Multiple Comparisons and Joint Significance in Panel Unit Root Testing with Evidence on International Interest Rate Linkage*

Uwe Hassler and Verena Werkmann**
Goethe University Frankfurt

JEL C12; C23; E43

Panel unit root test; uncovered interest rate parity; combination of p-values; multiple testing; bootstrap.

Received: 24.08.2012

Revision received: 05.03.2013

Accepted: 28.06.2013

Summary

This paper adds to the issue of inference regarding potentially nonstationary panels where units are correlated. Recently, it has been proposed to tackle this problem by computing individual p-values and combining them to an overall joint significance. We adopt and illustrate this fairly general approach allowing for competing means to account for cross-correlation when analyzing samples of $N = 10$ international interest rate differentials of different maturities.

Alternatively, we investigate the approach of multiple testing or multiple comparison that has rarely been employed in econometrics. The advantages are that the computation of p-values is not necessarily required, and that one may identify for which specific unit a null hypothesis of interest may be considered as violated while controlling the overall significance level of the multiple testing problem. This comes at the price of an increased computational burden.

1 Introduction

Within the last years, a number of problematic issues inherent in panel (unit root) testing techniques have been addressed in order to render their application more suitable for (macro-)econometric studies. Approaching the difficulty not to account for correlation across the units of the so-called first generation of panel unit root tests (cf. Levin et al. 2002; Im et al. 2003, for example), the second generation of tests featuring a variety of procedures (cf. Breitung/Das 2005; Moon/Perron 2004; Pesaran 2007, for example) has established itself in the panel unit root literature. What is more, modern panel techniques

[*] An earlier version of this paper was presented at the DIW Macroeconometric Workshop 2011 in Berlin.

[**] We thank the participants, and in particular In Choi, Horst Entorf, Christoph Hanck, Gebhard Kirchgässner, Benoit Perron, Peter Winker and Michael Wolf, and last but not least three anonymous referees for many helpful comments. This work further profited from Uwe Hassler's activities as a Research Associate at ZEW, Mannheim. Support through the project HA 3306/5-1 financed by the German Research Foundation (DFG) is gratefully acknowledged.

have become considerably more flexible with regard to individual specifics, e.g. they allow for a different lag length for each unit in the panel as well as for the estimation of cross-section specific slope parameters and intercepts. One example of these approaches is the transformation and combination of individual significance levels (p-values) to a joint test statistic. While these procedures are based on conducting individual-specific tests and hence can also be applied to unbalanced panels, they do not permit to make inference on the separate cross-section units. Hence, the user is still left in the dark about how many units and exactly which units can be assumed stationary when the joint unit root null is rejected.

A remedy to this issue is provided by so-called multiple testing procedures (MTP) that trace back to the Bonferroni method and the procedure by Holm (1979). Adjusting the overall significance level so as to account for the connectedness of the units in a panel, these methods do not only conduct separate tests but they also consider individual null hypotheses using each single p-value or test statistic without merging them into a joint test statistic. Thus, going beyond the 'skills' of p-value combination approaches, multiple testing procedures open up the possibility of making inference on the individual units in the panel without being at the risk of falsely rejecting true null hypotheses as one would be if conducting tests on the single times series without adequately adjusting the significance level. In particular, the joint resampling-based methods by Romano and Wolf (2005) and by Romano et al. (2008) do not even require the computation of p-values which makes them especially attractive when working with nonstandard distributions.

Within both approaches, some of the procedures are able to account for cross-correlation without further modification. In order to offer type I error control, they require, however, that particular (often worst case) assumptions on the dependence structure are fulfilled which often makes them conservative. Hence, it is also plausible that, generally, type I error control is not ensured for any dependence structure.

Other procedures necessitate a resampling approach in order to approximate the cross-sectional dependence structure. These methods are thus computationally more demanding but usually also more powerful than the former. The resampling approach needs to be chosen such that it matches the nature of the data, e.g. the dependence structure of the test statistics has to be preserved, nonstationarity has to be accounted for and so on.

As an empirical example, both types of testing procedures are employed for examining international interest rate linkage. As a necessary condition for the Uncovered Interest Parity to hold, the stability (or stationarity) of interest rate differentials has been an object of empirical research for quite some time. Due to the world-wide integration of macro-social forces such as economics and politics, it is also plausible to assume some kind of correlation between the countries. Hence, p-value combination procedures as described above qualify for conducting a meaningful analysis of whether the panel of interest rate differentials can be considered stationary. Multiple testing methods provide additional information on for how many and for which countries the interest rates can be deemed connected. The study reveals that, unsurprisingly, controlling for multiplicity can make a huge difference with respect to the number of null hypotheses rejected as spurious rejections are avoided.

For the procedures relying on resampling procedures, we employ a sieve bootstrap approach which, on the one hand conserves the dependence structure of the test statistics and on the other hand facilitates to account for nonstationarity of the data under the null.

We use interest rate data on 11 OECD countries including the US and Germany as 'Eurozone' representative in a time range from 1990.01 until 2012.10. Whereas a connection between US and OECD rates seems virtually nonexistent, fairly conclusive evidence in favor of the linkage to the Eurozone is found.

The structure of the paper is as follows. In the next section, the testing approaches used are introduced. In Section 3 some previous evidence on the connection of international interest rates is briefly reviewed and the empirical study is presented. Section 4 offers some conclusions.

2 Testing procedures

The tests for nonstationarity applied in this paper fall into two categories:[1] On the one hand, we discuss two procedures that (transform and) combine p-values obtained from individual test statistics to a joint test statistic. Such methods go back to the work by Fisher (1954) who is the first to suggest the transformation and combination of p-values p_i via the joint test statistic $F = -2\sum_{i=1}^{N}\ln(p_i)$ that follows a $\chi^2(2N)$ under H_0 if the units are cross-sectionally independent. Fisher's method is introduced into the panel literature by Maddala and Wu (1999) under independence; furthermore, they provide Monte Carlo evidence that a bootstrap version of the procedure works well in cross-correlated panels. Also Hanck (2009a) conducts a simulation study involving a bootstrap technique[2] which corroborates the findings by Maddala and Wu (1999) and adds results on the test's behavior under various scenarios. In Section 2.1, we discuss two more p-value combining techniques.

