TESTING FOR PREDICTABILITY IN PANELS WITH GENERAL PREDICTORS*

Joakim Westerlund†
Lund University
and
Financial Econometrics Group
Centre for Research in Economics and Financial Econometrics
Deakin University

Hande Karabiyik
Lund University

Paresh Narayan
Financial Econometrics Group
Centre for Research in Economics and Financial Econometrics
Deakin University

March 19, 2015

Abstract

The difficulty of predicting returns has recently motivated researchers to start looking for tests that are either robust or more powerful. Unfortunately, the way that these tests work typically involves trading robustness for power or vice versa. The current paper takes this as its starting point to develop a new panel-based approach to predictability that is both robust and powerful. Specifically, while the panel route to increased power is not new, the way in which the cross-section variation is exploited to achieve also robustness with respect to the predictor is. The result is two new tests that enable asymptotically standard normal and chi-squared inference across a wide range of empirical relevant scenarios in which the predictor may be stationary, unit root non-stationary, or anything in between. The cross-section dependence of the predictor is also not restricted, and can be weak, strong, or indeed anything in between. What is more, this generality comes at no cost in terms of test construction. The new tests are therefore very user-friendly.

*Westerlund and Karabiyik thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

†Corresponding author: Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

JEL Classification: C22; C23; G1; G12.

Keywords: Panel data; Predictive regression; Predictor persistency; Cross-section dependence.

1 Introduction

Consider the panel data variables $y_{i,t}$ and $x_{i,t}$, observable for $t = 1, ..., T$ time series and $i = 1, ..., N$ cross-section units. Recent years have witnessed an immense proliferation of research asking whether $y_{i,t}$ can be predicted using the one-period lagged value of some other variable, $x_{i,t-1}$ say. Examples of such situations are abound. The most common ones are found in finance. For example, if $y_{i,t}$ is (excess) stock returns, or the equity risk premium, then $x_{i,t}$ might be dividend yield, nominal interest rates, default or term spreads on bonds, inflation, valuation ratios, the consumption-wealth ratio, stock market volatility, labor income, aggregate output, output gap, or oil prices, just to mention a few (see Neely et al., 2012; Rapach and Zhou, 2013, and the references provided therein).

The conventional way in which earlier studies have been trying to test the predictability hypothesis is to first run a time series regression of $y_{i,t}$ onto a constant and $x_{i,t-1}$, and then to test whether the (predictive) slope on $x_{i,t-1}$ is zero by using a conventional t-test. The main finding of such tests is that the observed value of the t-statistic is typically greater than two. Most earlier studies, which tended to rely on normal critical values, were therefore able to reject the null hypothesis. However, it has since then become clear that the standard distribution theory for stationary processes can be quite misleading when testing for predictability, and that some of the rejections might actually be due to size distortions. The problem is that, although according to theory many of the predictors used should be stationary, or I(0), empirically most predictors are only slowly mean-reverting, and the evidence that they are not unit root non-stationary, or I(1), is weak (see, for example, Campbell and Yogo, 2006; Elliott and Stock, 1994; Lanne, 2002; for discussions and some confirmatory empirical results). Standard asymptotic theory, which presumes that $x_{i,t}$ is I(0), is therefore likely to be inappropriate. In fact, even if $x_{i,t}$ is known to be I(0), if the persistence is high enough, the standard asymptotic theory is likely to provide a poor approximation in small samples (Elliott and Stock, 1994). This observation has caused some researchers to consider an alternative framework based on the asymptotic theory for I(1) processes (see, for example, Cavanagh et al.,

1995; Elliott and Stock, 1994; Lanne, 2002; Westerlund and Narayan, 2015b). This theory has not only made it possible to study formally the effects of near I(1) predictors, but has also led to the development of a new class of tests designed specifically for the I(1) case.

But while the extension of the conventional asymptotic theory to the case with I(1) predictors has been beneficial in many ways, it has also revealed some rather disturbing facts about existing predictability tests. One such fact is that unless $x_{i,t}$ is exactly I(1) and/or exogenous, most, if not all, tests in the literature have limiting distributions that depend on (unestimable) nuisance parameters, thereby invalidating the use of asymptotic critical values. Many rejections of the no predictability null based on such asymptotic critical values are therefore again likely to be due to size distortions. In case of returns, the evidence of predictability is so strong that it has given rise to what has become known as the "stock return predictability puzzle", and some researchers have even characterized stock return predictability as a new stylized fact in finance (Cochrane, 1999).

The issue of size distortion and its implications for the stock return predictability puzzle has led to a widespread search for robust inference methods. There have been extensive efforts and some procedures such as Bonferroni type methods have received much attention (see, for example, Campbell and Yogo, 2006; Elliott and Stock, 1994; Westerlund and Narayan, 2015b). At present, Bonferroni type procedures represent the state-of-the-art in this literature, but they do have some undesirable properties. First, simulations are required to compute critical values since the limit distribution employed in the calculations is non-standard. These procedures are therefore rather unattractive from an applied point of view. Second, Bonferroni-based critical values are known to be crude, leading to conservative tests with low power. Third, many procedures, such as the one of Campbell and Yogo (2006), rely critically on $x_{i,t}$ being nearly I(1) and they break down if $x_{i,t}$ is I(0) (see Phillips, 2014). Fourth, many Bonferroni procedures require an initial estimator of the persistence of $x_{i,t}$, and such estimators are known to be rather imprecise. Finally, Bonferroni procedures are very difficult to extend to the case with multiple predictors. Hence, there is a need for alternative test procedures that are robust, yet do not suffer from the drawbacks of Bonferroni type procedures. Indeed, as Phillips (2014, pages 1191–1192) recently pointed out,

> The applied predictive regression literature is large and continues to grow rapidly in empirical finance and macroeconomics. Against this background of applied research, the ongoing interest in econometrics in uniform procedures

> of inference, the challenges presented by multiple predictors, and the pitfalls pointed out in the current contribution, there is substantial need for continuing econometric research on methods of inference that can cope with potential non-stationarities in many regressors, control size, and deliver good discriminatory power in detecting predictability. The credibility of applied research on this subject ultimately depends on the reliability of the inferential machinery available in econometrics. There is still much to do.

The current paper can be seen as a step in this direction. However, instead of considering one cross-section unit at a time, as is commonly done in the literature (see, for example, Ang et al., 2006; Polk et al., 2006), we consider a joint panel approach, which has a number of distinct advantages. First, testing cross-section units one at a time is wasteful in the sense that each test is conducted while ignoring the information contained in other units. Indeed, financial data are characterized not only by rich dynamics across time, but also by strong co-movements across assets and financial markets. Second, the use of panel rather than time series data not only increases the total number of observations and their variation, but also reduces the noise coming from the individual time series regressions. This is reflected in the power of the resulting panel predictability test, which is increasing in both N and T, as opposed to the time series approach where power is only increasing in T. Thus, from a power/precision point of view, a joint panel approach is typically preferred. Third, since power is increasing in both N and T, this means that in panels one can effectively compensate for a relatively small T by having a relatively large N, and vice versa.

Of course, the use of panel data to boost the power of tests for predictability is not new, but has been tried before (see Hjalmarsson, 2010; Kauppi, 2001; Westerlund and Narayan, 2015a). Unfortunately, the few panel-based studies that do exist are based on rather restrictive assumptions that are unlikely to be met in practice. The presence of cross-section dependence in $x_{i,t}$, for example, is typically ignored altogether, as is the fact that the dynamics of $x_{1,t}, ..., x_{N,t}$ are unlikely to be the same. In fact, most, if not all, panel-based tests assume that $x_{1,t}, ..., x_{N,t}$ are all I(1), which is again rather unrealistic, especially in large-N panels, where the cross-section units can be expected to be more heterogenous had N been smaller.

Our point of departure is a very general data generating process (DGP) in which $x_{i,t}$ is treated as a "black box". In fact, as long as the order of integration of $x_{i,t}$ is at most one, there are almost no restrictions. $x_{i,t}$ may be I(0), but it can also be I(1), or indeed anything

in between these two extremes. $x_{i,t}$ is also not restricted to be cross-section independent, but may be cross-section dependent in a very general fashion. $x_{i,t}$ may, for example, be driven in part by common factors that may potentially be I(1), which seems like a highly relevant scenario in practice. $x_{i,t}$ is also not restricted to be homoscedastic, but may be heteroscedastic across both time and cross-section units. Given this generality, one might think that corrections aimed at achieving asymptotically pivotal statistics are not really an option. However, this is not what we find. In fact, the statistics that we consider are not only asymptotically valid, but are also remarkably simple, requiring nothing but simple ordinary least squares (OLS) operations. Hence, in terms of the types of predictors that can be accommodated, the proposed tests are very robust indeed. The results from a small-scale Monte Carlo study show that our theoretical predictions are borne out well in small samples. In fact, the new tests have excellent small-sample properties. In our empirical illustration, we consider as an example the predictability of country-level stock returns.

The balance of the paper is organized as follows. In Section 2, we formalize the DGP considered, which is used in Section 3 to derive the asymptotic distributions of the test statistics considered. Sections 4 and 5 contain the Monte Carlo study and empirical illustration, respectively. Section 6 concludes. Proofs of important results are given in Appendix.

2 The model

In this paper, $y_{i,t}$ is a scalar and $x_{i,t}$ is an $m \times 1$ vector. The DGP of these variables is given by the following two equations:

$$y_{i,t} = \alpha_i + \beta_i' x_{i,t-1} + u_{i,t}, \tag{1}$$

$$u_{i,t} = \lambda_i f_t + \varepsilon_{i,t}, \tag{2}$$

where f_t is a common factor, λ_i is the associated factor loading and $\varepsilon_{i,t}$ is an idiosyncratic error term. This DGP is a panel extension of the prototypical predictive regression model that has been widely used in the time series literature, in which $x_{i,t}$ is a variable believed to be able to predict $y_{i,t}$. As we explain in more detail below, this earlier literature can be divided in two strands; one strand assumes that $x_{i,t}$ is I(0), while the other assumes that it is (nearly) I(1). This division is natural, because the results depend critically on the persistency of $x_{i,t}$. The main thrust of the present paper is that it makes almost no assumptions in this regard. The particular conditions that we are going to work under are summarized in Assumptions

EPS, F, LAM, IND and X, where the abbreviations "EPS", "F", "LAM", "IND" and "X" refer to $\varepsilon_{i,t}$, f_t, λ_i, the independence between $\varepsilon_{i,t}$, f_t and λ_i, and $x_{i,t}$, respectively. Throughout, $M < \infty$, $\text{tr}\, A$ and $||A|| = \sqrt{\text{tr}\,(A'A)}$ will be used to denote a generic positive constant, and the trace and Frobenius (Euclidean) norm of the matrix A, respectively. It is further convenient to denote by $\mathcal{F}_{i,t}$ ($\mathcal{F}_t$) the sigma-field generated by $\{\varepsilon_{i,n}\}_{n=1}^{t}$ ($\{f_n\}_{n=1}^{t}$).

Assumption EPS. $\varepsilon_{i,t}$ is independent across i with $E(\varepsilon_{i,t}|\mathcal{F}_{i,t-1}) = 0$, $E(\varepsilon_{i,t}^2) = \sigma_{\varepsilon,i}^2 > 0$ and $E(|\varepsilon_{i,t}|^8) \leq M$.

Assumption F. $E(f_t|\mathcal{F}_{t-1}) = 0$, $E(f_t^2) = \sigma_f^2 > 0$ and $E(|f_t|^8) \leq M$.

Assumption LAM. λ_i is either deterministic such that $|\lambda_i| \leq M$ or stochastic such that $E(|\lambda_i|^4) \leq M$. In both cases, $\overline{\lambda} = N^{-1}\sum_{i=1}^{N}\lambda_i \neq 0$ for all N, including $N \to \infty$.

Assumption IND. $\varepsilon_{i,t}$, f_t and λ_i are mutually independent.

Assumption X–I(0). $E(x_{i,t}) = \mu_{x,i}$, $E[(x_{i,t} - \mu_{x,i})^2] = \Sigma_{x,i}$, an $m \times m$ positive definite matrix, $E(||x_{i,t} - \mu_{x,i}||^4) \leq M$ and $T^{-1}\sum_{t=1}^{T}\sum_{n=1}^{T}||E[(x_{i,t} - \mu_{x,i})(x_{i,n} - \mu_{x,i})]|| \leq M$.

Assumption X–I(1). $T^{-1/2}(x_{i,\lfloor rT\rfloor} - \mu_{x,i}) \to_d B_{x,i}(r)$ as $T \to \infty$, where $\mu_{x,i} = E(x_{i,t})$ and $B_{x,i}(r)$ is a $m \times 1$ vector diffusion process with $E[dB_{x,i}(r)dB_{x,i}(r)'] = \Omega_{x,i}$, where $\Omega_{x,i}$ is positive definite.

The requirement that $E(\varepsilon_{i,t}|\mathcal{F}_{i,t-1}) = 0$ and $E(f_t|\mathcal{F}_{t-1}) = 0$ implies that $\varepsilon_{i,t}$ and f_t are serially uncorrelated. This is enough to ensure valid inference under Assumption X–I(0). Interestingly, under Assumption X–I(1) f_t need not be serially uncorrelated. In fact, it is enough that $T^{-1}\sum_{t=1}^{T}\sum_{n=1}^{T}|E(f_t f_n)| \leq M$, which means that f_t can be serially correlated in a very general manner. Hence, if $x_{i,t}$ is known to satisfy Assumption X–I(1), then Assumption F may be relaxed. The tests that we propose remain valid even in such cases and require no correction to account for serial correlation in f_t. Note how Assumptions EPS, F, LAM, IND and X place no restrictions on the correlation between $\varepsilon_{i,t}$ and f_t, on the one hand, and $x_{i,t}$, on the other hand. The predictor can therefore be endogenous. One way to relax the serial uncorrelated f_t assumption when $x_{i,t}$ is I(0) is to assume that $\varepsilon_{i,t}$ and $x_{i,t}$ are independent. In this paper, however, we follow the convention in both the panel and time series literatures and allow for endogeneity, but not serially correlated errors (see, for example, Campbell

and Yogo, 2006; Lanne, 2002; Stambaugh, 1999; Lewellen, 2004; Hjalmarsson, 2010; Kauppi, 2001; Westerlund and Narayan, 2015a, 2015b). As with the persistency, if $x_{i,t}$ is known to be exogenous, then the tests proposed here may be used in the presence of serial correlation in f_t without any correction.

