Macroeconomic Dynamics, **18**, 2014, 199–214. Printed in the United States of America.
doi:10.1017/S1365100512000338

MULTIPLE TESTING FOR OUTPUT CONVERGENCE

THOMAS DECKERS
Bonn Graduate School of Economics

CHRISTOPH HANCK
Rijksuniversiteit Groningen

This paper tests for output convergence across $n = 51$ economies, employing the definition of Pesaran [*Journal of Econometrics* 138, 312–355 (2007)]. The definition requires output gaps to be stationary around a constant mean. But when all $n(n-1)/2$ pairs of log per capita output gaps are considered, this results in more than 1,000 unit root tests to be conducted. Hence, because of the ensuing multiplicity of the testing problem, a nontrivial number of output gaps will be falsely declared to be stationary when each of the $n(n-1)/2$ hypotheses is tested at some conventional level like 5%. To solve the problem, we employ recent multiple testing techniques that allow us to bound the expected fraction of false rejections at a desired level. Monte Carlo results illustrate the usefulness of the techniques. The empirical results show that the data do not support the notion of output convergence after controlling for multiplicity.

Keywords: Growth Empirics, Multiple Testing, Convergence, Bootstrap

1. INTRODUCTION

A topic extensively researched in growth econometrics is output convergence of economies [Papageorgiou and Perez-Sebastian (2004)] or industries [Inklaar and Timmer (2009)]. A prominent notion of convergence across economies using a time-series definition is that of Pesaran (2007a). There, two economies converge only if a unit root test on the output gap of two economies, i.e., the difference of log per capita incomes, rejects. These pairwise tests with the null of no convergence are conducted for all different combinations of n countries. This results in $n(n-1)/2$ simultaneous tests. Given the large number of simultaneous tests, even if no country pair converges, one is bound to falsely reject the null of no convergence for many pairs.

We thank two anonymous referees, Jörg Breitung, Matei Demetrescu, Franz Palm, Stephan Smeekes and Jean-Pierre Urbain for useful comments that helped improve the paper. All errors are ours. The data and programs used in this paper are available upon request. Thomas Deckers gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft through the Bonn Graduate School of Economics (BGSE). Part of the paper was written while the authors were at Maastricht University, whose hospitality is also gratefully acknowledged. Address correspondence to: Thomas Deckers, Bonn Graduate School of Economics (BGSE). Department of Economics, Kaiserstrasse 1, 53113 Bonn, Germany; e-mail: thomas.deckers@uni-bonn.de.

1365-1005/12

When performing many simultaneous hypothesis tests, researchers often still rely on traditional significance tests. That is, one evaluates *each* hypothesis test individually at the same level α. Given some α, the probability of committing *at least one* type I error then is arbitrarily larger than α. To see this, note that the event of a rejection is a Bernoulli random variable with "success" probability α if the null is true. Assume (for illustration only) that all hypotheses are true and independent. Then P_l, the probability of finding l rejections in m tests (corresponding to, e.g., m country pairs), is the probability mass function of a binomial random variable,

$$P_l = \binom{m}{l}\alpha^l(1-\alpha)^{m-l}.$$

Therefore, the probability of at least one erroneous rejection for $\alpha = 0.05$ and $m = 100$ equals

$$P_{l\geq 1} = \sum_{j=1}^{100}\binom{100}{j}0.05^j(1-0.05)^{50-j} = 0.994.$$

Hence, the null of no convergence will be falsely rejected for several country pairs only because of the large number of tests performed.

We propose to tackle this problem using multiple testing techniques. These take the multiplicity of tests performed into account explicitly. One way to achieve such multiplicity control is to reject a hypothesis test only if its p-value satisfies $p_j \leq \alpha_j$ for some suitably chosen cutoff $\alpha_j \leq \alpha$. Such multiple testing techniques are routinely applied in many areas of applied statistics that involve multiple hypothesis testing, such as genomics [e.g., Dudoit and van der Laan (2007)]. The econometrics literature has so far paid little attention to solving the problem of multiplicity. For recent exceptions, see Hanck (2009), Moon and Perron (2009), or our companion paper [Deckers and Hanck (2011)], which deals with variable selection in cross-sectional growth regressions.

We focus on controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995). The FDR is defined as the expected number of falsely rejected hypotheses (here, falsely rejecting the null of no convergence) divided by the overall number of rejections. It is thus an extension of the notion of a type I error to the multiple testing situation. The power of different FDR controlling procedures might vary substantially [Romano et al. (2008a)]. A Monte Carlo study demonstrates the effectiveness of some popular procedures under different settings relevant to the present testing problems.

The empirical results show that output convergence between countries using a time series definition with the necessary condition of no unit root in the output gap of two economies does not seem to hold. This strengthens the findings of Pesaran (2007a).

Section 2 sketches the model used here and briefly surveys the literature. Section 3 describes the FDR-controlling procedures we employ. Section 4 assesses their

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

quality in a Monte Carlo study. Section 5 conducts FDR-controlling pairwise tests for output convergence. Section 6 concludes.

2. MODELS AND BACKGROUND

The literature discusses different definitions for output convergence and resulting tests; see, e.g., Islam (2003) for a survey. We study pairwise output convergence across countries and work with the time series model of Pesaran (2007a), to whom we refer for a more detailed exposition of the model. He assumes a common-factor model for the GDP of country i at time t,

$$y_{it} = c_i + g_i t + \boldsymbol{\theta}_i' \mathbf{f}_t + \epsilon_{it} + \eta_{it} \quad \text{for } i = 1, 2, \ldots, n, \tag{1}$$

where $\eta_{it} \sim I(0)$, but $\mathbf{f}_t$ and ϵ_{it} could be $I(1)$. His definition of convergence is based on the idea that for two countries i and j to converge, the output gap $d_{ijt} = y_{it} - y_{jt}$ should not fall outside a prespecified interval C with high probability π,

$$\Pr\left\{|y_{i,t+s} - y_{j,t+s}| < C|\mathcal{I}_t\right\} > \pi, \tag{2}$$

for all $s = 1, 2, \ldots, \infty$, where $\mathcal{I}_t$ is the information set at time t. For (2) to hold, the GDPs of two countries i and j should cointegrate and cotrend with vectors $(1, -1)'$. Pesaran (2007a) demonstrates that this condition can be checked by testing each output gap for a unit root and a linear trend. The absence of a unit root then becomes a necessary condition and the additional absence of a linear trend a sufficient one.

