

Robust block bootstrap panel predictability tests

Stephan Smeekes & Joakim Westerlund

To cite this article: Stephan Smeekes & Joakim Westerlund (2019) Robust block bootstrap panel predictability tests, *Econometric Reviews*, 38:9, 1089-1107, DOI: [10.1080/07474938.2018.1536102](https://doi.org/10.1080/07474938.2018.1536102)

To link to this article: <https://doi.org/10.1080/07474938.2018.1536102>



© 2018 Stephan Smeekes and Joakim Westerlund. Published with license by Taylor & Francis Group, LLC



Published online: 22 Nov 2018.



Submit your article to this journal [↗](#)



Article views: 895



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

Robust block bootstrap panel predictability tests*

Stephan Smeekes^a and Joakim Westerlund^{b,c}

^aMaastricht University, Maastricht, The Netherlands; ^bDepartment of Economics, Lund University, Lund, Sweden; ^cCentre for Financial Econometrics, Deakin University, Melbourne, Australia

ABSTRACT

This article develops two block bootstrap-based panel predictability test procedures that are valid under very general conditions. Some of the allowable features include cross-sectional dependence, heterogeneous predictive slopes, persistent predictors, and complex error dynamics, including cross-unit endogeneity. While the first test procedure tests if there is any predictability at all, the second procedure determines the units for which predictability holds in case of a rejection by the first. A weak unit root framework is adopted to allow persistent predictors, and a novel theory is developed to establish asymptotic validity of the proposed bootstrap. Simulations are used to evaluate the performance of our tests in small samples, and their implementation is illustrated through an empirical application to stock returns.

KEYWORDS

Block bootstrap; panel data; predictive regression; sequential testing; stock return predictability; weak unit roots

JEL CLASSIFICATION

C15; C22; C23; G1; G12


1. Introduction

A fundamental empirical issue in finance is whether future stock returns (or equity premiums) are predictable using publicly available information, and there is a huge literature on this (see Spiegel, 2008). The workhorse approach is to regress current returns onto a constant and one lag of some predictor, such as dividend yield, nominal interest rates, default or term spreads on bonds, or valuation ratios (see Rapach and Zhou, 2013), and then testing whether the predictive slope is zero by using a conventional *t*-test. Early studies rely on normal critical values, against which the zero slope restriction could typically be rejected. However, it has since then become clear that normal inference can be quite misleading in the standard situation when the predictor is both endogenous and persistent, and that some of the rejections might therefore be due to size distortions. This observation has attracted much attention among econometricians, so much so that there is by now a separate literature aimed at developing robust econometric tests for predictability (see Campbell and Yogo, 2006; Cavanagh et al., 1995; Elliott and Stock, 1994; Jansson and Moreira, 2006; Kostakis et al., 2015; Lewellen, 2004, to mention a few).

Parallel to this development, it has been recognized that many studies are not really using time series data, but rather panel data comprising time series observations on multiple cross-sectional units, such as firms, industries or countries. The standard approach to such data is to take any existing time series test, and to simply apply it to each unit in the sample (see, e.g., Ang and Bekaert, 2007; Driesprong et al., 2008; Polk et al., 2006; Rapach et al., 2013). This raises the issue of mass-significance, and the need to control the overall significance level of the approach. As a response, panel

CONTACT Joakim Westerlund  joakim.westerlund@nek.lu.se  Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden.

*Previous versions of the article were presented at a seminar at Tilburg University and at the NESG 2013.

 Supplemental data for this article can be accessed on the [the publisher's website](#).

© 2018 Stephan Smeekes and Joakim Westerlund. Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

econometric test procedures for predictability have been developed (see Hjalmarsson, 2010; Kauppi, 2001; Westerlund and Narayan, 2015; Westerlund et al., 2017). These procedures not only account for the multiplicity of the testing problem, but also increase the precision of the predictability test by taking the total number of observations and their variation into account.

But while certainly promising, existing panel test procedures suffer from a number of important drawbacks. The first drawback is the formulation of the hypothesis tested. In particular, while the null hypothesis is formulated as that there is no predictability, the alternative hypothesis is formulated as that there are at least some units for which predictability holds, which is too broad for any interesting economic conclusions. It could be that there is predictability for all units, but it could also be that there is only a small fraction of units for which predictability holds. Another drawback is the way in which cross-section dependence is accounted for. Specifically, a common factor structure is assumed, the effect of which is removed prior to implementation of the test for predictability. Hence, with this approach one is essentially testing for predictability in the remaining idiosyncratic component, thereby ignoring a potentially important source of predictive information, namely, the common one. Then there is also the fact that the assumed common factor structure need not be correct, leading to misleading conclusions. Finally, there is also the requirement that the number of cross-sectional units, N , should go to infinity with the number of time periods, T , which is certainly mistaken in practice. In particular, while T is relatively large, N is typically much smaller (see, e.g., Ang and Bekaert, 2007; Driesprong et al., 2008; Polk et al., 2006; Rapach et al., 2013).

In this article, we develop a set of procedures to ascertain the predictability of a panel. The point of departure is a very general data generating process (DGP) that allows, e.g., persistent predictors, general error serial and cross-sectional correlation, and endogeneity. In fact, except for some mild regulatory conditions, there are virtually no restrictions on the forms of serial and cross-sectional dependence that can be permitted. Given this generality, corrections aimed at achieving asymptotically pivotal statistics are not really an option. In this article, we therefore consider the block bootstrap as a means to obtain tests that are asymptotically valid. In doing so, we extend the work of Palm et al. (2011) for univariate unit root panels to a bivariate model.

Two test procedures based on the new block bootstrap are considered. The first test procedure is appropriate when wanting to test the hypothesis of a fully unpredictable panel versus at least some predictability, which is the same as the one considered previously in the literature. The second test procedure, which can be seen as an extension of the unit root test approach of Smeekes (2015), is designed to sequentially determine the units for which predictability holds. As already mentioned, while a non-rejection of the null of a fully unpredictable panel can be straightforwardly interpreted as that all the cross-section units are unpredictable, a rejection of the same null only supports the conclusion that there is at least one predictable unit. This begs the question: Which are the units that caused the rejection, i.e., which units are predictable? The two test procedures are therefore highly complementary and in fact form a complete panel predictability toolbox. GAUSS and R codes that implement this toolbox are available online at <http://researchers-sbe.unimaas.nl/stephansmeekes/code/>. Also, unlike existing tests, the procedures developed here are valid for any N , provided that T is large enough.

The rest of the article is organized as follows. Sections 2 and 3 introduce the model and the block bootstrap-based test procedures, whose asymptotic properties are analyzed in Section 4. Sections 5 and 6 are concerned with the small-sample implications of the asymptotic results, which are investigated using both simulated and real data. Section 7 concludes. All proofs are provided in the supplemental material.

2. The model

Consider the $N \times 1$ variables $y_t = (y_{1,t}, \dots, y_{N,t})$ and $x_t = (x_{1,t}, \dots, x_{N,t})'$. The DGP of these variables is given by

$$y_t = \alpha + \beta x_{t-1} + v_t, \tag{1}$$

$$x_t = \delta(I_N - \rho) + \rho x_{t-1} + w_t, \tag{2}$$

where $\alpha = (\alpha_1, \dots, \alpha_N)'$, $\beta = \text{diag}(\beta_1, \dots, \beta_N)$, $\delta = (\delta_1, \dots, \delta_N)'$ and $\rho = \text{diag}(\rho_1, \dots, \rho_N)$. As long as $x_{i,0} = O_p(1)$, the initialization does not affect the results. Therefore, in what remains, we assume that $x_0 = 0$. The resulting DGP is a panel extension of the prototypical time series predictive regression model, in which x_t is a variable believed to be able to predict y_t . The value of β_i determines the extent to which $x_{i,t}$ is able to predict $y_{i,t}$. If $\beta_i = 0$, there is no predictability, whereas if $\beta_i \neq 0$, then $y_{i,t}$ is predictable using $x_{i,t}$. As in the previous panel and time series literatures, we focus on the case when there is a single predictor.¹ If, as in the empirical application of Section 6, one has data on multiple predictors, the tests developed here can be applied in a one-predictor-at-the-time fashion, as is commonly done in the empirical literature.

The parameter ρ_i determines the persistence of $x_{i,t}$. Since in practice most predictors have been found to be highly persistent, yet not unit root nonstationary (see Lewellen, 2004), ρ_i is typically assumed to be “close” to but not exactly equal to one. Assumption 1 reflects this.

Assumption 1.

$$\rho = I_N + \frac{cm}{T},$$

where $c = \text{diag}(c_1, \dots, c_N) < 0$, and $m > 0$ is a scalar such that $m \rightarrow \infty$ and $m/T \rightarrow 0$ as $T \rightarrow \infty$.

Assumption 1 ensures that x_t is “weakly integrated” (see, e.g., Kostakis et al., 2015; Park, 2003, 2006; Phillips et al., 2010), although not local-to-unity, as when m is fixed. Hence, while $\rho \rightarrow I_N$, the rate at which this convergence takes place is slower than in the local-to-unity case, which makes a big difference. Indeed, while invalid in the local-to-unity case (see, e.g., Park, 2006, p. 640), as we show in Section 4, the block bootstrap considered here is valid in under Assumption 1. It will therefore be used in this article. The obvious drawback of the weak integration assumption is that the asymptotic approximation can be poor if x_t has a near unit root. In the supplemental material, we use Monte Carlo simulations as a mean to evaluate the effect of a violation of the weak integration assumption in small samples.

