Identifying Stationary Series in Panels: A Monte Carlo Evaluation of Sequential Panel Selection Methods

Mauro Costantini[†] and Claudio Lupi[‡⋆]

[†]*Department of Economics and Finance – Brunel University. Kingston Lane, Uxbridge, Middlesex UB8 3PH (UK). E-mail: Mauro.Costantini@brunel.ac.uk*

[‡]*Department of Economics – University of Molise. Via F. De Sanctis, I-86100 Campobasso (Italy). E-mail: lupi@unimol.it*

[⋆]*Corresponding author.*

Abstract

Sequential panel selection methods (SPSMs — procedures that sequentially use conventional panel unit root tests to identify $I(0)$ time series in panels) are increasingly used in the empirical literature. We check the reliability of SPSMs by using Monte Carlo simulations based on generating directly the individual asymptotic p values to be combined into the panel unit root tests, in this way isolating the classification abilities of the procedures from the small sample properties of the underlying univariate unit root tests. The simulations consider both independent and cross-dependent individual test statistics. Results suggest that SPSMs may offer advantages over time series tests only under special conditions.

Keywords: Panel unit root; Monte Carlo; p value distribution; ROC curve.
JEL codes: C12; C15; C23.

1 Introduction

Panel unit root (UR) tests are powerful tools to check the global null hypothesis that all the N series in a panel are $I(1)$, but are unsuitable to classify individual time series into nonstationary and stationary ones. With this aim, Chortareas and Kapetanios (2009) proposed using a sequential panel selection method (SPSM), based on the following steps:

1. Apply the panel UR test. If the global null is not rejected, do not reject the $I(1)$ hypothesis for all the series in the panel: the procedure stops.

2. If the global null is rejected then remove from the panel the series with the minimum individual Dickey-Fuller (DF) t-statistic: classify the removed series as $I(0)$.

3. Go to 1.

The result is a partition of the panel into two sets of $I(0)$ and $I(1)$ time series.

The procedure was originally conceived using DF tests jointly with Im et al.'s (2003) panel UR test, but different SPSMs can be obtained using different tests. However, the chosen panel test should be able to reject even in the presence of only one $I(0)$ series and should not be based on $N \to \infty$, given that it is applied sequentially over a decreasing number N of series. Furthermore, it should preferably be built by combining individual test statistics or p values, so to be consistent with the selection criterion used to eliminate from the panel one series at each iteration. For these reasons, beside Im et al.'s (2003) test, p value combination tests (Choi, 2001; Demetrescu et al., 2006) and Hanck's (2013) intersection test are natural candidates to be used within SPSMs. We label the resulting procedures as I-SPSM, C-SPSM, D-SPSM, and H-SPSM, respectively.

In this paper we investigate the performance of SPSMs as classification devices, as compared to standard time series UR tests and to Hommel's (1988) multiple testing procedure.[1] In particular, we intend to study the behaviour of the sequential procedures under the best theoretical conditions, so to obtain simulated upper bounds of the classification ability of these procedures. Other approaches have been recently proposed in the literature to determine the stationarity of individual time series in panels (see, e.g., Ng, 2008; Hanck, 2009; Moon and Perron, 2012; Smeekes, 2015); however, these procedures cannot be strictly labelled as SPSM (in the sense used by Chortareas and Kapetanios, 2009) and cannot be analyzed using our simulation method, specifically tailored on Chortareas and Kapetanios (2009).

We complement and extend Chortareas and Kapetanios's (2009) analysis along five directions:

1. we study the performance of the procedure using four different panel UR tests;

2. we use local-to-unit root alternatives;

3. our analysis covers the cases of independent and dependent test statistics;

[1] Hommel's procedure is a closed testing procedure related to Hanck (2013).

4. we focus on the classification performance of the procedure in a way that is not influenced by the finite-sample performance of the underlying individual UR tests;

5. we summarize the simulation results using ROC graphs, consistently with the literature on discrete classifiers.

Two by-products of this research may also be of interest to many researchers:

1. we offer a viable way to simulate dependent unit root and near-unit root test statistics and p values;

2. we report a response surface to compute the critical values of the *t-bar*$_{NT}$ test for any $N \in [2, 200]$ in this way generalizing Im et al. (2003, Table 2).