Pesaran tests for a unit root for all possible $n(n-1)/2$ country pairs. For large n and T, if no pair converges, the null of no convergence will be rejected for a fraction of pairs roughly equal to the level α of the individual tests. On the other hand, if all pairs converge, the fraction of pairs found convergent should tend to 1 for large n and T. Using Penn World Tables data, he only rejects the null for a fraction approximately equal to α. Hence, he finds no evidence for overall convergence. However, his approach does not make it possible to say whether the fraction of rejections consists entirely of type I errors or possibly does contain some correct rejections. Unlike Pesaran, through controlling the FDR, we can make statements about all individual country pairs rather than fractions of rejections.

Other studies using similar time series definitions of convergence also largely disfavor the overall convergence hypothesis. Bernard and Durlauf (1995) apply cointegration tests to a panel of 15 OECD countries. They fail to find convergence when testing for the cointegration vector $(1, -1)'$. Nevertheless they find cointegration relationships of the form $(1, -a)'$. This indicates, following their definition, conditional convergence, a concept less strict than the one employed here. Pesaran (2007a) shows that this approach can only handle a limited number of countries simultaneously. Earlier studies using univariate time series techniques [e.g., Campbell and Mankiw (1989); Quah (1990)] also fail to find evidence for overall convergence. A problem inherent in their approaches is the choice of

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

TABLE 1. Number of decisions made when testing m null hypotheses

	Declared nonsignificant	Declared significant	Total
True null hypotheses	$\mathbf{U}$	$\mathbf{F}$	m_0
Nontrue null hypotheses	$\mathbf{T}$	$\mathbf{S}$	$m - m_0$
	m-$\mathbf{R}$	$\mathbf{R}$	m

a “reference country” to which convergence of the other economies is tested. Investigating convergence of states within one country, Mello (2011) does find evidence for convergence for the more homogenous set of U.S. states.

3. CONTROLLING THE FALSE DISCOVERY RATE

The FDR as a desirable measurement of type I errors in multiple testing situations is introduced by Benjamini and Hochberg (1995). Adapting a notation similar to Benjamini and Hochberg (1995) and referring to Table 1, there are $m = n(n-1)/2$ hypotheses to be tested simultaneously, out of which m_0 are true. $\mathbf{R}$ is an observable random variable, whereas $\mathbf{U}$, $\mathbf{F}$, $\mathbf{S}$, and $\mathbf{T}$ are unobservable random variables. The proportion of falsely rejected null hypotheses can be described by $\mathbf{Q} = \mathbf{F}/(\mathbf{F}+\mathbf{S})$. Naturally, if $\mathbf{F} + \mathbf{S} = 0$, we take $\mathbf{Q} = 0$. The FDR is then defined as $\mathrm{E}(\mathbf{Q}) = \mathrm{E}[\mathbf{F}/(\mathbf{F}+\mathbf{S})] = \mathrm{E}(\mathbf{F}/\mathbf{R})$.

We now sketch the FDR-controlling procedures from Benjamini and Hochberg (1995) and Romano et al. (2008a) to be used later. We frequently refer to these as the BH method and the bootstrap method.

3.1. The BH Method

One first chooses a level γ at which to control the FDR. Let $\widehat{p}_{(1)} \leq \ldots \leq \widehat{p}_{(m)}$ be the sorted p-values and $H_{(1)}, \ldots, H_{(m)}$ the corresponding null hypotheses. For $1 \leq j \leq m$, let $\gamma_j = \frac{j}{m}\gamma$. Then BH rejects $H_{(1)}, \ldots, H_{(j^*)}$, where j^* is the largest j such that $\widehat{p}_{(j)} \leq \gamma_j$. If no such j exists, no hypothesis is rejected. This is a “step-up” method that considers the hypotheses sorted from most significant to least significant. It begins with the least significant hypothesis, accepts hypotheses up to a certain point, and rejects the remaining ones. Benjamini and Yekutieli (2001) show control of the FDR if positive regression dependency holds for the test statistics used for $H_1, \ldots, H_m$. Positive regression dependency describes very general forms of dependency of random variables. It includes, for example, but is much broader than and hence not limited to, multivariate normal distributions with nonnegative correlations, specific absolute-valued multivariate normal, multivariate gamma, and F distributions. Hence, crucially, the procedure can deal with the empirically relevant situation that test statistics and p-values are not independent. The Monte Carlo study in Section 4 shows that the FDR is also

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

controlled under plausible assumptions about the data-generating process (DGP) described in Section 2.

3.2. The Bootstrap Method

The bootstrap method is a step-down rather than a step-up procedure. Assume without loss of generality that a hypothesis H_i is rejected for large values of its test statistic T_i. Arrange the test statistics from smallest to largest, i.e., $T_1 \leq T_2 \leq \ldots \leq T_m$, and let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ be the corresponding hypotheses. A step-down procedure then compares the largest test statistic T_m with a suitable critical value c_m. If $T_m < c_m$ the procedure rejects no hypothesis; otherwise it rejects $H_{(m)}$ and steps down to T_{m-1}. The procedure continues until it either rejects $H_{(1)}$ or does not reject the current hypothesis. Hence, a step-down procedure rejects the hypotheses

$$H_{(m)}, H_{(m-1)}, \ldots, H_{(m-j^*)},$$

where j^* is the largest integer j satisfying

$$T_m \geq c_m, T_{m-1} \geq c_{m-1}, \ldots, T_{m-j} \geq c_{m-j}.$$

If no such j exists, the method does not reject any hypotheses.

