

# Out of Sample Predictability in Predictive Regressions with Many Predictor Candidates

Jesús Gonzalo\*

Department of Economics, Universidad Carlos III de Madrid  
and

Jean-Yves Pitarakis

Department of Economics, University of Southampton

February 17, 2022

## 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 global 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. An empirical application to key predictors of US economic activity illustrates the usefulness of our methods and highlights the important forward looking role played by the series of manufacturing new orders.

*Keywords:* Forecasting, Nested Models, High Dimensional Predictability.

---

\*The authors gratefully acknowledge financial support from the *Spanish Ministerio de Economía y Competitividad* through grant PID2019-104960GB-100 and the UK *Economic and Social Research Council* through grant ES/W000989/1

# 1 Introduction

Comparing the out of sample predictive accuracy of competing statistical models in data rich environments is an essential component of data science and a key step in the workflow that aims to produce reliable forecasts of an outcome of interest or discriminate between competing hypotheses. A vast body of statistics research over the past decade has been concerned with developing estimation and prediction techniques that can accommodate the availability of large datasets via regularisation techniques and sparsity assumptions on the underlying DGPs. An important objective driving this literature has been to obtain accurate out of sample predictions of a response variable via suitable estimation and covariate selection techniques.

The detection of predictability within linear regression settings has also been the subject of extensive research in the traditional 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, inference and prediction (e.g. persistent nature of many financial and economic predictors, endogeneity considerations, low signal to noise ratios, imbalance in the persistence properties of predictand and predictors). Unlike the above mentioned statistics literature however, predictive regressions as explored in econometrics have been mainly concerned with in-sample goodness of fit measures and traditional significance testing designed to explicitly accommodate these specificities, often in the context of single predictor settings (see Gonzalo and Pitarakis (2019) for a survey of this literature). Some of the early applied research did also highlight the importance of distinguishing between in-sample and out-of-sample predictability in the context of stock return predictability with valuation ratios and related predictors (see Pesaran and Timmermann (1995), Goyal and Welch (2008)).

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, possibly exceeding the available sample size. 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. We are thinking of an environment where one is confronted with not only a potentially large pool of predictors but also with these predictors allowed to display a mixture of dynamic characteristics, some (or all) being highly persistent and others noisier as it commonly occurs in economic and financial data. A macroeconomist interested in forecasts of GDP growth for instance faces hundreds of potentially useful predictors ranging from noisy indicators with very little memory such as financial returns to more persistent series such as interest rates. With an increased availability of unconventional predictor candidates that go beyond traditional macro and financial series (e.g., internet search related data, climate related predictors, etc.) and whose persistence properties are not well known, this aspect of our proposed methods that allows one to remain agnostic about the stochastic properties of predictors is particularly noteworthy.

Our operating 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 novel test statistic suitable for out of sample predictive accuracy comparisons in nested environments. This latter aspect is a particularly important contribution of this research as traditional test statistics for predictive accuracy comparisons are known to fail in nested environments. The resulting series of test statistics are subsequently reassembled into a single aggregate statistic allowing us to test the global 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. We refer to such a predictor as the key player. Although the identification of a single predictor may come across as providing only a limited picture of an underlying *true* specification, it is nevertheless a valuable picture. Parsimonious models can often achieve desirable levels of predictive strength by avoiding distortions due to overfitting for instance.

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). More generally, one is often faced with the need to compare the predictive accuracy of a simple model nested within a richer one. Nested models are one of the most commonly encountered setting in empirical research and help answer fundamental questions such as: does the inclusion of additional predictors significantly improve the predictive power of a smaller model or a non-predictability benchmark? Irrespective of whether one operates in a big data environment combined with heterogeneous predictor types or in a more idealised environment with few well behaved

predictors, conducting out of sample predictive accuracy comparisons between nested models raises many technical challenges that have also not been resolved in a satisfactory way despite a voluminous literature on the subject (e.g. the fact that two nested models collapse into the same specification under the hypothesis of equal predictive accuracy typically results in ill-defined test statistics with degenerate variances).

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 and is followed by a comprehensive local power analysis in Section 4. Section 5 introduces a theoretically supported power enhancing transformation to our proposed tests. Section 6 introduces our key player estimator and studies its asymptotic properties. Section 7 demonstrates the finite sample properties of our methods through a comprehensive simulation based exercise. Finally Section 8 illustrates the usefulness of our methods through an application to the predictability of US economic activity. Section 9 concludes. All technical proofs and further simulations are placed in a supplementary appendix.

## 2 Models and Theory

Let  $\{y_t\}$  denote a scalar random process. Given a sample of size  $n$  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  $\mathbf{x}_t = (x_{1t}, \dots, x_{pt})'$ . 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

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

where  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  and  $u_t$  is a random disturbance term. For later use we also define  $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\beta}')'$  and  $\mathbf{w}_t = (1, \mathbf{x}_t)'$  so that (1) can equivalently be expressed as  $y_{t+1} =$

$\theta'w_t + u_{t+1}$ . The predictors collected in  $\mathbf{x}_t$  can accommodate candidates with different degrees of persistence as commonly encountered in economic applications. 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

$$y_{t+1} = \theta_0 + u_{t+1}. \quad (2)$$

Given the above framework our main goal is to address the following questions. Suppose a researcher has access to a pool of predictors collected within  $\mathbf{x}_t$ . Is at least one of these predictors active relative to the benchmark model in (2)? In the affirmative, is it possible to accurately 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  $\hat{y}_{j,t+1|t}$ ,  $j = 1, \dots, p$ , 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  $t = 1, \dots, k_0$  and progressively expand the estimation window up to  $n - 1$ . Throughout this paper  $k_0$  is taken to be a given a fraction  $\pi_0$  of the sample size and we write  $k_0 = \lfloor n\pi_0 \rfloor$  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  $\hat{\theta}_{jt} = (\sum_{s=1}^t \tilde{w}_{j,s-1} \tilde{w}'_{j,s-1})^{-1} \sum_{s=1}^t \tilde{w}_{j,s-1} y_s$  for  $\tilde{w}_{jt} = (1, w_{jt})$  and  $w_{jt} \in \{x_{1t}, \dots, x_{pt}\}$  with forecasts obtained as  $\hat{y}_{j,t+1|t} = \hat{\theta}'_{jt} \tilde{w}_{jt}$  for  $t = k_0, \dots, n - 1$ . 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  $\{\hat{e}_{j,t+1|t}\}_{t=k_0}^{n-1}$  for  $j = 1, \dots, p$  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  $\beta = 0$  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  $E[\hat{e}_{0,t+1|t}^2 - \hat{e}_{j,t+1|t}^2] > 0$  for at least one  $j \in \{1, 2, \dots, p\}$ .

Addressing the two questions stated above raises four key challenges which the methods developed in this paper address. 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 methods to be computationally feasible to implement despite the availability of a potentially large pool of predictors. The third challenge is related to the need for inferences to remain reliable regardless of the persistence properties of the predictors. The fourth 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 the validity of most of these relies on assumptions that are not tenable in our time series environment with persistent predictors.

