le n}\left|\frac{\sum_{\ell=k_0}^{t}\hat{\beta}_{j\ell}^2x_{j\ell}^2}{\sqrt{n-k_0}} - \sqrt{1-\pi_0}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{\sigma_{v_iv_j}}{\sqrt{\sigma_{v_j}^2}}\frac{\sqrt{2c_j}}{c_i+c_j}\right)^2\right| = o_p(1),$

(iv) $\sup_{k_0\le t\le n}\left|\frac{\beta_i^*}{n^{(1+2\alpha)/4}}\frac{\sum_{\ell=k_0}^{t}\hat{\beta}_{j\ell}x_{i\ell}x_{j\ell}}{\sqrt{n-k_0}} - \sqrt{1-\pi_0}\frac{2c_j\sigma_{v_iv_j}}{(c_i+c_j)\sigma_{v_j}^2}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{\sigma_{v_iv_j}}{c_i+c_j}\right)\right| = o_p(1)$

PROOF OF LEMMA B1. For all four cases the results follow in an identical manner to Lemma A1(i)-(iv) with the use of Assumption 2A(iii) replaced with Assumption 2B(iii) and $n^{1/4}$ replaced with $n^{(1+2\alpha)/4}$.

PROOF OF PROPOSITION 2B. Using y_{t+1} as in (29) in $\hat{e}_{0,t+1|t} = y_{t+1}^2$ from (34) we have

$$\begin{aligned}\frac{\sum_{t=k_0}^{k_0+m_0-1}\hat{e}_{0,t+1|t}^2}{\sqrt{n-k_0}} &= \frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{\sum_{t=k_0}^{k_0+m_0-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2}{n^{(1+2\alpha)/2}\sqrt{n-k_0}} \\ &+ \frac{2}{n^{(1+2\alpha)/4}}\sum_{i\in\mathcal{I}^*}\beta_i^*\left(\frac{\sum_{t=k_0}^{k_0+m_0-1}x_{it}u_{t+1}}{\sqrt{n-k_0}}\right) \\ &= \frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{\sum_{t=k_0}^{k_0+m_0-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2}{n^{(1+2\alpha)/2}\sqrt{n-k_0}} + o_p(1) \qquad (47)\end{aligned}$$

and

$$\begin{aligned}\frac{\sum_{t=k_0+m_0}^{n-1}\hat{e}_{0,t+1|t}^2}{\sqrt{n-k_0}} &= \frac{\sum_{t=k_0+m_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{\sum_{t=k_0+m_0}^{n-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2}{n^{(1+2\alpha)/2}\sqrt{n-k_0}} \\ &+ \frac{2}{n^{(1+2\alpha)/4}}\sum_{i\in\mathcal{I}^*}\beta_i^*\left(\frac{\sum_{t=k_0+m_0}^{n-1}x_{it}u_{t+1}}{\sqrt{n-k_0}}\right) \\ &= \frac{\sum_{t=k_0+m_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{\sum_{t=k_0+m_0}^{n-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2}{n^{(1+2\alpha)/2}\sqrt{n-k_0}} + o_p(1). \qquad (48)\end{aligned}$$

Next, for $\hat{e}_{j,t+1|t}^2$ we have

$$\begin{aligned}\frac{\sum_{t=k_0}^{n-1}\hat{e}_{j,t+1|t}^2}{\sqrt{n-k_0}} &= \frac{\sum_{t=k_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{1}{n^{(1+2\alpha)/2}\sqrt{(n-k_0)}}\sum_{t=k_0}^{n-1}(\sum_{i\in\mathcal{I}^*}\beta_i^* x_{it})^2 \\ &- \sqrt{1-\pi_0}\,\frac{2c_j}{\sigma_{v_j}^2}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{\sigma_{v_iv_j}}{c_i+c_j}\right)^2 + o_p(1). \qquad (49)\end{aligned}$$

Using (47)-(49) in $\mathcal{D}_n(m_0, j)$ and rearranging gives

$$\begin{aligned}\mathcal{D}_n(m_0,j) &= \frac{1}{\omega(\mu_0)}\left(\frac{n-k_0}{m_0}\frac{\sum_{t=k_0}^{k_0+m_0-1}u_{t+1}^2}{\sqrt{n-k_0}} + \frac{n-k_0}{n-k_0-m_0}\frac{\sum_{t=k_0+m_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}}\right) - 2\frac{\sum_{t=k_0}^{n-1}u_{t+1}^2}{\sqrt{n-k_0}} \\ &+ 2\sqrt{1-\pi_0}\frac{1}{\omega(\mu_0)}\left(\sum_{i\in\mathcal{I}^*}\beta_i^*\frac{\sigma_{v_iv_j}}{\sqrt{\sigma_{v_j}^2}}\sqrt{\frac{2c_j}{(c_i+c_j)^2}}\right)^2 + o_p(1) \qquad (50)\end{aligned}$$

leading to the result in Proposition 2B.