The intuition behind the procedure is as follows. For any step-down procedure the FDR is

$$\begin{aligned} \text{FDR} = \text{E}\left[\frac{\mathbf{F}}{\max\{\mathbf{R}, 1\}}\right] &= \sum_{1\leq r\leq m} \frac{1}{r}\text{E}[\mathbf{F}|\mathbf{R} = r]P\{\mathbf{R} = r\} \\ &= \sum_{1\leq r\leq m} \frac{1}{r}\text{E}[\mathbf{F}|\mathbf{R} = r] \\ &\quad \times P\left(T_m \geq c_m, \ldots, T_{m-r+1} \geq c_{m-r+1}, T_{m-r} < c_{m-r}\right), \end{aligned} \qquad (\mathbf{3})$$

where the event $T_{m-r} < c_{m-r}$ is defined to be true when $r = m$.

Of course, (3) depends on m_0 through the number of falsely rejected hypotheses **F**. Because m_0 is unknown, one has to ensure that (3) is bounded above by γ for every possible m_0. That is exactly the condition used to recursively determine the critical values. It is quite straightforward [refer to Romano et al. (2008b) for details] to show that for $m_0 = 1$, for example, (3) simplifies to

$$\text{FDR} = \frac{1}{m}P\left(T_{1:1} \geq c_1\right),$$

where $T_{r:m_0}$ denotes the rth smallest test statistic of the m_0 true hypotheses. Hence, one can calculate the first critical value from

$$c_1 = \inf\left\{x \in \mathbf{R} : \frac{1}{m}P\left(T_{1:1} \geq x\right) \leq \gamma\right\}.$$

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

If $m\gamma > 1$, c_1 is set to $-\infty$. The remaining critical values $c_2, \ldots, c_m$ can be found using further steps of the recursion.

In practice these critical values are not available, because the measure P is unknown. We therefore approximate P using a bootstrap measure $\hat{P}$. Here, $\hat{P}$ is to estimate P such that T_j^*, the bootstrapped test statistic, is a good approximation of T_j whenever the corresponding null hypothesis is true [Romano et al. (2008a)]. Details on an appropriate bootstrap for the present problem are given in the Appendix. The key goal is to deal with unit roots under the null and to properly preserve the dependence structure of the series in order provide valid inference under correlation.

Note that neither multiple testing approach, unlike Pesaran's, requires $n \to \infty$ for valid inference. This is useful in and of itself, but also important because the resampling method we employ (see the Appendix) is only known to be valid for finite n.

Clearly, the active multiple testing literature has proposed other extensions of the type I error, such as the familywise error rate (FWER, defined as the probability of one or more false rejections, a stricter criterion than the FDR), as well as other FDR- and FWER-controlling procedures. We also studied the procedures of Storey et al. (2004) and Benjamini et al. (2006). For the latter, we find results very similar to those to be reported later, whereas Storey et al. (2004) often proved to be too liberal. Further examples include Storey (2002), Sarkar (2006), and Finner et al. (2009). We shall focus on the procedures described earlier, as these arguably are among the most popular ones, but also for brevity.

4. MONTE CARLO STUDY

We now shed some light on the performance of the FDR-controlling techniques described previously. We compare the procedures with each other and with the classical approach to hypothesis testing (i.e., rejecting H_i if $p_i \leq \alpha$). The first performance criterion is the average of the proportion of falsely rejected hypotheses. This will, as the number of simulations grows, converge to the FDR. The second criterion is the average number of right rejections divided by the number of false hypotheses, i.e., convergent pairs, times 100, or "power." As in Pesaran (2007a), we use the following DGP:

$$y_{it} = \gamma_i f_t + \epsilon_{it},$$

where

$$f_t = f_{t-1} + v_t, \quad v_t = \rho_v v_{t-1} + e_t, \quad e_t \sim \text{i.i.d. N}\left(0, 1 - \rho_v^2\right)$$

and

$$\epsilon_{it} = \rho_i \epsilon_{i,t-1} + \nu_{it}, \quad \nu_{it} \sim \text{i.i.d. N}\left(0, \sigma_{\nu i}^2\left(1 - \rho_i^2\right)\right), \quad \sigma_{\nu i}^2 \sim \text{i.i.d. uniform}[0.5, 1.5],$$

for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. We consider $n = 10$, and set $T = 50$, $T = 100$, and $T = 200$.[1] When generating the autoregressive processes we start at $t = -49$ and discard the first 50 draws. As in Pesaran (2007a), we take $\rho_v = 0.6$

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

and $\rho_i \sim$ i.i.d. uniform[0.2, 0.6]. We only consider one single pool of convergent countries; i.e., all countries within the pool converge with each other and all others do not. For all converging pairs we set $\gamma_i = \gamma_j = 1$. For all others, $\gamma_i \sim$ i.i.d. $\chi^2_{\kappa_i}$, where κ_i is drawn with replacement from integers 1 to 10. We consider three scenarios:

1. For $3/10$ of all countries $\gamma_i = 1$. For $n = 10$ this amounts to 3 convergent pairs.
2. For $1/2$ of all countries $\gamma_i = 1$. For $n = 10$ this amounts to 10 convergent pairs.
3. For all countries $\gamma_i \sim$ i.i.d. $\chi^2_{\kappa_i}$. This amounts to no convergent pairs.

When running the ADF unit root test regressions, we use deterministic lag length choices $p = \lfloor 5(T/100)^{1/4} \rfloor$, $p = \lfloor 6(T/100)^{1/4} \rfloor$, and $p = \lfloor 12(T/100)^{1/4} \rfloor$ as suggested in Schwert (1989). These rules are shown to work well in Demetrescu et al. (2011). To save computation time, we do not use model selection criteria such as the AIC here. [Also, Leeb and Pötscher (2008) show that data-driven lag-length selection may distort subsequent inference, although some unreported simulations suggest small differences in the present application.] We employ both OLS and GLS correction for deterministics in the ADF test (for details, see the Appendix), and consider both the cases of demeaning and detrending. To get accurate finite-sample p-values, we simulate MacKinnon-type finite sample distributions for both the OLS and GLS statistics.