2 Monte Carlo design

Since we focus on procedures that use panel UR tests based on p value combinations or on averaged test statistics, rather than simulating the individual time series constituting the panel, we directly simulate the asymptotic ($T \to \infty$) individual DF t statistics and p values under the UR null and under selected local alternatives. The simulated t statistics and p values are then used as the fundamental input to compute the panel UR tests outcomes. In so doing, the classification performance of the different procedures depends only on the procedures' properties, not on the specification and the finite-sample properties of the underlying individual UR tests. For this reason, the simulation outcomes can be interpreted as the best possible results attainable by each procedure: in no practical circumstance the examined procedures can be expected to do better on average than in our simulations.

The series in the panel are assumed to obey

$$y_{i,t} = \varrho_i y_{i,t-1} + \epsilon_{i,t} \tag{1}$$

$$\varrho_i = \exp\left(-\frac{\gamma_i}{T}\right) \approx 1 - \frac{\gamma_i}{T}\,. \tag{2}$$

with $i \in \mathbb{N}_N$, $t \in \mathbb{N}_T$, and $\mathbb{N}_k := \{1, 2, \ldots, k\}$. Under the global null, $\gamma_i = 0$ $\forall i \in \mathbb{N}_N$; under the alternative, $\gamma_i > 0$ $\forall i \in \mathcal{N}_{H_1} \subseteq \mathbb{N}_N$.

The simulation algorithm can be divided into 6 steps:

1. since $I(0)$ and $I(1)$ series are asymptotically uncorrelated, define the asymptotic correlation matrix among the test statistics as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{N_1} & \mathbf{O}_{N_1,N_0} \\ \mathbf{O}_{N_0,N_1} & \boldsymbol{\Sigma}_{N_0} \end{pmatrix}$$

where $\mathbf{O}_{N_j,N_k}$ $(j,k \in \{0,1\}, j \neq k)$ are null matrices and $\boldsymbol{\Sigma}_{N_j}$ $(j \in \{0,1\})$ is

$$\boldsymbol{\Sigma}_{N_j} = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix};$$

2. draw N values $\boldsymbol{n}$ from the N-variate normal $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$;

3. draw N uniform variables $\boldsymbol{u}$ as $\boldsymbol{u} = \Phi(\boldsymbol{n})$ and partition $\boldsymbol{u}$ as $\boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_0')'$, where $\boldsymbol{u}_1$ and $\boldsymbol{u}_0$ have N_1 and N_0 elements;

4. draw N_1 test statistics under the alternative hypothesis as $\boldsymbol{s}_1 = F_1^{-1}(\boldsymbol{u}_1)$ and N_0 test statistics under the null as $\boldsymbol{s}_0 = F_0^{-1}(\boldsymbol{u}_0)$, where $F_1(\cdot)$ and $F_0(\cdot)$ are the CDFs of the test statistics under the alternative and under the null hypothesis, respectively;

5. generate the p values under the alternative as $F_0(\boldsymbol{s}_1)$ and set the p values under the null as $\boldsymbol{u}_0$;

6. use the simulated individual p values and test statistics to carry out all the panel UR tests.

After defining the desired dependence among the test statistics[2] (step 1), steps 2–4 simulate dependent asymptotic UR and near-UR test statistics using a Gaussian copula with Dickey-Fuller and Phillips marginals (see, e.g, Asmussen and Glynn, 2007, p. 53). Then, step 5 derives the p values associated to the test statistics under the alternative (those under the null are uniformly distributed and are derived in step 3).

In order to solve steps 4 and 5, we simulate 200,000 values of the test statistics under the null and under each of the local alternatives and use the estimated CDF's $\hat{F}_0(\cdot)$ and $\hat{F}_1(\cdot)$. Under the null the t statistics are asymptotically distributed according to the DF distribution, and their simulation rises

[2]By simulating correlated individual test statistics we allow for rather general forms of dependence: any form of dependence that induces the presence of cross-correlation among the time series implies correlation of the individual test statistics and p values.

no special difficulty (see, e.g., Hatanaka, 1996, Chapter 7). The asymptotic distribution of the DF t statistic under the local-to-unit root alternative can be written in terms of functionals of the Ornstein-Uhlenbeck process that we simulate following Chan (1988), method II.