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)). The methods we develop in this paper rely on traditional standard normal asymptotics and do not require any simulation based inferences which we view as an important advantage from a practitioner’s perspective.

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), Shi (2015)). 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), McKeague and Qian (2015)). In McKeague and Qian (2015) 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, also assuming a finite pool of predictors. 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 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  $\hat{e}_{0,t+1|t}^2$  and  $\hat{e}_{j,t+1|t}^2$  will be identical under the null hypothesis, we propose to use a sample split estimator for the MSE of the benchmark model instead of the traditional sample mean. This is achieved by split-

ting the evaluation of the forecast errors associated with the benchmark model across two subsamples of size  $m_0$  and  $(n - k_0) - m_0$ . We formulate our test statistic as

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

A simple way to interpret (3) is by observing that it is based on the difference in out of sample MSEs of the benchmark and augmented models as it is the case for most predictive accuracy testing statistics. The key novelty here is that we estimate the out of sample MSE of the benchmark model via a sample-split estimator rather than the full sample mean used for the  $\hat{e}_{j,t}^2$ 's. In generic notation we are essentially estimating an unknown population mean with  $\bar{X}_{split,n} = (\bar{X}_{1n} + \bar{X}_{2n})/2$  rather than  $\bar{X}_n$  and within (3)  $m_0$  refers to the chosen location of the sample split. Throughout this paper we take  $m_0$  to be a user-defined parameter and express it as a fraction  $\mu_0$  of the effective sample size  $n - k_0$ , writing  $m_0 = \lfloor (n - k_0)\mu_0 \rfloor$ .

The motivation for proceeding this way is that the formulation in (3) avoids the variance degeneracy problems associated with nested model comparisons based on traditional DM type formulations. The basic intuition behind the usefulness of (3) is that even if the MSEs of the benchmark and alternative models are identical in the limit, the variances of  $\bar{X}_{split,n}$  and  $\bar{X}_n$  differ thus avoiding the degeneracy problem. More specifically, under the null hypothesis the numerator of (3) will have a non-degenerate positive limiting variance provided that  $\mu_0 \in (0, 1) \setminus \{1/2\}$ . The exclusion of the case  $\mu_0 = 1/2$  is due to the fact that for such a choice of the splitting location we would have  $\bar{X}_{split,n} \equiv \bar{X}_n$  bringing us back to the traditional DM type constructions which are not suitable for nested model comparisons. We may also think of (3) as a way of robustifying predictive accuracy inferences to the nestedness/non-nestedness dichotomy. The 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.

The normaliser  $\hat{\omega}_n$  in the denominator of (3) is understood to be a consistent estimator of the long run variance of the numerator. Strictly speaking, we may have wished to index it as  $\hat{\omega}_{j,n}$  to highlight the fact that it may be estimated using the residuals under the null or the residuals of the single predictor based augmented models (for each  $j$ ). This nuance is naturally inconsequential when it comes to the asymptotics of  $\mathcal{D}_n(m_0, j)$  under the null hypothesis whereby the benchmark model in (2) holds. It may however have important implications when it comes to the finite sample power properties of our proposed test, an issue we study in greater depth further below.

Given the sequences  $\{\hat{\epsilon}_{0,t+1|t}\}_{t=k_0}^{n-1}$ ,  $\{\hat{\epsilon}_{j,t+1|t}\}_{t=k_0}^{n-1}$  and suitable choices for  $\mu_0$  and  $\hat{\omega}_n$  the quantities in (3) can be trivially obtained for each possible predictor  $j$ , resulting in  $p$  such statistics which we aggregate into the following overall statistic

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

A large positive magnitude of  $\bar{\mathcal{D}}_n(m_0)$  is expected to indicate that at least one of the  $p$  predictors improves the predictability of  $y_{t+1}$  relative to the benchmark model.

REMARK 1: One may be tempted to view inferences based on the  $\mathcal{D}_n(m_0, j)$ 's through the lens of multiple comparison and *reality check* type settings and argue for alternative constructs to (4) such as taking the supremum of these  $\mathcal{D}_n(m_0, j)$ 's due for instance to power considerations. It is here important to note however that within our setting and under the null hypothesis of interest to us all  $\mathcal{D}_n(m_0, j)$ 's will be identical in the limit and in turn identical to  $\bar{\mathcal{D}}_n(m_0)$ . A useful analogy that illustrates this latter point is by thinking of the  $p$  components of  $\bar{\mathcal{D}}_n(m_0)$  under the *global* null hypothesis and in the limit as realisations from the same seed of a random number generator. If we were solely interested in null asymptotics and had no power concerns then any of the  $\mathcal{D}_n(m_0, j)$ 's could be viewed as inter-changeable.

In what follows our first objective is to establish the limiting behaviour of (4) under

the null hypothesis that there are no active predictors and which we refer to as the global null. We subsequently assess its local power properties against departures from (2) that are relevant to practitioners. This also allows us to formalise suitable choices for  $\mu_0$  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

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

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 and which we refer to as the *key player*. 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 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. Our 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  $u_t$ 's form a martingale difference sequence with respect to the natural filtration and the sequence  $\eta_t = u_{t+1}^2 - E[u_{t+1}^2]$  satisfies  $\sum_{t=k_0}^{[k_0+(n-k_0)r]} \eta_t / \sqrt{n - k_0} \xrightarrow{d} \phi W(r)$  for  $r \in [0, 1]$  with  $W(r)$  denoting a standard scalar Brownian Motion and  $\phi^2 =$

$\sum_{s=-\infty}^{\infty} \gamma_{\eta}(s) > 0$  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  $\sum_{t=[(n-k_0)r_a]}^{[(n-k_0)r_b]} (\hat{e}_{\ell, t+1|t}^2 - u_{t+1}^2) / \sqrt{n-k_0} = o_p(1) \forall \ell \in \{0, 1, 2, \dots, p\}$ . (iv)  $\mu_0$  satisfies  $\mu_0 \in (0, 1) \setminus \{1/2\}$ .

Assumption 1(i) rules out the presence of serial correlation in the  $u_t$ 's and requires the sequence of demeaned squared errors driving (2) to satisfy a suitable FCLT. We also note from the expression of  $\phi^2$  that the  $u_t$ 's while being serially uncorrelated could be either conditionally homoskedastic or conditionally heteroskedastic. Taking the  $u_t$ 's to be an m.d.s. with an ARCH type variance combined with mild existence of moments requirements would satisfy our environment in 1(i). Assumption 1(ii) requires that a consistent estimator of the long run variance associated with the  $\eta_t$ 's to be available. Letting  $\hat{\eta}_t$  denote a generic estimator of the  $\eta_t$ 's a trivial choice under conditional homoskedasticity would be  $\hat{\phi}^2 = \sum_t \hat{\eta}_t^2 / (n - k_0)$  while under conditional heteroskedasticity one may use a Newey-West type formulation as in Deng and Perron (2008). As we are operating under the global null, such an estimator is readily available using the residuals from the benchmark model in (2). Alternative formulations could also be based on the  $\hat{e}_{0, t+1}$ 's or the  $\hat{e}_{j, t+1}$ 's as also justified by Assumption 1(iii). The key point to make at this stage is that these options will have no bearing on the limiting *null* distribution of  $\bar{\mathcal{D}}_n(m_0)$ . However, such alternative choices may have an important influence on power, an issue we postpone to further below. 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  $\bar{\mathcal{D}}_n(\mu_0)$  to ensure that it has a non-degenerate asymptotic variance. To gain further intuition on this important point it is useful to explicitly evaluate