The assumptions placed on the dynamics of $x_{i,t}$ are very general. In particular, while Assumption X–I(0) is general enough to include the broad class of I(0) ARMA processes, Assumption X–I(1) include not only exact I(1) processes, but also processes that are nearly I(1). Hence, while under Assumption X–I(1) we refer to $x_{i,t}$ as "I(1)", we are in fact allowing also local deviations from the exact I(1) case. The tests that we consider are valid under both Assumption X–I(0) and X–I(1), which means that we can allow not only processes that are I(0) and (nearly) I(1), but also everything in between, including the class of moderate I(1) processes considered by Phillips et al. (2010). The types of dynamics that can be permitted under Assumptions X–I(0) and X–I(1) are therefore very general indeed. In fact, as long as it is at most I(1), the persistence of $x_{i,t}$ is virtually unrestricted. This is a major advantage as usually the asymptotic results depend directly on what is being assumed in this regard. If $x_{i,t}$ is I(0) as in, for example, Stambaugh (1999) and Lewellen (2004), then the conventional asymptotic theory for stationary processes is appropriate, whereas if $x_{i,t}$ is I(1), then the asymptotic theory for unit root processes apply. There also the locally I(1) case when yet another theory applies, as shown by, for example, Lanne (2002), Elliott and Stock (1994), and Cavanagh et al. (1995). As alluded to in the above, results reported in the current paper remain valid not only in these cases, but hold also when $x_{i,t}$ is moderately I(1), a case that has not been considered before in the predictability literature.

Assumptions X–I(0) and X–I(1) make no restrictions on the cross-sectional properties of $x_{i,t}$. Hence, unlike studies such as Hjalmarsson (2010), Kauppi (2001), and Westerlund and Narayan (2015a), in which the cross-section dependence in $x_{i,t}$ is either restricted or assumed to be absent altogether, in the present paper we place no restriction on the cross-section dependence of $x_{i,t}$. The dependence can be "weak", but it can also be "strong", as in the presence of common factors (see Chudik et al., 2011, for a detailed treatment of the concepts of weak and strong cross-section dependence). In fact, the factors can even contain unit roots, such that $x_{1,t}, ..., x_{N,t}$ are cointegrated, a possibility that we study in detail in the Monte Carlo analysis of Section 4.

As pointed out in studies such as Campbell and Yogo (2006), Westerlund (2014), and

Westerlund and Narayan (2015b), $x_{i,t}$ is likely to be heteroscedastic. The above DGP is very general in this regard, and does not exclude heteroscedasticity in $x_{i,t}$, which may be conditional but it can also be unconditional. In both cases, the heteroscedasticity is not restricted to a particular dimension but may appear over time as well as across the cross-section.

3 The tests

Consider the generic variable a_t. Define the "forwards" and "backwards" recursively demeaned versions of a_t as

$$a_t^* = a_t - (T-t+1)^{-1}\sum_{n=t}^{T} a_n, \qquad a_t^{**} = a_t - t^{-1}\sum_{n=1}^{t} a_n, \tag{3}$$

respectively. It is further convenient to define $M_A = I_{T-1} - A(A'A)^{-1}A'$ for any $(T-1)$-rowed matrix A, and to let $\overline{a} = N^{-1}\sum_{i=1}^{N} a_i$ for any a_i. The estimator of f_t that we will be considering is given by

$$\hat{f}_t = \overline{y}_t^*, \tag{4}$$

and the resulting estimator of β is

$$\hat{\beta} = \left(\sum_{i=1}^{N}(x_{i,-1}^*)'M_{\hat{f}}x_{i,-1}^{**}\right)^{-1}\sum_{i=1}^{N}(x_{i,-1}^{**})'M_{\hat{f}}y_i^*, \tag{5}$$

where $x_{i,-1}^* = (x_{i,1}^*,...,x_{i,T-1}^*)'$ and $x_{i,-1}^{**} = (x_{i,1}^{**},...,x_{i,T-1}^{**})'$ are $(T-1)\times m$, while $y_i^* = (y_{i,2}^*,...,y_{i,T}^*)'$ and $\hat{f} = (\hat{f}_2,...,\hat{f}_T)'$ are $(T-1)\times 1$.

Remark 1. The reason for the use of both forwards and backwards recursive detrending is taken from the panel unit root literature (see, for example, Westerlund, 2015), and ensures that there is no time overlap in $x_{i,t-1}^{**}$ and $y_{i,t}^*$. This, in turn, ensures that $\hat{\beta}$ is free of the otherwise so common "Stambaugh bias" (Stambaugh, 1999), which greatly complicates inference (see, for example, Hjalmarsson, 2008; Kauppi, 2001, for some discussions and results for the panel case).

Remark 2. A word on the choice of estimator of f_t. Under $H_0 : \beta_1 = ... = \beta_0 = 0_{m\times 1}$, $y_{i,t} = \alpha_i + u_{i,t} = \alpha_i + \lambda_i f_t + \varepsilon_{i,t}$, which is nothing but a static common factor model in stationary variables. This suggests that f_t can in principle be estimated using the principal components method, which has a long tradition in econometrics and statistics (see Bai, 2003,

Section 1, for a brief review of this literature). However, preliminary Monte Carlo results suggest that this estimator suffers from poor small-sample performance, especially in the empirically relevant case when $T > N$ (see also Westerlund and Urbain, 2015). In the present paper we therefore consider an alternative estimator that not only has good small-sample properties, but that is also computationally very convenient. In fact, it is difficult to think of a simpler estimator. The idea is to follow Pesaran (2006) and to use $\hat{f}_t = \overline{y}_t^*$ as an estimator of f_t^*, which under H_0 is "rotationally consistent" for f_t^*, in the sense that it converges to a scalar times f_t^*. Specifically, since $\varepsilon_{1,t}, ..., \varepsilon_{N,t}$ are independent, $\hat{f}_t = \overline{\lambda} f_t^* + \overline{\varepsilon}_t^* = \overline{\lambda} f_t^* + O_p(N^{-1/2})$. The fact that $\hat{f}_t$ estimates $\overline{\lambda} f_t^*$ rather than f_t^* itself is an issue when wanting to infer f_t^*, but not when just wanting to control for f_t^*. The reason is that, provided $\overline{\lambda} \neq 0$, we have $M_{\overline{\lambda} f^*} = M_{f^*}$. Hence, even if $\hat{f}_t$ estimates $\overline{\lambda} f_t^*$, $M_{\hat{f}}$ is still an "estimator" of M_{f^*}.

Remark 3. The estimator of β considered here is similar in spirit to the one of Hjalmarsson (2010), in which $x_{i,-1}$ and y_i are projected on $\overline{x}_{-1} = N^{-1} \sum_{i=1}^N x_{i,-1}$. Hence, with this approach one is essentially testing the predictive content of idiosyncratic part of $x_{i,t}$, thereby ignoring a potentially important source of predictive information, namely, the common one. Moreover, as Westerlund and Karabiyik (2015) show, said procedure is actually not valid in the sense that, while consistent, the resulting estimator is generally not asymptotically normal, thereby invalidating the use of the standard chi-squared-based inferential toolbox. The only exception is if $x_{i,t}$ is nearly I(1) and exogenous.

Let us introduce the following covariance matrices:

$$\Sigma_0 = \sum_{i=1}^N \sigma_{\varepsilon,i}^2 E(z_{i,-1}' z_{i,-1}), \tag{6}$$

$$\Sigma_x = \sum_{i=1}^N E[(x_{i,-1}^*)' x_{i,-1}^{**}], \tag{7}$$

where $z_{i,-1} = (z_{i,1}, ..., z_{i,T-1})'$ with $z_{i,t} = x_{i,t}^{**} - N^{-1} \sum_{i=1}^N \overline{\lambda}^{-1} \lambda_i x_{i,t}^{**}$. In Lemmas A.1 and A.2 of Appendix we show that while under Assumption X–I(1), $\sqrt{N}T(\hat{\beta} - \beta)$ is asymptotically normal with covariance matrix $NT^2 \Sigma_x^{-1} \Sigma_0 \Sigma_x^{-1}$, under Assumption X–I(0), $\sqrt{NT}(\hat{\beta} - \beta)$ is again asymptotically normal but now with covariance matrix $NT \Sigma_x^{-1} \Sigma_0 \Sigma_x^{-1}$. The fact that the former covariance matrix is just T times the latter means that we can construct test statistics based on $\hat{\beta}$ that are valid regardless of whether $x_{i,t}$ is I(0) or I(1). A natural candidate in this

regard is the Wald test, which in the current context is given by:

$$W = \hat{\beta}'(\Sigma_x \Sigma_0^{-1} \Sigma_x)\hat{\beta}. \tag{8}$$

The asymptotic distribution of this test statistic is provided in Theorem 1, which is our main result.

Theorem 1. *Under $H_0 : \beta_1 = ... = \beta_N = 0$, Assumptions EPS, F, LAM, IND and X–I(0) or X–I(1), as $N, T \to \infty$ with $N/T \to 0$,*

$$W \to_d \chi^2(m),$$

where $\to_d$ and $\chi^2(m)$ signify convergence in distribution and a chi-squared variate with m degrees of freedom, respectively.

In many cases, one is interested in testing not all but one of the predictors contained in $x_{i,t}$. This makes it possible to consider a t-statistic. The t-statistic for testing the n-th element of β is given by

$$TR = \delta_n' \hat{\beta} [\delta_n' (\Sigma_x \Sigma_0^{-1} \Sigma_x) \delta_n]^{1/2}, \tag{9}$$

where $\delta_n = (0, ..., 0, 1, 0, ..., 0)'$ is a $m \times 1$ vector with a one at position n and zeros elsewhere. In many cases, $x_{i,t}$ is just a scalar ($m = 1$), in which the formula for TR reduces to $TR = \hat{\beta} \Sigma_x \Sigma_0^{-1/2}$. The asymptotic distribution of TR is given in the following corollary to Theorem 1.

Corollary 1. *Under the conditions of Theorem 1,*

$$TR \to_d N(0, 1).$$

Remark 4. The fact that the asymptotic distributions of W and TR are the same regardless of whether $x_{i,t}$ is I(0) or I(1) is in stark contrast to the previous literature where the asymptotic results depend on what is being assumed regarding the persistency of $x_{i,t}$. If $x_{i,t}$ is I(0), then normality is usually possible (see Lewellen, 2004), whereas if $x_{i,t}$ is I(1), then the results involve functions of Brownian motion (see Elliott and Stock, 1994; Cavanagh et al., 1995). What is more, the results in the I(1) case depend critically on whether the root is exact or local. In particular, asymptotically pivotal results are only available in the exact I(1) case,

which means that existing tests typically suffer from a dependence on nuisance parameters. The single most popular approach to accommodate such dependencies is to use a two-step procedure in which the predictor is pretested for a unit root, and where the predictability test is implemented conditional on the outcome of the pretest. Unfortunately, this means loosing control of the overall significance level of the joint test. Therefore, in order to at least put an upper limit on the joint significance level, Campbell and Yogo (2006), Cavanagh et al. (1995), Lewellen (2004), and Torous et al. (2004), among others, have made use of the Bonferroni principle, which states that the significance level for the joint hypothesis that at least one of the tests end up in a rejection is less than or equal to the sum of the individual significance levels. Hence, if the individual unit root and predictability tests are performed at the 5% level, then their joint significance level cannot be larger than 10%. The obvious problem here is that the upper limit is very crude, and that the test is bound to be conservative. As a response to this, Westerlund and Narayan (2015b) propose subsampling. While this makes testing possible even in the locally I(1) case, there is also a cost to this in that the subsampling approach is both computationally demanding and not user friendly.

Remark 5. The secret behind the fact that the asymptotic distributions of W and TR are the same regardless the persistency of $x_{i,t}$ lies in the use of the cross-sectional variation of the data. Specifically, while in the time series case standard inference is typically only possible if $x_{i,t}$ is I(0), here we can use the cross-section to effectively smooth out the non-standard limiting time series distribution obtained when $x_{i,t}$ is I(1). Standard inference is therefore possible in both cases. Of course, the use of the cross-sectional variation alone does not automatically lead to standard asymptotic distributions. In fact, as far as we are aware, W and TR are unique in that they support standard chi-squared and normal inference regardless of whether $x_{i,t}$ is I(0) or I(1).

Remark 6. The advantage of not having to pretest for a unit root in $x_{i,t}$ cannot be overstated. Indeed, there are many reasons for avoiding the use of a pretest. The perhaps most obvious reason is that it is inconvenient, especially in the current panel context where appropriate accounting for cross-section dependence requires specialized tests with complicated construction. A related reason is that, in analogy to the time series literature, most panel unit root tests are constructed with a unit root under the null. As we illustrate in Section 5.2, this calls for careful interpretation of the test outcome in case of a rejection (see Westerlund

and Breitung, 2013, for a detailed discussion). Yet another reason is that the unit root test is typically correlated with the test for predictability, leading to size problems and hence misleading inference if this correlation is not appropriately accounted for in the predictability test (see Elliott and Stock, 1994).