Since the critical values of Im et al.'s (2003) $t\text{-}bar_{NT}$ test depend on N, in order to apply the test recursively we need the critical values for all possible values of $N \in \{1, \ldots, N_{\max}\}$. Therefore, we simulate $t\text{-}bar_{N\infty}$ under the null over 50,000 replications with $N \in \{1, 2, \ldots, 10, 15, 20, \ldots, 100, 120, 140, \ldots, 200\}$, compute the 100α-th percentile of each simulated distribution, $cv_{N,\alpha}$, and estimate a response surface of the form

$$cv_{N,\alpha} = \beta_0 + \beta_1 \ln(N) + \beta_2 \ln(N)^2 + \beta_3 \ln(N)^3 + \xi_{N,\alpha}\,. \tag{3}$$

The 5% critical values are computed as

$$\begin{aligned} \widehat{cv}_{N,0.05} &= \hat{\beta}_0 + \hat{\beta}_1 \ln(N) + \hat{\beta}_2 \ln(N)^2 + \hat{\beta}_3 \ln(N)^3 \\ &\approx -2.851 + 0.597 \ln(N) - 0.108 \ln(N)^2 + 0.007 \ln(N)^3 \\ &\quad (R^2 = 0.9998) \end{aligned} \tag{4}$$

for any $N \in \{1, \ldots, 200\}$: (4) generalizes and extends Table 2 in Im et al. (2003).

All experiments are simulated over 5,000 replications fixing the nominal significance level at $\alpha = 0.05$, the number of time series $N \in \{10, 20, 40, 80\}$, and the fraction of stationary alternatives $N_1/N \in \{0.20, 0.50, 0.80\}$. We consider four distinct local alternatives with $\gamma \in \{1, 5, 10, 20\}$, and the correlation among individual test statistics is $\rho \in \{0, 0.4, 0.8\}$. In order to maintain the paper within a reasonable length, we consider only individual DF tests with constant and no trend.

3 Monte Carlo results

We report the simulation results on the ROC space (for an introduction to ROC curves see, e.g, Fawcett, 2006; Swets and Pickett, 1982, chapter 1, and the supplementary material).[3] Standard quantities that are routinely computed in order to check classification performance are accuracy (*acc*, the fraction of correctly classified items), true positive rate (*tpr*, an estimate of the probability that a positive instance is correctly classified as positive), false positive rate (fpr, an estimate of the probability that a negative instance is mistakenly classified as positive), precision (*prec*, the fraction of correctly classified

[3]Theoretical issues that can help interpreting the results are offered in the supplementary material.

	DF	Hommel	C-SPSM	H-SPSM	I-SPSM
true positive rate	0.3319	0.0677	0.1884	0.0962	0.3244
false positive rate	0.0501	0.0037	0.0225	0.0053	0.1006
accuracy	0.6406	0.5348	0.6071	0.5520	0.6372
average recall	0.6409	0.5320	0.5829	0.5455	0.6119

Table 1: Average classification statistics over all the experiments.

positives among the items classified as positive), and average recall (*avrec*, the average of the *tpr* and the *tnr*, the true negative rate, that is an estimate of the probability that a negative instance is classified as negative). Accuracy is a commonly used indicator but, contrary to ROC curves (which are based on the *tpr* and fpr), is not invariant to changes in the composition of the panel in terms of $I(1)$ and $I(0)$ series, and tends to be misleading when the class distribution is unbalanced (which is often the case in the framework we are analyzing: see, e.g., Swets and Pickett, 1982; Fawcett, 2006). Therefore, we urge the reader to consider the analysis of the simulation results in terms of the more general ROC curves. In fact, by using ROC graphs we can jointly compare not only the *tpr* and fpr, but also the *prec* and *avrec* of each procedure. An optimal classifier would be represented on the ROC space by the point with coordinates $fpr = 0$ and $tpr = 1$. However, in general the ranking among classifiers depends on the researcher's loss function, which may weight differently the various performance measures. For example, a criterion could be that of fixing the maximum "acceptable" fpr, $fpr_{\max}$,[4] and to chose the classifier with the highest *tpr* among those with $fpr \leq fpr_{\max}$. Other possibilities could be those of preferring the classifier with the highest *avrec* or the highest *prec* (in the latter case one might probably prefer procedures that control the family-wise error rate, FWER, such as Hommel's). Finally, if the False Discovery Rate (FDR, Benjamini and Hochberg, 1995) enters the researcher's preferences, then one might observe that $\widehat{FDR} := 1 - \widehat{prec}$, so that one might prefer the classifier with the highest *tpr* among those whose *prec* is at least equal to the threshold that ensures that the FDR is below the desired level.