On the other hand, three multiple testing procedures are covered. Such procedures are more informative (and thus more useful for the empirical study at hand) than the p-value combinations as they allow for inference about each individual unit. Up to now, these methods have mostly been ignored in econometric applications which may be due to their reputation of being less powerful than so-called *per comparison* procedures that test each null individually at level α. However, the latter rather intuitive approach neglects the multiplicity of the testing problem which leads to an inflated number of false rejections of the individual unit root null hypotheses. In order to provide an illustration of this problem, we use the following textbook example:[3] Consider a panel of interest rate differentials of, for instance, $N = 10$ countries which are independent. Note that the independence assumption is only made to simplify the following computations. The unit root null is supposed to be true for each of the cross-section units. As $1/10 = 0.1 = 10\%$, one might think that, when testing each country separately, the probability of falsely rejecting at most one of the 10 nulls at $\alpha = 0.1$ is also 10%. Yet, from a statistical point of view, this reasoning is faulty as the event of a false rejection is a Bernoulli random variable with

[1] For a review on recent nonstationary panel techniques the reader is referred to Breitung and Pesaran (2008) or Kirchgässner et al. (2013: Ch. 7).

[2] In contrast to Maddala and Wu (1999) who apply Fisher's test to test for panel unit roots, Hanck (2009a) extends its use to panel cointegration.

[3] Cf. for example Kirchgässner et al. (2013: 257).

probability of success equal to 0.1. As a sum of N Bernoulli random variable is binomially distributed, the probability of at least one false rejection (or discovery) amounts to

$$\Pr\{\mathbf{V} \geq 1\} = \sum_{j=1}^{10} \binom{10}{j} 0.1^{j}(1-0.1)^{10-j} = 0.6513\,, \tag{2.1}$$

where $\mathbf{V}$ denotes the number of false rejections which is an unobserved random variable. Thus, when testing for interest rate linkage in the example panel, with probability 65.13% one tends to erroneously find evidence in favor of it. Obviously, adding more cross-sections to the panel even compounds this issue.

To eliminate this so-called 'multiplicity effect', testing procedures are called for that control the type I error rate for each individual hypothesis test at a pre-specified significance level. There exist several concepts of type I error control, two of which will be introduced in the following.

The probability of making at least one false discovery is called the *familywise error rate* (FWER). In the multiple testing methodology, one talks about controlling or protecting the FWER at significance level α when one ensures $\Pr\{\mathbf{V} \geq 1\} \leq \alpha$. However, holding down the number of false discoveries comes at the cost of a reduced ability to detect false nulls which has compromised the appeal of these techniques. If the number of individual null hypotheses (i.e. the number of cross-section units) is large, though, one might be willing to tolerate a larger rate of false rejections which, in turn, enhances the power of the multiple testing techniques. Also, the more rejections occur, the less relevant is the portion of false discoveries among them. Hence, Benjamini and Hochberg (1995) propose to control the *false discovery rate* (FDR):

$$\mathrm{FDR} = \mathrm{E}\left[\mathbf{Q}\right] \quad \text{with} \quad \mathbf{Q} = \begin{cases} \mathbf{V}/\mathbf{R} & \text{if} \quad \mathbf{R} > 0 \\ 0 & \text{if} \quad \mathbf{R} = 0 \end{cases}, \tag{2.2}$$

where $\mathbf{R}$ is the (observable) total number of rejections. Control of the FDR at α is achieved if FDR $\leq \alpha$. Another advantage of FDR control is that error rates using the proportion of false discoveries among all rejections stay stable with increasing N; obviously, this is not the case for error rates considering the absolute number of false discoveries like the FWER (Dudoit/van der Laan 2007: 145). Yet, it should be made clear that procedures controlling the FDR bound merely the average value of the proportion of false positives denoted by $\mathbf{Q}$ in (2.2). Hence, controlling the FDR causes difficulties for the interpretation of $\mathbf{Q}$ if the convergence rate of the proportion of false positives to the FDR is rather slow (Roquain 2011). For example, Romano and Wolf (2007) suggest a procedure controlling the upper-tail distribution of $\mathbf{Q}$ for a pre-specified proportion of false discoveries within the set of rejected null hypotheses. However, Roquain (2011) also remarks that FDR control remains applicable as it is a "simpler criterion for which the controlling methodology is (for now) much more developed" (Roquain 2011: 32).

Classical FWER controlling techniques are the Bonferroni method and the procedure by Holm (1979) both of which belong to the class of so-called *marginal multiple testing procedures*. This means they solely consider the marginal distributions of the test statistics and are thus considerably more conservative than procedures taking into account their true dependence structure (*joint multiple testing procedures*).[4] In Section 2.2, we review

[4] Cf. Dudoit and van der Laan (2007).

the FWER controlling marginal MTP by Hommel (1988), the FWER controlling joint bootstrap-based MTP by Romano and Wolf (2005) as well as the FDR controlling joint bootstrap-based MTP by Romano et al. (2008). The FDR controlling joint resampling-based MTP has also been applied by Deckers and Hanck (2013).