W and TR are infeasible, as they depend on Σ_x and Σ_0, which are unknown. Fortunately, constructing consistent estimates of these matrices is an easy task. Consider Σ_0. In order to construct an estimator of $z_{i,t}$, we need a consistent estimator of λ_i. Considering that under H_0, $y^*_{i,t} = \lambda_i f^*_t + \varepsilon^*_{i,t}$, the perhaps most obvious candidate is given by the OLS slope estimator in a regression of $y^*_{i,t}$ onto $\hat{f}_t$. That is, we use

$$\hat{\lambda} = \begin{bmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_N \end{bmatrix} = \sum_{t=2}^{T} y^*_t \hat{f}_t \left(\sum_{t=2}^{T} \hat{f}_t^2 \right)^{-1}. \tag{10}$$

However, for the same reason that $\hat{f}_t$ is consistent not for f_t but for $\overline{\lambda} f_t$, $\hat{\lambda}_i$ is consistent for $\overline{\lambda}^{-1}\lambda_i$. The estimator of $z_{i,t}$ is therefore given by $\hat{z}_{i,t} = x^{**}_{i,t} - N^{-1}\sum_{i=1}^{N} \hat{\lambda}_i x^{**}_{i,t}$. As an estimator of $\sigma^2_{\varepsilon,i}$, we use $\hat{\sigma}^2_{\varepsilon,i} = T^{-1}\hat{u}_i' M_{\hat{f}} \hat{u}_i$, where $\hat{u}_i = y^*_i - x^*_{i,-1}\hat{\beta}$. Estimators $\hat{\Sigma}_0$ and $\hat{\Sigma}_x$ of Σ_0 and Σ_x, respectively, can now be constructed as follows:

$$\hat{\Sigma}_0 = \sum_{i=1}^{N} \hat{\sigma}^2_{\varepsilon,i} \hat{z}'_{i,-1} \hat{z}_{i,-1}, \tag{11}$$

$$\hat{\Sigma}_x = \sum_{i=1}^{N} (x^*_{i,-1})' x^{**}_{i,-1}, \tag{12}$$

which, upon appropriate normalization by N and T, can be shown to be consistent. For ease of notation, in what follows we use W and TR to refer to the feasible test statistics.

4 Monte Carlo simulations

A Monte Carlo study was undertaken to investigate the small sample properties of the proposed tests. The DGP considered for this purpose is given by a simplified version of (1) and (2) that sets $\alpha_i = 1$ and $(f_t, \varepsilon_{i,t}) \sim N(0_{2\times 1}, I_2)$. To further ensure that $\overline{\lambda} \neq 0$, we make λ_i a draw from $N(1,1)$. Following the bulk of the literature, we focus on the case when there is a single predictor, although we also report some results for the case with two predictors. The following two-predictor DGP cover both cases:

$$x_{i,t} = \Phi x_{i,t-1} + v_{i,t}, \tag{13}$$

$$v_{i,t} = \Lambda_i F_t + \epsilon_{i,t}, \tag{14}$$

where

$$\Phi = \begin{bmatrix} \rho & 0 \\ 0 & 0.5 \end{bmatrix}, \quad \Lambda_i = \begin{bmatrix} \kappa \cdot \lambda_i & 0 \\ 0 & \delta_i \end{bmatrix}, \quad F_t = \begin{bmatrix} f_t \\ z_t \end{bmatrix}, \tag{15}$$

with $x_{1,0} = ... = x_{N,0} = 0_{2\times 1}$, $\epsilon_{i,t} \sim N(0_{2\times 1}, I_2)$, $\delta_i \sim N(0.5, 1)$, $z_t \sim N(0, 1)$ and $\kappa = 1$. In case of one predictor, we simply drop the second predictor from the above DGP. The presence of $\lambda_i f_t$ in $v_{i,t}$ means that the first predictor is correlated with the error in the equation for $y_{i,t}$. It is therefore endogenous. Moreover, since under $\rho = 1$, the equation for this predictor is given by $\kappa \cdot \lambda_i \sum_{n=1}^{t} f_n + \sum_{n=1}^{t} \epsilon_{i,n}$, $x_{1,t}, ..., x_{N,t}$ have a common stochastic trend. That is, $x_{i,t}$ is cross-section cointegrated. The following parameterizations of ρ will be considered:

R1. $\rho = 1$;

R2. $\rho = 1 - 2T^{-1}$;

R3. $\rho = 1 - 2T^{-9/10}$;

R4. $\rho = 0.8$,

which correspond the case when the first predictor is exactly I(1), locally I(1), moderately I(1) and I(0), respectively. In interest of space, we focus on the 5% size and power of TR and W, where the former test is constructed as double-sided. For the size simulations, $\beta_i = 0 \cdot 1_{2\times 1}$, where $1_{2\times 1} = (1, 1)'$, whereas for the power simulations, the following five parameterizations of β_i are considered:

P1. $\beta_i = 0.1 \cdot 1_{2\times 1}$;

P2. $\beta_i = 0.2 \cdot 1_{2\times 1}$;

P3. $\beta_i = T^{-1/2} \cdot 1_{2\times 1}$;

P4. $\beta_i = 0.2 \cdot 1_{2\times 1}$ for $i = 1, ..., N/5$ and $\beta_i = 0 \cdot 1_{2\times 1}$ for $i = N/5 + 1, ..., N$;

P5. $\beta_i \sim N(0_{2\times 1}, I_m)$ for $i = 1, ..., N/5$ and $\beta_i = 0 \cdot 1_{2\times 1}$ for $i = N/5 + 1, ..., N$.

While P1 and P2 cover the case of a fixed homogeneous alternative, in P3 alternative is local to the null. In P4 and P5 there is a mix of predictable and unpredictable units. Again, while the above specifications are for the case with two predictors, in the one predictor case we simply drop the equation corresponding to the second predictor. All results are based on

making 5,000 replications of samples of size $N \in \{10, 20, 30\}$ and $N \in \{50, 100, 200\}$, where N and T are chosen to reflect the theoretical requirement that N/T should go to zero.

The results are reported in Tables 1–4. In Tables 1 and 2, we report the size and power of TR and W, respectively, for the case with one predictor. Tables 3 and 4 contain the corresponding results in case of two predictors, where the TR results are obtained by testing the first predictor. The information content of all four tables can be summarized as follows.

- The size accuracy is very good in all cases considered, and it improves as N and in particular T increases. Note in particular how size accuracy tends to be quite satisfactory even when N and T are as small as 10 and 50, respectively. In fact, the distortions are generally not larger than that they can be attributed to simulation uncertainty.[1] The fact that size is almost completely flat in ρ (R1–R4) corroborates the theoretical result that TR and W are valid regardless of whether $x_{i,t}$ is I(0) or I(1) or in between. This is in stark contrast to existing tests, whose performance depend critically on the persistency of $x_{i,t}$ (see, for example, Lewellen, 2004; Westerlund and Narayan, 2015b).

- As already mentioned, in the simulated DGP the first predictor is endogenous. We experimented with different levels of endogeneity, as measured by κ, but since the results were almost identical, we only report the results for the case when $\kappa = 1$. According to the results, both tests considered are robust to endogeneity, which corroborates our asymptotic theory. This is again in stark contrast to existing simulation results, which tend to vary quite substantially depending on the endogeneity of $x_{i,t}$ (see, for example, Campbell and Yogo, 2006). What is more, this variation depends in an intricate way on the persistency of $x_{i,t}$. Existing tests therefore tend to be very sensitive in this regard, so much so that some authors have even recommended against their use (see Elliott and Stock, 1994; Westerlund and Narayan, 1015a, 2015b).

- As expected, power increases in both the sample size and in the deviation from the null, as measured by $||\beta_i||$. We also see that there is some variation in power depending on the value of ρ. Specifically, power seems to be decreasing in the persistency of $x_{i,t}$, being lowest when $\rho = 1$ and highest when $\rho = 0.8$. This is interesting, because for most existing tests it is the other way around, that is, power is increasing in $|\rho|$ (see, for

[1] With 5,000 replications the 95% confidence interval for the size of the 5% level tests studied here (in %) is [4.4, 5.6].

example, Lewellen, 2004; Westerlund and Narayan, 2015b).

- In the case of a mix of predictable and unpredictable units (P4 and P5) the tests are as powerful as when the null is violated for all units. This can be seen by comparing the results reported for P2 and P4, where the deviation from the null is the same but where there is a difference in how many predictable units there are. The fact that the presence of some unpredictable units under the alternative is not detrimental for power is of course a great advantage in practice.

In sum, the results reported in this section suggest that the new tests have good size accuracy across a broad range of empirically relevant DGPs and sample sizes. In fact, the tests seem to be remarkably robust in this regard. What is more, this robustness under the null does not seem to come at a cost in terms of power. Specifically, while there is some variation in the results depending on the persistency of $x_{i,t}$, power does not seem to be affected much by the presence of unpredictable units under the alternative. The new tests should therefore be a valuable addition to the already existing menu of panel predictability tests.

5 Empirical illustration

One of the most well-known facts about stock return predictability is that the assumed DGP of the predictor matters a lot, and that different assumptions can lead to very different results (see, for example, Campbell and Yogo, 2006; Goyal and Welch, 2003; Ferson et al., 2003; Westerlund and Narayan, 2015b). Welch and Goyal (2008) provide a comprehensive analysis of the empirical performance of a large collection of predictive models of the US equity premium using no less than 14 predictor candidates. Their conclusion is that most models considered are unstable or even spurious. As a solution, they recommend using more general econometric approaches, a recommendation that has since then received much support (see, for example, Rapach et al., 2010). Another well-known fact about stock return predictability is that existing time series tests based on US data suffer from low power, and that the use of the information contained in a panel of multiple countries can lead to more accurate results (see, for example, Hjalmarsson, 2010; Westerlund and Narayan, 2015a). Indeed, as Hjalmarsson (2010, page 50) points out, "Since the predictable component of stock returns must be small, if indeed one does exist, there seems to be little chance of reaching a

decisive conclusion using U.S. data alone, which effectively provides only one time series at the market level". Motivated by these facts, the purpose of the current section is to provide robust panel data evidence of predictability of stock returns.

The data set that we use is an update of the Hjalmarsson (2010) data set, which consists of country-level observations on excess stock returns (ER), and four predictors, namely, the dividend-price ratio (DP), earnings-price ratio (EP), short term interest rates (SR) and term spread (TS).[2] While the original data set ends in 2004, the updated data set includes 10 more years of data and ends in 2014. The number and choice of countries is the same as in the original sample. Unfortunately, the data coverage varies significantly among countries. We therefore ended up truncating the sample in order to make it balanced. The truncated sample stretches the 1988M4–2012M6 period and covers between 24 and 28 countries, depending on the choice of predictor. As in Hjalmarsson (2010), the full "global" panel is divided in two groups; emerging countries, and developed countries.[3] The list of the countries contained in each group is reported in Table 5.

5.1 Preliminary results

Before we discuss the results of the predictability tests, we first provide some preliminary results on the cross-section and serial correlation properties of the variables. As already mentioned, given the generality of the DGP considered here, such results are not really necessary. The purpose of this section is therefore mainly to provide a feeling for the appropriateness for the DGP laid out in Section 2.

The extent of cross-correlation in each variable can be inferred by considering average of the pair-wise correlation coefficients across all country pairs, and the associated CD test of Pesaran (2004). The results reported in Table 6 show that for all three variables the null hypothesis of no cross-correlation is rejected for all panels. The fact that the cross-correlation is present in all panels suggests that one needs to use a method that is robust to cross-correlation. The non-parametric treatment of the present paper is therefore a great advan-

[2]DP is defined as the ratio between the sum of total dividend payments in the past year and the current price per share. EP is the ratio between the current price per share and the latest 12 months of available earnings. SR is based on three-month treasury bills, private discount rates or interbank rates, depending on the availability of the data. TS is the logarithmic difference between the long term interest rates and the short term interest rates. Long rates are measured by the yield on 10-year bonds or a maturity that is close to 10-year government bonds, depending again on data availability. ER is defined as the ratio of the returns in local currency and the local short rate. ER, DP and EP are all log-transformed. All data are downloaded from the Global Financial Data database.

[3]The grouping is based on the Morgan Stanley capital international (MSCI) classification.

tage, for it does not impose any restrictions on the cross-correlation properties of the predictors and imposes minimal restrictions on the cross-correlation properties of the dependent variable. We also see that there is a high correlation between the extent of cross-correlation in ER, on the one hand, and in DP, EP, SR, and TS, on the other hand, suggesting that much of the correlation in ER can be accounted for by conditioning on DP, EP, SR and/or TS. The allowance of a common factor in the error driving $y_{i,t}$ should therefore be enough to capture any remaining cross-correlations, which again gives credence to the model that we are considering.

The above results have implications not only for the test for predictability, but also when it comes to unit root pretesting, as it invalidates the use of so-called "first generation" panel unit root tests, which presume that the data are independent across the cross-section. In this section we therefore focus on the cross-section augmented IPS (CIPS) test of Pesaran (2007), which can be seen as a cross-correlation robust version of the first-generation test of Im et al. (IPS, 2003). The implementation of this test requires a choice of lag order that should be enough to account for the cross-section and serial correlation in the data. Here we follow the usual convention in the literature and employ the Bayesian information criterion (BIC). The null hypothesis is that of a unit root.

The results reported in Table 7 are for the case with a constant and trend. The results for the constant-only specification were very similar and can be obtained upon request from the corresponding author. It is seen that in most of the cases the unit root null is rejected. The only exceptions are SR for emerging and developed countries, where the null hypothesis is not rejected at the 1% level, and TS (SR and TS) for the developed (global) panel, where the null is not rejected at the 5% level. The bulk of the evidence is therefore against the unit root null. This requires careful interpretation, as a rejection of this test does not provide evidence of stationarity for all countries, but only that there is at least some countries for which stationarity holds. A rejection could therefore be due to stationarity for the panel as a whole, but it could also be due to only a few I(0) series in an otherwise I(1) panel. As a measure of the overall persistency of the variables, we look at the average estimated autoregressive roots obtained as a by-product of the CIPS test. The results reported in Table 7 suggest that while the unit root null tends to be rejected, the predictors are still very persistent. Hence, while ER seems to be I(0), we cannot rule out the possibility that the predictors contain a mix of I(0) and I(1) series, which seems like a very plausible scenario in general.

The discussion in the above paragraph illustrates quite clearly one of the problems involved when wanting to test for predictability and using a test that is conditional on the outcome of a pretest for a unit root. The proposed tests do not require such pretests, and are therefore ideal in situations such as this one when the pretest is inconclusive.