Remark 1. Except for the relatively slow rate of shrinking, Assumption 1 is very general when it comes to the types of persistency that can be permitted. Note in particular how the elements of c may differ, which means that the extent of the persistency may vary across the cross-section. In fact, we could even allow m to vary across i , suggesting that the rate at which $\rho_i \rightarrow 1$ need not be the same, provided of course that Assumption 1 is still met.

Most predictors are not only persistent but also endogenous. For example, if y_t is stock returns and x_t is the dividend–price ratio, then an increase in the stock price will lower dividends and raise returns. We therefore assume that

$$u_t = \Psi(L)\varepsilon_t, \tag{3}$$

where $u_t = (v_t', w_t')'$ and $\Psi(z) = \sum_{j=0}^{\infty} \Psi_j z^j$ with $\Psi_0 = I_{2N}$.

Assumption 2.

- (a) $\sum_{j=0}^{\infty} j \|\Psi_j\| < \infty$ and all the rows of $\Psi(1)$ are nonzero;
- (b) ε_t is independently and identically distributed (iid) with $\mathbb{E}\varepsilon_t = 0$, $\mathbb{E}\varepsilon_t \varepsilon_t' = \Sigma_{\varepsilon\varepsilon}$ and $\mathbb{E}\|\varepsilon_t\|^\kappa < \infty$ for some $\kappa \geq 4$.

¹The only study known to us that considers multiple predictors is that of Phillips et al. (2010).

Under Assumption 2, the long-run covariance matrix of u_t is given by

$$\Omega = \lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \sum_{s=1}^T \mathbb{E} u_t u_s' = \begin{bmatrix} \Omega_{vv} & \Omega_{vw} \\ \Omega_{wv} & \Omega_{ww} \end{bmatrix} = \Psi(1) \Sigma_{\varepsilon\varepsilon} \Psi(1)' = \Sigma + \Lambda + \Lambda',$$

where

$$\Sigma = \mathbb{E} u_t u_t' = \sum_{j=0}^{\infty} \Psi_j \Sigma_{\varepsilon\varepsilon} \Psi_j',$$

$$\Lambda = \mathbb{E} \sum_{k=1}^{\infty} u_t u_{t+k}' = \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \Psi_j \Sigma_{\varepsilon\varepsilon} \Psi_{j+k}'$$

are the contemporaneous and one-sided long-run covariance matrices of u_t , respectively, which are partitioned conformably with Ω . In what follows, Ω_{vv} and $\Lambda_{ww} = \mathbb{E} \sum_{k=1}^{\infty} w_t w_{t+k}'$ are going to be particularly important, and we are therefore going to use $\omega_{v,i}^2 = [\Omega_{vv}]_{ii}$ and $\lambda_{wv,i} = [\Lambda_{ww}]_{ii}$, respectively, to denote their diagonal elements.

Remark 2. The types of cross-sectional dependencies that can be accommodated within the current DGP are more general than those that have been considered earlier in the literature (see Hjalmarrsson, 2010; Kauppi, 2001; Westerlund and Narayan, 2015; Westerlund et al., 2017), and include Granger causality, common factors and even “weak cointegration” between units (as in Phillips and Magdalinos, 2009). Moreover, since the blocks of Ω need not be diagonal, the units of v_t and w_t are not only allowed to be both serially and cross-sectionally correlated in a general fashion, but they can also be correlated with each other. The types of endogeneity that can be permitted here is therefore very general indeed.

3. The test procedures

Denote by p the number of units for which $y_{i,t}$ can be predicted using $x_{i,t-1}$, i.e., p is the number of units for which $\beta_i \neq 0$. The purpose of this article is to make inference regarding p and to determine which units are predictable. As will be explained later, our procedures will allow us to do both simultaneously. Let us denote by $0 = p_1 < \dots < p_K < N$ a set of K user-defined numbers, representing the number of predictable units to be considered in the testing; how to select these numbers will be explained later. Let $H_0(p_k)$ denote the null hypothesis that $p = p_k$, where $k = 1, \dots, K$, and let $H_1(p_{k+1})$ denote the alternative hypothesis that $p \geq p_{k+1}$. The test statistic for testing $H_0(p_k)$ versus $H_1(p_{k+1})$ is henceforth going to be written in a general notation as $\tau(p_k, p_{k+1})$. As mentioned in Section 1, the idea is to begin by testing $H_0(0)$ versus $H_1(1)$. If $H_0(0)$ is not rejected, then all the cross-section units are unpredictable and so the testing is stopped. If, however, $H_0(0)$ is rejected, then there is at least one unit for which predictability holds, and therefore the testing continues by sequentially considering $H_0(p_k)$ versus $H_1(p_{k+1})$ for $k = 1, \dots, K$. The testing stops when the null hypothesis cannot be rejected anymore.

The construction of $\tau(p_k, p_{k+1})$ depends on where in the testing sequence we are. If $H_0(0)$ is true, all the units of the panel are unpredictable, which enables full panel pooling, whereas if $H_0(0)$ is false, then this type of pooling is no longer possible. The basic building block in constructing $\tau(p_k, p_{k+1})$ is given by the following bias corrected t -statistic for testing $\beta_i = 0$:

$$\theta_i = \frac{\sum_{t=2}^T x_{i,t-1}^d y_{i,t}^d - T \hat{\lambda}_{wv,i}}{\hat{\omega}_{v,i} \sqrt{\sum_{t=2}^T (x_{i,t-1}^d)^2}}, \tag{4}$$

with $x_{i,t-1}^d = x_{i,t-1} - T^{-1} \sum_{s=2}^T x_{i,s-1}$ and an analogous definition of $y_{i,t}^d$. In order to describe $\hat{\lambda}_{wv,i}$ and $\hat{\omega}_{v,i}$, we need to introduce some notation. We begin by defining $\hat{u}_t = (\hat{v}_t', \hat{w}_t')'$, where \hat{v}_t and

\hat{w}_t are the residuals obtained by applying OLS to (1) and (2), respectively. The estimated long-run covariance matrix of \hat{u}_t is given by

$$\hat{\Omega} = \begin{bmatrix} \hat{\Omega}_{vv} & \hat{\Omega}_{vw} \\ \hat{\Omega}_{wv} & \hat{\Omega}_{ww} \end{bmatrix} = \hat{\Sigma} + \hat{\Lambda} + \hat{\Lambda}',$$

where

$$\hat{\Sigma} = T^{-1} \sum_{t=2}^T \hat{u}_t \hat{u}_t' \tag{5}$$

$$\hat{\Lambda} = \sum_{j=1}^{J-1} K(j/J) T^{-1} \sum_{t=j+1}^T \hat{u}_{t-j} \hat{u}_t' \tag{6}$$

with $K(x) = (1-|x|)1(|x| \leq 1)$ being the Bartlett kernel and $J > 0$ is the associated kernel bandwidth parameter. In this notation, $\hat{\omega}_{v,i}^2 = [\hat{\Omega}_{vv}]_{ii}$ and $\hat{\lambda}_{wv,i} = [\hat{\Lambda}_{wv}]_{ii}$. These are the only elements of $\hat{\Omega}_{vv}$ and $\hat{\Lambda}_{wv}$ that will be used in the testing.

Remark 3. In many cases, it is not necessary to bootstrap the t -statistic but one can also bootstrap the OLS estimator itself.² However, in the present case the variance correction in the numerator is necessary to account for the endogeneity, and so we can just as well bootstrap the t -statistic.

Remark 4. A nonzero λ_{wv} implies that past values of w_t are correlated with (the current value of) v_t . One may therefore interpret the endogeneity bias that arises from a non-zero λ_{wv} as evidence for predictability in itself. Such predictability is, however, weaker than the predictability arising from a non-zero β_b , as $x_{i,t}$ is more persistent than $v_{i,t}$.³

3.1. Pooled tests for testing $p = 0$ versus $p \geq 1$

In this section, we consider the relatively simple problem of testing $H_0(0)$ (no predictability) versus $H_1(1)$ (there is at least one predictable unit). The purpose is to determine if there is any predictability at all. It is therefore important that the test is powerful enough, and for this reason we only consider pooled test statistics. The two most common ways to construct such statistics are to use either “panel” pooling, or “group mean” pooling. The particular panel and group mean test statistics considered in this article are given by

$$\tau_P(0, 1) = \frac{\sum_{i=1}^N \left(\sum_{t=2}^T x_{i,t-1}^d y_{i,t}^d - T \hat{\lambda}_{wv,i} \right)}{\sqrt{\sum_{i=1}^N \hat{\omega}_{v,i}^2 \sum_{t=2}^T \left(x_{i,t-1}^d \right)^2}}, \tag{7}$$

$$\tau_{GM}(0, 1) = \sum_{i=1}^N \theta_i, \tag{8}$$

respectively. It should be noted that while there is a dependence on the number of predictable units under the null, $\tau_P(0, 1)$ and $\tau_{GM}(0, 1)$ do not really depend on the number of predictable units under the alternative. The reason for still writing the test statistics as a function of the latter

²For example, when testing for a unit root in $x_{i,t}$, rather than bootstrapping the associated t -statistic, one may bootstrap $T(\hat{\rho}_i - 1)$, where $\hat{\rho}_i$ is the OLS estimator of ρ_i in the i -th equation of (2).

³If one is interested in this weaker form of predictability, one may use the same test statistic as before but without the variance correction in the numerator. It should be pointed out that this change will lead to reduced rate of divergence under the alternative of predictability (see Lemma 1 in the [supplemental material](#)), and hence relatively low power of the test.

is to emphasize that in case of a rejection the appropriate conclusion is that there is at least one unit for which predictability holds. For easy reference, however, whenever possible, we write τ_p (τ_{GM}) for $\tau_p(0, 1)$ ($\tau_{GM}(0, 1)$).