2.1 *p*-value combinations

The two procedures presented in this section can be used with any (unit root) test for which *p*-values p_i, $i = 1, \ldots, N$, are available which makes them more flexible than original panel techniques like the method by Pesaran (2007) employed in the working paper by Hassler and Werkmann (2012). Moreover, they can be applied to unbalanced panels without modification, and their use of standard distributions is a further advantage.[5]

In this paper, we use the common augmented Dickey-Fuller (ADF) test[6]. Consider N equations containing each T observations over time for each unit generated by an autoregressive model. In the simplest modeling framework each of these sequences is given by

$$\Delta y_{i,t} = \mu_i + \beta_i y_{i,t-1} + \sum_{j=1}^{k_i} g_{i,j} \Delta y_{i,t-j} + \varepsilon_{i,t}\,, \quad i = 1, \ldots, N\,, \tag{2.3}$$

where k_i denotes the number of lagged changes included in the regression, which may vary with the units. Furthermore, $y_{i,0}$ is given, $\beta_i \leq 0$ and the errors $\varepsilon_{i,t}$ are identically, independently distributed (iid) across t with $E(\varepsilon_{i,t}) = 0$, $E(\varepsilon_{i,t}^2) = \sigma_i^2 < \infty$ and $E(\varepsilon_{i,t}^4) < \infty$. The individual null hypothesis is as usual $H_{0,i}$: $\beta_i = 0$, and it is tested with the *t*-ratio on β_i. The null hypothesis that all the units have a unit root is given by the intersection:

$$H_0 = \bigcap_{i=1,\ldots,N} H_{0,i}\,.$$

Rejection of H_0 does not imply that all units are stationary, and in particular, a rejection does not identify the stationary units.

Choi (2001) suggests to apply the inverse normal method to independent panels which is modified by Hartung (1999) to account for cross-correlation (see also Demetrescu et al. 2006). With the standard normal distribution function $\Phi(\cdot)$, it involves the transformation and combination of *p*-values computed for N statistics. Hartung's idea comprises robustifying the inverse normal method against correlation under normality. He assumes constant correlation across the cross-section units:

$$\begin{aligned} Cov(\tau_i, \tau_j) &= \rho, \quad \text{for } i \neq j, \\ Cov(\tau_i, \tau_j) &= 1, \quad \text{for } i = j, \quad i, j = 1, \ldots, N\,, \end{aligned}$$

[5] Cf. Werkmann (2013) for a simulation study investigating the performance of *p*-value combinations in an unbalanced panels setting.

[6] An alternative choice would be the Phillips-Perron test which, however, is shown to be rather conservative in some cases (see Hanck 2013).

where the $\Phi^{-1}(p_i) =: \tau_i$ label the standard normal probits $\tau_1, \ldots, \tau_N$ and where ρ is consistently estimated by $\widehat{\rho}^* = \max(-1/(N-1), \widehat{\rho})$ with

$$\widehat{\rho} = 1 - \frac{1}{N-1} \sum_{i=1}^{N} \left(\tau_i - N^{-1} \sum_{i=1}^{N} \tau_i \right)^2 .$$

The combined test statistic is given by

$$H = \frac{\sum_{i=1}^{N} \Phi^{-1}(p_i)}{\sqrt{N + N(N-1)\left[\widehat{\rho}^* + \kappa \sqrt{\frac{2}{N+1}}(1 - \widehat{\rho}^*)\right]}} , \tag{2.4}$$

where the parameter κ is introduced by Hartung (1999) for regulating the small sample behavior of the test statistic. When applying the procedure to the interest rate differentials in the study at hand, κ is 0.2 according to Hartung (1999). The test statistic by Choi (2001) under independence is obtained for $\widehat{\rho}^* = 0$ and $\kappa = 0$. If $\widehat{\rho}^*$ is consistent as $N \longrightarrow \infty$ and a multivariate normal distribution of $\tau_1, \ldots, \tau_N$ can be assumed, $H \xrightarrow{d} \mathcal{N}(0,1)$ under the null hypothesis.

Examining the approach by Hartung (1999), Demetrescu et al. (2006) contribute three aspects: First, they reveal that the test is robust to a wider class of correlation patterns allowing for random correlation models, and they demonstrate experimentally that the normal approximation of H is a valid guideline in finite samples without constant correlation. Second, they provide a necessary and sufficient condition for limiting normality of H and third, they perform a simulation study in which they show that the modified inverse normal method can be sensibly applied to ADF tests in the presence of cross-correlation, even for small N.

Modifying the Bonferroni inequality, Simes (1986) proposes the following procedure (S test):

1. Compute the p-values p_i, $i = 1, \ldots, N$ and arrange them from smallest to largest such that $p_{(1)} \leq \cdots \leq p_{(N)}$.

2. Reject H_0 at α if and only if there is at least one sufficiently small p-value $p_{(j)}$, such that

$$p_{(j)} \leq j \cdot \alpha / N \quad \text{for some} \quad j = 1, \ldots, N , \tag{2.5}$$

where the p_i are the p-values corresponding to the test statistics $t_1, \ldots, t_N$. The adjusted significance level $j \cdot \alpha/N =: \alpha_j^{[s]}$ is also referred to as the cut-off value for the Simes test. This inequality constrains the overall significance level α such that the probability of rejecting at least one hypothesis in case that all are true does not exceed α, i.e. it controls the FWER in the sense of (2.1) at significance level α under independence (the case of independence is addressed below). In such a case, i.e. if FWER $\leq \alpha$ for N true nulls, one talks about controlling the FWER in the weak sense. Of course, it is more desirable that a method controls the FWER for any possible configuration of true and false nulls which is known as FWER control in the strong sense. Intuitively, Simes' testing procedure seems to be more informative regarding the (non)stationarity of the individual units than the inverse normal method as it examines separate p-values instead of combining them to a joint test statistic; however, the S test does not protect the FWER in the strong sense which renders test decisions with respect to the individual null hypotheses invalid. For curing this drawback, Hommel (1988) proposes a multiple testing procedure based on Simes' inequality which makes it possible to detect the stationary units while preserving (strong) FWER control at α (see Section 2.2).