In order to access the significance of the problem of predictor endogeneity, we test if $\varepsilon_{i,t}$ is correlated with the innovations driving $x_{i,t}$ (see, for example, Lewellen, 2004; Westerlund and Narayan, 2012, 2015b, for similar approaches). As an estimator of ε_i, we take $\hat{\varepsilon}_i = M_{\hat{f}}\hat{u}_i$, where $M_{\hat{f}}$ and $\hat{u}_i$ are as in Section 3. The estimator of the innovations driving x_i, denoted $\hat{\eta}_i$, is obtained by taking the OLS residual in a regression of x_i onto $x_{i,-1}$. Our test of endogeneity involves regressing $\hat{\varepsilon}_i$ onto $\hat{\eta}_i$, and then testing the significance of $\hat{\eta}_i$. The results are presented in Table 8. We see that in all cases but one (for the emerging countries when using TS as a predictor) the null hypothesis of no endogeneity is rejected at the 1% level or better. We also see that both the sign and magnitude of the estimated slope on $\hat{\eta}_i$ can differ quite substantially across the predictors, suggesting that a parametric approach might potentially require different models for different predictors.

5.2 Predictability test results

In testing for predictability, while we could also consider joint tests, in this section we follow the bulk of the previous literature and test one predictor at a time (see, for example, Campbell and Yogo, 2006; Hjalmarsson, 2010; Lewellen, 2004; Westerlund and Narayan, 2012, 2015a, 2015b). We focus on the t-statistic, TR, the results of which are reported in Table 9. In interest of comparison, the conventional fixed effects estimator and the estimator of Hjalmarsson (2010) are also considered. The Hjalmarsson (2010) estimator requires that the predictors are nearly I(1) and, as discussed in Remark 3 of Section 3, cannot handle predictors that are endogenous. Hence, given the preliminary results reported in Section 5.1, this estimator is not expected to work very well in the present context. The same is true for the fixed effects estimator, which suffers from the same drawbacks.

The first thing to note is that the results differ quite markedly across the three estimators considered. Specifically, while the results obtained by using the Hjalmarsson (2010) and fixed effects estimators tend to be very similar, these results tend to be quite different from the ones obtained by using the proposed estimator. As an example, consider DP. While according to the proposed estimator, DP is significant at the 5% level for the emerg-

ing economies, according to the Hjalmarsson (2010) and fixed effects estimators, DP is insignificant at all conventional significance levels. Similarly, while according to the proposed estimator, DP is insignificant for the developed countries, according to the competition, it is significant at the 10% level. In fact, it is only for EP and SR that all three estimators lead to the same conclusion regardless of the grouping. This is true not only when looking at the significance, but also when considering the sign and magnitude of the estimated predictive slopes. We believe that said variation in the results illustrates quite clearly the need for robust methods, a conclusion that is supported by the Monte Carlo study of Westerlund and Karabiyik (2015), showing that the Hjalmarsson (2010) estimator can be highly misleading. The proposed estimator is superior in this regard and is therefore the preferred choice.

In view of the discussion of the previous paragraph, in what remains we focus on the results obtained by using the proposed estimator. One observation is that there is only one predictor that is able to significantly predict returns in all groups, namely, SR. The good performance of this predictor is in agreement with studies such as Ang and Bekaert (2001), and Rapach et al. (2005). In a recent working paper, Rapach et al. (2015) look specifically at the predictive ability of SR to predict aggregate US returns. According to their results, SR is the strongest predictor of the equity risk premium identified to date. Our results suggest that this conclusion is true not only for the US, but that it applies also to the wide range of countries considered in the present study. Another observation is that whenever significance is found, DP, EP and TS always enter with their expected positive sign (see Hjalmarsson, 2010, for a discussion). The fact that for DP, EP and TS the evidence of predictability is very weak suggests that the previously obtained evidence of stock return predictability should not be taken at fact value, as the possibility remains that it is due in part to the use of unsuitable econometric techniques.

6 Conclusion

The difficulty of predicting (stock) returns using time series data, typically for the US, has recently motivated researchers to consider panel data as a means to increase the power of conventional (time series) tests. Indeed, since the predictable component of returns is bound to be small, if indeed one does exist, there seems to be little chance of reaching a decisive conclusion based on US data alone. Unfortunately, the few panel data tests that do exist

are based on very restrictive assumptions that are unlikely to hold in applied work. In the present paper we take this as our starting point to develop two new tests that can be used to infer panel predictive regressions under a wide range of empirically relevant specifications of the predictor. In fact, the predictor can have arbitrary serial and cross-section correlation structures, and still the tests are extremely simple and user-friendly. There is also no need for any pretesting of the predictor, which is a great advantage in practice, especially in heterogenous panels where the properties of the predictor can vary quite substantially across the cross-section. In spite of this generality, the tests support asymptotically standard normal and chi-squared inference, a result that is verified in finite samples using Monte Carlo simulation. In fact, the new tests have excellent finite-sample properties and seem to be working well even in relatively small samples.

References

Ang, A., and G. Bekaert (2001). Stock Return Predictability: Is it there? National Bureau of Economic Research Working Paper No. 8207.

Ang, A., Hodrick, R. J., Xing, Y., Zhang, X. (2006). The Cross-Section of Volatility and Expected Returns. *Journal of Finance* **LXI**, 259–299.

Campbell, J. Y., and M. Yogo (2006). Efficient Tests of Stock Return Predictability. *Journal of Financial Economics* **81**, 27–60.

Cavanagh, C., G. Elliott and J. Stock (1995). Inference in Models with Nearly Integrated Regressors. *Econometric Theory* **11**, 1131–1147.

Chakraborty, A., and G. W. Evans (2008). Can Perpetual Learning Explain the Forward-Premium Puzzle? *Journal of Monetary Economics* **55**, 477–490.

Chudik, A. M., H. Pesaran and E. Tosetti (2011). Weak and Strong Cross Section Dependence and Estimation of Large Panels. *Econometrics Journal* **14**, C45–C90.

Cochrane, J. H. (1999). New Facts in Finance. Working paper 7169. National Bureau of Economic Research.

Elliott, G., and J. Stock (1994). Inference in Time Series Regression when the Order of Integration of a Regressor is Unknown. *Econometric Theory* **10**, 672–700.

Ferson, W., S. Sarkissian and T. Simin (2003). Spurious Regression in Financial Economics? *Journal of Finance* **58**, 1393–1413.

Hai, W., Mark, N.C., and Y. Wu (1997) Understanding Spot and Forward Exchange Rate Regressions. *Journal of Applied Econometrics* **12**, 715-734.

Hjalmarsson, E. (2008). The Stambaugh Bias in Panel Predictive Regressions. *Finance Research Letters* **5**, 47–58.

Hjalmarsson, E. (2010). Predicting Global Stock Returns. *Journal of Financial and Quantitative Analysis* **45**, 49–80.

Im, K., M. H. Pesaran and Y. Shin (2003). Testing for Unit Root in Heterogeneous Panels. *Journal of Econometrics* **115**, 53–74.

Kauppi, H. (2001). Panel Data Limit Theory and Asymptotic Analysis of a Panel Regression with Near Integrated Regressors. In Baltagi, B. H., T. B. Fomby and R. C. Hill (Eds.), *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, Advances in Econometrics, Volume 15, 239–274. Emerald Group Publishing Limited.

Lanne, M. (2002). Testing the Predictability of Stock Returns. *The Review of Economics and Statistics* **84**, 407–415.

Lewellen, J. (2004). Predicting Returns with Financial Ratios. *Journal of Financial Economics* **74**, 209–235.

Neely, C. J., D. E. Rapach, J. Tu and G. Zhou (2012). Forecasting the Equity Risk Premium: The Role of Technical Indicators. Federal Reserve Bank of St. Louis Working Paper 2010-008E, Federal Reserve Bank of St. Louis.

Pesaran, M. H. (2004). General Diagnostic Tests for Cross Section Dependence in Panels. CESifo Working Papers. No. 1233. 967–1012.

Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica* **74**, 967–1012.

Pesaran, H. M. (2007). A Simple Panel Unit Root Test in the Presence of Cross Section Dependence. *Journal of Applied Econometrics* **22**, 265–312.

Phillips, P. C. B. (2014). On Confidence Intervals for Autoregressive Roots and Predictive Regression. *Econometrica* **82**, 1177–1195.

Phillips, P. C. B., T. Magdalinos and L. Giraitis (2010). Smoothing Local-to-Moderate Unit Root Theory. *Journal of Econometrics* **158**, 274–279.

Polk, C., S. Thompson and T. Vuolteenaho (2006). Cross-Sectional Forecasts of the Equity Premium. *Journal of Financial Economics* **81**, 101–141.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *Review of Financial Studies* **23**, 821–862.

Rapach, D. E., M. E. Wohar, and J. Rangvid (2005). Macro Variables and International Stock Return Predictability. International *Journal of Forecasting* **21**, 137–166.

Rapach, D. E. and G. Zhou (2013). Forecasting Stock Returns. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Part A, 328–383. Elsevier.

Rapach, D. E., M. Ringgenberg, and G. Zhou (2015). Short Interest and Aggregate Stock Returns. WFA - Center for Finance and Accounting Research Working Paper No. 14/002.

Stambaugh, R. F. (1999). Predictive Regressions. *Journal of Financial Economics* **54**, 375–421.

Torous, W., R. Valkanov and S. Yan (2004). On Predicting Stock Returns with Nearly Integrated Explanatory Variables. *Journal of Business* **77**, 937–966.

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

Westerlund, J., and P. Narayan (2012). Does the Choice of Estimator Matter when Forecasting Returns? *Journal of Banking and Finance* **36**, 2632–2640.

Westerlund, J., and J. Breitung (2013). Lessons From a Decade of IPS and LLC. *Econometric Reviews* **32**, 547–591.

Westerlund, J. (2014). Heteroskedasticity Robust Panel Unit Root Tests. *Journal of Business & Economic Statistics* **32**, 112–135.

Westerlund, J. (2015). The Effect of Recursive Detrending on Panel Unit Root Tests. *Journal of Econometrics* **185**, 453–467.

Westerlund, J., and P. K. Narayan (2015a). A Random Coefficient Approach to the Predictability of Stock Returns in Panels. Forthcoming in *Journal of Financial Econometrics*.

Westerlund, J., and P. K. Narayan (2015b). Testing for Predictability in Conditionally Heteroskedastic Stock Returns. Forthcoming in *Journal of Financial Econometrics*

Westerlund, J., and J.-P. Urbain (2015). Cross-Sectional Averages versus Principal Components. *Journal of Econometrics* **185**, 372–377.

Westerlund, J., and H. Karabiyik (2015). On the Hjalmarsson Estimator of Predictive Panel Regressions. Unpublished manuscript.

Appendix: Proofs

Lemma A.1. *Under $H_0 : \beta_1 = ... = \beta_N = 0$, Assumptions EPS, F, LAM, IND and X–I(1), as $N, T \to \infty$ with $N/T \to 0$,*

$$\sqrt{NT}\hat{\beta} \to_d \left(\lim_{N,T\to\infty} NT^2 \Sigma_x^{-1} \Sigma_0 \Sigma_x^{-1} \right)^{1/2} N(0_{m\times 1}, I_m),$$

where

$$\begin{aligned} \Sigma_0 &= \sum_{i=1}^{N} \sigma_{\varepsilon,i}^2 E(z'_{i,-1} z_{i,-1}), \\ \Sigma_x &= \sum_{i=1}^{N} E[(x^*_{i,-1})' x^{**}_{i,-1}], \end{aligned}$$

with $z_{i,-1} = x^{**}_{i,-1} - N^{-1} \sum_{i=1}^{N} \overline{\lambda}^{-1} \lambda_i x^{**}_{i,-1}$.

Proof of Lemma A.1.

Let $d_t = \hat{f}_t - \overline{\lambda} f_t^*$, where $\hat{f}_t = \overline{y}_t^*$. Under $H_0 : \beta_1 = ... = \beta_N = 0$,

$$\overline{y}_t^* = \overline{u}_t^* = \overline{\lambda} f_t^* + \overline{\varepsilon}_t^*,$$

implying

$$d_t = \hat{f}_t - \overline{\lambda} f_t^* = \overline{\varepsilon}_t^* = O_p(N^{-1/2}). \tag{A1}$$

This is a point-wise result, but it can be shown to hold also uniformly in t.