the limiting variance, say  $\omega^2$ , 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 (see Lemma A1 in the appendix) results in

$$\omega^2 = \frac{(1 - 2\mu_0)^2}{4\mu_0(1 - \mu_0)}\phi^2 \quad (6)$$

so that the availability of a consistent estimator for  $\phi^2$  also ensures that  $\omega^2$  can be estimated consistently provided that  $\mu_0$  satisfies Assumption 1(iv). Expression (6) also highlights the well known variance degeneracy problem one would face in this context if we had instead used the full sample mean associated with the forecast errors of the benchmark model by setting  $\mu_0 = 1/2$ .

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

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

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

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.

## 4 Asymptotic Local 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 interested in exploring are the influence of the persistence properties of predictors on power and the role played by the choice of  $\mu_0$  in  $\overline{\mathcal{D}}_n(\mu_0)$ . We analyse local power within the following parameterisation

$$y_{t+1} = \beta_n' \mathbf{x}_t + u_{t+1} \quad (8)$$

with  $\beta_n = n^{-\gamma} \beta^*$  for  $\beta^* = (\beta_1^*, \dots, \beta_p^*)'$ , noting that within (8) all but one of the  $\beta_i^*$ 's may be zero. We let

$$\mathcal{I}^* = \{1 \leq j \leq p: \beta_j^* \neq 0\} \quad (9)$$

denote the set of active predictors with  $|\mathcal{I}^*| = q \geq 1$  referring to the size of the true model.

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  $\mathbf{x}_t$  to be stationary and ergodic processes (scenario **A**). We then focus on the case where the  $\mathbf{x}_t$ 's are parameterised as persistent processes (scenario **B**). Finally our last scenario sets  $\mathbf{x}_t = (\mathbf{x}_{1,t}, \mathbf{x}_{2,t})'$  with  $\mathbf{x}_{1,t}$  and  $\mathbf{x}_{2,t}$  containing non-persistent and persistent predictors respectively (scenario **C**). In this latter case (8) takes the following form

$$y_{t+1} = \beta_{1n}' \mathbf{x}_{1t} + \beta_{2n}' \mathbf{x}_{2t} + u_{t+1} \quad (10)$$

with  $\mathbf{x}_{1t} = (x_{1,t}, \dots, x_{p_1,t})$  and  $\mathbf{x}_{2t} = (x_{p_1+1,t}, \dots, x_{p,t})$  so that the pool of  $p$  predictors is subdivided into two types. The slope parameter vectors are in turn specified as  $\beta_{1n} = n^{-\gamma_1} \beta_1^*$  for  $\beta_1^* = (\beta_{1,1}^*, \dots, \beta_{1,p_1}^*)'$  and  $\beta_{2n} = n^{-\gamma_2} \beta_2^*$  for  $\beta_2^* = (\beta_{2,p_1+1}^*, \dots, \beta_{2,p}^*)'$ . This mixed environment requires us to also modify the formulation of the active set of predictors included in the DGP. For this purpose we let

$$\mathcal{I}_1^* = \{1 \leq j \leq p_1: \beta_{1,j}^* \neq 0\} \quad (11)$$

$$\mathcal{I}_2^* = \{p_1 + 1 \leq j \leq p: \beta_{2,j}^* \neq 0\} \quad (12)$$

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

Assumption 2A 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  $\sup_{\lambda \in [0,1]} |\sum_{t=1}^{[n\lambda]} x_{it}x_{jt}/n - \lambda E[x_{it}x_{jt}]| = o_p(1)$  and  $\sum_{t=1}^{[n\lambda]} x_{it}u_{t+1}/\sqrt{n} = O_p(1)$  for  $i, j = 1, \dots, p$ .

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. Another important point to make here is that under the local to the null specification in (8) the residual variance estimated from the benchmark model will continue to remain consistent so that Assumption 1(ii) requiring  $\hat{\phi}^2 \xrightarrow{p} \phi^2$  continues to hold when estimated using the residuals from the benchmark model or any of the alternative choices mentioned earlier.

Regarding the scenario with persistent predictors we parameterise these as mildly integrated processes via

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

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  $\Sigma_{vv}$  denote the  $p \times p$  covariance of the  $v'_{jt}$ s and refer to its diagonal components as  $\sigma_{v_j}^2$  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  $\sum_{t=1}^{[n\lambda]} x_{it}x_{jt}/n^{1+\alpha} \xrightarrow{p} \lambda\sigma_{v_i v_j}/(c_i + c_j)$ ,  $\sum_{t=1}^{[n\lambda]} x_{jt}^2/n^{1+\alpha} \xrightarrow{p} \lambda\sigma_{v_j}^2/(2c_j)$  and  $\sum_{t=1}^{[n\lambda]} x_{jt}u_{t+1}/n^{\frac{1+\alpha}{2}} = O_p(1)$  for  $i, j = 1, \dots, p$ .

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  $n \rightarrow \infty$  we have

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

where

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

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. The expression in (14) and its counterparts under persistence presented further below offer novel insights on the asymptotic behaviour of predictive accuracy comparisons not explored in the existing literature. We can also observe that power is monotonic as the non-centrality component in (14) is non-decreasing as the slope parameters of the active predictors increase. Another implication of (14) is that under fixed alternatives  $\bar{\mathcal{D}}_n(\mu_0) \rightarrow \infty$  and more specifically

$$\bar{\mathcal{D}}_n(\mu_0) \stackrel{H_1}{\equiv} O_p(\sqrt{n}). \quad (16)$$

REMARK 2. 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 due to the way the test statistic  $\bar{\mathcal{D}}_n(\mu_0)$  has been constructed or to our inference framework in general. 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

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

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 pseudo-signal 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  $\rho_{a,a}^2 = 1$  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, all other things kept equal.

Another important implication of (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 all other things being equal the presence of persistent predictors in the pool will improve the detection ability of our test. We can also note that the role of persistence may manifest itself not only via  $E[x_{at}^2]$  but also via  $\rho_{a,j}^2$  due to the spurious correlation phenomenon characterising persistent processes. These issues are explored in the next proposition.

***Local Power under Persistence (scenario B)***

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

parameters  $\beta_i = n^{-(1+2\alpha)/4}\beta_i^*$  for  $i \in \mathcal{I}^*$  and as  $n \rightarrow \infty$  we have

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

with  $g(\mu_0, \pi_0, \phi)$  as in (15).