The global Simes test relies on the following test statistic

$$S = \sum_{j=1}^{N} 1_{\{p_{(j)} \leq \alpha_j^{[s]}\}} \tag{2.6}$$

with $1_{\{A\}}$ being the indicator of event A. The decision rule then reads: Reject the joint null if $S > 0$, i.e. if at least one p-value is smaller than the adjusted significance level $\alpha_j^{[s]}$. Simes (1986) argues that the (global) S test achieves FWER control at α for independent test statistics and uniformly distributed p-values on the interval $[0, 1]$. Naturally, in practice the assumption of independent units is often implausible. Sarkar and Chang (1997) and Sarkar (1998), however, prove that this restriction can be relaxed considerably while type I error control is maintained.[7]

The combination of p-values has recently been shown to perform favorably relative to competing panel unit root tests by Hanck (2013).

2.2 Multiple testing

As mentioned above, the Simes test is a global testing procedure, i.e. it is not suitable for testing the N null hypotheses individually. Hommel (1988: 384) argues that when Simes' method is used for conducting individual hypothesis tests, relying on (2.5) may yield an overrejection of true null hypotheses in certain cases, i.e. type I error control is not preserved for any combination of true and nontrue nulls in a panel. Thus, based on the Simes inequality, Hommel (1988) suggests a multiple testing procedure[8] which allows for detecting the number as well as the identity of the significant null hypotheses in a panel. Briefly, the algorithm can be summarized as follows:

1. If $p_{(N)} \leq \alpha$, where $p_{(N)}$ is the largest p-value obtained from the test statistic on β_i in (2.3), reject all $H_{0,i}$ and terminate the algorithm.
2. Otherwise, compute

$$j^{[h]} = \max\left\{i = 1, \ldots, N : p_{(N-i+k)} > k\alpha/i \quad \text{for} \quad k = 1, \ldots, i\right\}, \tag{2.7}$$

and reject all $H_{0,i}$ with $p_i \leq \alpha/j^{[h]}$ where $\alpha/j^{[h]} =: \alpha^{[h]}$.

[7] More precisely, Sarkar (1998) shows that the test statistics have to be multivariate totally positive of order 2 (MTP$_2$). This condition is fulfilled for a rather large class of distributions, e.g. multivariate normal distribution with nonnegative correlations (see Sarkar 1998, for a detailed definition of the MTP$_2$ property).

[8] Also Hochberg (1988) suggests a multiple testing procedure built on Simes' inequality. While the Hochberg algorithm is somewhat less complicated and hence more popular with practitioners, it is less powerful than Hommel's method.

This procedure can also be written down as a so-called *step-up* procedure which starts with considering the least significant null (i.e. the one with the largest p-value). If a null is not rejected, the procedure 'steps up' to the next smaller p-value until a particular null hypothesis is rejected. Then, the remaining nulls are implicity rejected as well.[9] Note that, especially for small T, this procedure may underestimate the true fraction of stationary units, cf. Hanck (2013).

In the next paragraphs, two joint multiple testing procedures are introduced. In contrast to marginal multiple testing procedures, which often assume independence or a 'worst case' dependence structure, these methods use resampling (here: bootstrap) techniques in order to approximate the true dependence structure of the test statistics and are therefore in general less conservative than the former. Moreover, the methods covered in the following directly use the test statistics t_i which saves the user the computation of p-values. Therefore, unlike the joint significance procedures in Section 2.1 and Hommel's method, they remain feasible even if one cannot obtain p-values for the underlying test. Romano and Wolf (2005) recommend to employ studentized test statistics for several reasons; principally, however, they argue that if the standard deviations of the test statistics are not the same, one cannot readily compare basic test statistics to one another. Thus, for both procedures, $t_i = \widehat{\beta}_i / s.e.(\widehat{\beta}_i)$, $i = 1, \ldots, N$. Furthermore, both methods are *step-down* procedures, i.e. both consider the most significant (i.e. for a Dickey-Fuller test the most negative) test statistic first and if the corresponding null is rejected they 'step down' to the next larger test statistic; the algorithm continues in that way until no further null is rejected. However, for both procedures assume without loss of generality that a particular null hypothesis is rejected for large positive values of the corresponding test statistic, i.e. assume right-tailed hypothesis tests. We adopt this convention by multiplying the test statistics t_i of the original (left-tailed) ADF test by (-1). Hence, for both methods the test statistics have to be arranged in descending order where the test statistic with the largest value corresponds to the most significant null hypothesis.

The decision rule for the FWER controlling testing procedure by Romano and Wolf (2005) (RW procedure) is the following: Denote the ordered test statistics by $t_{r_1} \geq \cdots \geq t_{r_N}$ where $r_1, \ldots, r_N$ denote the index of the ordered test statistics. Reject the $\Delta R_1 = R_1 - R_0$ (with $R_0 = 0$) most significant null hypotheses $H_{0,r_{n_1}}$, $n_1 = 1, \ldots, R_1$, if it holds for the associated test statistic $t_{r_{n_1}}$ that $t_{r_{n_1}} > d_1$ where d_1 is a critical value chosen in such a way that a joint asymptotic coverage probability of $1 - \alpha$ of the following joint rectangular confidence region is guaranteed:

$$\left[\widehat{\beta}_{r_1} - s.e.(\widehat{\beta}_{r_1}) \cdot d_1,\ \infty\right) \times \cdots \times \left[\widehat{\beta}_{r_N} - s.e.(\widehat{\beta}_{r_N}) \cdot d_1,\ \infty\right) . \tag{2.8}$$

Obviously, an equivalent decision rule is to reject a particular null $H_{0,r_{n_1}}$ if $0 \notin \left[t_{r_{n_1}} - d_1,\ \infty\right)$. If no null hypothesis is rejected, the algorithm stops; otherwise, construct a second confidence region for the $N - R_1$ remaining nulls with labels $r_{R_1+1}, \ldots, r_N$:

$$\left[\widehat{\beta}_{r_{R_1+1}} - s.e.(\widehat{\beta}_{r_{R_1+1}}) \cdot d_2,\ \infty\right) \times \cdots \times \left[\widehat{\beta}_{r_N} - s.e.(\widehat{\beta}_{r_N}) \cdot d_2,\ \infty\right) , \tag{2.9}$$

and reject the next $\Delta R_2 = R_2 - R_1$ nulls $H_{0,r_{n_2}}$, $n_2 = R_1 + 1, \ldots, R_2$ if $t_{r_{n_2}} > d_2$. Repeat this procedure until no further hypotheses are rejected. Following Romano and Wolf (2005), the critical value d_k (where k denotes the number of steps until the algorithm