Let us now consider $\hat{\beta}$. Under H_0,

$$\sqrt{NT}\hat{\beta} = \left(\frac{1}{NT^2} \sum_{i=1}^{N} (x^*_{i,-1})' M_{\hat{f}} x^{**}_{i,-1} \right)^{-1} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' M_{\hat{f}} u_i^*,$$

where

$$\begin{aligned} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' M_{\hat{f}} u_i^* &= \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' M_{\overline{\lambda} f^*} u_i^* - \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' (M_{\overline{\lambda} f^*} - M_{\hat{f}}) u_i^* \\ &= \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' M_{\overline{\lambda} f^*} \varepsilon_i^* - \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' (M_{\overline{\lambda} f^*} - M_{\hat{f}}) f^* \lambda_i \\ &- \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} (x^{**}_{i,-1})' (M_{\overline{\lambda} f^*} - M_{\hat{f}}) \varepsilon_i^* \\ &= R_1 - R_2 - R_3. \end{aligned} \tag{A2}$$

Consider R_3. From the definitions of $M_{\overline{\lambda}f^*}$ and $M_{\hat{f}}$,

$$\begin{aligned} M_{\overline{\lambda}f^*} - M_{\hat{f}} &= d(\hat{f}'\hat{f})^{-1}d' + d(\hat{f}'\hat{f})^{-1}\overline{\lambda}(f^*)'^*\overline{\lambda}(\hat{f}'\hat{f})^{-1}d'^*\overline{\lambda}[(\hat{f}'\hat{f})^{-1} \\ &- (\overline{\lambda}^2(f^*)'^*)^{-1}]\overline{\lambda}(f^*)', \end{aligned}$$

implying

$$\begin{aligned} &||T^{-1}(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})\varepsilon^*_i|| \\ &\leq N^{-1}||\sqrt{N}T^{-1}(x^{**}_{i,-1})'d||\,||(T^{-1}\hat{f}'\hat{f})^{-1}||\,||\sqrt{N}T^{-1}d'\varepsilon^*_i|| \\ &+ (NT)^{-1/2}||\sqrt{N}T^{-1}(x^{**}_{i,-1})'d||\,||(T^{-1}\hat{f}'\hat{f})^{-1}||\,|\overline{\lambda}|\,||T^{-1/2}(f^*)'\varepsilon^*_i|| \\ &+ N^{-1/2}||T^{-1}(x^{**}_{i,-1})'f^*||\,|\overline{\lambda}|\,||(T^{-1}\hat{f}'\hat{f})^{-1}||\,||\sqrt{N}T^{-1}d'\varepsilon^*_i|| \\ &+ ||T^{-1}(x^{**}_{i,-1})'f^*||\,|\overline{\lambda}|^2\,T||(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}||||T^{-1}(f^*)'\varepsilon^*_i||. \end{aligned}$$

It is easy to see that $||T^{-1/2}(f^*)'\varepsilon^*_i||$ and $||T^{-1}(x^{**}_{i,-1})'f^*||$ must be $O_p(1)$. By imposing H_0 we can further show that

$$\begin{aligned} ||\sqrt{N}T^{-1}(x^{**}_{i,-1})'d|| &= ||T^{-1}(x^{**}_{i,-1})'\sqrt{N}\overline{\varepsilon}^*|| = O_p(1), \\ ||\sqrt{N}T^{-1}(\varepsilon^*_i)'d|| &= ||T^{-1}(\varepsilon^*_i)'\sqrt{N}\overline{\varepsilon}^*|| \\ &\leq N^{-1/2}||T^{-1}(\varepsilon^*_i)'\varepsilon^*_i|| + T^{-1/2}\left|\left|T^{-1/2}(\varepsilon^*_i)'\frac{1}{\sqrt{N}}\sum_{n\neq i}^{N}\varepsilon^*_n\right|\right| \\ &= O_p(T^{-1/2}) + O_p(N^{-1/2}), \\ ||\sqrt{N}T^{-1/2}(f^*)'^{-1/2}(f^*)\sqrt{N}\overline{\varepsilon}^*|| &= O_p(1), \\ ||NT^{-1}d'^{-1}(\sqrt{N}\overline{\varepsilon}^*)'\sqrt{N}\overline{\varepsilon}^*|| &= O_p(1). \end{aligned}$$

By using these results and

$$\hat{f}'\hat{f} = (f^*\overline{\lambda} + d)'^*\overline{\lambda} + d) = \overline{\lambda}^2(f^*)'^* + d'^*\overline{\lambda} + \overline{\lambda}(f^*)'d + d'd,$$

we can show that

$$\begin{aligned} T^{-1}||\hat{f}'\hat{f} - \overline{\lambda}^2(f^*)'^*|| &\leq 2(NT)^{-1/2}||\sqrt{N}T^{-1/2}d'^*||\,|\overline{\lambda}| + N^{-1}||NT^{-1}d'd|| \\ &= O_p((NT)^{-1/2}) + O_p(N^{-1}), \end{aligned}$$

which in turn implies

$$\begin{aligned} &T||(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'f^*)^{-1}|| \\ &\leq ||(T^{-1}\hat{f}'\hat{f})^{-1}||T^{-1}||\hat{f}'\hat{f} - \overline{\lambda}^2(f^*)'^*||||(\overline{\lambda}^2T^{-1}(f^*)'^*)^{-1}|| \\ &= O_p((NT)^{-1/2}) + O_p(N^{-1}). \end{aligned}$$

Hence, by adding the results,

$$||T^{-1}(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})\varepsilon^*_i|| = O_p((NT)^{-1/2}) + O_p(N^{-1}), \tag{A3}$$

and therefore,

$$\begin{aligned} ||R_3|| &= \left|\left|\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})\varepsilon^*_i\right|\right| \leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{N}||T^{-1}(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})\varepsilon^*_i|| \\ &= O_p(T^{-1/2}) + O_p(N^{-1/2}). \end{aligned} \tag{A4}$$

Consider R_2. In analogy to the analysis of R_3, we have

$$\begin{aligned} &(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})f^*\lambda_i \\ &= (x^{**}_{i,-1})'d(\hat{f}'\hat{f})^{-1}d'^*\lambda_i + (x^{**}_{i,-1})'d(\hat{f}'\hat{f})^{-1}\overline{\lambda}(f^*)'^*\lambda_i + (x^{**}_{i,-1})'f^*\overline{\lambda}(\hat{f}'\hat{f})^{-1}d'^*\lambda_i \\ &+ (x^{**}_{i,-1})'f^*\overline{\lambda}[(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}]\overline{\lambda}(f^*)'^*\lambda_i \\ &= (x^{**}_{i,-1})'d(\hat{f}'\hat{f})^{-1}d'^*\lambda_i + (x^{**}_{i,-1})'d\overline{\lambda}^{-1}\lambda_i \\ &+ (x^{**}_{i,-1})'d[(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}]\overline{\lambda}(f^*)'^*\lambda_i + (x^{**}_{i,-1})'f^*\overline{\lambda}(\hat{f}'\hat{f})^{-1}d'^*\lambda_i \\ &+ (x^{**}_{i,-1})'f^*\overline{\lambda}[(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}]\overline{\lambda}(f^*)'^*\lambda_i, \end{aligned}$$

from which we obtain

$$\begin{aligned} &T^{-1}||(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})f^*\lambda_i - (x^{**}_{i,-1})'d\overline{\lambda}^{-1}\lambda_i|| \\ &\leq N^{-1}T^{-1/2}||\sqrt{N}T^{-1}(x^{**}_{i,-1})'d||||(T^{-1}\hat{f}'\hat{f})^{-1}||||\sqrt{N}T^{-1/2}d'^*||\,|\lambda_i| \\ &+ N^{-1/2}||\sqrt{N}T^{-1}(x^{**}_{i,-1})'d||T||(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}|||\overline{\lambda}|||T^{-1}(f^*)'^*|||\lambda_i| \\ &+ (NT)^{-1/2}||T^{-1}(x^{**}_{i,-1})'f^*|||\overline{\lambda}|||(T^{-1}\hat{f}'\hat{f})^{-1}||||\sqrt{N}T^{-1/2}d'^*|||\lambda_i| \\ &+ ||T^{-1}(x^{**}_{i,-1})'f^*|||\overline{\lambda}|^2T||(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}||||T^{-1}(f^*)'^*|||\lambda_i| \\ &= O_p((NT)^{-1/2}) + O_p(N^{-1}). \end{aligned} \tag{A5}$$

For $T^{-1}(x^{**}_{i,-1})'d\lambda^{-1}\lambda_i$, we use that

$$\begin{aligned} &\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x^{**}_{i,-1})'d \\ &= N^{-1}\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x^{**}_{i,-1})'\varepsilon^*_i + \frac{1}{NT}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x^{**}_{i,-1})'\frac{1}{\sqrt{N}}\sum_{n\neq i}^{N}\varepsilon^*_n \\ &= \frac{1}{NT}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x^{**}_{i,-1})'\frac{1}{\sqrt{N}}\sum_{n\neq i}^{N}\varepsilon^*_n + O_p(N^{-1}), \end{aligned} \tag{A6}$$

where the remaining term is mean zero. As for the variance of this term, we use

$$\begin{aligned} E[(\varepsilon_{i,t}^*)^2] &= E\left[\left(\varepsilon_{i,t} - \frac{1}{T-t+1}\sum_{n=t}^{T}\varepsilon_{i,n}\right)^2\right] \\ &= E(\varepsilon_{i,t}^2) - \frac{2}{T-t+1}\sum_{n=t}^{T}E(\varepsilon_{i,n}\varepsilon_{i,t}) + \frac{1}{(T-t+1)^2}\sum_{n=t}^{T}\sum_{m=t}^{T}E(\varepsilon_{i,n}\varepsilon_{i,m}) \\ &= \sigma_{\varepsilon,i}^2[1-(T-t+1)^{-1}], \end{aligned}$$

and

$$\begin{aligned} E(\varepsilon_{i,t}^*\varepsilon_{i,s}^*) &= E\left[\left(\varepsilon_{i,t} - \frac{1}{T-t+1}\sum_{n=t}^{T}\varepsilon_{i,n}\right)\left(\varepsilon_{i,s} - \frac{1}{T-s+1}\sum_{m=s}^{T}\varepsilon_{i,m}\right)\right] \\ &= E(\varepsilon_{i,t}\varepsilon_{i,s}) - \frac{1}{T-t+1}\sum_{n=t}^{T}E(\varepsilon_{i,n}\varepsilon_{i,s}) - \frac{1}{T-s+1}\sum_{m=s}^{T}E(\varepsilon_{i,m}\varepsilon_{i,t}) \\ &+ \frac{1}{(T-t+1)(T-s+1)}\sum_{n=t}^{T}\sum_{m=s}^{T}E(\varepsilon_{i,n}\varepsilon_{i,m}) \\ &= -\frac{1}{T-s+1}E(\varepsilon_{i,t}^2) + \frac{1}{(T-t+1)(T-s+1)}\sum_{n=t}^{T}E(\varepsilon_{i,n}^2) = 0 \end{aligned}$$

for $t > s$. It follows that

$$E[\varepsilon_n^*(\varepsilon_n^*)'] = \sigma_{\varepsilon,i}^2\text{diag}[1-(T-1)^{-1}, 1-(T-2)^{-1}, ..., 1-1^{-1}] = \sigma_{\varepsilon,i}^2(I_{T-1} - A_{T-1}),$$

where $A_{T-1} = \text{diag}[(T-1)^{-1}, (T-2)^{-1}, ..., 1^{-1}]$. Therefore, letting $\sigma_\varepsilon^2 = N^{-1}\sum_{i=1}^{N}\sigma_{\varepsilon,i}^2$ and

$$\Sigma_2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{n=1}^{N}\sigma_{\varepsilon,n}^2 E[\bar{\lambda}^{-2}\lambda_i\lambda_j(x_{i,-1}^{**})'x_{j,-1}^{**}],$$

we have

$$
\begin{aligned}
& E\left(\frac{1}{N^3T^2}\sum_{i=1}^{N}\sum_{n\neq i}^{N}\sum_{m\neq j}^{N}\sum_{j=1}^{N}\overline{\lambda}^{-1}\lambda_i(x_{i,-1}^{**})'\varepsilon_n^*(\varepsilon_m^*)'\overline{\lambda}^{-1}\lambda_j x_{j,-1}^{**}\right) \\
&= E\left(\frac{1}{N^3T^2}\sum_{i\neq m}^{N}\sum_{n\neq i}^{N}\sum_{m\neq j}^{N}\sum_{j\neq n}^{N}\overline{\lambda}^{-2}\lambda_i\lambda_j(x_{i,-1}^{**})'E[\varepsilon_n^*(\varepsilon_m^*)'|x_{i,-1}^{**},x_{j,-1}^{**}]x_{j,-1}^{**}\right) \\
&+ E\left(\frac{1}{N^3T^2}\sum_{i=1}^{N}\sum_{n\neq i}^{N}\sum_{m\neq n}^{N}\overline{\lambda}^{-2}\lambda_i\lambda_n(x_{i,-1}^{**})'\varepsilon_n^*E[(\varepsilon_m^*)'|x_{i,-1}^{**},x_{n,-1}^{**}]x_{n,-1}^{**}\right) \\
&+ E\left(\frac{1}{N^3T^2}\sum_{n\neq m}^{N}\sum_{m\neq j}^{N}\sum_{j\neq n}^{N}\overline{\lambda}^{-2}\lambda_m\lambda_j(x_{m,-1}^{**})'E(\varepsilon_n^*|x_{m,-1}^{**},x_{j,-1}^{**})(\varepsilon_m^*)'x_{j,-1}^{**}\right) \\
&= E\left(\frac{1}{N^3T^2}\sum_{i\neq m}^{N}\sum_{n\neq i}^{N}\sum_{m\neq j}^{N}\sum_{j\neq n}^{N}\overline{\lambda}^{-2}\lambda_i\lambda_j(x_{i,-1}^{**})'E[\varepsilon_n^*(\varepsilon_m^*)'|x_{i,-1}^{**},x_{j,-1}^{**}]x_{j,-1}^{**}\right) \\
&= E\left(\frac{1}{N^3T^2}\sum_{i\neq n}^{N}\sum_{n\neq i}^{N}\sum_{j\neq n}^{N}\overline{\lambda}^{-2}\lambda_i\lambda_j(x_{i,-1}^{**})'E[\varepsilon_n^*(\varepsilon_n^*)'|x_{i,-1}^{**},x_{j,-1}^{**}]x_{j,-1}^{**}\right) \\
&= \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}T^{-2}E[\overline{\lambda}^{-2}\lambda_i\lambda_j(x_{i,-1}^{**})'\sigma_\varepsilon^2(I_{T-1}-A_{T-1})x_{j,-1}^{**}] \\
&= \sigma_\varepsilon^2\frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T}T^{-2}[1-(T-t+1)^{-1}]E[\overline{\lambda}^{-2}\lambda_i\lambda_j x_{i,t-1}^{**}(x_{j,t-1}^{**})'] \\
&= \sigma_\varepsilon^2\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T}T^{-2}E[\overline{\lambda}^{-2}\lambda_i\lambda_j x_{i,t-1}^{**}(x_{j,t-1}^{**})'] + O_p(T^{-1}) \\
&\to_p \lim_{N,T\to\infty}\sigma_\varepsilon^2\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}T^{-2}E[\overline{\lambda}^{-2}\lambda_i\lambda_j(x_{i,-1}^{**})'x_{j,-1}^{**}] = \lim_{N,T\to\infty}N^{-3}T^{-2}\Sigma_2 \qquad \text{(A7)}
\end{aligned}
$$

as $N,\ T\to\infty$. By using this and the fact that $x_{i,t-1}^{**}$ and $\varepsilon_{n,t}^*$ are independent for all $i\neq n$, we obtain

$$
\begin{aligned}
\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x_{i,-1}^{**})'d &= \frac{1}{NT}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x_{i,-1}^{**})'\frac{1}{\sqrt{N}}\sum_{n\neq i}^{N}\varepsilon_n^* + O_p(N^{-1}) \\
&\to_d \left(\lim_{N,T\to\infty}N^{-3}T^{-2}\Sigma_2\right)^{1/2}N(0_{m\times 1},I_m), \qquad \text{(A8)}
\end{aligned}
$$

as $N,\ T\to\infty$, and subsequently,

$$
\begin{aligned}
R_2 &= \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x_{i,-1}^{**})'(M_{\overline{\lambda}f^*}-M_{\hat{f}})f^*\lambda_i \\
&= \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x_{i,-1}^{**})'d\overline{\lambda}^{-1}\lambda_i + O_p(T^{-1/2}) + O_p(N^{-1/2}) \\
&\to_d \left(\lim_{N,T\to\infty}N^{-3}T^{-2}\Sigma_2\right)^{1/2}N(0_{m\times 1},I_m). \qquad \text{(A9)}
\end{aligned}
$$