Table 2 shows the results for $n = 10$, so 45 hypotheses are tested and 10 pairs are convergent, i.e., contain no unit root (scenario 2). For all p, as expected, classical testing severely violates the FDR for both OLS or GLS tests. The multiple testing procedures control the FDR for sufficiently high p (we even find that all FDR-controlling procedures also control the FWER). A large number of lags is plausible for this DGP, as the d_{ijt} are sums of $AR(1)$ processes, which generally follow an $ARMA(2, 1)$ [Granger and Morris (1976)], or $AR(\infty)$, process, which has to be approximated with a large number of lags.[2] Moreover, that FDR control can only be attained for high p is in line with Pesaran (2007a), who finds size distortions of the individual tests when $p < 4$. Given the required p for each FDR-controlling procedure, the bootstrap is most powerful. For the BH method, GLS yields somewhat higher power. As the power properties of the univariate unit root tests translate into the multiple testing procedures, their power is unsurprisingly higher under demeaning than under detrending.

Varying T and the number of false hypotheses leaves the results qualitatively the same as regards FDR control. Classical testing results in a violation of the FDR in all scenarios. The main difference for $T = 100$ and $T = 200$ is that FDR control of the BH and bootstrap procedure is only attained for sufficiently large choices of p, i.e., for the rules $p = \lfloor 6(T/100)^{1/4} \rfloor$ and $p = \lfloor 12(T/100)^{1/4} \rfloor$. Unsurprisingly, the procedures become more powerful with T.[3]

The third scenario only considers the "size," i.e., FDR control of the procedures. Table 3 reports some results for $T = 50$ and demeaned tests. We find that both the bootstrap method and BH control the FDR for sufficiently high p, which is then equivalent to the FWER.

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

TABLE 2. Simulation results

	OLS demeaning		GLS demeaning		OLS detrending		GLS detrending	
	FDR ×100	Power	FDR ×100	Power	FDR ×100	Power	FDR ×100	Power
				$T = 50$				
$p = 4$								
Classical	15.8	51.0	17.2	49.4	20.7	32.3	20.8	30.2
BH	5.1	8.7	6.4	10.8	3.0	2.1	2.8	2.1
Bootstrap	4.5	9.9	4.8	9.8	3.7	3.7	3.1	3.5
$p = 5$								
Classical	17.5	42.4	18.3	42.4	22.7	25.8	22.4	23.8
BH	4.7	5.6	6.4	7.7	2.1	1.5	2.5	1.5
Bootstrap	3.9	6.5	2.8	5.9	2.4	2.1	2.5	2.1
$p = 10$								
Classical	21.6	15.9	23.2	17.3	24.8	7.7	27.7	8.1
BH	2.9	0.7	5.0	1.5	1.1	0.2	1.9	0.3
Bootstrap	7.8	1.1	0.9	0.8	1.1	0.2	1.3	0.2
				$T = 100$				
$p = 5$								
Classical	11.4	92.5	12.7	70.3	14.5	76.8	15.6	69.0
BH	7.4	61.6	7.6	41.0	6.4	28.7	7.1	22.2
Bootstrap	6.7	65.5	7.7	44.2	7.5	37.2	8.0	28.7
$p = 6$								
Classical	11.1	84.8	12.7	62.8	13.7	65.4	14.6	57.6
BH	6.6	41.2	6.9	28.6	4.4	14.6	4.7	12.0
Bootstrap	6.8	46.9	5.0	30.3	5.5	22.5	5.4	16.9
$p = 12$								
Classical	11.1	42.5	15.2	31.5	15.8	2.3	17.8	18.5
BH	2.2	4.9	4.8	4.4	0.6	1.1	1.6	0.8
Bootstrap	1.5	5.5	1.1	3.2	1.3	2.2	1.3	1.1
				$T = 200$				
$p = 5$								
Classical	11.6	100.0	12.2	86.8	13.3	99.9	13.7	95.1
BH	7.6	99.9	7.4	70.9	8.8	98.7	8.5	84.4
Bootstrap	8.8	99.9	8.6	74.4	10.2	99.3	9.8	87.2
$p = 7$								
Classical	10.1	99.8	11.1	75.4	13.0	98.9	13.5	8.9
BH	6.0	97.9	6.5	54.4	7.6	89.2	7.5	62.5
Bootstrap	7.4	98.9	7.1	58.2	9.3	93.6	8.5	68.0
$p = 14$								
Classical	7.7	87.7	11.5	45.0	9.4	66.6	11.2	43.5
BH	3.2	42.2	3.9	17.4	2.0	14.8	2.0	6.5
Bootstrap	4.1	55.2	3.9	19.9	2.9	24.8	2.5	9.7

Note: Monte Carlo results for the DGP from Section 4, with 1,000 replications. Power is defined as the number of right rejections divided by the number of false hypotheses, i.e., convergent pairs, times 100. Tests were conducted for $\alpha = \gamma = 0.05$. p is chosen according to the rules $p = \lfloor 5(T/100)^{1/4} \rfloor$, $p = \lfloor 6(T/100)^{1/4} \rfloor$, and $p = \lfloor 12(T/100)^{1/4} \rfloor$. The FDR-controlling procedures applied are described in Section 3. The bootstrap method uses 500 resamples.