3.2. Sequential test procedure

The tests considered in the previous section are appropriate if one wishes to infer whether there is actually any predictability at all. In many cases, however, one would like to go further than just concluding that there is some predictability in case of a rejection. In this section, we therefore consider a sequential test that determines the units for which predictability holds. In so doing, we will assume that the testing numbers, p_1, \dots, p_K , are known and that p belongs to this set; later on we discuss how to proceed in general when p is permitted to lie between test numbers.

The sequential test requires that the units can be ranked according to significance. Any unit-specific test statistic can be used for this purpose, provided that (i) a larger value provides more evidence for predictability, and (ii) the marginal distribution is the same across units (see Remark 2 of Smeekes, 2015, for a discussion). Here we use $|\theta_i|$. Let us therefore denote by $|\theta|_{(1)} \geq \dots \geq |\theta|_{(N)}$ the “reverse” order statistics associated with $|\theta_1|, \dots, |\theta_N|$. The test statistic to be used in the sequential testing, denoted $\tau_{SQ}(p_k, p_{k+1})$, is given by the order statistic corresponding to the alternative hypothesis to be tested;

$$\tau_{SQ}(p_k, p_{k+1}) = |\theta|_{(p_{k+1})}, \tag{9}$$

and is appropriate for testing $H_0(p_k)$ versus $H_1(p_{k+1})$.⁴ The sequential test procedure considered in this article is based on repeated use of this test statistic, and is similar to the procedure used by Smeekes (2015) to determine the proportion of stationary units in a panel. The procedure does not just yield an estimate of p , but in fact estimates the set of predictable units. The algorithm for determining p is given below.

3.2.1. Search algorithm

1. Test $H_0(p_1)$ against $H_1(p_2)$ using the test statistic $\tau_{SQ}(p_1, p_2)$. Reject $H_0(p_1)$ if the p -value is lower than the chosen significance level α .
2. If $H_0(p_1)$ is not rejected, set $\hat{p} = p_1$, whereas if $H_0(p_1)$ is rejected, use $\tau_{SQ}(p_2, p_3)$ to test $H_0(p_2)$ against $H_1(p_3)$.
3. Keep testing until $H_0(p_k)$ cannot be rejected anymore, and set $\hat{p} = p_k$. If all null hypotheses up until and including $H_0(p_K)$ are rejected, set $\hat{p} = N$.

Let $\mathbb{S}_x = \{i : |\theta_i| \geq |\theta|_{(x)}\}$ be the set of x units for which the null hypothesis of no predictability has been rejected. The estimated set of predictable units is simply given by $\mathbb{S}_{\hat{p}}$.

The above search algorithm is based on the assumption that the researcher knows beforehand which numbers p_1, \dots, p_K to test, which of course need not be the case in practice. The most natural approach is to simply add the units one-by-one such that $p_k = k-1$ for $k = 1, \dots, N$. This approach has the advantage that all possible numbers are tested and the set of predictable units can be determined exactly. The drawback is that it is likely to suffer from low power, especially when N is “large.” The drawback of taking the numbers further apart is that if p lies between these values, the method will be unable to detect it. While in Section 4, we elaborate on the interpretation of the results when p is in between tested numbers, in Sections 5 and 6, we discuss the selection of p_1, \dots, p_K .

⁴The intuition for taking the order statistic corresponding to $H_1(p_{k+1})$ is as follows. Suppose for simplicity that the units are added one-by-one, which amounts to setting $p_k = k-1$ for $k = 1, \dots, N$. In this case, it is quite clear that when $H_0(p_1)$ is tested against $H_1(p_2)$, one has to take $|\theta|_{(p_2)} = |\theta|_{(1)}$ as a test statistic, as $|\theta|_{(p_1)} = |\theta|_{(0)}$ is undefined.

3.3. The bootstrap

Define the following bias corrected estimators of β_i and ρ_i :

$$\tilde{\beta}_i = \frac{\sum_{t=2}^T x_{i,t-1}^d y_{i,t}^d - T \hat{\lambda}_{ww,i}}{\sum_{t=2}^T (x_{i,t-1}^d)^2}, \tag{10}$$

$$\tilde{\rho}_i = \frac{\sum_{t=2}^T x_{i,t-1}^d x_{i,t}^d - T \hat{\lambda}_{ww,i}}{\sum_{t=2}^T (x_{i,t-1}^d)^2}, \tag{11}$$

where $\hat{\lambda}_{ww,i} = [\hat{\Lambda}_{ww}]_{ii}$. The next algorithm describes how $\tilde{\beta}_i$ and $\tilde{\rho}_i$ are used in testing $H_0(p_k)$ versus $H_1(p_{k+1})$.

3.4.1. Bootstrap algorithm

1. Let $z_t = (x_t', y_t')'$. Obtain $z_t^d = (y_t^d, x_t^d)' = z_t - T^{-1} \sum_{s=1}^T z_s$.
2. Let $\tilde{\gamma} = (\tilde{\beta}', \tilde{\rho}')$, where $\tilde{\beta} = \text{diag}(\tilde{\beta}_1, \dots, \tilde{\beta}_N)$, $\tilde{\rho} = \text{diag}(\tilde{\rho}_1, \dots, \tilde{\rho}_N)$, $z_0 = 0$, and calculate $\tilde{u}_t = (\tilde{v}_t', \tilde{w}_t')' = z_t^d - \tilde{\gamma} x_{t-1}^d$ for $t = 2, \dots, T$.
3. Choose a block length ℓ . Draw $\mathcal{I}_1, \dots, \mathcal{I}_k$ iid from the uniform distribution on $\{1, 2, \dots, T-\ell\}$, where $k = \lceil (T-1)/\ell \rceil$ is the number of blocks.
4. Construct $u_t^* = (v_t^{*'}, w_t^{*'})' = \tilde{u}_{\mathcal{I}_k + s_t} - (T-\ell)^{-1} \sum_{\tau=1}^{T-\ell} \tilde{u}_{\tau + s_t}$, where $t = 2, \dots, T$, $k_t = \lceil t/\ell \rceil$ and $s_t = t - (k_t - 1)\ell$.
5. Let $y_t^* = v_t^*$ and $x_t^* = \tilde{\rho} x_{t-1}^* + w_t^*$ for $t = 2, \dots, T$ with $x_1^* = x_1^d$ and $y_1^* = y_1^d$.
6. Let $\hat{u}_t^* = (\hat{v}_t^{*'}, \hat{w}_t^{*'})'$, where \hat{v}_t^* and \hat{w}_t^* are the residuals obtained by applying OLS to the bootstrap versions of (1) and (2), respectively. The bootstrap versions of $\hat{\omega}_{v,i}^2$ and $\hat{\lambda}_{vw,i}$, denoted $\hat{\omega}_{v,i}^{*2}$ and $\hat{\lambda}_{vw,i}^*$ respectively, are as before but with $\hat{\Omega}$ and $\hat{\Lambda}$ replaced by

$$\hat{\Omega}^* = \frac{1}{T} \left(\sum_{m=1}^{k-1} \sum_{s=1}^{\ell} \sum_{j=1}^{\ell} \hat{u}_{(m-1)\ell+s}^* \hat{u}_{(m-1)\ell+j}^{*'} + \sum_{s=1}^{T-(k-1)\ell} \sum_{j=1}^{T-(k-1)\ell} \hat{u}_{(k-1)\ell+s}^* \hat{u}_{(k-1)\ell+j}^{*'} \right),$$

$$\hat{\Lambda}^* = \frac{1}{T} \left(\sum_{m=1}^{k-1} \sum_{s=1}^{\ell} \sum_{j=1}^{s-1} \hat{u}_{(m-1)\ell+s-j}^* \hat{u}_{(m-1)\ell+s}^{*'} + \sum_{s=1}^{T-(k-1)\ell} \sum_{j=1}^{s-1} \hat{u}_{(k-1)\ell+s-j}^* \hat{u}_{(k-1)\ell+s}^{*'} \right),$$

respectively.

7. Calculate

$$\theta_i^* = \frac{\sum_{t=2}^T x_{i,t-1}^{*d} y_{i,t}^{*d} - T \hat{\lambda}_{ww,i}^*}{\hat{\omega}_{v,i}^* \sqrt{\sum_{t=2}^T (x_{i,t-1}^{*d})^2}},$$

with $x_{i,t-1}^{*d} = x_{i,t-1}^* - T^{-1} \sum_{s=2}^T x_{i,s-1}^*$ and an analogous definition of $y_{i,t}^{*d}$.

8. Obtain θ_i^* for all $i \in \mathbb{S}_{p_k}^c$, where $\mathbb{S}_{p_k}^c$ is the complement of \mathbb{S}_{p_k} , the set of units for which the no predictability null has been rejected in previous steps. The bootstrap test statistic is given by

$$\tau_{SQ}^*(p_k, p_{k+1}) = \theta_{(p_{k+1}-p_k):\mathbb{S}_{p_k}^c}^*,$$

i.e., $\tau_{SQ}^*(p_k, p_{k+1})$ is the $(p_{k+1}-p_k)$ -th largest value of $\theta_i^* \in \mathbb{S}_{p_k}^c$.⁵

⁵Note that one has to take the $(p_{k+1}-p_k)$ -th largest value of θ_i^* calculated for the “remaining” units. This might seem at odds with the calculation of the original test statistic as the p_{j+1} -th smallest value of θ_i for all units. However, one should note that p_k units have already been dropped in previous steps and are therefore not included in the bootstrap sample for this step. Therefore, one should not take the p_{k+1} -th smallest value of θ_i^* for the bootstrap test statistic, but the $(p_{k+1}-p_k)$ -th smallest to account for the p_k units that have already been dropped.