[9] For a presentation of the algorithm as a step-wise decision rule see Dmitrienko et al. (2010).

stops) should ideally be computed as the $1-\alpha$ quantile of the sampling distribution of the maximum of $\left(\frac{\widehat{\beta}_{r_i}}{s.e.(\widehat{\beta}_{r_i})}\right)$:

$$d_k := d_k(1-\alpha, P) = \inf\left\{x : \Pr\left\{\max_{R_{k-1}+1\le i\le N}\left(\frac{\widehat{\beta}_{r_i}}{s.e.(\widehat{\beta}_{r_i})}\right)\le x\right\}\ge 1-\alpha\right\},$$

where P is the true probability mechanism and $\Pr(A)$ is the probability of event A given P. As P is not known in practice, it is infeasible to compute the ideal critical value. Hence, P and thereby the critical values d_k have to be approximated by means of a bootstrap technique which can frequently be done consistently while preserving FWER control (Romano/Wolf 2005, Theorems 3.1 and 4.1). The estimated probability mechanism and critical values are then labelled $\widehat{P}$ and $\widehat{d}_k$, respectively.

As before, for the procedure by Romano et al. (2008) (RSW procedure) the test statistics are arranged in ascending order such that $t_{(1)} \le \ldots \le t_{(N)}$ where $t_{(N)}$ is the largest test statistic which now corresponds to the most significant null hypothesis $H_{0,(N)}$. According to Romano et al. (2008), for any step-down procedure the FDR is given as

$$\begin{aligned}\mathrm{FDR} = \mathrm{E}\left[\frac{\mathbf{V}}{\max\{\mathbf{R},1\}}\right] &= \sum_{1\le r\le N}\frac{1}{r}\mathrm{E}\left[\mathbf{V}|\mathbf{R}=r\right]P\{\mathbf{R}=r\} \qquad (2.10)\\ &= \sum_{1\le r\le N}\frac{1}{r}\mathrm{E}\left[\mathbf{V}|\mathbf{R}=r\right]\\ &\quad\times P\{t_{(N)}\ge c_N,\ldots,t_{(N-r+1)}\ge c_{N-r+1},t_{(N-r)}<c_{N-r}\},\end{aligned}$$

with $\mathbf{V}$, $\mathbf{R}$ defined as stated above and r is the realization of $\mathbf{R}$, i.e. the number of rejected null hypotheses. As before, P denotes the true probability measure. The expression for the FDR in (2.10) hinges on N_0, the number of true null hypotheses. As N_0 is unknown, (2.10) should be bounded from above by α for every combination of true and false null hypotheses. Romano et al. (2008) make use of this condition to compute the critical values $c_1, \ldots, c_N$ through recursion. If, e.g. the number of true nulls is $N_0 = 1$, (2.10) is reduced to

$$\mathrm{FDR} = \frac{1}{N}P\{t_{1:1}\ge c_1\},$$

where $t_{r:N_0}$ labels the r-th largest test statistic of the N_0 true nulls. This leads to the determination of the first critical value as

$$c_1 = \inf\left\{x\in\mathbb{R}\colon \frac{1}{N}P\{t_{1:1}\ge x\}\le\alpha\right\},$$

with $c_1 = -\infty$ if $N\alpha > 1$. The remaining critical values c_l, $l = 2, \ldots, N$, are computed via additional steps of the recursion for $N_0 = l$. As in the procedure by Romano and Wolf (2005), the true probability distribution P is not known in practice and hence, the critical values from above are not readily obtainable. Rather, a resampling technique is called for that allows for approximating the true probability measure via a bootstrap measure $\widehat{P}$. Romano et al. (2008) point out that a valid construction of $\widehat{P}$ has to be conducted in such a fashion that, whenever $H_{0,j}$ is true, the associated test statistic t_j is well approximated by its bootstrapped counterpart t_j^*. The associated critical values computed under $\widehat{P}$ are then denoted by $\widehat{c}_j$.

Of course, the choice of the bootstrap approach depends on the data. In the present case of testing for international interest rate linkage, the appropriate bootstrap technique needs to satisfy two requirements: First, the unit roots (i.e. nonstationary data) under the null hypothesis have to be taken into account and second, the dependence structure of the countries has to be conserved such that meaningful inferences regarding the (non)stationarity of the data can be made. In line with Deckers and Hanck (2013), we employ the sieve bootstrap to address the issues quoted above:[10] First, the bootstrap technique generates asymptotically valid bootstrap augmented Dickey-Fuller test statistics as shown by Chang and Park (2003) and Swensen (2003), i.e. when employing the α quantile of the empirical distribution of bootstrap ADF test statistics as a critical value for testing, one asymptotically obtains a procedure with size α. As the minimum mapping is continuous, this also holds true for the empirical distribution of the minimum of the bootstrap test statistics. Second, the empirical dependence structure between the units is conserved by resampling entire residual vectors at one point in time t.

Hanck (2009a) provides experimental evidence that the sieve bootstrap is also able to robustify a test against cross-unit cointegration in the sense that the size is controlled at α and that the power is consistent compared to the case where cross-unit cointegration is absent. An alternative to the sieve bootstrap would be the block bootstrap as employed by Moon and Perron (2012).

Note that, apart from the two multiple testing procedures by Romano and Wolf (2005) and Romano et al. (2008), also the Fisher-type test introduced in Section 2.1 is used as a bootstrap version to make it robust against cross-correlation.[11]

3 Empirical results

3.1 International interest rate linkage revisited

Stability of international interest rate differentials is often considered as a necessary condition for the uncovered interest parity (UIP) to hold. This is one reason why the international interest rate linkage has been a sphere of interest in empirical research for a long time. A lot of studies using interest rates with different maturities have been performed. Some of the typically mixed evidence is reviewed in the following.