TABLE 3. Size study: For all i, $\gamma_i \sim$ i.i.d. $\chi^2(\kappa_i)$ (κ_i is drawn with replacement from the integers 1 to 10), hence no pair converges

	OLS demeaning			GLS demeaning		
	$p=4$	$p=5$	$p=10$	$p=4$	$p=5$	$p=10$
Classical	0.509	0.455	0.366	0.540	0.494	0.402
BH	0.047	0.037	0.027	0.072	0.039	0.036
Bootstrap	0.067	0.049	0.027	0.079	0.039	0.018

Note: The FDR is equivalent to the FWER in this size study. Thus, the procedures also control the FWER whenever they control the FDR at the indicated levels. We set $T = 50$ and $n = 10$, so there are 45 hypothesis tests to conduct, of which no hypothesis is false. See also notes to Table 2.

5. OUTPUT CONVERGENCE REVISITED

We now employ the FDR-controlling techniques described in Section 3 to a data set of $n = 51$ countries ranging from 1950 to 2003, so $T = 54$. The resulting number of different country pairs of output gap series d_{ijt} is 1,275. The data is from the Penn World Tables, Version 6.2 [Heston et al. (2006)] and includes all countries for which data on per capita output were available for the indicated time span, cf. Table 4. We apply standard ADF tests and ADF-GLS tests to d_{ijt}. We choose the lag length p according to $p = \lfloor 5(T/100)^{1/4} \rfloor$ and $p = \lfloor 6(T/100)^{1/4} \rfloor$, as those yielded the largest number of right rejections although still controlling the FDR for the bootstrap procedure when $T = 50$, which is close to the actual T. For the present data this results in $p = 4$ and $p = 5$. (Results are similar for other choices of p.) We additionally select p with the modified AIC criterion of Ng and Perron (2001). An explorative data analysis shows that nearly all output gap series show a trending pattern. Therefore, we include time trends in the unit root tests, also following Pesaran (2007a). Critical values and p-values are adjusted to sample size.

Table 5 shows the results of the pairwise convergence tests. We corroborate the results of Pesaran (2007a) that the null of no convergence is only rejected for a fraction of pairs less than or equal to the individual significance level of the tests. This result is robust over all selection procedures for p, with the MAIC finding fewer rejections than the rules of thumb. It may be surprising that OLS rejects more

TABLE 4. Countries included

Argentina	Costa Rica	Honduras	Luxembourg	Norway	Sri Lanka	United States
Australia	Denmark	Iceland	Mauritius	Pakistan	Sweden	Uruguay
Austria	Egypt	India	Mexico	Panama	Switzerland	Venezuela
Belgium	El Salvador	Ireland	Morocco	Peru	Thailand	
Bolivia	Ethiopia	Israel	Netherlands	Philippines	Trinidad & Tobago	
Brazil	Finland	Italy	New Zealand	Portugal	Turkey	
Canada	France	Japan	Nicaragua	South Africa	Uganda	
Colombia	Guatemala	Kenya	Nigeria	Spain	United Kingdom	

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

TABLE 5. Pairwise tests for output convergence

	OLS detrending		GLS detrending	
	# pairs	% of pairs	# pairs	% of pairs
		$p = 4$		
$\alpha = \gamma = 0.01$				
Classical approach	2	0.16	3	0.24
BH	0	0	0	0
Bootstrap	0	0	0	0
$\alpha = \gamma = 0.05$				
Classical approach	40	3.13	23	1.80
BH	0	0	0	0
Bootstrap	0	0	0	0
$\alpha = \gamma = 0.1$				
Classical approach	83	6.51	47	3.68
BH	0	0	0	0
Bootstrap	0	0	0	0
		$p = 5$		
$\alpha = \gamma = 0.01$				
Classical approach	10	0.78	8	0.63
BH	0	0	0	0
Bootstrap	0	0	0	0
$\alpha = \gamma = 0.05$				
Classical approach	45	3.53	39	3.06
BH	0	0	0	0
Bootstrap	0	0	0	0
$\alpha = \gamma = 0.1$				
Classical approach	96	7.53	73	5.49
BH	0	0	0	0
Bootstrap	0	0	0	0
		$p(MAIC)$		
$\alpha = \gamma = 0.01$				
Classical approach	9	0.71	0	0
BH	0	0	0	0
Bootstrap	0	0	0	0
$\alpha = \gamma = 0.05$				
Classical approach	40	3.13	9	0.71
BH	0	0	0	0
Bootstrap	0	0	0	0
$\alpha = \gamma = 0.1$				
Classical approach	83	6.51	27	2.12
BH	0	0	0	0
Bootstrap	0	0	0	0

Note: "# pairs" shows the number of country pairs for which the null of a unit root is rejected. "% of pairs" shows the proportion of rejected pairs compared to the total number of pairs. The procedures are described in Section 3. We use 5,000 bootstrap resamples.

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

frequently than the GLS-based test, as some conventional wisdom has it that the latter is more powerful. However, Müller and Elliott (2003) show that the power ranking reverses when the process under investigation has a large initial condition, i.e., is far away from the deterministic part of the series at the beginning of the sample. The present results suggest this is the case for many output gap series. This is intuitive, as large deviations from the output gap at the beginning of the sample are quite plausible for many country pairs, in particular when developing countries are considered.

When we account for the multiplicity of tests performed, for no level γ do we find any rejection of the null; this also holds in particular for the most powerful FDR-controlling procedure, the bootstrap method using either OLS or GLS detrending.[4] This also implies that all rejections found with classical testing are to be attributed to multiple type I errors, not convergence. This possibly surprising finding clarifies the results in Pesaran (2007a). His approach does not address whether rejection of the null for some individual pairs might be an indication of "true convergence." From finding no converging pairs when employing a suitable testing framework for individual tests (rather than for fractions of rejections), the confidence in Pesaran's no time series–convergence finding is strengthened.[5]

Some authors argue that many time series encountered in international macroeconomics are better described by nonlinearly adjusting processes under H_1; see, e.g., Kim and Moh (2010) or Norman (2010) for recent applications to real exchange rates, and Chong et al. (2008) and Ucar and Omay (2009) for studies investigating the present topic of income convergence. Linear ADF-type tests are potentially less powerful against such deviations from the null hypothesis. We therefore also test for income convergence using the nonlinear unit root test of Kapetanios et al. (2003), again controlling the FDR using the BH and bootstrap method. We again use detrended series. Our results are, however, qualitatively very similar to those reported previously. Specifically, the bootstrap method still identifies no pair to be convergent, whereas the BH method finds 0, 1, and 1 converging pairs at $\gamma = 0.01$, $\gamma = 0.05$, $\gamma = 0.10$ for $p = 4$, and 0, 3, and 13 converging pairs for $p = 5$. Hence, evidence of stationary output gaps is still very weak at best, with at most 1% ($\approx 100 \times 13/1{,}275$) of the pairs found to be convergent. We therefore view our results as robust to the presence of possibly nonlinear adjustment to output gaps.