For R_1, note that $M_{\overline{\lambda} f^*} = M_{f^*}$. This result, together with the independence of $x^{**}_{i,t-1}$ and $\varepsilon^*_{n,t}$, and the fact that

$$\begin{aligned} ||T^{-1}(x^{**}_{i,-1})'(I_{T-1} - M_{f^*})\varepsilon^*_i|| &= ||T^{-1}(x^{**}_{i,-1})'P_{f^*}\varepsilon^*_i|| \\ &\leq T^{-1/2}||T^{-1}(x^{**}_{i,-1})'f^*||||(T^{-1}(f^*)'^*)^{-1}||||T^{-1/2}(f^*)'\varepsilon^*_i|| \\ &= O_p(T^{-1/2}), \end{aligned}$$

implies

$$\begin{aligned} R_1 &= \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x^{**}_{i,-1})'\varepsilon^*_i - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}T^{-1}(x^{**}_{i,-1})'(I_{T-1} - M_{f^*})\varepsilon^*_i \\ &= \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x^{**}_{i,-1})'\varepsilon^*_i + O_p(\sqrt{N}T^{-1/2}). \end{aligned} \tag{A10}$$

The first term on the right-hand side of this equation is clearly mean zero. Moreover, by the same steps used for deriving Σ_2, the variance can be shown to be given by

$$\begin{aligned} &E\left(\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x^{**}_{i,-1})'\varepsilon^*_i(\varepsilon^*_j)'x^{**}_{j,-1}\right) \\ &= E\left(\frac{1}{NT^2}\sum_{i=1}^{N}(x^{**}_{i,-1})'E[\varepsilon^*_i(\varepsilon^*_i)'|x^{**}_{i,-1}]x^{**}_{i,-1}\right) \\ &\to_p \lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}T^{-2}\sigma^2_{\varepsilon,i}E[(x^{**}_{i,-1})'x^{**}_{i,-1}] = \lim_{N,T\to\infty}N^{-1}T^{-2}\Sigma_1, \end{aligned} \tag{A11}$$

where

$$\Sigma_1 = \sum_{i=1}^{N}\sigma^2_{\varepsilon,i}E[(x^{**}_{i,-1})'x^{**}_{i,-1}].$$

This implies

$$R_1 = \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x^{**}_{i,-1})'M_{\overline{\lambda} f^*}\varepsilon^*_i \to_d \left(\lim_{N,T\to\infty}N^{-1}T^{-2}\Sigma_1\right)^{1/2}N(0_{m\times 1}, I_m) \tag{A12}$$

as N, $T \to \infty$ with $N/T \to 0$.

R_1 and R_2 are not independent, but correlated random variables. Letting

$$\Sigma_{12} = \sum_{j=1}^{N}\sum_{n=1}^{N}\sigma^2_{\varepsilon,n}E[\overline{\lambda}^{-1}\lambda_j(x^{**}_{j,-1})'x^{**}_{n,-1}],$$

the covariance between them is given by

$$\begin{aligned}
&\frac{1}{(NT)^2}E\left(\sum_{j=1}^{N}\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'\sum_{n\neq j}^{N}\varepsilon_n^*\sum_{i=1}^{N}(\varepsilon_i^*)'x_{i,-1}^{**}\right)\\
&= \frac{1}{(NT)^2}E\left(\sum_{j=1}^{N}\sum_{i\neq j}^{N}\sum_{n\neq j}^{N}\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'E[\varepsilon_n^*(\varepsilon_i^*)'|x_{j,-1}^{**},x_{i,-1}^{**}]x_{i,-1}^{**}\right)\\
&+ \frac{1}{(NT)^2}E\left(\sum_{j=1}^{N}\sum_{n\neq j}^{N}\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'E(\varepsilon_n^*|x_{j,-1})(\varepsilon_j^*)'x_{j,-1}^{**}\right)\\
&= \frac{1}{(NT)^2}E\left(\sum_{j=1}^{N}\sum_{i\neq j}^{N}\sum_{n\neq j}^{N}\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'E[\varepsilon_n^*(\varepsilon_i^*)'|x_{j,-1}^{**},x_{i,-1}^{**}]x_{i,-1}^{**}\right)\\
&= \frac{1}{(NT)^2}E\left(\sum_{j=1}^{N}\sum_{n\neq j}^{N}\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'E[\varepsilon_n^*(\varepsilon_n^*)'|x_{j,-1}^{**},x_{n,-1}^{**}]x_{n,-1}^{**}\right)\\
&= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{n\neq j}^{N}T^{-2}\sigma_{\varepsilon,n}^2E[\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'x_{n,-1}^{**}]+O_p(T^{-1})\\
&\to_p \lim_{N,T\to\infty}\frac{1}{N^2}\sum_{j=1}^{N}\sum_{n=1}^{N}T^{-2}\sigma_{\varepsilon,n}^2E[\overline{\lambda}^{-1}\lambda_j(x_{j,-1}^{**})'x_{n,-1}^{**}]=\lim_{N,T\to\infty}(NT)^{-2}\Sigma_{12}. \qquad \text{(A13)}
\end{aligned}$$

It follows that

$$\lim_{N,T\to\infty}\text{var}(R_1-R_2)=N^{-1}T^{-2}\Sigma_1+N^{-3}T^{-2}\Sigma_2-2(NT)^{-2}\Sigma_{12}, \qquad \text{(A14)}$$

where the right-hand side is identically the asymptotic variance of $N^{-1/2}T^{-1}\sum_{i=1}^{N}z_{i,-1}'\varepsilon_i^*$, where $z_{i,-1}=x_{i,-1}^{**}-N^{-1}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i x_{i,-1}^{**}$. In fact, by using the results provided for R_1 and R_2, it is not difficult to show that

$$\begin{aligned}
R_1-R_2 &= \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x_{i,-1}^{**})'(\varepsilon_i^*-d\overline{\lambda}^{-1}\lambda_i)+O_p(\sqrt{N}T^{-1/2})\\
&= \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}z_{i,-1}'\varepsilon_i^*+O_p(\sqrt{N}T^{-1/2}). \qquad \text{(A15)}
\end{aligned}$$

Hence, if we let

$$\Sigma_0=\text{var}\left(\sum_{i=1}^{N}z_{i,-1}'\varepsilon_i^*\right)=\sum_{i=1}^{N}E(z_{i,-1}'E[\varepsilon_i^*(\varepsilon_i^*)'|x_{1,-1}^{**},...,x_{N,-1}^{**}]z_{i,-1})=\sum_{i=1}^{N}\sigma_{\varepsilon,i}^2E(z_{i,-1}'z_{i,-1}),$$

then

$$\lim_{N,T\to\infty}\text{var}(R_1-R_2)=N^{-1}T^{-2}\Sigma_0. \qquad \text{(A16)}$$

By combining the results obtained for R_1, R_2 and R_3, the limit of the numerator of $\sqrt{N}T(\hat{\beta}-\beta)$ can be written as

$$\begin{aligned}\frac{1}{\sqrt{N}T}\sum_{i=1}^{N}(x_{i,-1}^{**})'M_{\hat{f}}u_i^* &= R_1+R_2+R_3 = R_1+R_2+o_p(1)\\ &\to_d \left(\lim_{N,T\to\infty}N^{-1}T^{-2}\Sigma_0\right)^{1/2}N(0_{m\times 1},I_m) \qquad \text{(A17)}\end{aligned}$$

which holds provided that N, $T\to\infty$ with $N/T\to 0$.

Let us now consider the denominator of $\sqrt{N}T(\hat{\beta}-\beta)$, which we write in the following fashion:

$$\frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**}=\frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\overline{\lambda}f^*}x_{i,-1}^{**}-\frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'(M_{\overline{\lambda}f^*}-M_{\hat{f}})x_{i,-1}^{**},$$

where

$$\begin{aligned}&||T^{-2}(x_{i,-1}^{*})'(M_{\overline{\lambda}f^*}-M_{\hat{f}})x_{i,-1}^{**}||\\ &= (NT)^{-1}||\sqrt{N}T^{-1}(x_{i,-1}^{*})'^{2}||(T^{-1}\hat{f}'\hat{f})^{-1}||\\ &+ 2N^{-1/2}T^{-1}||\sqrt{N}T^{-1}(x_{i,-1}^{*})'^{-1}\hat{f}'\hat{f})^{-1}||||\overline{\lambda}||||T^{-1}(f^*)'x_{i,-1}^{**}||\\ &+ T^{-1}||T^{-1}(x_{i,-1}^{*})'^{*}|||\overline{\lambda}|^2T||(\hat{f}'\hat{f})^{-1}-(\overline{\lambda}^2(f^*)'^{*})^{-1}||||T^{-1}(f^*)'x_{i,-1}^{**}||\\ &= O_p((NT)^{-1}),\end{aligned}$$

giving

$$\begin{aligned}\left|\left|\frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'(M_{\overline{\lambda}f^*}-M_{\hat{f}})x_{i,-1}^{**}\right|\right| &\le \frac{1}{N}\sum_{i=1}^{N}||T^{-2}(x_{i,-1}^{*})'(M_{\overline{\lambda}f^*}-M_{\hat{f}})x_{i,-1}^{**}||\\ &= O_p((NT)^{-1}).\end{aligned}$$

Hence, since

$$\begin{aligned}&||T^{-2}(x_{i,-1}^{*})'(I_{T-1}-M_{\overline{\lambda}f^*})x_{i,-1}^{**}||\\ &= ||T^{-2}(x_{i,-1}^{*})'(I_{T-1}-M_{f^*})x_{i,-1}^{**}|| = ||T^{-2}(x_{i,-1}^{*})'P_{f^*}x_{i,-1}^{**}||\\ &= T^{-1}||T^{-1}(x_{i,-1}^{*})'^{*}||||(T^{-1}(f^*)'f^*)^{-1}||||T^{-1}(f^*)'x_{i,-1}^{**}||\\ &= O_p(T^{-1}),\end{aligned}$$

we can show that

$$\begin{aligned}\frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**} &= \frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\overline{\lambda}f^{*}}x_{i,-1}^{**} - \frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'(M_{\overline{\lambda}f^{*}} - M_{\hat{f}})x_{i,-1}^{**} \\ &= \frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'x_{i,-1}^{**} + O_p(T^{-1}) \\ &\to_p \lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}T^{-2}E[(x_{i,-1}^{*})'x_{i,-1}^{**}] = \lim_{N,T\to\infty}N^{-1}T^{-2}\Sigma_x, \quad \text{(A18)}\end{aligned}$$

with an obvious definition of Σ_x.

Hence, by putting the results together,

$$\begin{aligned}\sqrt{NT}\hat{\beta} &= \left(\frac{1}{NT^2}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**}\right)^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'M_{\hat{f}}u_i^{*} \\ &\to_d [(N^{-1}T^{-2}\Sigma_x)^{-1}N^{-1}T^{-2}\Sigma_0(N^{-1}T^{-2}\Sigma_x)^{-1}]^{1/2}N(0_{m\times 1}, I_m) \quad \text{(A19)}\end{aligned}$$

as N, $T \to \infty$ with $N/T \to 0$. ■

Lemma A.2. *Under $H_0 : \beta_1 = ... = \beta_N = 0$, Assumptions EPS, F, LAM, IND and X–I(0), as N, $T \to \infty$ with $N/T \to 0$,*

$$\sqrt{NT}\hat{\beta} \to_d (NT\Sigma_x^{-1}\Sigma_0\Sigma_x^{-1})^{1/2}N(0_{m\times 1}, I_m). \quad \text{(A20)}$$

Proof of Lemma A.2.