A quick preview of the simulation results is offered in Table 1, which reports the main average classification indicators over all the experiments carried out in the paper. The table highlights that the SPSMs do not show on average a superior classification ability with respect to standard time series tests. On the other hand, closed multiple testing procedures are not partic-

[4]In what follows we consider acceptable $fpr \approx 0.05$, much less so $fpr > 0.10$.

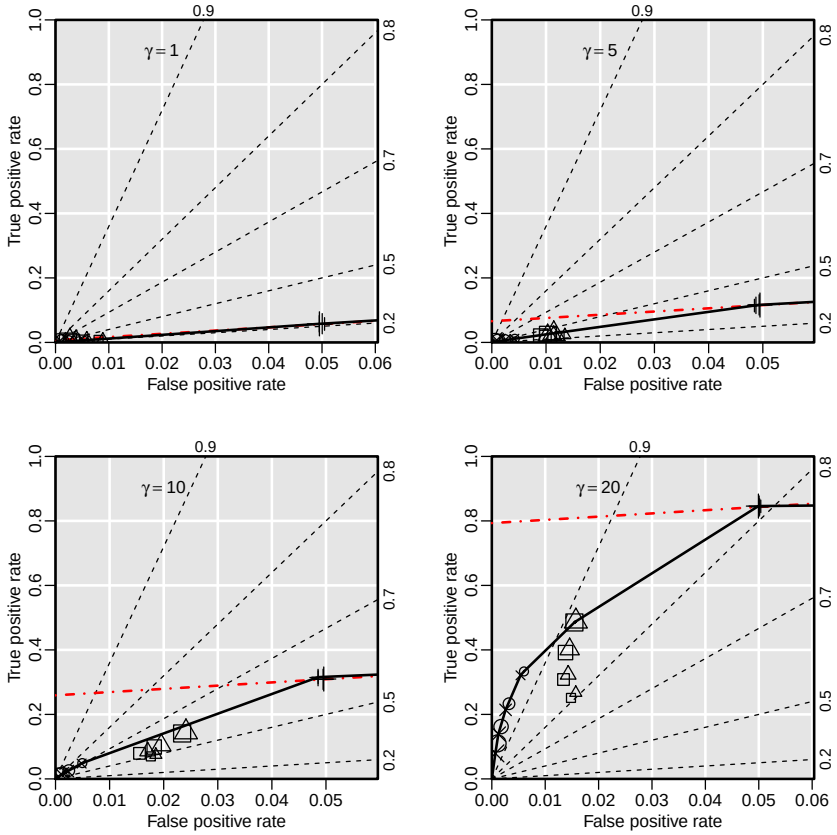

Figure 1: $\rho = 0$, $N_1/N = 0.2$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The dot-dashed line connects classifiers with the same average recall as the worst case among the standard DF. The broken solid line is the ROC convex hull. $\square$ = C-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

ularly well suited for classification purposes, unless one wishes to control the FWER. Of course, the results vary across experiments, as we can appreciate from the analysis of the ROC curves plotted in Figures 1–4.

We consider first the case $\rho = 0$. To save space, we report only the results relative to $N_1/N \in \{0.2, 0.8\}$.[5]

The I-SPSM and C-SPSM procedures have very similar performances. When the fraction of stationary time series is small (Figure 1), either procedure does not provide advantages over the standard time series approach. Irrespective of the distance of the local alternative from the UR null hypothesis, these two

[5]Results with $N_1/N = 0.5$ are available in the supplementary material.