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

$$\bar{\mathcal{D}}_n(\mu_0) \stackrel{H_1}{\equiv} O_p(n^{\frac{1+2\alpha}{2}}). \quad (19)$$

If we were to restrict all predictors to have the same non-centrality parameter, say  $c_i = c \forall i = 1, \dots, p$ , (18) reduces to

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

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

### **Local Power under Mixed Predictors (scenario C)**

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 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  $n \rightarrow \infty$*

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

with  $g(\mu_0, \pi_0, \phi)$  as in (15).

REMARK 3. The expressions in (14), (18), (21) provide useful insights on suitable choices of  $\mu_0$  when constructing our test statistic. As  $\mu_0$  affects local power via  $g(\mu_0, \pi_0, \phi)$  a choice in the vicinity of 0.5 is expected to lead to the most favourable power outcomes.

## 5 A Power enhancing modification of $\overline{\mathcal{D}}_n(\mu_0)$

We here consider a modification of  $\overline{\mathcal{D}}_n(\mu_0)$  designed to enhance its power without affecting its null distribution. Our proposal is in the spirit of Fan, Liao and Yao (2015) and involves augmenting the 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  $d_{nj} = \sum_{t=k_0}^{n-1} (\hat{\epsilon}_{0,t+1|t} - \hat{\epsilon}_{j,t+1|t})^2 / (n - k_0)$ . Within our nested context and under the null of the benchmark model we have  $d_{nj} = O_p(n^{-1/2}) \forall j = 1, \dots, p$  while under the alternative whereby the true model contains at least one active predictor we have  $\tilde{d}_{nj} \equiv \sqrt{n - k_0} d_{nj} / \hat{\omega}_n = O_p(1) \forall j = 1, \dots, p$  with the associated limiting random variable being strictly positive. This prompts us to propose the following augmentation to  $\overline{\mathcal{D}}_n(m_0)$

$$\overline{\mathcal{D}}_n^d(m_0) = \frac{1}{p} \sum_{j=1}^p (\mathcal{D}_n(m_0, j) + \tilde{d}_{nj}) \quad (22)$$

which we expect to be power enhancing while being size-neutral. Noting that

$$\overline{\mathcal{D}}_n^d(m_0) - \overline{\mathcal{D}}_n(m_0) = \frac{1}{p} \sum_{j=1}^p \tilde{d}_{nj} \quad (23)$$

Proposition 3 below formalises these observations.

**Proposition 3.** (i) Under Assumptions 1(i)-(iv) and the null hypothesis we have  $\overline{\mathcal{D}}_n^d(m_0) - \overline{\mathcal{D}}_n(m_0) \xrightarrow{p} 0$  as  $n \rightarrow \infty$ . (ii) Under Assumptions 2A, 2B or 2C we have  $\overline{\mathcal{D}}_n^d(m_0) - \overline{\mathcal{D}}_n(m_0) \xrightarrow{p} Q_\ell > 0$ ,  $\ell = A, B, C$  with  $Q_\ell$  given by the second component in the right hand side of (14), (18) and (21) respectively.

A key implication of Proposition 3 (ii) is that a test of size  $\alpha$  based on  $\mathcal{D}_n^d(\mu_0)$  will be strictly preferable in terms of local power to a test of the same size based on  $\mathcal{D}_n(\mu_0)$ . A more formal comparison using Pitman's asymptotic relative efficiency is also informative as it takes a particularly simple form in the present context. Indeed, our local power results in (14), (18), (21) combined with Proposition 3(ii) above imply that

$$\text{ARE}(\mathcal{D}_n(\mu_0), \mathcal{D}_n^d(\mu_0)) = 1/2 < 1 \quad (24)$$

irrespective of any model specific parameters.

REMARK 4: The above power enhancing transformation based on  $\tilde{d}_{nj}$  is analogous to implementing an adjustment to the forecast errors associated with the larger models. More specifically, implementing our main test statistic in (3) with the  $\hat{e}_{j,t+1}^2$ 's replaced with say  $\tilde{e}_{j,t+1}^2 = \hat{e}_{j,t+1}^2 - (\hat{e}_{0,t+1} - \hat{e}_{j,t+1})^2$  results in a formulation that is algebraically identical to (22). It is now interesting to observe that these  $\tilde{e}_{j,t+1}^2$ 's are essentially adjusting the  $\hat{e}_{j,t+1}^2$ 's for estimation noise coming from the estimation of the larger model when its true parameters are zero. This is precisely the motivation behind the well known Clark and West adjustment to equal predictive accuracy tests proposed in Clark and West (2007). Unlike the setting in Clark and West (2007) however our proposed test statistics result in formal normal limits rather than approximate ones.

## 6 Detecting the key player

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  $\hat{j}_n \in \arg \max_{j=1,\dots,p} \mathcal{D}_n(m_0, j)$  and  $\hat{j}_n^d \in \arg \max_{j=1,\dots,p} \mathcal{D}_n^d(m_0, j)$  which select 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 these estimators 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 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 these key player estimators is summarised in the following Proposition.

**Proposition 4.** (i) Under Assumptions 2A or 2B and as  $n \rightarrow \infty$  we have  $\{\hat{j}_n, \hat{j}_n^d\} \xrightarrow{p} j_0 \in \mathcal{I}^*$  for  $q \geq 1$ . (ii) Under Assumptions 2C and as  $n \rightarrow \infty$  we have  $\{\hat{j}_n, \hat{j}_n^d\} \xrightarrow{p} j_0 \in \mathcal{I}_1^* \cup \mathcal{I}_2^*$ .

When the DGP consists solely of a single predictor (stationary or persistent), part (i) of Proposition 4 implies that  $\hat{j}_n$  or  $\hat{j}_n^d$  will be consistent for that true predictor asymptotically. When there are multiple predictors of the same type the same result implies that  $\hat{j}_n$  or  $\hat{j}_n^d$  remain consistent for one of the  $q > 1$  active predictors i.e.  $\hat{j}_n$  or  $\hat{j}_n^d$  is consistent for one of the true components in  $\mathcal{I}^*$ . Part (ii) of Proposition 4 relates to a scenario with mixed active predictors and states that in such a mixed setting  $\hat{j}_n$  or  $\hat{j}_n^d$  will continue to point to one of the true predictors which may come from either of the two sets.

Using the results provided in the proof of Proposition 4 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 4(ii) clearly applies and implies that  $\hat{j}_n$  or  $\hat{j}_n^d$  will asymptotically point to either  $x_{at}$  or  $x_{bt}$ . More specifically (see proof of Proposition 4), we have that  $\hat{j}_n$  or  $\hat{j}_n^d$  will asymptotically point to  $x_{bt}$  (the persistent predictor) if

$$(\beta_b^*)^2 > \frac{E[x_{at}^2]}{(\sigma_{v_b}^2/2c_b)}(\beta_a^*)^2 \quad (25)$$

and to  $x_{at}$  otherwise. It is now interesting to observe from (25) that  $\hat{j}_n$  or  $\hat{j}_n^d$  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  $x_{bt}$ . More generally, these results suggest that the key player selected by our proposed methods will depend on the relative magnitude of its associated slope combined with its relative variance (relative to the remaining active predictors). Thus when it comes to where  $\hat{j}_n$  or  $\hat{j}_n^d$  point asymptotically, there will be a

trade-off between slope strength and variance dominance.