9. Repeat steps 3–8 B times and calculate the bootstrap p -value as $B^{-1} \sum_{b=1}^B I[\tau_{SQ}^{*,b}(p_k, p_{k+1}) > \tau_{SQ}(p_k, p_{k+1})]$, where $I(x)$ is the indicator function and $\tau_{SQ}^{*,b}(p_k, p_{k+1})$ is the b -th bootstrap test statistic.

To perform τ_P (τ_{GM}), simply replace $\tau_{SQ}^*(p_k, p_{k+1})$ in step 8 of the algorithm by τ_P^* (τ_{GM}^*), the statistic based on the bootstrap sample. In addition, as τ_P (τ_{GM}) is a two-sided test, the bootstrap p -value in step 9 is calculated as

$$2 \min \left\{ \frac{1}{B} \sum_{b=1}^B I \left[\tau_P^{*,b}(p_k, p_{k+1}) > \tau_P(p_k, p_{k+1}) \right], \frac{1}{B} \sum_{b=1}^B I \left[\tau_P^{*,b}(p_k, p_{k+1}) \leq \tau_P(p_k, p_{k+1}) \right] \right\}.$$

Remark 5. The bootstrap variance correction explicitly takes into account the known block-wise structure of the bootstrap process. The effects of the method of studentization of block bootstrap statistics in a stationary setting have been extensively researched (see, e.g., Härdle et al., 2003, Section 3). Götze and Künsch (1996) find that the type of correction used in step 6 works best in terms of both small-sample properties and asymptotic refinements, and our (unreported) simulations results confirm this.⁶

Remark 6. According to our Monte Carlo results, the small-sample precision of $\hat{\Omega}$ and $\hat{\Lambda}$ can be greatly improved by iteration. In this case, we define $\tilde{\Omega} = \tilde{\Sigma} + \tilde{\Lambda} + \tilde{\Lambda}'$, where $\tilde{\Sigma}$ and $\tilde{\Lambda}$ are defined as in (5), but with \hat{u}_t replaced by the bias corrected residuals \tilde{u}_t defined in step 2 of the bootstrap algorithm. The bias correction is then carried out using $\tilde{\Omega}$ and $\tilde{\Lambda}$ in place of $\hat{\Omega}$ and $\hat{\Lambda}$, respectively.

Following the bulk of the previous literature, the tests developed here are designed for in-sample testing. Alternatively, we may use out-of-sample evaluation, either directly, for instance through the out-of-sample R^2 (Campbell and Thompson, 2008), or indirectly by measuring the economic benefit for investors, for instance through the evaluation of the performance of trading strategies (Marquering and Verbeek, 2004). With only minor modifications our bootstrap method can be used for such out-of-sample evaluation.

To fix ideas, consider the generic out-of-sample predictability statistic τ_{OOS} . This could be a formal statistical test, but it can also be R^2 or some measure of economic benefit, such as total return or the Sharpe ratio. We assume that there is a total of $T+h$ observations on z_t available, where the first (last) T (h) observations form the in-sample (out-of-sample) period. The two sample periods play very distinctive roles; with the in-sample data the investor estimates the model and makes the prediction for the future, or decides on the weights in the investment portfolio. The prediction or trading strategy is then evaluated using the out-of-sample data, by for instance calculating the out-of-sample R^2 for the predictions, or a performance measure for the investment strategy.

To implement the bootstrap, note that since both test and bootstrap are used for ex post evaluation, we can use the full sample up to $T+h$ to generate the bootstrap sample. Therefore, in steps 1 and 2 of the bootstrap algorithm all $T+h$ observations available are used in the estimation, while in steps 3–5 $T+h$ bootstrap observations $\{z_t^*\}_{t=1}^{T+h}$ are generated. In the final steps of the bootstrap algorithm, the bootstrap statistic τ_{OOS}^* is calculated, again distinguishing between the in-sample and out-of-sample periods. As we always bootstrap under the null hypothesis of no predictability, τ_{OOS}^* is generated under the null hypothesis irrespective of the specific statistic considered. Asymptotic validity follows from the same arguments used in Section 4 for the in-sample tests considered here.

⁶One disadvantage of using the correction in step 6 is that it does not guarantee that $\hat{\Omega}^*$ and $\hat{\Lambda}^* Z \hat{\Lambda}^*$ are positive definite, which, if required, can be remedied by weighting with a lag window.

4. Asymptotic distributions

The results reported in this section are based on keeping N fixed and sending T (and also m) to infinity, which means that in practice what matters for accuracy is that T is sufficiently large. The fixed- N , large- T requirement is not only consistent with the typical data set in the literature, but is in fact necessary for the validity of the unit-by-unit sequential approach, as $p_k = k-1$ is indistinguishable from $p_{k+1} = k$ when $N \rightarrow \infty$.

4.1. The sample statistics

We begin by reporting the asymptotic distributions of the test statistics when applied to the sample data. The results are summarized in Theorem 1.

Theorem 1. *Suppose that Assumptions 1 and 2 hold, $m = o(T^{1/2-1/\kappa})$ and $J = o(m^{-1}\sqrt{T})$. Then the following hold as $m, J, T \rightarrow \infty$:*

- (i) Under $H_0(0)$,

$$\begin{aligned} \tau_{GM}(0, 1) &\rightarrow_d \sum_{i=1}^N X_i, \\ \tau_P(0, 1) &\rightarrow_d \frac{\sum_{i=1}^N Y_i}{\sqrt{\text{tr}(\Omega_{vv} \odot C \odot \Omega_{ww})}}, \end{aligned}$$

where $Y = (Y_1, \dots, Y_N)' =_d (C \odot \Omega_{ww} \odot \Omega_{vv})^{1/2} Z$, with $Z \sim N(0, I_N)$, \rightarrow_d and $=_d$ signify equality in distribution and convergence in distribution, respectively, \odot is the Hadamard product, $\text{tr}(A)$ is the trace of the matrix A , C is a symmetric $N \times N$ matrix with typical element $[C]_{ij} = -1/(c_i + c_j)$, $X_i = \sqrt{-2c_i} Y_i / \omega_{v,i} \omega_{w,i}$, and $\omega_{w,i}^2 = [\Omega_{ww}]_{ii}$.

- (ii) Under $H_1(1)$, $\tau_{GM}(0, 1), \tau_P(0, 1) = O_p(T/\sqrt{m})$.
- (iii) Under $H_0(p)$,

$$\begin{aligned} \tau_{SQ}(p_k, p_{k+1}) &= O_p(T/\sqrt{m}) \quad \text{if } p_{k+1} \leq p, \\ \tau_{SQ}(p_k, p_{k+1}) &\rightarrow_d X_{(p_{k+1}-p; \beta_i=0)} \quad \text{if } p_{k+1} > p, \end{aligned}$$

where $X_{(p_{k+1}-p; \beta_i=0)}$ indicates the $(p_{k+1}-p)$ -th largest value of the set of X_i variables for which $\beta_i = 0$.

The rate of divergence of $\tau_{SQ}(p_k, p_{k+1})$ under the alternative implies that our tests have power against local alternatives of the form $\beta_i = b_i \sqrt{m}/T$ for some $b_i \neq 0$. In the [supplemental material](#), we derive the relevant asymptotic distribution theory for such alternatives, which is used in [Section 5](#) to evaluate the small sample behavior of our tests.

Remark 7. The requirement that $m = o(T^{1/2-1/\kappa})$ is the same as in Park (2006). The required expansion rate of J , which is stricter than the usual $o(\sqrt{T})$ rate, can be explained in the following way. As m increases, the convergence rate of $\tilde{\beta}_i$ and $\tilde{\rho}_i$ is reduced. Therefore, \tilde{u}_t is a poor estimator of u_t , and as a result $\hat{\Omega}$ and $\hat{\Lambda}$ are poor estimators of Ω and Λ , respectively. To compensate, J must be set as a decreasing function of m .

Remark 8. The corrections applied to the numerators and denominators of the test statistics ensure that the unit-specific nuisance parameters are eliminated, but not those arising from the cross-sectional dependence. In the time series case ($N=1$), it is easy to see that $X_1 = \sqrt{-2c_1} Y_1 / \omega_{v,1} \omega_{w,1} = \sqrt{-2c_1} / \omega_{v,1} \omega_{w,1} (\omega_{v,1} \omega_{w,1} Z_1 / \sqrt{-2c_1}) = Z_1$. Therefore, θ_1 has a limiting $N(0, 1)$ distribution. If, however, $N > 1$, then X_1, \dots, X_N are correlated, which means that the panel statistics are not asymptotically pivotal. The only exception is the empirically irrelevant case when there is no cross-section dependence.

4.2. The bootstrap statistics

Theorem 2. *Suppose that Assumptions 1 and 2 hold, $m = o(T^{1/2-1/\kappa})$ and $\ell = o(m^{-1}\sqrt{T})$. Then the following hold as $m, \ell, T \rightarrow \infty$:*

(i) *Under $H_0(0)$ and $H_1(p)$ for any $0 < p \leq N$,*

$$\begin{aligned} \tau_{GM}^*(0, 1) &\rightarrow_{d^*} \sum_{i=1}^N X_i \text{ in probability,} \\ \tau_P^*(0, 1) &\rightarrow_{d^*} \frac{\sum_{i=1}^N Y_i}{\sqrt{\text{tr}(\Omega_{VV} \odot C \odot \Omega_{WW})}} \text{ in probability,} \end{aligned}$$

where \rightarrow_{d^*} signifies convergence in distribution conditional on the realization of the original sample.