A survey on short-term interest rates is conducted by Cumby and Mishkin (1986) with the general outcome that the evidence is neither favorable for all interest rate spreads being stationary between the US and 7 OECD countries nor to the hypothesis that there is no linkage at all.

Alexius (2001) argues that "before elevating the empirical failure of UIP to a stylized fact, *long[-term]* interest rates should also be studied". Thus, she examines quarterly long-term bond yields for 13 OECD countries and the US in a sample period beginning in 1957.Q1 to 1997.Q4. Her conclusion is that although being "tentative" due to data problems[12],

[10] A description of the bootstrap procedure can be found in the Appendix Section B.

[11] Cf. Hanck (2009a).

[12] Alexius (2001) points out two sources of possible measurement errors inherent in the data: 1) Imperfect information about the maturities of the bonds and 2) coupon payments on long-term government bonds.

the results obtained may hint at the fact that UIP is a valid concept in the analysis of long-term interest rates. Taking up this notion, Chinn and Meredith (2004) perform a study on long and short-term interest rates using monthly and quarterly data for the G-7 countries from 1980 to 2000. They confirm the absence of an interest rate linkage over short horizons as well as the finding by Alexius (2001) that the performance is more favorable over long horizons.

Considering unit root and cointegration testing, Kirchgässner and Wolters (1995) provide evidence on a strong, while not total, German influence on the development in other European countries using 3 month market rates for 7 European countries and the US for periods from 1974 to 1994. Wolters (2002) and Brüggemann and Lütkepohl (2005) focus on just two economies, the US and the Eurozone represented by Germany. Brüggemann and Lütkepohl (2005) obtain results being supportive of a stable interest rate linkage in long-term interest rates using monthly 10 year bond rates for the period 1985.01 to 2004.12. In contrast, Wolters (2002) finds no such evidence in a similar study based on monthly data from 1994 to 2001; he argues that this may be due to an insufficient sample size.

In the following section the results of the present empirical study are presented.

3.2 Joint significance results

International interest rate linkage is investigated using monthly data for short-term, medium-term, and long-term interest rates for eleven OECD countries, i.e. government bond yields with 3 months, 5 years and 10 years to maturity. The countries considered are Australia (au), Canada (ca), Denmark (dk), Germany (de), Japan (jp), New Zealand (nz), Norway (no), Sweden (se), Switzerland (ch), United Kingdom (uk) and the United States (us). The sample starts in 1990.01 and ends in 2012.10 ($T = 274$ observations). More detailed information on the data used can be found in the appendix. All test statistics and p-values are computed for augmented Dickey-Fuller (ADF) regressions for which the lag lengths are selected according to the modified Akaike information criterion (MAIC) with a maximum lag length of 15 chosen with the data-dependent rule $[12 \cdot (T/100)^{0.25}]$. The regressions contain intercepts but no linear trends. As stationarity of international interest rate differentials is a necessary condition for the uncovered interest rate parity (UIP) to hold true, we apply the tests discussed in the previous section to the panels of differentials between 10 OECD countries and Germany and the US, respectively ($N = 10$). Our choice of Germany as the reference country representing the Eurozone is motivated by the fact that databases, e.g. *Global Financial Data* (GFD), also use German interest rates as a proxy for Euro interest rate data. All p-values presented are according to MacKinnon (1996).

When investigating international interest rate linkages, Kirchgässner and Wolters (1995) found that Germany seems to influence the development of other countries in the short-run. The US, however, seem to have "at best a very weak direct influence" (Kirchgässner/Wolters 1995, Abstract).

First, consider the outcome obtained by the p-value combination procedures (see Tables 1 and 2). The bootstrap version of the Fisher-type panel test finds very conclusive evidence of interest rate linkage at all conventional significance levels as well as for all bond maturities for both, the panel of German differentials and the panel of US differentials. Hartung's modification of the inverse normal method, however, confirms this unambiguous result for the German panel only; regarding the US panel, Hartung's test concludes that some evidence for stationary differentials can be found for medium and long-term yields whereas for a short horizon, the OECD interest rates seem to be unattached to the US rates. In view of these partly conflicting results, which test decision should we trust? To begin with, keep in mind that although both procedures use individual p-values, they are not considered separately when the test decision is made; they are rather combined to one joint test statistic, i.e. both procedures might not be able to reject even though there are stationary units in the panel. Hence, it needs to be contemplated under which circumstances the procedures are able to detect sufficiently strong evidence of stationarity for rejecting the joint null: From the p-value combination rules by Fisher (1954) and by Hartung (1999) (see Section 2), one can infer that both testing procedures are most powerful when there are many sufficiently small p-values, i.e. many false null hypothe-

Table 1 United States, p-value combinations

	value of test statistic	**p-value**	**critical value**		
			1%	**5%**	**10%**
10 year bonds					
Fisher	48.914	0.000	41.810	31.797	27.374
Hartung	–1.551	0.060	–2.326	–1.645	–1.282
5 year bonds					
Fisher	46.143	0.005	42.883	33.507	28.797
Hartung	–1.931	0.027	–2.326	–1.645	–1.282
3 month bonds					
Fisher	38.232	0.010	36.856	29.079	26.260
Hartung	–1.203	0.115	–2.326	–1.645	–1.282

Notes: For the procedure by Hartung (1999) p-values and critical values are taken from the standard normal distribution. For the Fisher-type test p-values and critical values are obtained from the sieve bootstrap. Shading in gray refers to critical values for which the corresponding joint unit root null is rejected.