## 7 Implementation and Experimental Properties

This section aims to document the empirical properties of our proposed test and key player estimator in finite samples. Given our theoretical analysis about the superior local power properties of  $\mathcal{D}_n^d(\mu_0)$  versus  $\mathcal{D}_n(\mu_0)$  we concentrate our discussion on  $\mathcal{D}_n^d(\mu_0)$  which we then follow with experiments documenting the correct decision frequencies associated with the proposed key player estimators. A **supplementary appendix** accompanying this paper provides additional simulation based illustrations of the finite sample behaviour of our proposed tests.

**Implementation:** The implementation of our test statistics in (3) and (4) requires the availability of an estimator of the long run variance associated with the numerator of (3) which we generically referred to as  $\hat{\omega}_n^2$ . The asymptotic outcomes documented above operated under the assumption that an estimator satisfying  $\hat{\omega}_n^2 \xrightarrow{P} \omega^2$  was available. Given the expression of  $\omega^2$  obtained in (6) and the fact that we operate under given  $\mu_0$  it is also clear that a consistent estimator of  $\phi^2$  would also ensure the availability of a consistent estimator of  $\hat{\omega}_n^2$ .

We may consider two alternative estimators of the long run variance in (3) using either residuals from the null model or the marginal regressions considered under the alternative. From Assumption 1(iii) it is straightforward to note that these will be asymptotically equivalent under the null hypothesis of interest but they may result in potentially important differences in finite samples, when it comes to power in particular. For conditionally homoskedastic  $\eta_t$ 's we consider

$$\hat{\omega}_n^{2,a} = \frac{(1 - 2\mu_0)^2}{4\mu_0(1 - \mu_0)} \frac{\sum_{t=k_0}^{n-1} (\hat{e}_{0,t+1}^2 - \overline{\hat{e}_0^2})^2}{n - k_0}, \quad (26)$$

$$\hat{\omega}_n^{2,b} = \frac{(1 - 2\mu_0)^2}{4\mu_0(1 - \mu_0)} \frac{\sum_{t=k_0}^{n-1} (\hat{e}_{j,t+1}^2 - \overline{\hat{e}_j^2})^2}{n - k_0}. \quad (27)$$

which can also be readily adapted to accommodate conditionally heteroskedastic  $u_t^i s$  (equivalently, serial correlation in the  $\eta_t^i s$ ) using Newey-West type formulations. Letting  $\hat{\eta}_t$  denote a generic estimator of  $\eta_t$ , such an estimator for the counterpart to  $\hat{\omega}_n^{2,a}$  above would be given by

$$\tilde{\omega}_n^{2,a} = \frac{(1 - 2\mu_0)^2}{4\mu_0(1 - \mu_0)} \sum_{s=-m}^m \left(1 - \left|\frac{s}{n - k_0}\right|\right) \hat{\gamma}_\eta(s) \quad (28)$$

where  $\hat{\gamma}_\eta(s) = \sum \hat{\eta}_t \hat{\eta}_{t-s} / n - k_0$  for  $\hat{\eta}_t = \hat{u}_{0,t+1}^2 - \hat{\sigma}_{0,u}^2$  and similarly for (27). Note that (28) specialises to (26) under conditional homoskedasticity, whereby  $\gamma_\eta(s) = 0 \forall s \neq 0$  and  $m$  refers to a suitable bandwidth for which one may consider the rule of thumb  $m = m_n = 0.75(n - k_0)^{1/3}$ . In the simulations presented below we analyse finite sample size and power using the variance normaliser in (27).

**Empirical Size:** The DGP is given by the benchmark specification in (2). The pool of  $p$  predictors is taken to follow the VAR(1) process  $\mathbf{x}_t = \mathbf{\Phi} \mathbf{x}_{t-1} + \mathbf{v}_t$  with  $\mathbf{v}_t \sim N(0, \mathbf{\Sigma}_{vv})$  which we parameterise in ways that can distinguish between uncorrelated, weakly correlated and strongly correlated predictors. We consider the following scenarios for the persistence properties of the predictors: (A)  $\mathbf{\Phi} = 0.50 \mathbf{I}_p$ , (B)  $\mathbf{\Phi} = 0.95 \mathbf{I}_p$  and (C)  $\mathbf{\Phi} = (\mathbf{\Phi}_1, \mathbf{\Phi}_2)$ ,  $\mathbf{\Phi}_1 = 0.5 \mathbf{I}_{p_1}$ ,  $\mathbf{\Phi}_2 = 0.95 \mathbf{I}_{p-p_1}$ . Letting  $\mathbf{\Omega}$  denote the covariance matrix of  $(u_t, \mathbf{v}_t)'$  we write

$$\mathbf{\Omega}_{p+1 \times p+1} = \begin{pmatrix} \sigma_u^2 & \boldsymbol{\sigma}'_{uv} \\ \boldsymbol{\sigma}_{uv} & \mathbf{\Sigma}_{vv} \end{pmatrix} \quad (29)$$

where  $\boldsymbol{\sigma}_{uv} = (\sigma_{uv_1}, \sigma_{uv_2}, \dots, \sigma_{uv_p})'$  collects the covariances between the shocks to  $y_t$  and the shocks to individual predictors and  $\mathbf{\Sigma}_{vv}$  is the  $p \times p$  covariance matrix of the  $p$  predictors. Our experiments involving either purely stationary, purely persistent or mixed predictors are conducted across three configurations of  $\mathbf{\Omega}$ : (i)  $\mathbf{\Omega}_0$ :  $\sigma_u^2 = 1, \boldsymbol{\sigma}_{uv} = \mathbf{0}_{p \times 1}, \mathbf{\Sigma}_{vv} = \mathbf{I}_p$ , (ii)  $\mathbf{\Omega}_1$ :  $\sigma_u^2 = 1, \boldsymbol{\sigma}_{uv} = \mathbf{0}_{p \times 1}, \mathbf{\Sigma}_{vv} = [0.5^{|i-j|}]_{i,j}$  and (iii)  $\mathbf{\Omega}_2$ :  $\sigma_u^2 = 1, \boldsymbol{\sigma}_{uv} = [(-0.5)^j]_j, \mathbf{\Sigma}_{vv} = [0.5^{|i-j|}]_{i,j}$ . Accordingly we label these size related DGPs as **(A-i)-(A-iii)**, **(B-i)-(B-iii)** and **(C-i)-(C-iii)**. Scenario (i) forces all  $p$  predictors to be uncorrelated between



themselves. It also requires the shocks to the predictors and the shocks to the predictand to be uncorrelated as does scenario (ii). In this latter case predictors are now allowed to be correlated. Finally scenario (iii) allows for the shocks to predictand and predictors to be contemporaneously correlated.