We stress that convergence was tested using a strict definition [Islam (2003)]. Hence, our findings should not be read as ruling out convergence, using for example conditional definitions, nor do we make any statements about convergence of single countries to a steady state output. We do claim that the data do not support the notion of convergence across economies, using a time series definition with the necessary condition of no unit root in the log per capita output gap of two economies. Clearly, several factors contribute to this finding. On one hand, output gaps may simply not converge, according to the present notion employed. On the other hand, the procedures used here may not be able to detect the convergent pairs, i.e., lack power. As shown in Section 4, the procedures are indeed only capable of

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

identifying a fraction of the stationary series. This is because the multiple testing procedures inevitably inherit the often weak power properties of the underlying unit root tests. That said, Section 4 does reveal the procedures to be capable of finding at least some stationary series when they are present in the panel under study, and the fact that none are found in this section casts substantial doubt on the prevalence of the convergence notion used here for the data investigated.

Before concluding, we remark (we thank an anonymous referee for raising this point) that it may prove valuable to exploit the logical dependence structure among the hypotheses in other applications. Concretely, suppose we had found that $y_{1,t} - y_{2,t} \sim I(0)$ and $y_{1,t} - y_{3,t} \sim I(0)$. This implies that $y_{2,t} - y_{3,t} \sim I(0)$, so that conducting the third test is redundant. Hence, the total number of comparisons to be performed can be reduced, with potential advantages for the power of the multiple testing procedures. Some progress in this direction might be achieved by extending results from Shaffer (1986). She considers modifying further rounds of the Holm (1979) procedure for comparing n means to remove redundant hypotheses after an initial rejection has been found. This is not an issue for the present application, as we do not find an initial rejection. It is hence—and also because extending her idea would likely be challenging to our case of more than 1,000 tests—beyond the scope of the present paper. We believe, however, that it does deserve further attention.

6. CONCLUSION

This paper highlights the importance of accounting for the multiplicity of tests performed when testing for output convergence. Multiplicity control was achieved by controlling the expected fraction of erroneously rejected hypotheses to the total number of rejections, i.e., the FDR. Among others, this was done using a bootstrap approach that takes the dependence structure of the test statistics into account and thus has high power. We investigate cross-country convergence using pairwise unit-root tests. Controlling the FDR, we find no evidence of this type of convergence. This strengthens the results of Pesaran (2007a), whose framework considers only the fraction of rejected pairs and is not concerned with single-country pairs.

There are more literatures in applied econometrics where accounting for multiplicity is important. For example, some forecasting exercises involve a large number of candidate explanatory variables. Relatedly, Lim and Brooks (2010) compare deviations from the random walk benchmark of stock indices from many countries. As we have shown in our application, classical significance tests produce a non-negligible number of hypothesis tests that appear to be only spuriously significant because of the large number of tests performed. Thus, the techniques studied here may prove fruitful in many other literatures.

NOTES

[1] Because of the large computation time required to find the bootstrap critical values in the Algorithm from the Appendix we only conduct limited experiments for larger n, for which we find a qualitatively similar pattern. Results are available upon request. Note, however, that the empirical

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

application of Section 5, for which $n = 51$ and hence $n(n-1)/2 = 1{,}275$, runs in a few minutes on a standard PC and hence does not suffer from a serious curse of dimensionality. We moreover experimented with a version of Pesaran's DGP in which the common factor is stationary and the factor loadings γ_i are heterogeneous. Detailed results, which were qualitatively similar, are available.

2. As unreported but available in a results document, choosing a lower p resulted in some upward size distortion.

3. Results for scenario 1 are available upon request.

4. Like Pesaran (2007a), we also perform analogous exercises for subgroups such as European or American countries. Again, our findings coincide with his in that we do not find individual converging country pairs.

5. As we have shown in the Monte Carlo study, we also control the FWER at γ. Hence, the probability of even a single false rejection is bounded by γ.

6. Some exploratory simulations suggested a similar performance as for the procedure described below, provided certain block lengths were used. We here advocate the sieve in view of well-established guidelines to choose p in practice.

REFERENCES

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

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli (2006) Adaptive linear step-up procedure that control the false discovery rate. *Biometrika* 93, 491–507.

Benjamini, Yoav and Daniel Yekutieli (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165–1188.

Bernard, Andrew B. and Steven N. Durlauf (1995) Convergence in international output. *Journal of Applied Econometrics* 10, 97–108.

Burridge, Peter and A.M. Robert Taylor (2004) Bootstrapping the HEGY seasonal unit root tests. *Journal of Econometrics* 123, 67–87.

Campbell, Y.J. and N.G. Mankiw (1989) International evidence on the persistence of economic fluctuations. *Journal of Monetary Economics* 23, 319–333.

Chang, Yoosoon (2004) Bootstrap unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* 120, 263–293.

Chong, Terence Tai-Leung, Melvin J. Hinich, Venus Khim-Sen Liew, and Kian-Ping Lim (2008) Time series test of nonlinear convergence and transitional dynamics. *Economics Letters* 100, 337–339.

Deckers, Thomas and Christoph Hanck (2011) Variable Selection via Multiple Testing with an Application to Growth Econometrics. Technical report, University of Groningen.