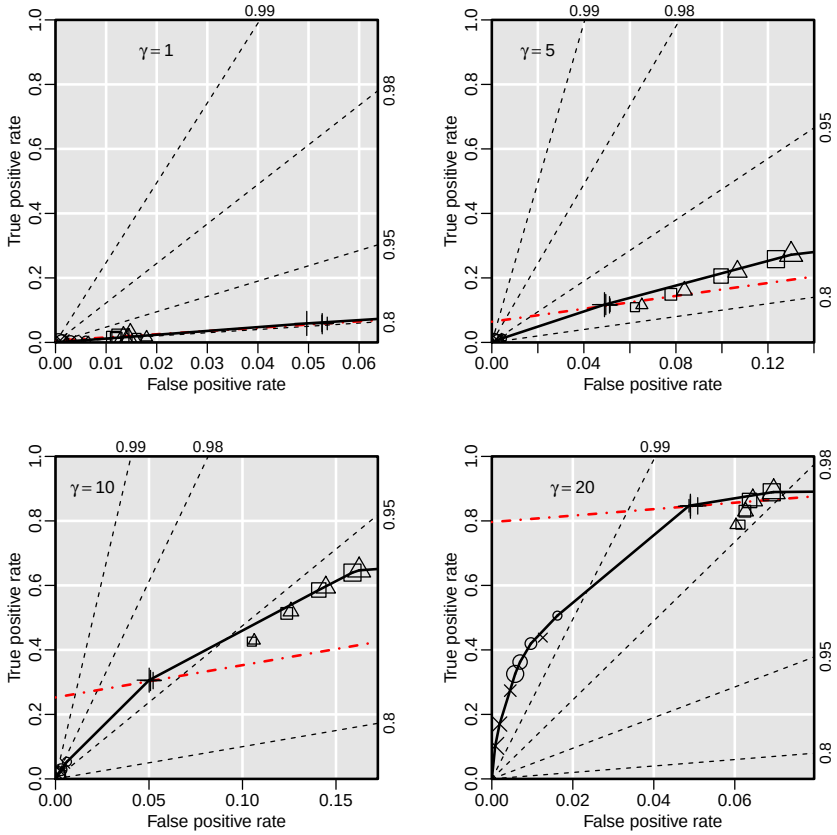

Figure 2: $\rho = 0$, $N_1/N = 0.8$. See Figure 1 for the explanation of the different symbols.

procedures tend to be sub-optimal, especially for small N. When $\gamma = 20$, the effect of panel dimension (N) on the classification performance is evident. In this case, the I-SPSM and the C-SPSM can reach higher *prec* than univariate DF tests for large panels, but with lower *tpr* and *avrec*. The H-SPSM and Hommel's method are very conservative and of scarce practical use under these conditions: in a classification perspective, control of the family-wise error rate is excessive, giving rise to an overall weak classification criterion.

When most of the series in the panel are $I(0)$ (Figure 2) the results are more articulated. When the alternative hypothesis is very close to the null ($\gamma = 1$), then no method is significantly better than random guessing. When the null and the alternative hypotheses are more separated ($\gamma = 5$, $\gamma = 10$), both the *tpr* and the fpr of the I-SPSM and of the C-SPSM increase with N. In particular, the fpr exceeds 0.15 for $\gamma = 10$ and $N > 40$. On the other hand, when $\gamma = 10$, and especially for medium to large panels, the I-SPSM and the C-SPSM have larger *avrec* than the classification based on standard