Empirical size outcomes are obtained for  $p \in \{10, 50, 500\}$  and samples of size  $n = 500$  with  $\pi_0 = 0.25$  used as the starting point for generating recursive forecasts i.e.  $n - k_0 = 375$ . For the sample-split location of our test statistic we experiment with  $\mu_0 \in \{0.35, 0.40, 0.45\}$ . Results are collected in Table 1 below using 5000 Monte-Carlo replications and a nominal size of 10%. It is worth pointing out that the case  $p = 500$  implies an environment where the number of predictors exceeds the effective sample size of  $n - k_0 = 375$ . Recalling that our main result in Proposition 1 is obtained under  $n \rightarrow \infty$ , our chosen parameterisations of the pair  $(n, p)$  is meant to illustrate the finite sample adequacy of our asymptotics even when  $n$  lies below  $p$  or is near  $p$ .

From Table 1 we can note that the *one predictor at a time* approach based on  $\mathcal{D}_n^d(\mu_0)$  appears to be robust to the dimension of the predictor pool with almost identical size estimates obtained across  $p = 10$ ,  $p = 50$  and  $p = 500$  predictors. This highlights the excellent approximation provided by our asymptotics even when  $n - k_0$  is smaller than  $p$ . Outcomes can also be seen to be robust to predictor persistence as expected from our result in Proposition 1. Equally importantly, we can highlight the fact that the chosen sample split location  $\mu_0$  has very little influence on outcomes with almost identical empirical sizes obtained across all chosen magnitudes of  $\mu_0$ . This is particularly important as our earlier local power analysis suggested that choosing  $\mu_0$  in the vicinity of 0.5 should result in better power outcomes, all other things being equal.

**Empirical Power:** We consider predictive regressions with up to four active predictors parameterised as

$$y_{t+1} = \beta_{an}x_{a,t} + \beta_{bn}x_{b,t} + \beta_{cn}x_{c,t} + \beta_{dn}x_{d,t} + u_{t+1} \quad (30)$$

Table 1: Empirical Size of  $\mathcal{D}_n^d(\mu_0)$  (10% Nominal)

$\mu_0$	p=10	p=50	p=500	p=10	p=50	p=500	p=10	p=50	p=500
	A(i)			A(ii)			A(iii)		
0.35	0.106	0.102	0.103	0.107	0.103	0.103	0.105	0.103	0.103
0.40	0.108	0.103	0.104	0.109	0.105	0.106	0.108	0.105	0.106
0.45	0.109	0.093	0.100	0.115	0.093	0.100	0.114	0.094	0.100
	B(i)			B(ii)			B(iii)		
0.35	0.105	0.103	0.103	0.106	0.102	0.103	0.107	0.102	0.103
0.40	0.114	0.105	0.105	0.114	0.104	0.104	0.114	0.104	0.104
0.45	0.116	0.099	0.101	0.119	0.100	0.103	0.122	0.100	0.103
	C(i)			C(ii)			C(iii)		
0.35	0.106	0.103	0.103	0.105	0.102	0.102	0.105	0.102	0.102
0.40	0.110	0.104	0.103	0.109	0.106	0.103	0.109	0.106	0.103
0.45	0.115	0.093	0.100	0.119	0.098	0.102	0.118	0.099	0.102

with  $\beta_{an} = \beta_a^*/n^{0.25}$ ,  $\beta_{bn} = \beta_b^*/n^{0.25}$ ,  $\beta_{cn} = \beta_c^*/n^{0.675}$ , and  $\beta_{dn} = \beta_d^*/n^{0.675}$ . The two predictors labelled as  $\{a, b\}$  are chosen to be non-persistent while the predictors labelled as  $\{c, d\}$  will have more persistence. Accordingly it will be understood that the pool of predictors to which  $\{x_{at}, x_{bt}, x_{ct}, x_{dt}\}$  belong is generated from a VAR(1) parameterised as in (C-iii) above. The active predictors  $\{x_{at}, x_{bt}\}$  belong to the first set of  $p_1$  predictors and  $\{x_{ct}, x_{dt}\}$  belong to the second set of  $p - p_1$  predictors. We take  $x_{at} = x_{1t}$ ,  $x_{bt} = x_{2t}$  and  $x_{ct} = x_{p_1+1,t}$ ,  $x_{dt} = x_{p_1+2,t}$ . The *local* parameterisations of the slope parameters are chosen in a way to be compatible with our earlier local power analysis where we documented local departures of  $n^{-0.25}$  under the stationary setting and  $n^{-(1+2\alpha)/4}$  under mild integratedness. For this latter case, setting  $\alpha = 0.85$  in the mildly integrated process in (3) results in  $(1+2\alpha)/4 = 0.675$ . As these persistent predictors are driven by the VAR(1) component with slopes 0.95 we may also note that this corresponds roughly to  $(1 - c/n^\alpha) = (1 - 10/500^{0.85}) \approx 0.95$ .

We consider 3 DGP configurations: **(i)**  $\beta_a^* \in \{2, 3, 4, 5\}$ ,  $\beta_b^* \in \{5, 6, 7, 8\}$ ,  $\beta_c^* = 0$ ,  $\beta_d^* = 0$ , **(ii-a)**  $\beta_a^* = 0$ ,  $\beta_b^* = 0$ ,  $\beta_c^* \in \{2, 3, 4, 5\}$ ,  $\beta_d^* \in \{5, 6, 7, 8\}$  and **(ii-b)**  $\beta_a^* = 0$ ,  $\beta_b^* = 0$ ,  $\beta_c^* \in \{5, 6, 7, 8\}$ ,  $\beta_d^* \in \{8, 9, 10, 11\}$ . Scenario (i) involves two active predictors  $x_{at}$  and  $x_{bt}$  that are not persistent (i.e., selected from the pool of  $p_1$  predictors that follow autoregressive processes with slopes equal to 0.5). Scenarios (ii-a) and (ii-b) involve two active predictors  $x_{ct}$  and  $x_{dt}$  selected from the remaining pool of  $p - p_1$  persistent predictors. Relative to (ii-a), DGP (ii-b) is characterised by a stronger signal to noise ratio.

It is also useful to point out that with  $n = 500$  the chosen slope parameterisations translate into  $\beta_{an} \in \{0.422, 0.634, 0.845, 1.057\}$  and  $\beta_{bn} \in \{1.057, 1.269, 1.480, 1.692\}$  for scenario (i),  $\beta_{cn} \in \{0.030, 0.045, 0.060, 0.075\}$  and  $\beta_{dn} \in \{0.075, 0.090, 0.106, 0.121\}$  for scenario (ii-a),  $\beta_{cn} \in \{0.075, 0.090, 0.106, 0.121\}$  and  $\beta_{dn} \in \{0.121, 0.136, 0.151, 0.166\}$  for scenario (ii-b). These highlight the fact that power is evaluated as the DGP moves further away from the null under a fixed sample size set at  $n = 500$ . We also note that the above slope magnitudes span low, medium and high signal to noise ratios.

Our power experiments are implemented using a predictor pool of  $p = 100$  predictors with the first half consisting of autoregressive processes with slopes set at 0.50 and the second half having slopes of 0.95. All experiments are implemented using the covariance structure labelled as  $\mathbf{\Omega}_2$  above. Outcomes associated with  $\mathcal{D}_n^d(\mu_0)$  are collected in Table 2.