Demetrescu, Matei, Uwe Hassler, and Vladimir Kuzin (2011) Pitfalls of post-model-selection testing: Experimental quantification. *Empirical Economics* 40, 359–372.

Dudoit, Sandrine and Mark J. van der Laan (2007) *Multiple Testing Procedures and Applications to Genomics*, Springer Series in Statistics, Berlin: Springer.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock (1996) Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–836.

Finner, Helmut, Thorsten Dickhaus, and Markus Roters (2009) On the false discovery rate and an asymptotically optimal rejection curve. *Annals of Statistics* 37, 596–618.

Granger, Clive W. J. and Matthew J. Morris (1976) Time series modelling and interpretation. *Journal of the Royal Statistical Society. Series A (General)* 139, 246–257.

Hanck, Christoph (2009) For which countries did PPP hold? A multiple testing approach. *Empirical Economics* 37, 93–103.

Heston, Alan, Robert Summers, and Bettina Aten (2006) Penn World Table, Version 6.2. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September.

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

Holm, Sture (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

Inklaar, Robert and Marcel P. Timmer (2009) Productivity convergence across industries and countries: The importance of theory-based measurement. *Macroeconomic Dynamics* 13(S1), 218–240.

Islam, Nazrul (2003) What have we learnt from the convergence debate? *Journal of Economic Surveys* 17, 312–355.

Kapetanios, George, Yongcheol Shin, and Andy Snell (2003) Testing for a unit root in the nonlinear STAR framework. *Journal of Econometrics* 112, 359–379.

Kim, Hyeongwoo and Young-Kyu Moh (2010) A century of purchasing power parity confirmed: The role of nonlinearity. *Journal of International Money and Finance* 29, 1398–1405.

Leeb, Hannes and Benedikt M. Pötscher (2008) Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24, 338–367.

Lim, Kian-Ping and Robert D. Brooks (2010) Why do emerging stock markets experience more persistent price deviations from a random walk over time? A country-level analysis. *Macroeconomic Dynamics* 14(S1), 3–41.

Mello, Marcelo (2011) Stochastic convergence across U.S. states. *Macroeconomic Dynamics* 15, 160–183.

Moon, Hyungsik R. and Benoit Perron (2009) Beyond Panel Unit Root Tests: Using Multiple Testing to Determine the Non Stationarity Properties of Individual Series in a Panel. Technical report, Université de Montréal.

Müller, Ulrich K. and Graham Elliott (2003) Tests for unit roots and the initial condition. *Econometrica* 71, 1269–1286.

Ng, Serena and Pierre Perron (2001) Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69, 1519–1554.

Norman, Stephen (2010) How well does nonlinear mean reversion solve the PPP puzzle? *Journal of International Money and Finance* 29, 919–937.

Palm, Franz C., Stephan Smeekes, and Jean-Pierre Urbain (2011) Cross-sectional dependence robust block bootstrap panel unit root tests. *Journal of Econometrics* 163, 85–104.

Papageorgiou, Chris and Fidel Perez-Sebastian (2004) Can transition dynamics explain the international output data? *Macroeconomic Dynamics* 8, 466–492.

Paparoditis, Efstathios and Dimitris N. Politis (2003) Residual-based block bootstrap for unit root testing. *Econometrica* 71, 813–855.

Pesaran, M. Hashem (2007a) A pair-wise approach to testing for output and growth convergence. *Journal of Econometrics* 138, 312–355.

Pesaran, M. Hashem (2007b) A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Econometrics* 22, 265–312.

Phillips, Peter Charles Bonest (1987) Towards a unified asymptotic theory for autoregression. *Biometrika* 74, 535–547.

Quah, Danny (1990) International Patterns of Growth: I. Persistence in Cross-Country Disparities. Working paper, MIT.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf (2008a) Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test* 17, 417–442.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf (2008b) Formalized data snooping based on generalized error rates. *Econometric Theory* 24, 404–447.

Sarkar, Sanat K. (2006) False discovery and false nondiscovery rates in single-step multiple testing procedures. *Annals of Statistics* 34, 394–415.

Schwert, G. William (1989) Tests for unit roots: A Monte Carlo investigation. *Journal of Businnes and Economic Statistics* 7, 5–17.

Shaffer, Juliet Popper (1986) Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* 81, 826–831.

Smeekes, Stephan (in press) Detrending bootstrap unit root tests. *Econometric Reviews*.

Storey, John D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* 64, 479–498.

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

Storey, John D., Jonathan E. Taylor, and David Siegmund (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B* 66, 187–205.

Ucar, Nuri and Tolga Omay (2009) Testing for unit root in nonlinear heterogeneous panels. *Economics Letters* 104, 5–8.

APPENDIX: BOOTSTRAP PROCEDURES

Previous applications of the Romano et al. (2008a) approach tested stationary variables. We now describe how to extend their approach to the present testing problem on nonstationary time series. We apply the following semiparametric sieve bootstrap unit root test procedure developed in Smeekes (in press). Other resampling schemes that properly mimic the null distribution of the unit root tests statistics, such as the block bootstrap [see, e.g., Paparoditis and Politis (2003)] may be used here, too.[6]