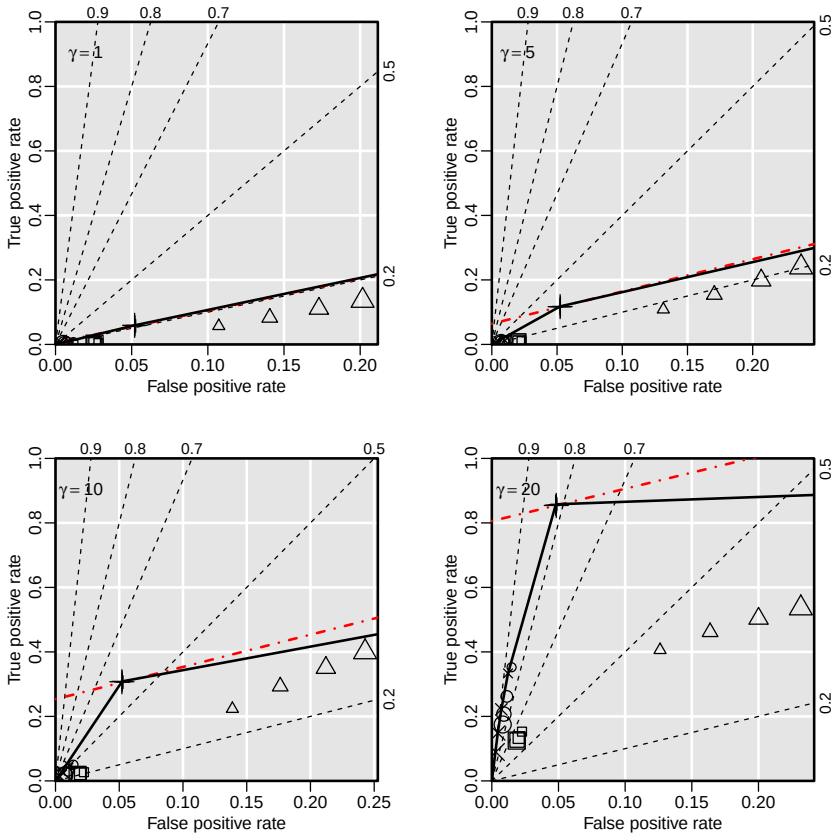

Figure 3: $\rho = 0.8$, $N_1/N = 0.2$, $\square$ = D-SPSM. See Figure 1 for the explanation of the remaining symbols.

DF tests: provided one is willing to accept larger fpr and smaller $prec$, the two SPSM procedures may offer some advantages in terms of classification over the standard time series tests. When the hypotheses are well separated ($\gamma = 20$), then the fpr reduces, but since the power of the DF test increases, there is no real gain in using either the I-SPSM or the C-SPMS, especially in the presence of relatively small panels ($N \leq 20$).[6]

The other procedures remain very conservative. In fact, the H-SPSM and the related closed testing procedure (Hommel, 1988) never reach a better result than a half of the tpr that can be attained using conventional DF tests, with only a very limited increase in $prec$.

When $\rho \neq 0$, the panel UR tests proposed by Choi (2001) and Im et al. (2003) are biased. We substitute the C-SPSM with the D-SPSM, based on

[6]The observed behaviour is not related to the effect described in Hanck (2008): see supplementary material.

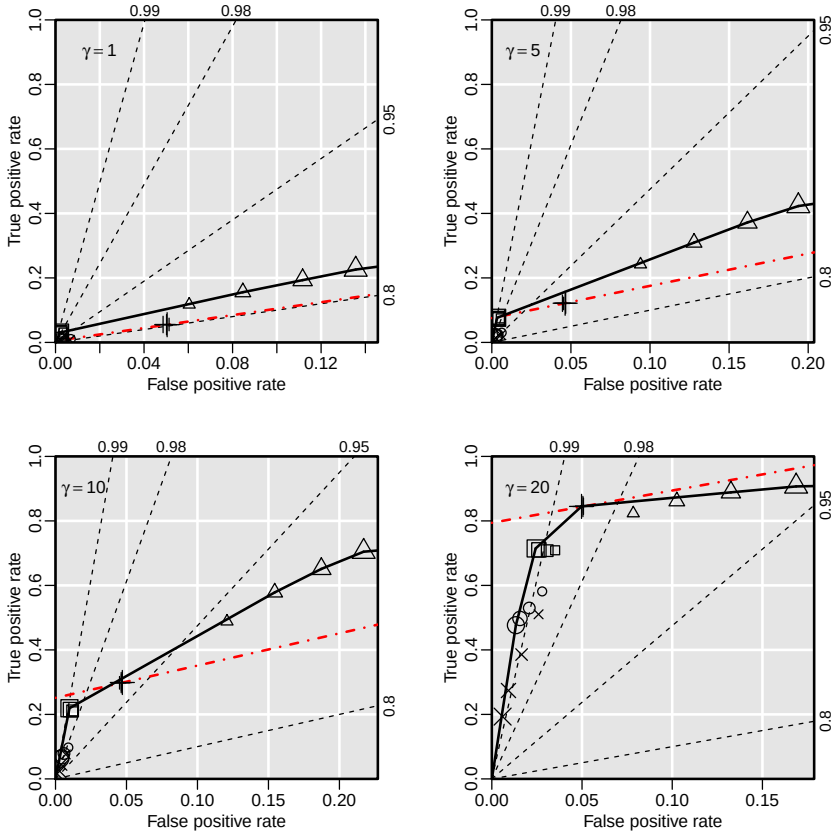

Figure 4: $\rho = 0.8$, $N_1/N = 0.8$, $\square$ = D-SPSM. See Figure 1 for the explanation of the remaining symbols.