We note that power increases towards 100% as the slope parameters move away from the null (each column of Table 2 corresponds to one pair of slopes, and moving rightwards along the table illustrates power progression for larger departures from the null). Under DGP(i) empirical powers lie in the vicinity of 100% for  $\mu_0 \geq 0.40$  and across all slope parameterisations. DGP(ii-a) is associated with much weaker signal to noise ratios which as expected translate into much less favourable empirical powers. Nevertheless we do note powers as high as 75% even in this context. Here it is important to recall that this DGP consists solely of persistent predictors. Had we used the slope magnitudes  $(\beta_{an}, \beta_{bn})$

instead of  $(\beta_{cn}, \beta_{dn})$  in this context all empirical powers would have resulted in 100% or nearly 100% correct decision frequencies. This can also be inferred from the outcomes based on DGP (ii-b) which uses larger slope magnitudes for the same persistent predictor scenario. Under  $(\beta_{cn}, \beta_{dn}) = (0.121, 0.166)$  which are much lower than the most favourable slope pairs considered in DGP(i) for instance we note empirical power outcomes in excess of 90%.

Table 2: Empirical Power of  $\mathcal{D}_n^d(\mu_0)$  under DGPs (i)-(ii)

DGP (i)				
$\beta_{an}$	0.423	0.634	0.846	1.057
$\beta_{bn}$	1.057	1.269	1.480	1.692
$\mu_0 = 0.35$	0.829	0.936	0.976	0.991
$\mu_0 = 0.40$	0.977	0.997	1.000	1.000
$\mu_0 = 0.45$	1.000	1.000	1.000	1.000
DGP (ii-a)				
$\beta_{cn}$	0.030	0.045	0.060	0.075
$\beta_{dn}$	0.075	0.090	0.106	0.121
$\mu_0 = 0.35$	0.174	0.231	0.304	0.346
$\mu_0 = 0.40$	0.217	0.296	0.391	0.464
$\mu_0 = 0.45$	0.349	0.500	0.641	0.747
DGP (ii-b)				
$\beta_{cn}$	0.075	0.090	0.106	0.121
$\beta_{dn}$	0.121	0.136	0.151	0.166
$\mu_0 = 0.35$	0.337	0.399	0.464	0.492
$\mu_0 = 0.40$	0.455	0.533	0.604	0.655
$\mu_0 = 0.45$	0.742	0.827	0.885	0.924

***Finite Sample Properties of the key player estimator:*** In this last set of experiments we illustrate the result in Proposition 3 by documenting the behaviour of the proposed *key player* estimator as  $n$  is allowed to grow. We base our evaluation of  $\hat{j}_n^d$  on its

ability to point to one of the true underlying active predictors. Although this estimator would be typically implemented upon rejection of the null hypothesis our empirical results below are understood to be unconditional in the sense that the correct decision frequencies associated with  $\hat{j}_n^d$ 's are averaged across all realisations rather than the ones associated with rejections of the null. This is justified here on the basis that  $\hat{j}_n^d$  may still point to the true predictor (or one of the true predictors) even if the underlying inferences based on  $\mathcal{D}_n^d(\mu_0)$  do not result in a rejection of the null. Differently put, the realised magnitude of  $\mathcal{D}_n^d(\mu_0)$  could still be the largest amongst the  $j = 1, \dots, p$  realisations even if it falls below the null hypothesis rejection threshold.

The DGPs parallel the specifications labelled as (i)-(ii) in our earlier power analysis. The number of predictors is set at  $p = 100$  throughout. The specific parameterisations are: (i)  $(\beta_{an}, \beta_{bn}) = (0.634, 1.269)$ ,  $x_{at} = x_{1t}$ ,  $x_{bt} = x_{2t}$ , (ii-a)  $(\beta_{cn}, \beta_{dn}) = (0.075, 0.121)$ ,  $x_{ct} = x_{51,t}$ ,  $x_{dt} = x_{52,t}$ , (ii-b)  $(\beta_{cn}, \beta_{dn}) = (0.106, 0.151)$ ,  $x_{ct} = x_{51,t}$ ,  $x_{dt} = x_{52,t}$ . We implement the above experiments across samples of size  $n = 100$  and  $n = 200$ . In DGP(i) the active predictors are  $j_0 \in \{1, 2\}$ . In DGPs (ii-a) and (ii-b) we have  $j_0 \in \{51, 52\}$ . Results are collected in Table 3 which displays relevant correct decision frequencies. An important primary observation we can make is the similarity of outcomes across the different sample split locations (i.e., choices for  $\mu_0$ ) whose choice clearly does not matter much. For DGP (i) our estimator picks up  $j_0 = 2$  in excess of 99% of the times even under  $(n = 100, p = 100)$ . The preference for  $j_0 = 2$  over  $j_0 = 1$  which is also included as an active predictor is due to the much larger slope associated with  $x_{2t}$  combined with the fact that these predictors have the same variance. For DGP (ii-a) which is driven by the two active predictors  $x_{51,t}$  and  $x_{52,t}$  we note that the bulk of  $\hat{j}_n^d$ 's decision frequencies converge towards  $x_{52,t}$  (e.g., about 70% under  $n = 200$ ). This can again be explained by the fact that  $x_{52,t}$  has a stronger signal due to its larger slope parameter. Similar outcomes also characterise DGP (ii-b).

Table 3: Key Player Estimators: Frequency of detection of active predictors

	$\mu_0 = 0.35$	$\mu_0 = 0.40$	$\mu_0 = 0.45$	$\mu_0 = 0.35$	$\mu_0 = 0.40$	$\mu_0 = 0.45$
	n=100			n=200		
<i>DGP – (i), <math>j_0 \in \{1, 2\}</math></i>						
$\hat{j}_n^d = 1$	0.009	0.009	0.009	0.000	0.000	0.000
$\hat{j}_n^d = 2$	0.991	0.991	0.991	1.000	1.000	1.000
<i>DGP – (ii – a), <math>j_0 \in \{51, 52\}</math></i>						
$\hat{j}_n^d = 51$	0.184	0.182	0.183	0.235	0.235	0.234
$\hat{j}_n^d = 52$	0.413	0.414	0.413	0.703	0.702	0.703
$\hat{j}_n^d \neq \{51, 52\}$	0.403	0.404	0.404	0.062	0.063	0.063
<i>DGP – (ii – b), <math>j_0 \in \{51, 52\}</math></i>						
$\hat{j}_n^d = 51$	0.252	0.253	0.252	0.270	0.271	0.270
$\hat{j}_n^d = 52$	0.508	0.508	0.508	0.705	0.705	0.705
$\hat{j}_n^d \neq \{51, 52\}$	0.239	0.239	0.239	0.025	0.025	0.025