1. Calculate the output gap $d_{ijt} = y_{it} - y_{jt}$ for each country pair, i.e., for $i = 1, \ldots, n-1$, $j = 2, \ldots, n$, and $t = 1, \ldots, T$. Do the next steps simultaneously for all $i = 1, \ldots, n-1$ and $j = 2, \ldots, n$.
2. Following Pesaran (2007a), detrend the output gap d_{ijt}. We consider two detrending schemes. One is to calculate $d^d_{ijt} = d_{ijt} - \hat{\boldsymbol{\phi}}' \boldsymbol{z}_t$, where $\boldsymbol{z}_t = (1, t)'$. Here, $\hat{\boldsymbol{\phi}}$ is the usual OLS estimator, $\hat{\boldsymbol{\phi}} = (\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t')^{-1} \times (\sum_{t=1}^{T} \boldsymbol{z}_t d_{ijt})$. We also consider GLS detrending. Elliott et al. (1996) show that GLS detrending may result in higher power of the ADF test against local alternatives of the form $\rho = 1 + \bar{c}T^{-1}$. Let $\boldsymbol{z}_{1\bar{c}} = \boldsymbol{z}_1$ and $\boldsymbol{z}_{t\bar{c}} = \boldsymbol{z}_t - (1 + \bar{c}T^{-1})\boldsymbol{z}_{t-1}$ for $t = 1, 2, \ldots, T$. Likewise, define $d_{ij1\bar{c}} = d_{ij1}$ and $d_{ijt\bar{c}} = d_{ijt} - (1 + \bar{c}T^{-1})d_{ij,t-1}$ for $t = 1, 2, \ldots, T$. Then the other is to calculate
$$\hat{\boldsymbol{\phi}}_{\bar{c}} = \left(\sum_{t=1}^{T} \boldsymbol{z}_{t\bar{c}} \boldsymbol{z}_{t\bar{c}}' \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{z}_{t\bar{c}} d_{ijt\bar{c}} \right).$$
Finally, we obtain $d^d_{ijt} = d_{ijt} - \hat{\boldsymbol{\phi}}_{\bar{c}}' \boldsymbol{z}_t$.
3. Estimate an ADF regression of order p for d^d_{ijt} and calculate the residuals as
$$\hat{\epsilon}_{ijt} = \Delta d^d_{ijt} - \hat{\alpha} d^d_{ij,t-1} - \sum_{j=1}^{p} \hat{\psi}_j \Delta d^d_{ij,t-j}.$$
If the estimated AR process is explosive, we impose a root bound as in Burridge and Taylor (2004). Demean the residuals, $\tilde{\epsilon}_{ijt} = \hat{\epsilon}_{ijt} - \frac{1}{n-p-1} \sum_t \hat{\epsilon}_{ijt}$. Also calculate the ADF statistic $t_{\hat{\alpha}}$ and $\text{ADF}_{-1} = (-1)t_{\hat{\alpha}}$, in order to reject for large values as assumed in the derivation of critical values (following).
4. Resample $\tilde{\epsilon}_{ijt}$ nonparametrically with replacement to obtain the bootstrap residuals ϵ^*_{ijt}. Here, the $\tilde{\epsilon}_{ijt}$ are resampled jointly across country pairs in order to suitably preserve the cross-sectionality in the panel.
5. Build u^*_{ijt} recursively as $u^*_{ijt} = \sum_{j=1}^{p} \hat{\psi}_j u^*_{ij,t-j} + \epsilon^*_{ijt}$. Then build $d^*_{ijt} = d^*_{ij,t-1} + u^*_{ijt}$. [Following Smeekes (in press), we do not add deterministic components to the bootstrapped series, for simplicity.]
6. Detrend d^*_{ijt} as in step 2 to obtain d^{d*}_{ijt}.

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press

7. Estimate by OLS the ADF regression

$$\Delta d_{ijt}^{d*} = \hat{\alpha}^* d_{ij,t-1}^{d*} + \sum_{j=1}^{p} \hat{\psi}_j \Delta d_{ij,t-j}^{d*} + \hat{\epsilon}_{ijt}^*$$

   and calculate the ADF test statistic for $\hat{\alpha}^*$ and ADF_{-1}^*. Repeat steps 2–7 B times.
8. Given $\hat{P}$, the critical values are defined recursively as follows: Having determined $\hat{c}_1, \ldots, \hat{c}_{j-1}$, the jth critical value is determined using the minimization rule [Romano et al. (2008a)]:

$$\hat{c}_j = \inf\left\{c \in \mathbf{R} : \sum_{m-j+1 \le r \le m} \frac{r-m+j}{r} \right.$$

$$\left. \times \hat{P}\left(T_{j:j}^* \ge c, \ldots, T_{m-r+1:j}^* \ge \hat{c}_{m-r+1}, T_{m-r:j}^* < \hat{c}_{m-r}\right) \le \gamma \right\}. \qquad \textbf{(A.1)}$$

   Note the meaning of $T_{r:t}^*$. The index t stems from the ordering of the original test statistics, whereas r corresponds to the bootstrapped test statistics. So $T_{r:t}^*$ says: Out of the t smallest original test statistics, pick the rth smallest of the corresponding bootstrap test statistics.
9. Apply the rule (A.1) with $m = n(n-1)/2$ and ADF_{-1}^* to calculate the critical values.
10. Use the critical values from (8) and compare them to ADF_{-1} from (3) using the procedure (A.1).

The bootstrap method is consistent, i.e., satisfies $\limsup_{T\to\infty}$ FDR $\leqslant \alpha$ under a set of weak conditions [see Romano et al. (2008a, Theorem 1)]. These are (i) continuous marginal distributions of the test statistics, (ii) connected support of the joint distribution of the test statistic, (iii) test statistics forming an exchangeable sequence, (iv) availability of consistent estimators of the standard errors of coefficient estimators, e.g., $\hat{\alpha}$, and (v) weak convergence of bootstrap distributions to the true one as $T \to \infty$. We do not provide a formal proof of the validity of these conditions in the present application. However, some heuristics are useful. Conditions (i) and (iv) are well known to hold in the unit root literature [Phillips (1987)], whereas (ii) is a regularity condition [see Romano et al. (2008a, Remark 6)]. Pesaran (2007b) shows (iii) to hold in a very similar panel unit root–testing framework. Bootstrap consistency results for nonstationary panels (v) follows from extending the univariate setup of Smeekes (in press) to the joint distribution of the m test statistics along the lines of, e.g., Chang (2004) and Palm et al. (2011), who provide bootstrap panel consistency results for the sieve and block bootstrap, respectively.

https://doi.org/10.1017/S1365100512000338 Published online by Cambridge University Press