Demetrescu et al. (2006), in order to take into account dependence in the p value combination test, and we leave the I-SPSM for reference. The H-SPSM and Hommel's procedure remain valid in the presence of positively dependent test statistics. To save space, we report only the results relative to $\rho = 0.8$ and $N_1/N \in \{0.2, 0.8\}$.[7]

When the fraction of stationary time series is small (Figure 3), apart from the expected large values of the *fpr* for the I-SPSM, the results are qualitatively similar to those obtained under independence. However, the *tpr* and *prec* of the panel-based classification procedures are lower than those obtained in the presence of independent test statistics.

When the fraction of stationary series is large (see Figure 4) the excess of rejections disappears, except of course for the I-SPSM. This is due to the

[7]Results with $\rho = 0.4$ and with $N_1/N = 0.5$ are are available in the supplementary material.

fact that dependence makes the p values under the null and under the alternative more clearly separated. If the null and the alternative are sufficiently separated, the D-SPSM can reach higher *prec* than the classification based on standard DF tests, but at the cost of lower *tpr* and *avrec*. The I-SPSM, which is based on a biased panel test, can have larger *avrec* when $\gamma = 5$ or $\gamma = 10$. However, this is obtained at the cost of huge increases in the fpr and lower *prec*.

Summarizing, our results show that SPSMs are sensitive to the composition of the panel in terms of cross-section dimension, the fraction of time series under the null and under the alternative, the distance of the null and the alternative, and the existence and strength of dependence across individual tests. SPSMs might offer advantages, in terms of *avrec*, over standard time series tests only in a few special cases. However, in general it is problematic to assess in advance if an empirical setting reflects the favourable conditions under which SPSMs can be safely applied.

Given the general nature of our Monte Carlo analysis, we expect the overall conclusion to remain unchanged if more powerful individual unit root tests and/or different panel unit root tests are used.

Acknowledgements

Preliminary versions of this paper were presented at the CFE2013 and the IAAE2014 Conferences. We are grateful to all participants for their comments and suggestions. We owe special thanks to G.Becheri, M.Demetrescu, G.Kapetanios, R.Kruse, X.Sheng, and J.Westerlund for discussion. We thank an anonymous referee and an associate editor for their constructive criticisms that helped us in improving the paper upon a previous version. The usual disclaimer applies.

References

Asmussen, S., Glynn, P. W., 2007. Stochastic Simulation: Algorithms and Analysis. Vol. 57 of Stochastic Modelling and Applied Probability. Springer, New York, NJ.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57 (1), 289–300.

Chan, N. H., 1988. The parameter inference for nearly nonstationary time series. Journal of the American Statistical Association 83 (403), 857–862.

Choi, I., 2001. Unit root tests for panel data. Journal of International Money and Finance 20 (2), 249–272.

Chortareas, G., Kapetanios, G., 2009. Getting PPP right: Identifying mean-reverting real exchange rates in panels. Journal of Banking and Finance 33 (2), 390–404.

Demetrescu, M., Hassler, U., Tarcolea, A.-I., 2006. Combining significance of correlated statistics with application to panel data. Oxford Bulletin of Economics and Statistics 68 (5), 647–663.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

Hanck, C., 2008. The error-in-rejection probability of meta-analytic panel tests. Economics Letters 101 (1), 27–30.

Hanck, C., 2009. For which countries did ppp hold? a multiple testing approach. Empirical Economics 37 (1), 93–103.

Hanck, C., 2013. An intersection test for panel unit roots. Econometric Reviews 32 (2), 183–203.

Hatanaka, M., 1996. Time-Series-Based Econometrics: Unit Roots and Cointegration. Advanced Texts in Econometrics. Oxford University Press, Oxford.

Hommel, G., 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75 (2), 383–386.

Im, K.-S., Pesaran, M. H., Shin, Y., 2003. Testing for unit roots in heterogeneous panels. Journal of Econometrics 115 (1), 53–74.

Moon, H., Perron, B., 2012. Beyond panel unit root tests: Using multiple testing to determine the nonstationarity properties of individual series in a panel. Journal of Econometrics 169 (1), 29–33.

Ng, S., 2008. A simple test for nonstationarity in mixed panels. Journal of Business and Economic Statistics 26 (1), 113–127.

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.

Swets, J. A., Pickett, R. M., 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York, NJ.