(ii) *Under $H_0(p)$,*

$$\tau_{SQ}^*(p_k, p_{k+1}) \rightarrow_{d^*} X_{(p_{k+1}-p_k:\mathbb{S}_{p_k}^c)} \text{ in probability.}$$

Theorem 2 establishes the asymptotic validity of the pooled bootstrap statistics, and establishes the asymptotic properties of the sequential bootstrap statistics in a single step. The consequences for the properties of the sequential approach as a whole are given in the following corollary to Theorem 2.

Corollary 1. *Under the assumptions of Theorem 2,*

$$\begin{aligned} \lim_{T \rightarrow \infty} \mathbb{P}(\hat{p} = p_k) &= 0 && \text{if } p_{k+1} \leq p, \\ \lim_{T \rightarrow \infty} \mathbb{P}(\hat{p} = p_k) &\in [0, 1] && \text{if } p_k < p < p_{k+1}, \\ \lim_{T \rightarrow \infty} \mathbb{P}(\hat{p} = p_k) &= 1 - \alpha && \text{if } p_k = p, \\ \lim_{T \rightarrow \infty} \mathbb{P}(\hat{p} = p_k) &\in [\alpha, 1] && \text{if } p_{k-1} < p < p_k, \\ \limsup_{T \rightarrow \infty} \mathbb{P}(\hat{p} = p_k) &\leq \alpha && \text{if } p_{k-1} \geq p, \end{aligned}$$

where α is the chosen significance level and $\mathbb{P}(A)$ is the probability of the event A . In addition, \mathbb{S}_p will equal the true set of predictable units with probability tending to one.

Corollary 1 says that if p is among the numbers to be tested, the sequential method is asymptotically valid in the sense that $\lim_{T \rightarrow \infty} \mathbb{P}(\hat{p} < p_{k-1}) \leq \alpha$ and $\lim_{T \rightarrow \infty} \mathbb{P}(\hat{p} > p_{k+1}) = 0$. In addition, the correct units are identified as predictable. The corollary also sheds light on what is likely to happen if the true proportion is in between selected numbers. Specifically, assuming that $\hat{p} = p_k$, we have

$$\mathbb{P}(p \in [p_{k-1}, p_{k+1}]) = 1 - \mathbb{P}(p < p_{k-1}) - \mathbb{P}(p > p_{k+1}) \geq 1 - \alpha. \tag{13}$$

Hence, if the units are not added one-by-one, so that there is a possibility that p lies between the numbers considered in the testing, then the finding that $\hat{p} = p_k$ is best interpreted as providing evidence of $p \in [p_{k-1}, p_{k+1}]$.

5. Monte Carlo simulations

5.1. Setup

In this section, we investigate briefly the performance of the proposed tests in small samples. The DGP used for this purpose is given by a restricted version of (1) and (2) that sets $\alpha = \delta = 0$

(although we do not assume knowledge of this in the construction of the tests) and $m = T^{1-\gamma}$, such that $\rho = 1 + c/T^\gamma$. We similarly set $\beta = b/T^\gamma$, such that under the alternative the predictability is weak. Also,

$$u_t = \lambda f_t + \varepsilon_t, \quad (14)$$

where $\lambda = (\lambda'_v, \lambda'_w)'$, $\lambda_v = (\lambda_{v,1}, \dots, \lambda_{v,N})'$ with a similar definition of λ_w , and $f_t = \phi f_{t-1} + \varepsilon_t$ with $\phi = 0.5$ and $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ independent of $\varepsilon_t \sim N(0, I_{2N})$. Hence, in the DGP considered here u_t is assumed to have a common factor representation, which is by no means necessary under our assumptions. Moreover, both the serial and cross-sectional dependence are assumed to originate with f_t , which is also not necessary. However, it is convenient, not only in terms of transparency, but also because it facilitates a straightforward comparison with the test of Westerlund et al. (2017), as we explain in detail below. As for b_i , the i -th diagonal element of b , $b_i \sim U[5, 15]$ for $i = 1, \dots, p$, and $b_i = 0$ for $i = p + 1, \dots, N$, which is consistent with the parameterizations considered in the time series literature (see, e.g., Campbell and Yogo, 2006; Cavanagh et al., 1995; Elliott and Stock, 1994; Jansson and Moreira, 2006). We further set $p = \lfloor qN \rfloor$, where $q = 0, 0.2, 0.5, 0.9$.

The parameters that matters for the persistence of x_t are c , T and γ . The i -th diagonal element of c , c_i , is made a draw from $U[-5, 0]$, which is again consistent with previous studies. As for γ , we consider two specifications; $\gamma = 0.9$ and $\gamma = 1$. In interest of space, however, we focus on the case when $\gamma = 0.9$, and place the $\gamma = 1$ results in the [supplemental material](#). Two values of each of N and T are considered; $N = 10, 30$ and $T = 100, 250$, which corresponds roughly to the empirical sample sizes considered in [Section 6](#). This means that when $T = 100$, which is the value that makes ρ_i furthest away from one, $\rho_i \sim U[0.93, 1]$ with an average of 0.96. Hence, even in this case x_t highly persistent.

In our setup, the extent of serial and cross-sectional correlation in $v_{i,t}$ ($w_{i,t}$) is determined by $\lambda_{v,i}$ ($\lambda_{w,i}$), ϕ and σ_ε^2 , as is clear from $\mathbb{E}v_{i,t}v_{j,t-h} = \mathbb{E}\lambda_{v,i}\lambda_{v,j}\sigma_\varepsilon^2\phi^h/(1-\phi^2)$ for $h > 0$ or $i \neq j$ and $\mathbb{E}v_{i,t}^2 = \mathbb{E}\lambda_{v,i}^2\sigma_\varepsilon^2/(1-\phi^2) + 1$. Moreover, since $\mathbb{E}v_{i,t}w_{i,t-h} = \mathbb{E}\lambda_{v,i}\lambda_{w,i}\sigma_\varepsilon^2\phi^h/(1-\phi^2)$, the extent of endogeneity is determined by $\lambda_{v,i}$, $\lambda_{w,i}$, ϕ and σ_ε^2 . Four factor loading cases are considered:

- L1. $\lambda_{v,i} = \lambda_{w,i} = 0$ and $\sigma_\varepsilon^2 = 1$.
- L2. $\lambda_{v,i} \sim U[-1, 3]$, $\lambda_{w,i} = 0$ and $\sigma_\varepsilon^2 = 1$.
- L3. $\lambda_{v,i} = -\lambda_{w,i} \sim U[-1, 3]$ and $\sigma_\varepsilon^2 = 1$.
- L4. $\lambda_{v,i} = -\lambda_{w,i} \sim U[2, 5]$ and $\sigma_\varepsilon^2 = 2$.

L1–L4 are intended to demonstrate the flexibility of the proposed tests. L1 is most restrictive. Here $v_{i,t}$ and $w_{i,t}$ are independent, both cross-sectionally and across time, and there is no endogeneity. L2 is more general than L1 in that while $w_{i,t}$ is again uncorrelated, $v_{i,t}$ is both serially and cross-sectionally correlated. The correlation between $v_{i,t}$ and $v_{j,t-h}$ can be computed using the above formula and the known properties of the uniform distribution. It is given by $0.32 \cdot \phi^h$ for $i \neq j$ and $0.76 \cdot \phi^h$ for $i = j$. L3 is in turn more general than L2, and allows both $v_{i,t}$ and $w_{i,t}$ to be serially and cross-sectionally uncorrelated, and also correlated with each other (endogeneity). The correlation between $v_{i,t}$ and $w_{i,t}$ is -0.76 , which is in the range considered by, e.g., Campbell and Yogo (2006), Elliott and Stock (1994), and Jansson and Moreira (2006). This correlation in L4 is -0.97 , which is even higher than in L3, and would have to be considered an extreme case.

A word on the tests considered. As a benchmark for τ_p and τ_{GM} , we include the test of Westerlund et al. (2017), henceforth denoted τ_{WKN} , which is arguably the most general existing competitor. However, while relatively general when compared to existing tests, τ_{WKN} is still rather restrictive. In particular, as alluded to in [Section 1](#), while $x_{i,t}$ is permitted to be both serially and cross-sectionally correlated in a very general fashion, the cross-correlation in $v_{i,t}$ is assumed to be made up of a single common factor. The type of endogeneity that can be permitted is also highly restrictive, but is satisfied in this section. Moreover, while in general both the idiosyncratic and

Table 1. Empirical rejection frequencies of τ_P , τ_{GM} , and τ_{WKN} .

N	T	Case	$q=0$			$q=0.2$			$q=0.5$			$q=0.9$			
			τ_P	τ_{GM}	τ_{WKN}	τ_P	τ_{GM}	τ_{WKN}	τ_P	τ_{GM}	τ_{WKN}	τ_P	τ_{GM}	τ_{WKN}	
10	100	L1	0.016	0.014	0.049	0.425	0.366	0.218	0.971	0.971	0.638	1.000	1.000	0.912	
		L2	0.025	0.018	0.035	0.099	0.147	0.165	0.410	0.570	0.581	0.825	0.939	0.854	
		L3	0.059	0.040	0.041	0.402	0.510	0.422	0.791	0.905	0.735	0.965	0.995	0.875	
		L4	0.065	0.067	0.055	0.151	0.115	0.573	0.684	0.674	0.629	0.997	0.998	0.698	
	250	L1	0.017	0.016	0.046	0.557	0.533	0.250	0.994	0.996	0.728	1.000	1.000	0.933	
		L2	0.026	0.025	0.040	0.135	0.200	0.215	0.507	0.704	0.621	0.894	0.975	0.899	
		L3	0.044	0.037	0.051	0.455	0.597	0.499	0.826	0.946	0.811	0.982	0.999	0.907	
		L4	0.061	0.066	0.043	0.137	0.099	0.678	0.735	0.751	0.746	0.999	1.000	0.800	
	30	100	L1	0.011	0.009	0.037	0.853	0.816	0.437	1.000	1.000	0.941	1.000	1.000	0.993
			L2	0.025	0.022	0.036	0.245	0.354	0.410	0.825	0.950	0.921	0.992	1.000	0.993
			L3	0.054	0.042	0.043	0.476	0.729	0.682	0.895	0.989	0.869	0.994	1.000	0.927
			L4	0.051	0.056	0.039	0.132	0.118	0.591	0.735	0.711	0.670	1.000	0.999	0.783
250		L1	0.024	0.020	0.047	0.934	0.935	0.489	1.000	1.000	0.960	1.000	1.000	0.998	
		L2	0.029	0.024	0.042	0.313	0.451	0.475	0.897	0.983	0.960	1.000	1.000	0.997	
		L3	0.059	0.049	0.043	0.542	0.778	0.743	0.922	0.995	0.919	0.997	1.000	0.947	
		L4	0.059	0.063	0.042	0.134	0.111	0.700	0.813	0.812	0.789	1.000	1.000	0.839	

Notes: q refers to the fraction of predictable units and γ is such that $m = T^{1-\gamma}$. The results in the table are based on setting $\gamma = 0.9$.