Table 2 Germany, p-Value Combinations

	value of test statistic	**p-value**	**critical value**		
			1%	**5%**	**10%**
10 year bonds					
Fisher	58.001	0.000	33.803	28.884	26.583
Hartung	–2.659	0.004	–2.326	–1.645	–1.282
5 year bonds					
Fisher	59.678	0.000	35.632	28.805	26.249
Hartung	–4.802	0.000	–2.326	–1.645	–1.282
3 month bonds					
Fisher	78.254	0.000	36.841	29.981	26.053
Hartung	–3.064	0.001	–2.326	–1.645	–1.282

Notes: See Table 1.

ses. However, Littell and Folks (1971) establish that the Fisher-type test is asymptotically more powerful than the inverse normal method: One situation fulfilling these requirements is the test of a simple null hypothesis which coincides with the unit root null when using the Dickey-Fuller test. Also, in a recent Monte Carlo study by Werkmann (2013), it is shown that, when the underlying test is the ADF test, the Fisher-type test is more powerful than the test by Hartung (1999) for almost every configuration of true and false nulls considered.

Another argument in favor of the Fisher-type procedure would be the fact that the true dependence structure of the test statistics is approximated via a bootstrap estimate. In contrast to Hartung's method, this approach works for any type of cross-sectional dependence, hence it is more exact with respect to the true correlation pattern than the modified inverse normal method when the true dependence structure is unknown.

3.3 Multiple comparison results

In Tables 3 and 4, multiple comparison results at the 5% significance level are displayed. For results at the 1% and the 10% level the reader is referred to the online appendix at www.jbnst.de/en.

Although the cut-off values for the Simes test are also collected in the tables presenting the outcome of the multiple testing procedures, keep in mind that for this procedure one may not reject the individual hypotheses for which the corresponding cut-off values are highlighted if FWER control is to be maintained. The Simes procedure tests the global null exclusively, i.e. if one or more cut-off(s) are shaded in gray, one can merely conclude that the joint (unit root) null is rejected. Based on Simes' inequality, the step-up procedure by Hommel (1988) is able to indicate which of the individual differentials can be deemed stationary. Hence, the results of these two methods are connected and will be reported together. For the panel of US differentials, neither the Simes test nor Hommel's method finds any evidence of interest rate linkage to the US.

Considering the German panel, the Simes test rejects the panel unit root null only for short-term rates at the 1% level; at 5% and 10%, the joint null can be rejected for all maturities examined. The outcome of Hommel's procedure, however, qualifies the Simes result considerably: Only one to three (out of ten) individually stationary differentials can be detected at these significance levels.

For both resampling-based procedures note that we report the values for the original ADF test statistics in the tables, not those multiplied by -1. Having a look at the results for these two methods, it is again obvious that there is at best very weak evidence in favor of interest rate linkage between the ten OECD countries and the US: Solely for short-run rates and at the 10% level both methods are able to reject the individual unit root null for one country and two countries, respectively.

Regarding interest linkage to Germany (i.e. Eurozone), the RW procedure finds favorable evidence at the 10% level for long as well as medium and short maturity horizons. This result is corroborated by the RSW procedure at both the 5% and the 10% level. Clearly, the RSW procedure rejects the null hypotheses also deemed significant by the RW procedure and in several cases, the former rejects more nulls than the latter. This is due to the fact that multiple testing procedures controlling the FDR (as the RSW procedure does) are able to detect more false null hypotheses, i.e. they are more powerful than the ones controlling the FWER (see Section 2); however, this benefit comes at the cost of an

Table 3 United States, Multiple Testing Procedures, 5% Significance Level

	t_j	p_j	$\alpha_j^{[s]}$	$\alpha^{[h]}$	$\widehat{d}_k$	$\widehat{c}_j$
10 year bonds						
uk	−3.186	0.022	0.005	0.0050	−3.426	−3.429
nz	−3.161	0.023	0.010	0.0050		−3.220
au	−2.977	0.038	0.015	0.0050		−3.065
dk	−2.728	0.071	0.020	0.0050		−2.923
se	−2.683	0.078	0.025	0.0050		−2.781
no	−2.655	0.083	0.030	0.0050		−2.669
ca	−2.441	0.131	0.035	0.0050		−2.563
de	−2.284	0.178	0.040	0.0050		−2.365
ch	−2.163	0.221	0.045	0.0050		−2.180
jp	−1.547	0.508	0.050	0.0050		−1.463
5 year bonds						
nz	−3.207	0.021	0.005	0.0050	−3.491	− 3.494
au	−3.101	0.028	0.010	0.0050		− 3.271
no	−2.854	0.052	0.015	0.0050		− 3.135
se	−2.768	0.064	0.020	0.0050		− 2.991
dk	−2.760	0.065	0.025	0.0050		− 2.821
ca	−2.753	0.067	0.030	0.0050		− 2.712
uk	−2.493	0.118	0.035	0.0050		− 2.532
de	−2.106	0.242	0.040	0.0050		− 2.389
ch	−1.544	0.510	0.045	0.0050		− 2.184
jp	−0.925	0.779	0.050	0.0050		− 1.477
3 month bonds						
nz	−3.418	0.011	0.005	0.0050	−3.538	−3.541
ca	−3.018	0.034	0.010	0.0050		−3.245
uk	−2.626	0.089	0.015	0.0050		−3.017
de	−2.296	0.174	0.020	0.0050		−2.869
au	−2.151	0.225	0.025	0.0050		−2.748
no	−2.141	0.229	0.030	0.0050		−2.588
jp	−2.022	0.277	0.035	0.0050		−2.434
dk	−1.895	0.335	0.040	0.0050		−2.317
se	−1.878	0.342	0.045	0.0050		−2.097
ch	−1.541	0.511	0.050	0.0050		−1.373

Notes: t_j is the augmented Dickey-Fuller test statistic and p_j is the corresponding p-value. $\alpha_j^{[s]}$ and $\alpha^{[h]}$ denote the cut-off values for the procedures by Simes (1986) and Hommel (1988), respectively. $\widehat{d}_k$ and $\widehat{c}_j$ label the critical values computed for the left-tailed versions of the methods by Romano and Wolf (2005) and Romano et al. (2008), respectively. Except for the Simes test, shading in gray refers to cut-off values or critical values for which the corresponding individual unit root null is rejected. Hommel's $j^{[h]} = 10$ for all maturities.

increased probability of false rejections. Thus, it depends on the user's conservativeness which method he or she is willing to use.