## 8 Application: Predictability of Economic Activity

We apply our methods to the predictability of US economic activity and the monthly growth rate in industrial production in particular. The predictor pool consists of 130 lagged monthly series drawn from the FRED-MD database whose detailed constituents are discussed in McCracken and Ng (2016). These series have also been transformed and outliers processed as documented in McCracken and Ng (2016). FRED-MD consists of eight groups of time series and is closely aligned with the early Stock and Watson dataset (Stock and Watson (2002)): (1) output and income, (2) labor market, (3) housing, (4) consumption, (5) money and credit, (6) interest and exchange rates, (7) prices and (8) stock-market. Our selection of 130 predictors follows Giannone, Lenza and Primiceri (2021, GLP2021) using the data vintage provided by the authors and the same sample range of February 1960 to December 2014. This allows us to make comparisons between the inferences developed in this paper and existing findings in the literature. Recalling that

our forecasts are generated recursively we set the starting point of the first recursion at the 165<sup>th</sup> month (i.e.,  $k_0 = \lceil 658(0.25) \rceil = 165$ ). We implement our inferences across  $\mu_0 \in \{0.30, 0.35, 0.40, 0.45\}$  and consider both the “raw”  $\mathcal{D}_n(\mu_0)$  statistic and its power enhanced counterpart  $\mathcal{D}_n^d(\mu_0)$ . We also implement these two test statistic formulations using variance normalisers based on residuals obtained under the null and the alternative. Accordingly we label these four versions as  $\mathcal{D}_n(\mu_0)_0$ ,  $\mathcal{D}_n^d(\mu_0)_0$ ,  $\mathcal{D}_n(\mu_0)_1$  and  $\mathcal{D}_n^d(\mu_0)_1$ .

Table 4 presents the p-values associated with testing the global null of no predictability. We note strong rejections of the null across all implementations. US economic activity is clearly predictable using past macroeconomic information. More importantly our key-player estimator based on either of the four formulations of our test statistic identifies the same series given by the ISM: New Orders Index (coded as NAPMNOI and numbered as 61 in FRED-MD). This is a monthly index published by the Institute for Supply Management and informing about the number of surveyed businesses reporting increased customer orders relative to the previous month. It is interesting that our key-player estimator has pointed to this predictor as NAPMNOI is the first forward looking indicator made public early each month. This variable has been found to be an important predictor of US recessions in Liu and Moench (2016) but perhaps more interestingly here, NAPMNOI can also be seen to be the most important predictor detected through the Bayesian methods developed in GLP2021. In their Figure 6 (GLP2021, p. 2425) this variable can be seen to be amongst the ones having the highest probabilities of inclusion. The picture is even clearer in Fava and Lopes (2020, FL2020) who have reconsidered GLP2021’s study by evaluating its sensitivity to the chosen priors. Figure 7 in FL2020 clearly points to the 61<sup>th</sup> predictor as the one having a 100% probability of inclusion as a predictor of the US growth rate in industrial production.

Although going beyond the detection of a key-player is beyond our scope it is interesting to evaluate leading predictors beyond the argmax of  $\mathcal{D}_n^d(\mu_0)$ . Table 5 isolates the top 6

Table 4: p-values

$\mu_0$	0.30	0.35	0.40	0.45		0.30	0.35	0.40	0.45
$\mathcal{D}_n(\mu_0)_0$	0.000	0.001	0.003	0.001	$\mathcal{D}_n(\mu_0)_1$	0.000	0.001	0.002	0.000
$\mathcal{D}_n^d(\mu_0)_0$	0.000	0.000	0.000	0.000	$\mathcal{D}_n^d(\mu_0)_1$	0.000	0.000	0.000	0.000

predictors leading to the highest magnitudes of  $\mathcal{D}_n^d(\mu_0)$ . We note a cluster of interest rate related predictors and a labour market indicator. It is again interesting to point out that predictors 92 and 39 also appear amongst the predictors with the highest probabilities of being included in both GLP2021 and FL2020. The heatmap presented in Figure 6 of GLP2021 (p. 2425) does also show a clustering of active predictors with IDs in the 90s range as in our Table 5.

Table 5: Key Player and Top 5 predictors

Fred-MD ID	Fred-MD Code	Description
61	NAPMNOI	ISM: New Orders Index
96	T1YFFM	1-Year Treasury C Minus FEDFUNDS
95	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS
92	BAA	Moodys Seasoned Baa Corporate Bond Yield
93	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS
39	NDMANEMP	All Employees: Nondurable goods

## 9 Conclusions

We proposed a method for 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. An important novelty of our approach is its robustness to the dynamic properties of predictors which can be noisy, persistent or a mixture of both. In addition, our approach is able to accommodate a large number of predictors at little computational cost and has been shown to be very reliable even in contexts where the effective sample size is smaller than the available pool of predictors and this despite an asymptotic theory that



operates solely under  $n \rightarrow \infty$ . As argued in McKeague and Qian (2008) who developed a theory of marginal screening in high dimensional regressions assuming a fixed pool of predictors, extending our own asymptotic framework that accommodates persistence and predictor correlatedness to environments in which  $p$  is allowed to grow with  $n$  would also raise formidable technical challenges. Nevertheless our local power analysis and simulation based results have clearly highlighted the suitability and accuracy of our asymptotic regime in large  $p$  environments.

In a wide range of applications one is often interested in whether a series is best described as a mean independent process as opposed to being predictable with some predictor belonging to a large information set. One 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. Although our primary focus is not about uncovering a true model our framework does allow us to detect a *key player* which can be valuable information in itself. It could for instance be used in a model further augmented with diffusion index type factors. It could also be useful in contexts where predictors consist solely of principal components as in such environments PC type factors are typically obtained while being agnostic about how they relate to the predictand. As PCs are typically obtained using a pool of very diverse predictors in terms of their persistence properties, the robustness of our methods to such characteristics also makes them particularly suitable.

## REFERENCES

- Berenguer-Rico, V. and B. Nielsen (2020). Cumulated Sum of Squares Statistics for Nonlinear and Nonstationary Regressions, *Econometric Theory* 36(1), 1-47.
- Clark, T.E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* 138 (1), 291-311.
- Clark, T.E. and M. W. McCracken (2013). Advances in Forecast Evaluation, in *Hand-*

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

Fava, B. and H. F. Lopes (2020). The Illusion of the Illusion of Sparsity: An exercise in prior sensitivity, *arXiv:2009.14296v1*, September 2020.

Ghysels, E., J. B. Hill, 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.

Giannone, D., M. Lenza, and G. E. Primiceri (2021). Economic Predictions with Big Data: The Illusion of Sparsity, *Econometrica*, 89, 2409-2437.

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

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

Liu, W. and E. Moench (2016). What predicts US recessions?, *International Journal of Forecasting*, 32, 1138-1150.

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.

McCracken, M. and S. Ng (2016). FRED-MD: A Monthly Database for Macroeconomic Research, *Journal of Business and Economic Statistics*, 34, 574-589.

Pesaran, M. H. and A. Timmermann (1995). Predictability of Stock Returns: Robustness and Economic Significance, *Journal of Finance*, 50, 1201-1228.

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.

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.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic Forecasting Using Diffusion Indexes, *Journal of Business and Economic Statistics* 20, 147-162.

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.

## SUPPLEMENTARY MATERIAL

**Title:** Technical Proofs and further simulations. (pdf)