Cases L1–L4 differ with respect to the extent of serial and cross-sectional correlation, and endogeneity.

common components of $v_{i,t}$ are assumed to be serially uncorrelated, with weakly integrated predictors, f_t may be serially correlated. The DGP considered here is therefore tailored to meet the requirements of τ_{WKN} , and it is important to keep this in mind when interpreting the results.

As a benchmark for τ_{SQ} , since there is no other test like it in the literature, we consider the naive testing strategy that consists of classifying the units one-by-one using the individual θ_i statistics. Of course, since the asymptotic distributions of these statistics are generally unknown (see Theorem 1), the p -values still have to be bootstrapped, and for this purpose we use the same algorithm as in Section 3.3. This means that while the serial- and cross-sectional dependence is accounted for, the multiplicity of the testing problem is not. The resulting test is henceforth denoted τ_I .

A word on the implementation of the proposed tests. The sequential procedure is based on setting $p_k - p_{k-1} = N/10$ for $k = 1, \dots, K$, which means that the spacing between the number of predictable units to be tested is increasing in N . As we illustrate in Section 5.2 below, the fact that $p_k - p_{k-1}$ is allowed to increase with N is important for the performance of τ_{SQ} . The block length and bandwidth are set equal to $\ell = J = \lfloor 1.75T^{1/3} \rfloor$, a value that was also used by Palm et al. (2011).⁷ All other implementation issues, including kernel and bootstrap variance estimator, are dealt with as explained in Section 3. We focus on the tests based on the iterative bias correction procedure described in Remark 6, although in the supplemental material we also report some results for the non-iterative tests. For τ_P , τ_{GM} , and τ_{WKN} , we report 5% size and raw power, whereas for τ_{SQ} , we report the average proportion of units incorrectly classified as predictable (PIC) and the average proportion of units correctly classified as predictable (PCC). Here PIC can be loosely interpreted as size, while PCC can be loosely interpreted as power. All results are based on 2,000 simulations and 399 bootstrap replications.

5.2. Results

The empirical rejection frequencies for τ_P , τ_{GM} , and τ_{WKN} are reported in Table 1. While for $q=0$ these values represent size, for $q>0$ they represent power. We begin by considering the

⁷For the variance estimation, we also considered automatic bandwidth selection and pre-whitening, but this did not affect the results.

former set of results. With 2,000 replications the 95% confidence interval for the size of the 5% level tests studied here is [0.04, 0.06]. With this in mind, we see that all three tests tend to perform really well in all four cases considered. Of course, size accuracy is not perfect and there are some distortions. However, most distortions go in the “right” downwards direction, leading to conservative tests. The proposed tests generally suffer most with sizes that can be down to 0.009 when the nominal level is 5%. The reason for these distortions is the bias correction, which is not perfect unless T is really large (see Remarks 4 and 6 for discussions). The τ_{WKN} test is also somewhat undersized, although the distortions are, as already pointed out, relatively small, as might be expected given that the DGP considered is a perfect match for this test. Of course, once the conditions of the DGP are relaxed there are no assurances of continued good performance. For example, as explained earlier, τ_{WKN} presumes that the cross-section dependence in $v_{i,t}$ is generated by a single common factor. The performance of this test is therefore likely to deteriorate once we move away from the single factor DGP considered here, and we have unreported results that confirm this.

The power of the tests increases with q , N and T , which is a reflection Theorem 1 (ii). The performances of τ_P and τ_{GM} are very similar, although the overall impression is that the latter test is slightly more powerful. The relatively high power for τ_{GM} is not uniform, however, and there are cases where τ_P is more powerful. The least powerful test is typically given by τ_{WKN} . The main exception is in L4 when $q = 0.2$, in which case τ_{WKN} tends to be more powerful than the proposed tests.

As pointed out in Section 4, in the supplemental material we report the asymptotic distribution theory that applies under the kind of weak predictability considered here. The fact that powers of τ_P and τ_{GM} go down as q becomes smaller is just as expected given this theory. Unfortunately, since Westerlund et al. (2017) only provide the asymptotic distribution of their test under the null hypothesis, theory cannot explain why τ_P and τ_{GM} are in general more sensitive to variations in q than τ_{WKN} . One possibility is that the relative performance is due to differences in test construction. A major difference when compared to τ_P and τ_{GM} is that τ_{WKN} does not require bias correction, which is due to the fact that the type of endogeneity that can be permitted in this other test is relatively restrictive. Moreover, owing again to the difference in generality, while the proposed tests must be bootstrapped, τ_{WKN} does not. Both the bias correction and the bootstrap are needed to ensure that the tests are asymptotically correctly sized under the null hypothesis, and such size-preserving measures are known to be quite costly in terms of power. This could explain why τ_P and τ_{GM} are sometimes less powerful than τ_{WKN} .

Let us now consider Table 2, which contain the results for τ_{SQ} and τ_I . The first thing to note is the very low PIC values for τ_{SQ} . Most values are below 0.01, and are substantially smaller than those reported for τ_I . In fact, in a majority of cases the PIC values for τ_I are more than 10 times larger than those reported for τ_{SQ} . Hence, τ_I misclassifies more unpredictable units as predictable, which is to be expected given that this test does not account for the multiplicity of the testing problem.

In terms of PCC, while τ_I tends to dominate quite markedly in L1–L3, in L4 the difference in performance is very small. Hence, again, power is gained by letting go of size control. Interesting, the gain in power when compared to τ_{SQ} is not even near enough to offset the increase in size distortion. We also see that while the performance of τ_I only increases in T , for τ_{SQ} the PCC is increasing in q , T and N . For the effect of N to kick in it is crucial to allow for larger gaps between p_{k-1} and p_k when N increases, as we do here. Unreported simulations show that the PCC of the unit-by-unit approach based on setting $p_k = k-1$ deteriorates as N increases, which is to be expected given the discussion in Section 3.2. Of course, this effect is not unique to τ_{SQ} , but is there for most (if not all) sequential multiple testing procedures of this type. Indeed, as Smeekes (2015) shows in the context of selecting the number of (non-)stationary units in a panel, his sequential procedure generally compares favorably when compared to existing procedures,

Table 2. Average proportions of correctly and incorrectly selected predictable units using τ_{SQ} and τ_I .

N	T	Case	q = 0		q = 0.2				q = 0.5				q = 0.9						
			PIC		PCC		PIC		PCC		PIC		PCC		PIC		PCC		
			τ_{SQ}	τ_I	τ_{SQ}	τ_I	τ_{SQ}	τ_I	τ_{SQ}	τ_I	τ_{SQ}	τ_I	τ_{SQ}	τ_I	τ_{SQ}	τ_I	τ_{SQ}	τ_I	
10	100	L1	0.000	0.007	–	–	0.000	0.008	0.284	0.607	0.000	0.006	0.289	0.605	0.000	0.006	0.305	0.608	
		L2	0.000	0.013	–	–	0.001	0.012	0.097	0.281	0.001	0.013	0.089	0.264	0.001	0.010	0.094	0.271	
		L3	0.000	0.012	–	–	0.001	0.010	0.433	0.617	0.001	0.010	0.443	0.606	0.001	0.008	0.467	0.615	
		L4	0.000	0.006	–	–	0.000	0.006	0.865	0.951	0.001	0.007	0.876	0.951	0.003	0.005	0.908	0.949	
	250	L1	0.000	0.015	–	–	0.001	0.018	0.511	0.750	0.001	0.019	0.526	0.753	0.002	0.019	0.553	0.751	
		L2	0.001	0.025	–	–	0.001	0.024	0.170	0.367	0.001	0.024	0.176	0.371	0.001	0.027	0.183	0.367	
		L3	0.001	0.022	–	–	0.001	0.022	0.527	0.671	0.001	0.023	0.544	0.674	0.004	0.028	0.558	0.669	
		L4	0.002	0.021	–	–	0.003	0.020	0.935	0.979	0.004	0.021	0.944	0.976	0.025	0.022	0.966	0.980	
	30	100	L1	0.000	0.007	–	–	0.001	0.007	0.341	0.606	0.002	0.006	0.463	0.606	0.004	0.006	0.561	0.611
			L2	0.000	0.012	–	–	0.001	0.012	0.051	0.273	0.002	0.012	0.123	0.269	0.002	0.014	0.160	0.270
			L3	0.000	0.010	–	–	0.002	0.010	0.384	0.625	0.004	0.010	0.498	0.623	0.007	0.011	0.548	0.615
			L4	0.000	0.006	–	–	0.000	0.006	0.750	0.946	0.001	0.007	0.840	0.950	0.004	0.005	0.909	0.949
250		L1	0.000	0.018	–	–	0.004	0.017	0.541	0.754	0.007	0.017	0.672	0.757	0.028	0.017	0.771	0.755	
		L2	0.001	0.024	–	–	0.004	0.023	0.139	0.369	0.007	0.024	0.227	0.364	0.008	0.024	0.272	0.367	
		L3	0.000	0.022	–	–	0.004	0.023	0.459	0.673	0.005	0.022	0.568	0.676	0.015	0.022	0.626	0.674	
		L4	0.001	0.020	–	–	0.002	0.022	0.878	0.979	0.004	0.020	0.916	0.978	0.022	0.019	0.954	0.978	

Notes: PIC and PCC refer to the average proportion of units correctly and incorrectly classified as predictable, respectively. See Table 1 for an explanation of the rest.

such as those of Moon and Perron (2012), Romano and Wolf (2005), and Romano et al. (2008). As mentioned in Section 3.2, the sequential procedure considered here is very similar to the one of Smeekes (2015). It is therefore expected to perform relatively well when compared to these other procedures if used in the present more general context.