The proof of Lemma A.2 follows from simple manipulations of that of Lemma A.1. In terms of the notation of Proof of Lemma A.1, under H_0,

$$\begin{aligned}\sqrt{NT}\hat{\beta} &= \left(\frac{1}{NT}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**}\right)^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'M_{\hat{f}}u_i^{*} \\ &= \left(\frac{1}{NT}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**}\right)^{-1}\sqrt{T}(R_1 - R_2 - R_3).\end{aligned}$$

Consider $\sqrt{T}R_3$. Clearly, all the results of Lemma A.1 not depending on $x_{i,t-1}$ hold also in the present case. One change is that now

$$||\sqrt{N}T^{-1/2}(x_{i,-1}^{**})'d|| = ||T^{-1/2}(x_{i,-1}^{**})'\sqrt{N}\overline{\varepsilon}^{*}|| = O_p(1).$$

Also, since f_t is serially unclorrelated, $||T^{-1/2}(x^{**}_{i,-1})'f^*||$ is of the same order. It follows that

$$\begin{aligned}
&T^{-1/2}(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})\varepsilon^*_i \\
&\leq \; N^{-1}\sqrt{N}T^{-1/2}(x^{**}_{i,-1})'d(T^{-1}\hat{f}'\hat{f})^{-1}\sqrt{N}T^{-1}d'\varepsilon^*_i \\
&+ \; \overline{\lambda}(NT)^{-1/2}\sqrt{N}T^{-1/2}(x^{**}_{i,-1})'d(T^{-1}\hat{f}'\hat{f})^{-1}T^{-1/2}(f^*)'\varepsilon^*_i \\
&+ \; \overline{\lambda}(NT)^{-1/2}T^{-1/2}(x^{**}_{i,-1})'f^*(T^{-1}\hat{f}'\hat{f})^{-1}\sqrt{N}T^{-1}d'\varepsilon^*_i \\
&+ \; \overline{\lambda}^2T^{-1/2}T^{-1/2}(x^{**}_{i,-1})'f^*T[(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}]T^{-1/2}(f^*)'\varepsilon^*_i \\
&= \; \overline{\lambda}(NT)^{-1/2}T^{-1/2}(x^{**}_{i,-1})'f^*(T^{-1}\hat{f}'\hat{f})^{-1}\sqrt{N}T^{-1}d'\varepsilon^*_i + O_p(N^{-1}) + O_p((NT)^{-1/2}).
\end{aligned}$$

Here

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} T^{-1/2}(x^{**}_{i,-1})'f^*(T^{-1}\hat{f}'\hat{f})^{-1}\sqrt{N}T^{-1}d'\varepsilon^*_i \\
&= \; \frac{1}{N^{3/2}}\sum_{i=1}^{N} T^{-1/2}(x^{**}_{i,-1})'f^*(T^{-1}\hat{f}'\hat{f})^{-1}T^{-1}(\varepsilon^*_i)'\varepsilon^*_i \\
&+ \; \frac{1}{N\sqrt{T}}\sum_{i=1}^{N} T^{-1/2}(x^{**}_{i,-1})'f^*(T^{-1}\hat{f}'\hat{f})^{-1}T^{-1/2}(\varepsilon^*_i)'\frac{1}{\sqrt{N}}\sum_{n\neq i}^{N}\varepsilon^*_n \\
&= \; O_p(N^{-1/2}) + O_p(T^{-1/2}),
\end{aligned}$$

implying that

$$||\sqrt{T}R_3|| = \left|\left|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})\varepsilon^*_i\right|\right| = O_p(T^{-1/2}) + O_p(N^{-1/2}). \tag{A21}$$

For $\sqrt{T}R_2$, we use

$$\begin{aligned}
&T^{-1/2}||(x^{**}_{i,-1})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})f^*\lambda_i - (x^{**}_{i,-1})'d\overline{\lambda}^{-1}\lambda_i|| \\
&\leq \; N^{-1}T^{-1/2}||\sqrt{N}T^{-1/2}(x^{**}_{i,-1})'d||||(T^{-1}\hat{f}'\hat{f})^{-1}||||\sqrt{N}T^{-1/2}d'^*||\,|\lambda_i| \\
&+ \; N^{-1/2}||\sqrt{N}T^{-1/2}(x^{**}_{i,-1})'d||T||(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}|||\overline{\lambda}|||T^{-1}(f^*)'^*|||\lambda_i| \\
&+ \; (NT)^{-1/2}||T^{-1/2}(x^{**}_{i,-1})'f^*|||\overline{\lambda}|||(T^{-1}\hat{f}'\hat{f})^{-1}||||\sqrt{N}T^{-1/2}d'^*|||\lambda_i| \\
&+ \; ||T^{-1/2}(x^{**}_{i,-1})'f^*|||\overline{\lambda}|^2T||(\hat{f}'\hat{f})^{-1} - (\overline{\lambda}^2(f^*)'^*)^{-1}||||T^{-1}(f^*)'^*|||\lambda_i| \\
&= \; O_p((NT)^{-1/2}) + O_p(N^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x^{**}_{i,-1})'d \;&=\; \frac{1}{N\sqrt{T}}\sum_{i=1}^{N}\overline{\lambda}^{-1}\lambda_i(x^{**}_{i,-1})'\frac{1}{\sqrt{N}}\sum_{n\neq i}^{N}\varepsilon^*_n + O_p(N^{-1}) \\
&\to_d \; \left(\lim_{N,T\to\infty} N^{-2}T^{-1}\Sigma_2\right)^{1/2} N(0_{m\times 1}, I_m),
\end{aligned}$$

leading to the following limit as $N,\ T \to \infty$:

$$\begin{aligned}
\sqrt{T}R_2 &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'(M_{\overline{\lambda}f^*} - M_{\hat{f}})f^*\lambda_i \\
&= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'d\overline{\lambda}^{-1}\lambda_i + O_p(T^{-1/2}) + O_p(N^{-1/2}) \\
&\to_d \left(\lim_{N,T\to\infty} N^{-2}T^{-1}\Sigma_2\right)^{1/2} N(0_{m\times 1}, I_m) \qquad \text{(A22)}
\end{aligned}$$

Moreover,

$$\begin{aligned}
\sqrt{T}R_1 &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'M_{\overline{\lambda}f^*}\varepsilon_i^* \\
&= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'\varepsilon_i^* - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}T^{-1}(x_{i,-1}^{**})'(I_{T-1} - M_{f^*})\varepsilon_i^* \\
&= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'\varepsilon_i^* + O_p(\sqrt{N}T^{-1/2}) \\
&\to_d \left(\lim_{N,T\to\infty}(NT)^{-1}\Sigma_1\right)^{1/2} N(0_{m\times 1}, I_m) \qquad \text{(A23)}
\end{aligned}$$

as $N,\ T \to \infty$ with $N/T \to 0$.

We can similarly show that the covariance between $\sqrt{T}R_1$ and $\sqrt{T}R_2$ is given by $T\Sigma_{12}$. Hence,

$$\lim_{N,T\to\infty} T \cdot \text{var}(R_1 - R_2) = (NT)^{-1}\Sigma_1 + N^{-3}T^{-1}\Sigma_2 - 2N^{-2}T^{-1}\Sigma_{12} = (NT)^{-1}\Sigma_0,$$

where, similarly to before,

$$\begin{aligned}
\sqrt{T}(R_1 - R_2) &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(x_{i,-1}^{**})'(\varepsilon_i^* - d\overline{\lambda}^{-1}\lambda_i) + O_p(\sqrt{N}T^{-1/2}) \\
&= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}z_{i,-1}'\varepsilon_i^* + O_p(\sqrt{N}T^{-1/2}).
\end{aligned}$$

The above results plus

$$\frac{1}{NT}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**} = \frac{1}{NT}\sum_{i=1}^{N}(x_{i,-1}^{*})'x_{i,-1}^{**} + O_p(T^{-1}) \to_p \lim_{N,T\to\infty}(NT)^{-1}\Sigma_x, \qquad \text{(A24)}$$

implies that

$$\begin{aligned}
\sqrt{NT}\hat{\beta} &= \left(\frac{1}{NT}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**}\right)^{-1}\sqrt{T}(R_1 - R_2 - R_3) \\
&= \left(\frac{1}{NT}\sum_{i=1}^{N}(x_{i,-1}^{*})'M_{\hat{f}}x_{i,-1}^{**}\right)^{-1}\sqrt{T}(R_1 - R_2) + o_p(1) \\
&\to_d ([(NT)^{-1}\Sigma_x]^{-1}(NT)^{-1}\Sigma_0[(NT)^{-1}\Sigma_x]^{-1})^{1/2}N(0_{m\times 1}, I_m) \qquad \text{(A25)}
\end{aligned}$$

as N, $T \to \infty$ with $N/T \to 0$. ■

Proof of Theorem 1.

This proof is a simple consequence of Lemmas A.1 and A.2. In particular, from Lemma A.1 we have that under Assumption X–I(1),

$$W = \hat{\beta}'(\Sigma_x \Sigma_0^{-1} \Sigma_x)\hat{\beta} = \sqrt{N}T\hat{\beta}'^{-1}T^{-2}\Sigma_x(N^{-1}T^{-2}\Sigma_0)^{-1}N^{-1}T^{-2}\Sigma_x](\sqrt{N}T\hat{\beta}) \to_d \chi^2(m)$$

as N, $T \to \infty$ with $N/T \to 0$. Under Assumption X–I(0), on the other hand, according to Lemma A.2,

$$W = \sqrt{NT}\hat{\beta}'^{-1}\Sigma_x((NT)^{-1}\Sigma_0)^{-1}(NT)^{-1}\Sigma_x](\sqrt{NT}\hat{\beta}) \to_d \chi^2(m).$$

Therefore, $W \to_d \chi^2(m)$ under both Assumption X–I(1) and X–I(0). ■

Proof of Corollary 1.

This proof from the same arguments as in Theorem 1. It is therefore omitted. ■

Table 1: Simulated 5% size and power of TR in the case of a single predictor.

N	T	Size	P1	P2	P3	P4	P5
R1: $\rho = 1$							
10	50	2.8	8.8	25.9	15.6	30.9	40.7
10	100	3.0	23.6	50.4	23.6	51.5	53.5
10	200	3.5	49.6	72.5	36.9	68.5	66.2
20	50	3.5	9.5	28.1	16.9	33.1	42.3
20	100	3.6	24.7	52.6	24.7	54.0	57.1
20	200	3.5	51.0	74.3	36.5	72.3	70.0
30	50	4.0	9.9	28.8	16.7	34.0	45.4
30	100	3.9	26.2	55.6	26.2	56.6	61.1
30	200	3.7	54.7	76.8	40.4	74.9	72.4
R2: $\rho = 1 - 2T^{-1}$							
10	50	4.7	20.4	53.6	34.9	47.8	50.9
10	100	4.3	52.1	82.1	52.1	71.2	66.1
10	200	4.8	82.6	93.1	70.5	84.8	78.9
20	50	4.6	31.0	70.5	50.5	60.9	49.4
20	100	4.9	69.9	91.0	69.9	81.5	66.9
20	200	4.4	92.5	97.3	85.1	92.0	80.0
30	50	4.7	39.8	81.5	62.6	68.9	55.8
30	100	4.2	81.9	95.5	81.9	87.8	74.0
30	200	4.5	95.3	97.4	91.0	94.5	83.8
R3: $\rho = 1 - 2T^{-9/10}$							
10	50	5.0	23.7	61.6	41.2	53.5	53.1
10	100	4.7	61.5	90.4	61.5	78.3	70.3
10	200	4.8	90.9	97.1	80.4	90.9	83.5
20	50	4.7	37.8	80.8	59.8	68.8	52.1
20	100	5.1	80.0	96.1	80.0	87.8	70.6
20	200	4.6	97.1	99.1	92.8	96.1	83.5
30	50	5.1	48.0	89.2	72.3	77.5	59.1
30	100	4.9	89.8	98.3	89.8	93.5	77.1
30	200	4.6	98.4	99.3	96.5	97.4	87.7
R4: $\rho = 0.8$							
10	50	5.7	29.1	75.3	50.8	66.3	59.5
10	100	5.3	67.8	97.8	67.8	88.0	76.8
10	200	5.3	93.9	100.0	76.3	94.8	86.6
20	50	4.9	47.6	92.3	73.6	83.2	59.6
20	100	5.4	87.3	99.9	87.3	96.2	77.2
20	200	5.0	99.4	100.0	93.4	98.7	86.8
30	50	5.5	60.6	97.6	86.1	90.7	70.6
30	100	4.7	96.0	100.0	96.0	98.8	86.1
30	200	4.5	99.9	100.0	98.5	99.8	92.9

Notes: ρ refers to the autoregressive root of the predictor. Let β_i denote the predictive slope. The reported rejection frequencies for P1–P5 represent power. "P1"–"P3" refer to the case when $\beta_i = 0.1 \cdot 1_{2\times 1}$, $\beta_i = 0.2 \cdot 1_{2\times 1}$ and $\beta_i = T^{-1/2} \cdot 1_{2\times 1}$, respectively. "P4" refers to the case when $\beta_i = 0.2 \cdot 1_{2\times 1}$ for $i = 1, ..., N/5$ and $\beta_i = 0 \cdot 1_{2\times 1}$ for $i = N/5 + 1, ..., N$, whereas "P5" refers to the case when $\beta_i \sim N(0_{2\times 1}, I_m)$ for $i = 1, ..., N/5$ and $\beta_i = 0 \cdot 1_{2\times 1}$ for $i = N/5 + 1, ..., N$.

Table 2: Simulated 5% size and power of W in the case of a single predictor.

N	T	Size	P1	P2	P3	P4	P5
R1: $\rho = 1$							
10	50	4.4	8.2	23.3	13.4	27.4	37.1
10	100	4.5	19.3	46.6	19.3	47.6	50.5
10	200	4.1	45.1	69.5	32.0	65.4	63.6
20	50	5.0	9.1	25.4	14.9	29.8	38.1
20	100	4.2	21.3	48.2	21.3	50.7	53.7
20	200	3.8	46.0	72.0	31.8	69.5	66.9
30	50	5.2	9.0	25.6	14.4	30.5	41.5
30	100	4.7	22.0	52.1	22.0	53.0	57.8
30	200	4.6	50.4	74.4	35.6	72.7	70.2
R2: $\rho = 1 - 2T^{-1}$							
10	50	5.0	15.4	46.9	28.3	42.2	45.9
10	100	5.1	45.1	79.2	45.1	67.8	62.4
10	200	4.4	79.4	92.1	64.9	82.7	76.7
20	50	4.9	24.4	64.9	43.3	55.4	44.5
20	100	5.0	64.2	89.7	64.2	78.9	63.5
20	200	4.5	90.7	96.9	81.6	90.8	77.9
30	50	5.0	32.3	76.5	54.8	63.6	51.1
30	100	4.8	77.3	94.8	77.3	85.8	70.9
30	200	4.8	94.4	97.2	89.0	93.8	81.6
R3: $\rho = 1 - 2T^{-9/10}$							
10	50	5.0	18.4	54.8	33.4	48.1	48.5
10	100	5.2	54.4	88.1	54.4	74.6	66.6
10	200	5.1	88.8	96.7	75.9	89.7	81.4
20	50	4.7	29.7	75.1	51.9	64.0	47.3
20	100	5.4	74.7	95.4	74.7	85.7	66.8
20	200	4.6	96.3	98.9	90.6	95.6	81.5
30	50	5.1	39.6	85.8	64.9	73.1	55.4
30	100	4.6	86.5	97.8	86.5	92.5	74.0
30	200	5.0	98.0	99.2	95.5	97.2	86.4
R4: $\rho = 0.8$							
10	50	5.4	21.5	68.3	41.1	59.7	54.4
10	100	5.5	58.8	96.7	58.8	85.3	73.8
10	200	5.3	90.6	99.9	67.9	93.9	84.6
20	50	5.5	37.8	88.9	65.6	78.8	53.9
20	100	5.0	81.3	99.9	81.3	95.1	73.8
20	200	5.5	99.0	100.0	89.5	98.5	84.6
30	50	4.9	51.3	96.4	80.4	87.6	66.6
30	100	4.7	93.4	100.0	93.4	98.4	83.9
30	200	4.5	99.9	100.0	97.6	99.8	91.9

Notes: See Table 1 for a description.