LEMMA C1. Under Under Assumption 2C, $\hat{\beta}_{jt}$ as in (33) and $\forall j \in \{1,\ldots,p\}$ we have as $n \to \infty$

(i) $\sup_{r\in[\pi_0,1]}\left|n^{1/4}\hat{\beta}_{j,[nr]}-\frac{1}{E[x_{jt}^2]}\sum_{i\in\mathcal{I}_1^*}\beta_i^*E[x_{it}x_{jt}]\right|=o_p(1)\quad\text{for } j\in\mathcal{I}_1^*$

(ii) $\sup_{r\in[\pi_0,1]}\left|n^{(1+2\alpha)/4}\hat{\beta}_{j,[nr]}-\sum_{i\in\mathcal{I}_2^*}\beta_i^*\,\frac{\sigma_{v_iv_j}}{\sigma_{v_j}^2}\left(\frac{2c_j}{c_i+c_j}\right)\right|=o_p(1)\quad\text{for } j\in\mathcal{I}_2^*.$

PROOF OF LEMMA C1. (i) and (ii) are obtained following the same derivations as LEMMA A1(i) and LEMMA B1(i) and details are therefore omitted. It is here useful to note the distinct behaviour of the slope estimates obtained from the *one predictor at a time* regressions depending on whether the fitted predictor belongs to $\mathcal{I}_1^*$ or $\mathcal{I}_2^*$. This result is driven by the well known phenomenon of asymptotic independence between persistent and stationary predictors.

PROOF OF PROPOSITION 2C. The result in (21) is obtained following identical derivations to (14) and (18) and details are therefore omitted.

PROOF OF PROPOSITION 3. For part (i) of Proposition 3 we focus solely on the case of a single stationary active predictor in the DGP as the remaining scenarios follow identical lines. It is useful to first note that the argmax of $\mathcal{D}_n(m_0,j)$ will be equivalent to $\arg\min_j\mathcal{S}_n(j)$ where

$$\mathcal{S}_n(j)=\frac{\sum_{t=k_0}^{n-1}(\hat{e}_{j,t+1|t}^2-u_{t+1}^2)}{\sqrt{n-k_0}}\quad j=1,\ldots,p. \tag{51}$$

The main result now follows by establishing that $\mathcal{S}_n(j)$ converges to a deterministic limit that is uniquely minimized at $j=j_0$. We continue to operate under the DGP given by (28) with $|\mathcal{I}^*|=1$ (i.e. there is a single active predictor) and with no loss of generality we set that predictor to be x_{1t}. Recalling that $\hat{e}_{j,t+1|t}=y_{t+1}-\hat{\beta}_{jt}x_{jt}$ and using Lemma A1 it immediately follows that for $j=j_0=1$ we have $\mathcal{S}_n(j=1)\xrightarrow{p}0$ while for $j\neq j_0=1$ and using Lemma A1 we have

$$\mathcal{S}_n(j)\xrightarrow{p}(\beta_1^*)^2\sqrt{1-\pi_0}E[x_{1t}^2](1-\rho_{1j}^2)\quad\forall j\neq j_0 \tag{52}$$

which is strictly positive for any predictor different from x_{1t}, thus leading to the required result. (ii) For part (ii) of Proposition 3 we consider the DGP given by (30) and that consists

of predictors with mixed persistence properties. We operate with a pool of p_1 stationary predictors and $p-p_1 \equiv p_2$ persistent predictors and with no loss of generality take $j = 1, \ldots, p_1$ to index the stationary predictors and $j = p_1 + 1, \ldots, p$ the persistent predictors. We assume two active predictors given by $x_{at} = x_{1t}$ and $x_{bt} = x_{p_1+1,t}$ respectively. Using the results in Lemmas A1, B1 and C1 and standard algebra gives

$$\mathcal{S}_n(j = 1) \xrightarrow{p} \sqrt{1-\pi_0}(\beta^*_{p_1+1})^2 \frac{\sigma^2_{v_{p_1+1}}}{2c_{p_1+1}} \tag{53}$$

$$\mathcal{S}_n(j \in \{2, \ldots, p_1\}) \xrightarrow{p} \sqrt{1-\pi_0}(\beta^*_{p_1+1})^2 \frac{\sigma^2_{v_{p_1+1}}}{2c_{p_1+1}} + \sqrt{1-\pi_0}(\beta^*_1)^2 E[x^2_{1t}](1-\rho^2_{1j}) \tag{54}$$

$$\mathcal{S}_n(j = p_1 + 1) \xrightarrow{p} \sqrt{1-\pi_0}(\beta^*_1)^2 E[x^2_{1t}] \tag{55}$$

$$\begin{aligned}\mathcal{S}_n(j \in \{p_1 + 2, \ldots, p\}) \xrightarrow{p} & \sqrt{1-\pi_0}(\beta^*_1)^2 E[x^2_{1t}] + \sqrt{1-\pi_0}(\beta^*_{p_1+1})^2 \frac{\sigma^2_{v_{p_1+1}}}{2c_{p_1+1}} \\ & \times \left(1 - \frac{(\sigma_{p_1+1,j}/(c_{p_1+1} + c_j))^2}{(\sigma^2_{p_1+1}/2c_{p_1+1})(\sigma^2_j/2c_j)}\right)\end{aligned} \tag{56}$$

Comparing (53) with (54) and (55) with (56) implies that $\hat{j}_n$ will asymptotically point to either $j = 1$ or $j = p_1 + 1$ (i.e. one of the two true predictors) as stated.