The PCC of τ_{SQ} is markedly lower in L2 than in the other cases. This is in accordance with theory. In particular, as we show in the supplemental material, under the weak predictability parametrization considered here,

$$\theta_i \rightarrow_d \frac{b_i \omega_{w,i}}{\sqrt{-2c_i \omega_{v,i}}} + X_i$$

as $T \rightarrow \infty$. Since X_i does not depend on b_i power is driven by the first term on the right, which represents a drift in the asymptotic distribution. By using again the above given formulas for the variances and autocovariances of $v_{i,t}$ and $w_{i,t}$, $\omega_{v,i}^2$ and $\omega_{w,i}^2$ and hence also the drift can be computed. What we find is that while in L1, L3, and L4 $\omega_{v,i}^2 = \omega_{w,i}^2$, in L2 $\omega_{v,i}^2 = 57/9 \approx 6.33$ and $\omega_{w,i}^2 = 1$. The drift is therefore relatively small in L2, which is suggestive of low PCC. We also see that power is typically somewhat higher in L3 and L4, than in L1 and L2. In the supplemental material, we offer an explanation for this.

All-in-all, we find that our asymptotic results tend to provide a reasonable approximation to the observed behavior in small samples. In particular, we find that while τ_P and τ_{GM} can sometimes be oversized when the predictors are highly endogenous and close to unit root non-stationary, iterative correction seems to provide a very effective means by which these distortions can be removed. Another finding is that τ_{SQ} seems to be very good at controlling the PIC. The PCC can sometimes be low, but is generally acceptable, and it improves with increases in q , T , and N .

6. Empirical application

One of the advantages of using stock-level panel data when predicting returns is that since the behavior of individual stocks is relatively uninteresting little is lost by taking an overall panel perspective. However, many studies of return predictability use cross-country panels (see, e.g., Ang and Bekaert, 2007; Driesprong et al., 2008; Hjalmarsson, 2010; Polk et al., 2006; Rapach et al.,

Table 3. Regression-specific sample sizes.

Regression	Emerging		Global		Developed	
	<i>N</i>	<i>T</i>	<i>N</i>	<i>T</i>	<i>N</i>	<i>T</i>
ER–EP	9	297	30	297	21	297
ER–DP	8	296	28	296	20	296
ER–SR	9	299	31	299	22	299
ER–TS	4	292	24	292	20	292

2005, 2013), in which case the unit of observation is certainly not without interest. The bootstrap tests developed in the present article are ideally suited for panel data of this type and we will use them here as an illustration of their usefulness from an applied point of view.

The data set is an update of the one of Hjalmarsson (2010). While the original data set ends in 2004, the updated data set include 10 more years of data and stretches the period 1988M3–2013M1. 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. Since the aim here is to maximize the total number of observations (subject to the balanced panel restriction), the truncation is done separately for each regression. The resulting truncated sample size for each regression is reported in Table 3. We see that while *N* is quite small, *T* is much larger, which is appropriate given the fixed-*N*, large-*T* asymptotic approach used in Section 4. The data set contains five variables; the dependent variable, excess returns (ER), and four predictors, the dividend-price (DP) ratio, the earnings-price (EP) ratio, the short term interest rate (SR) and the term spread (TS). A complete description of these variables and of the data source is provided in Hjalmarsson (2010). Also, as in this other article, for each of the four ER–predictor combinations we consider three subsamples; the global sample containing all countries, the developed countries, and the emerging market countries.

6.1. Preliminary results

Before, we come to the predictability test results we report some preliminary evidence on the extent of serial and cross-section correlation. As a measure of the cross-section correlation, we compute the pair-wise correlation coefficients of the first-differenced variables (to safeguard against possible non-stationarity) of each predictive regression. The simple average of these correlation coefficients across all pairs of countries, together with the associated CD test discussed in Pesaran et al. (2008), are reported in Table 4. The average correlation coefficient for all variables and subsamples ranges between 0.04 and 0.55, and the CD statistic is highly significant in all cases. Hence, as expected, the countries are not uncorrelated with each other. We also see that the correlation is highest among the developed countries, which is partly as expected.

As a second preliminary we test the variables for unit roots. However, because of the cross-correlations, we cannot use the conventional “first-generation” approach of just combining individual time series unit root tests as if they were independent. For this purpose, we employ the block bootstrap τ_p and τ_{gm} test statistics of Palm et al. (2011), which are very similar in spirit to the τ_p and τ_{GM} statistics considered here. The tests are constructed with a common unit root under the null hypothesis and possibly heterogeneous autoregressive roots under the alternative, suggesting that a rejection of the null should be taken as evidence in favor of stationarity for at least one unit. The block length is set as in Section 5 to $\lfloor 1.75T^{1/3} \rfloor$, and the number of bootstrap replications is 9,999. All tests are implemented while allowing for a country-specific intercepts and time trends. The results reported in Table 5 suggest that for ER the evidence against the unit root null is very strong, as to be expected. The test values for the predictors are smaller (in absolute value), but a large majority is still significant at the 1% level, and at the 10% level there is no

Table 4. Cross-correlation results.

Regression	Variable	Emerging			Global			Developed		
		CORR	CD	<i>p</i> -value	CORR	CD	<i>p</i> -value	CORR	CD	<i>p</i> -value
ER-EP	EP	0.090	9.283	0.000	0.072	25.823	0.000	0.078	19.401	0.000
	ER	0.232	23.996	0.000	0.405	145.463	0.000	0.518	129.463	0.000
ER-DP	DP	0.093	8.437	0.000	0.246	82.327	0.000	0.334	79.181	0.000
	ER	0.218	19.820	0.000	0.403	134.953	0.000	0.518	122.809	0.000
ER-SR	SR	0.038	3.904	0.000	0.076	28.274	0.000	0.120	31.416	0.000
	ER	0.182	18.926	0.000	0.387	144.484	0.000	0.518	136.198	0.000
ER-TS	TS	0.132	5.545	0.000	0.136	38.575	0.000	0.163	38.380	0.000
	ER	0.274	11.475	0.000	0.480	136.394	0.000	0.547	128.879	0.000

Notes: "CORR" and "CD" refer to the average pair-wise correlation coefficient, and the Pesaran et al. (2008) test of the null hypothesis of no cross-section correlation, respectively. The results are for the first-differenced variables.

evidence in favor of the unit root null at all. This is reflected in the estimated largest autoregressive roots, which lie in the interval [0.851, 0.959]. This means that the predictors are less persistent than those considered in the Monte Carlo simulations, suggesting that the proposed predictability tests should perform well with no major violations of the weak integration assumption.

In order to get a rough feeling for the extent of endogeneity, which we have seen can create a problem when coupled with near-unit root predictors, we computed the correlation between \hat{v}_t and \hat{w}_t , the OLS residuals from (1) and (2), respectively. The correlations lie between -0.04 and 0.137 with an average of 0.182 , suggesting that endogeneity is not a big problem. Based on this and the Monte Carlo evidence reported in the [supplemental material](#) suggesting that iterated bias correction can be costly in terms of power, in the empirical analysis we do not iterate.

6.2. Predictability test results

We begin by considering the results obtained by applying τ_P and τ_{GM} , which are implemented in the same way as in [Section 5](#). The only differences are that we detrend rather than demean the data, and that we set the number of bootstrap replications to 9,999.⁸ The results reported in [Table 6](#) lead to the following conclusions. First, except possibly for the emerging market subsample, where τ_{GM} is significant at the 1% level, EP has no real predictive ability. This finding is in agreement with those of Hjalmarsson (2010), who reports some evidence of predictability for EP, but only for the emerging market economies. Second, for DP, SR and TS there is strong evidence to suggest that ER is in fact predictable, which is in agreement with the general consensus in the literature (Rapach and Zhou, 2013, p. 372).

The results reported so far suggests that for three out of four predictors there is evidence of predictability for at least some countries. In view of this, a natural question is: Which are the countries for which ER can be predicted? To answer this question we employ the sequential test procedure, the results of which are reported in [Table 7](#). As mentioned in [Section 5](#), with N large, sequentially adding units one-by-one ($p_k - p_{k-1} = 1$) is likely to lead to underestimation of p . To compensate for this tendency, in this section we use different sequences for different subsamples. According to [Table 3](#), in the developed and global samples N is about two and three times larger than in the emerging market subsample, respectively. We therefore set $p_k - p_{k-1} = 1, 2, 3$ in the emerging, developed and global samples, respectively. The data are again detrended and the number of bootstrap replications is again set to 9,999, but, apart from this, the implementation is exactly as in [Section 5](#). The first thing to note from the results is that the number rejections is very few, if any. The only exception is for SR where we count six rejections for the global sample,

⁸The constant-only results are reported in the [supplemental material](#).