Table 3: Simulated 5% size and power of TR in the case of two predictors but testing only the first.

N	T	Size	P1	P2	P3	P4	P5
R1: $\rho = 1$							
10	50	7.0	33.5	54.6	43.8	33.1	49.7
10	100	6.0	54.6	72.7	54.6	51.4	60.0
10	200	5.7	73.2	85.3	65.4	68.4	68.5
20	50	6.8	42.6	62.2	53.2	39.4	49.6
20	100	6.2	62.2	76.4	62.2	58.8	63.4
20	200	6.3	77.9	87.4	72.3	72.7	73.1
30	50	7.0	48.8	66.6	58.3	44.2	54.3
30	100	6.4	68.3	80.6	68.3	63.3	66.5
30	200	5.8	80.3	89.5	75.8	76.2	75.0
R2: $\rho = 1 - 2T^{-1}$							
10	50	6.5	54.0	81.7	70.5	48.4	61.6
10	100	5.3	83.3	93.7	83.3	74.9	75.9
10	200	5.9	94.0	97.5	89.5	88.2	83.6
20	50	6.0	75.4	92.0	86.3	69.7	65.4
20	100	5.7	92.7	96.8	92.7	88.6	79.3
20	200	5.3	97.0	98.7	95.4	94.7	87.7
30	50	6.1	84.7	94.2	91.6	79.0	75.5
30	100	5.2	95.1	97.8	95.1	92.5	85.7
30	200	5.2	97.9	99.2	97.0	96.9	91.2
R3: $\rho = 1 - 2T^{-9/10}$							
10	50	6.1	60.6	87.2	76.7	54.5	66.0
10	100	5.4	89.8	97.0	89.8	81.9	81.9
10	200	5.5	97.7	99.0	95.3	93.6	89.5
20	50	5.9	82.7	95.5	91.6	76.4	70.9
20	100	5.6	96.6	98.7	96.6	93.9	85.0
20	200	5.3	99.1	99.7	98.7	98.1	92.4
30	50	5.9	90.7	96.9	95.3	86.6	81.9
30	100	5.3	97.9	98.9	97.9	96.7	91.5
30	200	5.3	99.4	99.8	99.0	99.0	96.0
R4: $\rho = 0.8$							
10	50	6.2	72.9	96.9	89.2	67.0	81.0
10	100	5.9	97.0	99.9	97.0	92.4	95.5
10	200	5.3	99.7	100.0	98.8	98.4	98.7
20	50	5.6	93.6	99.7	98.7	89.9	84.4
20	100	5.5	99.9	100.0	99.9	99.5	96.8
20	200	5.4	100.0	100.0	100.0	100.0	99.3
30	50	6.3	98.4	99.9	99.6	96.8	94.9
30	100	5.2	100.0	100.0	100.0	100.0	99.7
30	200	5.6	100.0	100.0	100.0	100.0	99.9

Notes: See Table 1 for a description.

Table 4: Simulated 5% size and power of W in case of two predictors.

N	T	Size	P1	P2	P3	P4	P5
R1: $\rho = 1$							
10	50	5.6	54.7	87.2	74.8	49.5	49.3
10	100	5.1	83.9	95.6	83.9	74.2	65.1
10	200	5.2	94.3	98.2	89.0	88.1	76.9
20	50	5.7	81.6	96.0	91.6	73.6	78.1
20	100	5.3	95.8	98.3	95.8	91.3	87.9
20	200	4.7	98.0	99.3	96.8	96.1	94.0
30	50	4.7	92.5	97.7	96.7	87.1	63.0
30	100	5.1	97.8	99.0	97.8	96.1	77.3
30	200	5.1	98.8	99.6	98.4	98.0	87.0
R2: $\rho = 1 - 2T^{-1}$							
10	50	5.5	71.6	96.6	89.7	65.2	65.1
10	100	5.1	95.3	99.2	95.3	89.5	81.1
10	200	5.4	98.5	99.6	96.7	96.3	89.7
20	50	5.2	93.6	99.1	98.2	89.0	90.8
20	100	5.2	98.9	99.7	98.9	97.9	96.4
20	200	5.5	99.5	99.9	99.2	99.2	98.6
30	50	4.6	98.1	99.4	99.0	96.2	82.4
30	100	4.7	99.4	99.9	99.4	99.3	91.3
30	200	5.1	99.9	100.0	99.6	99.7	96.1
R3: $\rho = 1 - 2T^{-9/10}$							
10	50	5.7	76.5	97.8	92.3	69.7	69.8
10	100	5.1	97.7	99.7	97.7	93.3	86.7
10	200	5.3	99.5	99.9	98.8	98.6	93.1
20	50	5.6	95.8	99.6	99.1	92.4	93.0
20	100	5.2	99.5	99.8	99.5	99.2	97.7
20	200	5.5	100.0	100.0	99.8	99.8	99.3
30	50	4.8	99.0	99.8	99.6	97.9	87.6
30	100	4.7	99.8	99.9	99.8	99.8	94.9
30	200	5.4	99.9	100.0	99.9	99.9	98.4
R4: $\rho = 0.8$							
10	50	5.6	84.8	99.8	97.6	80.1	83.1
10	100	5.4	99.5	100.0	99.5	98.4	96.9
10	200	5.3	100.0	100.0	99.8	99.9	99.3
20	50	5.8	99.1	100.0	100.0	97.8	98.3
20	100	5.1	100.0	100.0	100.0	100.0	100.0
20	200	5.3	100.0	100.0	100.0	100.0	100.0
30	50	5.4	99.9	100.0	100.0	99.8	97.4
30	100	4.6	100.0	100.0	100.0	100.0	99.9
30	200	5.6	100.0	100.0	100.0	100.0	100.0

Notes: See Table 1 for a description.

Table 5: List of countries.

Group	Predictor	N	Countries
Emerging	DP, EP, SR	8	Brazil, Chile, India, Mexico, Malaysia, Philippines, Thailand, Turkey
	TS	4	India, Malaysia, Thailand, South Africa
Developed	DP, EP, SR	20	Australia, Austria, Belgium, Canada, Switzerland, Germany, Denmark, Estonia, Finland, France, Greece, Hong Kong, Italy, Japan, Netherlands, New Zealand, Portugal, Singapore, UK, US
	TS	20	Australia, Austria, Belgium, Canada, Switzerland, Germany, Denmark, Estonia, Finland, France, Ireland, Italy, Japan, Netherlands, New Zealand, Portugal, Sweden, Singapore, UK, US
Global	DP, EP, SR	28	Australia, Austria, Belgium, Brazil, Canada, Switzerland, Chile, Germany, Denmark, Estonia, Finland, France, Greece, Hong Kong, India, Italy, Japan, Mexico, Malaysia, Netherlands, New Zealand, Philippines, Portugal, Singapore, Thailand, Turkey, UK, US
	TS	24	Australia, Austria, Belgium, Canada, Switzerland, Germany, Denmark, Estonia, Finland, France, Ireland, India, Italy, Japan, Malaysia, Netherlands, New Zealand, Portugal, Sweden, Singapore, Thailand, UK, US, South Africa

Notes: "DP", "EP", "SR" and "TS" refer to the dividend-price ratio, the earnings-price ratio, the short rate and the term spread, respectively.

Table 6: CD test results.

Group	Variable	Correl	CD	p-value
Emerging	ER	0.289	26.033	0.000
	DP	0.273	24.565	0.000
	EP	0.158	14.217	0.000
	SR	0.471	42.406	0.000
	TS	0.127	5.292	0.000
Developed	ER	0.580	136.151	0.000
	DP	0.440	103.372	0.000
	EP	0.311	73.057	0.000
	SR	0.727	170.769	0.000
	TS	0.572	134.343	0.000
Global	ER	0.457	151.157	0.000
	DP	0.336	111.265	0.000
	EP	0.235	77.701	0.000
	SR	0.618	204.536	0.000
	TS	0.428	121.105	0.000

Notes: "Correl" refers to the average of the pair-wise cross-correlation coefficients, and "CD" refers to the Pesaran (2004) test of the null hypothesis of no cross-section correlation.

Table 7: CIPS unit root test results for the case with a constant and trend.

Group	Variable	AR	CIPS	5% crit	1% crit
Emerging	ER	−0.051	−17.910	−2.83	−3.03
	DP	0.895	−3.951	−2.83	−3.03
	EP	0.902	−3.918	−2.83	−3.03
	SR	0.945	−2.833	−2.83	−3.03
	TS	0.928	−3.133	−2.83	−3.03
Developed	ER	−0.002	−16.919	−2.70	−2.85
	DP	0.924	−3.071	−2.70	−2.85
	EP	0.930	−3.103	−2.70	−2.85
	SR	0.958	−2.786	−2.70	−2.85
	TS	0.917	−2.163	−2.70	−2.85
Global	ER	0.021	−16.556	−2.65	−2.77
	DP	0.924	−3.167	−2.65	−2.77
	EP	0.923	−3.304	−2.65	−2.77
	SR	0.960	−2.519	−2.65	−2.77
	TS	0.919	−2.342	−2.65	−2.77

Notes: "AR" refers to the average of the estimated autoregressive roots in the fitted unit root test regression, "CIPS" refers to the panel unit root test of Pesaran (2007), and "crit" refer to the critical values, which are taken from Table II (c) in Pesaran (2007).

Table 8: Endogeneity test results.

Group	Predictor	Slope	SE	t-ratio	p-value
Emerging	DP	−0.249	0.024	−10.414	0.000
	EP	0.121	0.046	2.603	0.009
	SR	−0.012	0.004	−3.220	0.001
	TS	0.031	0.026	1.210	0.226
Developed	DP	−0.161	0.018	−8.739	0.000
	EP	0.078	0.019	4.238	0.000
	SR	−0.009	0.001	−11.682	0.000
	TS	0.006	0.002	2.679	0.007
Global	DP	−0.257	0.033	−7.843	0.000
	EP	0.117	0.019	6.018	0.000
	SR	−0.016	0.002	−7.022	0.000
	TS	0.009	0.003	3.270	0.001

Notes: "Slope" refers to the pooled OLS slope estimator in a regression of $\hat{\varepsilon}_i$ onto $\hat{\eta}_i$, where $\hat{\varepsilon}_i = M_{\hat{f}}\hat{u}_i$ and $\hat{\eta}_i$ is the OLS residual in a regression of x_i onto $x_{i,-1}$. "SE" refers to the estimated standard errors, which are robust to heteroskedasticity and serial correlation.

Table 9: Predictability test results.

Group	Predictor	$\hat{\beta}$	SE	t-ratio	p-value
The proposed estimator					
Emerging	DP	0.016	0.007	2.275	0.023
	EP	0.007	0.007	0.948	0.343
	SR	−0.022	0.006	−3.900	0.000
	TS	0.009	0.010	0.864	0.387
Developed	DP	−0.003	0.003	−0.949	0.343
	EP	0.000	0.002	0.160	0.873
	SR	−0.012	0.003	−3.635	0.000
	TS	0.013	0.008	1.580	0.114
Global	DP	−0.004	0.005	−0.723	0.470
	EP	0.003	0.003	1.092	0.275
	SR	−0.026	0.007	−3.680	0.000
	TS	0.016	0.009	1.693	0.090
The Hjalmarsson (2006) estimator					
Emerging	DP	0.014	0.009	1.498	0.134
	EP	−0.014	0.008	−1.720	0.085
	SR	−0.013	0.004	−3.743	0.000
	TS	0.018	0.015	1.225	0.221
Developed	DP	0.007	0.004	1.845	0.065
	EP	−0.002	0.003	−0.575	0.565
	SR	−0.015	0.002	−6.915	0.000
	TS	0.004	0.003	1.339	0.181
Global	DP	0.003	0.008	0.422	0.673
	EP	−0.005	0.004	−1.129	0.259
	SR	−0.013	0.002	−6.700	0.000
	TS	0.004	0.003	1.287	0.198
The fixed effects estimator					
Emerging	DP	0.010	0.009	1.101	0.271
	EP	−0.007	0.013	−0.556	0.578
	SR	−0.007	0.001	−9.850	0.000
	TS	0.001	0.009	0.133	0.894
Developed	DP	0.004	0.002	2.014	0.044
	EP	−0.001	0.001	−0.987	0.324
	SR	−0.009	0.000	−21.565	0.000
	TS	0.001	0.001	1.294	0.196
Global	DP	0.003	0.005	0.561	0.575
	EP	−0.003	0.003	−0.776	0.438
	SR	−0.008	0.001	−16.242	0.000
	TS	0.001	0.001	0.815	0.415

Notes: "$\hat{\beta}$" refer to the estimated predictive slope, and "SE" refers to the estimated standard error.