Table 5. Unit root test results.

Regression	Subsample	Predictor					ER				
		AR	τ_p	p -value	τ_{gm}	p -value	AR	τ_p	p -value	τ_{gm}	p -value
ER-EP	Emerging	0.851	-117.651	0.000	-45.569	0.000	0.113	-262.201	0.000	-271.451	0.000
	Global	0.874	-103.842	0.000	-39.723	0.000	0.126	-258.955	0.000	-262.111	0.000
	Developed	0.878	-76.727	0.000	-37.218	0.000	0.141	-254.736	0.000	-258.107	0.000
ER-DP	Emerging	0.899	-60.526	0.000	-35.341	0.000	0.120	-259.996	0.000	-267.027	0.000
	Global	0.926	-39.113	0.000	-25.068	0.000	0.128	-257.319	0.000	-260.076	0.000
	Developed	0.936	-21.605	0.000	-20.959	0.002	0.131	-253.767	0.000	-257.296	0.000
ER-SR	Emerging	0.946	-11.491	0.092	-16.428	0.000	0.114	-264.992	0.000	-273.639	0.000
	Global	0.947	-14.464	0.036	-16.332	0.000	0.130	-260.511	0.000	-262.701	0.000
	Developed	0.948	-70.548	0.000	-16.292	0.000	0.136	-254.916	0.000	-258.226	0.000
ER-TS	Emerging	0.927	-20.105	0.002	-21.478	0.004	0.015	-271.645	0.000	-274.448	0.000
	Global	0.953	-13.993	0.044	-15.952	0.004	0.116	-255.652	0.000	-255.948	0.000
	Developed	0.959	-13.851	0.054	-14.847	0.016	0.136	-248.877	0.000	-252.248	0.000

Notes: "AR" refers to the average of the estimated largest autoregressive roots for each country. τ_p and τ_{gm} are the panel unit root tests of Palm et al. (2011).

Table 6. Pooled predictability test results.

Regression	Subsample	τ_p	p -value	τ_{GM}	p -value
ER-EP	Emerging	-0.257	0.810	-1.609	0.012
	Global	-0.303	0.836	-0.544	0.356
	Developed	-0.283	0.974	-0.088	0.907
ER-DP	Emerging	3.295	0.098	2.423	0.000
	Global	3.686	0.088	1.208	0.050
	Developed	1.720	0.670	0.722	0.564
ER-SR	Emerging	-4.959	0.001	-1.716	0.001
	Global	-5.336	0.000	-2.143	0.000
	Developed	-3.704	0.029	-2.318	0.002
ER-TS	Emerging	-0.327	0.568	0.165	0.803
	Global	6.214	0.061	1.490	0.045
	Developed	6.399	0.054	1.754	0.030

Table 7. Rejections from the sequential test procedure.

Regression	Subsample	Rejections
ER-EP	Emerging	None
	Global	None
	Developed	None
ER-DP	Emerging	None
	Global	None
	Developed	None
ER-SR	Emerging	Brazil
	Global	Belgium, Brazil, Denmark, France, New Zealand, South Africa
	Developed	Belgium, Denmark, France, New Zealand
ER-TS	Emerging	None
	Global	None
	Developed	None

Notes: The rejections are obtained by applying τ_{SQ} at the 5% level.

of which four (two) are due to developed (emerging) economies. This finding is in agreement with studies such as Ang and Bekaert (2007), Hjalmarsen (2010), Rapach et al. (2013), and Rapach et al. (2005), who all find strong evidence in favor of predictability based on SR. The poor performance of DP and EP corroborates the previous findings of Ang and Bekaert (2007), and Rapach et al. (2005). The fact that for these predictors many of the rejections at the overall panel level seem to be driven by single countries illustrates the risk of using a pooled panel approach. It is possible that some countries have been incorrectly classified as unpredictable. However, even if we were to account for this possibility, a majority of countries would still be unpredictable.

7. Concluding remarks

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 stock 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. The few panel data tests that do exist are, however, not only based on restrictive assumptions, but are also somewhat uninformative in the sense that they cannot be used to identify the units for which returns can be predicted. In the present article we take this as our starting point to develop a block bootstrap algorithm that can be used to infer panel predictive regressions under very general conditions. Three tests based on this bootstrap are proposed, denoted τ_P , τ_{GM} , and τ_{SQ} . While the first two are suitable when testing the null hypothesis of no predictability versus the general alternative, the third can be used to identify the units for which predictability holds. The asymptotic validity of the tests is proven for the case when N is fixed and $T \rightarrow \infty$, and is investigated in finite samples using Monte Carlo simulations. What we find is that while τ_P and τ_{GM} can sometimes be oversized, τ_{SQ} can run into problems with low PCC. As expected, however, these are small-sample problems that disappear with increases in T . In our real data application, we consider as an example a cross-country data set on stock returns and four predictors.

Acknowledgments

The authors would like to thank seminar and conference participants, and in particular Esfandiari Maasoumi (Editor) and three anonymous referees, for their very constructive comments. Smeekes thanks the Netherlands Organization for Scientific Research (NWO) for financial support. Westerlund thanks the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship, and the Jan Wallander and Tom Hedelius Foundation for financial support under research grant number P2014–0112:1.

References

- Ang, A., Bekaert, G. (2007). Stock return predictability: Is it there? *Review of Financial Studies* 20(3):651–707.
- Campbell, J. Y., Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* 21(4):1509–1531.
- Campbell, J. Y., Yogo, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics* 81(1): 27–60.
- Cavanagh, C. L., Elliott, G., Stock, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory* 11(5):1131–1147.
- Driesprong, G., Jacobsen, B., Maat, B. (2008). Striking oil: Another puzzle? *Journal of Financial Economics* 89(2): 307–327.
- Elliott, G., Stock, J. H. (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory* 10(3–4):672–700.
- Götze, F., Künsch, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Annals of Statistics* 24:1914–1033.
- Härdle, W., Horowitz, J. L., Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review* 71(2):435–459.
- Hjalmarsson, E. (2010). Predicting global stock returns. *Journal of Financial and Quantitative Analysis* 45(1):49–80.
- Jansson, M., Moreira, M. J. (2006). Optimal inference in regression models with nearly integrated regressors. *Econometrica* 74(3):681–714.
- Kauppi, H. (2001). Panel data limit theory and asymptotic analysis of a panel regression with near integrated regressors. In: Baltagi, B. H., Fomby, T. B., Hill, R. C., eds., *Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Volume 15 of Advances in Econometrics*. Bingley: Emerald Group Publishing Limited, pp. 239–274.
- Kostakis, A., Magdalinos, T., Stamatiogiannis, M. P. (2015). Robust econometric inference for stock return predictability. *The Review of Financial Studies* 28(5):1506–1553.

- Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics* 74(2):209–235.
- Marquering, W., Verbeek, M. (2004). The economic value of predicting stock index returns and volatility. *Journal of Financial and Quantitative Analysis* 39(2):407–429.
- Moon, H. R., Perron, B. (2012). Beyond panel unit root tests: Using multiple testing to determine the non stationarity properties of individual series in a panel. *Journal of Econometrics* 169(1):29–33.
- Palm, F. C., Smeekes, S., Urbain, J.-P. (2011). Cross-sectional dependence robust block bootstrap panel unit root tests. *Journal of Econometrics* 163(1):85–104.
- Park, J. Y. (2003). Weak unit roots. Working paper 2003–17, Department of Economics, Rice University.
- Park, J. Y. (2006). A bootstrap theory for weakly integrated processes. *Journal of Econometrics* 133(2):639–672.
- Pesaran, H., Ullah, A., Yamagata, T. (2008). A bias-adjusted LM test of error cross section independence. *The Econometrics Journal* 11(1):105–127.
- Phillips, P. C. B., Magdalinos, T. (2009). Limit theory for cointegrated systems with moderately integrated and moderately explosive regressors. *Econometric Theory* 25(2):482–526.
- Phillips, P. C. B., Magdalinos, T., Giraitis, L. (2010). Smoothing local-to-moderate unit root theory. *Journal of Econometrics* 158(2):274–279.
- Polk, C., Thompson, S., Vuolteenaho, T. (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics* 81(1):101–141.
- Rapach, D. E., Strauss, J. K., Zhou, G. (2013). International stock return predictability: What is the role of the United States? *Journal of Finance* 68(4):1633–1662.
- Rapach, D. E., Wohar, M. E., Rangvid, J. (2005). Macro variables and international stock return predictability. *International Journal of Forecasting* 21(1):137–166.
- Rapach, D. E., Zhou, G. (2013). Forecasting stock returns. In: Elliott G., Timmermann A., eds., *Handbook of Economic Forecasting*, Volume 2, Part A. Amsterdam: Elsevier, pp. 328–383.
- Romano, J. P., Shaikh, A. M., Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Econometric Theory* 24:404–447.
- Romano, J. P., Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica* 73(4):1237–1282.
- Smeekes, S. (2015). Bootstrap sequential tests to determine the order of integration of individual units in a time series panel. *Journal of Time Series Analysis* 36(3):398–415.
- Spiegel, M. (2008). Forecasting the equity premium: Where we stand today. *Review of Financial Studies* 21(4):1453–1454.
- Westerlund, J., Karabiyik, H., Narayan, P. (2017). Testing for predictability in panels with general predictors. *Journal of Applied Econometrics* 32(3):554–574.
- Westerlund, J., Narayan, P. K. (2015). A random coefficient approach to the predictability of stock returns in panels. *Journal of Financial Econometrics* 13(3):605–664.