Note that these two procedures approximate the true dependence structure of the test statistics. Such resampling-based multiple testing procedures are hence in general more

Table 4 Germany, Multiple Testing Procedures, 5% Significance Level

	t_j	p_j	$\alpha_j^{[s]}$	$\alpha^{[h]}$	$\widehat{d}_k$	$\widehat{c}_j$
10 year bonds						
au	−4.137	0.001	0.005	0.0056	−3.501	−3.502
se	−3.403	0.012	0.010	0.0056	−3.475	−3.239
no	−3.281	0.017	0.015	0.0056		−3.033
ca	−3.025	0.034	0.020	0.0056		−2.924
nz	−2.964	0.040	0.025	0.0056		−2.797
uk	−2.364	0.153	0.030	0.0056		−2.641
us	−2.284	0.178	0.035	0.0056		−2.516
dk	−2.016	0.280	0.040	0.0056		−2.329
ch	−2.009	0.283	0.045	0.0056		−2.190
jp	−1.663	0.449	0.050	0.0056		−1.492
5 year bonds						
au	−3.838	0.003	0.005	0.0063	−3.377	−3.379
se	−3.628	0.006	0.010	0.0063	−3.377	−3.180
no	−3.547	0.007	0.015	0.0063	−3.377	−3.040
nz	−3.041	0.032	0.020	0.0063	−3.307	−2.897
ca	−2.955	0.041	0.025	0.0063		−2.765
uk	−2.934	0.043	0.030	0.0063		−2.658
ch	−2.219	0.200	0.035	0.0063		−2.518
us	−2.106	0.242	0.040	0.0063		−2.354
dk	−1.749	0.405	0.045	0.0063		−2.129
jp	−0.918	0.781	0.050	0.0063		−1.511
3 month bonds						
ca	−4.380	$<$ 0.001	0.005	0.0063	−3.673	−3.674
ch	−4.271	0.001	0.010	0.0063	−3.673	−3.339
uk	−3.545	0.008	0.015	0.0063	−3.556	−3.060
dk	−3.188	0.022	0.020	0.0063		−2.916
no	−3.046	0.032	0.025	0.0063		−2.762
se	−2.910	0.045	0.030	0.0063		−2.618
au	−2.750	0.067	0.035	0.0063		−2.393
nz	−2.685	0.078	0.040	0.0063		−2.283
us	−2.296	0.174	0.045	0.0063		−2.018
jp	−2.246	0.191	0.050	0.0063		−1.437

Notes: See Table 3. Hommel's $j^{[h]} = 9$ for 10 year bonds and $j^{[h]} = 8$ for 5 year and 3 month bonds.

powerful than marginal multiple testing procedures like Hommel's method which solely makes use of the marginal distribution of the test statistics.

Interestingly, except for the RW procedure, the MTP outcome indicates that interest rate linkage is stronger for medium to short horizons.

As expected, the application of multiple testing procedures leads to considerably fewer rejections of the unit root null than when testing each time series separately at an unadjusted significance level α. In particular for the US panel controlling for multiplicity makes a substantial difference: Using the per comparison approach, one would (possibly spuriously) reject two and three null hypotheses at 5% for long and medium-term rates; at 10% significance, it would even be possible to make six more, maybe false discoveries.

Also, we examined the interest rates with respect to possible cointegration relationships. For this purpose, we regress the OECD rates on the reference country and perform ADF tests on the unrestricted residuals. Without imposing the $(1, -1)$ restriction, we essentially find the same results as obtained by unit root testing: There is strong evidence in favor of a linkage to Germany (i.e. the Eurozone) while for the US almost no connection can be detected.

4 Conclusion

The study at hand has analyzed the interest rate relation between 10 OECD countries and Germany and the US using government bonds with different maturities in a sample period from 1990.01 to 2012.10 employing *p*-value combinations and multiple testing techniques. While the former provide information on the panel as a whole, the latter enable us to identify the individual countries for which the interest rates are connected while accounting for the multiplicity issue which is present when conducting individual tests for each country in the panel. The probability of spurious rejections is explicitly controlled (at a user-specified level).

For the US differentials mixed evidence is found. While the Fisher-type test reproduces the results obtained for German differentials, the test by Hartung (1999) finds conclusive evidence in favor of interest linkage only for medium-term maturities. The Simes test and the multiple testing procedures rule out stationarity of the interest rate differentials even more explicitly. The only exception from this result is the panel of short-term rate differentials for which the procedures by Romano and Wolf (2005) and by Romano et al. (2008), respectively, can reject the individual unit root null at the 10% level for New Zealand, and New Zealand and Canada, respectively.

As intended, multiplicity control achieves fewer – possibly less spurious – rejections of the unit root null than the per comparison approach.

Considering the results on German linkage, a different picture arises: Strong evidence in favor of stationarity is found. Both, tests relying on the computation of individual or combined significance levels and multiple testing procedures are indicative of stationary relations between Germany and the OECD countries studied. The evidence is weaker for bonds with long maturities. In particular, merely a weak linkage between the US and Germany can be found. Therefore, the favorable results on German linkage and the almost nonexistent linkage between the US and the other OECD countries does not result in a contradiction. This insight is gained from the multiple testing procedures since they enable the user to learn the exact identity of the differentials deemed (non)stationary.

Note that, prior to the introduction of the Euro, the European Exchange Rate Mechanism (ERM) as a part of the European Monetary System (EMS) interconnected the exchange rates of the EEC countries in order to preserve economic and social stability; more precisely, the exchange rates were forced to stay within a relatively tight band around the European Currency Unit (ECU) also known as the 'European Snake'. After the Euro was introduced, the currencies outside the Eurozone have been linked to it, again for the benefit of stability and to prepare them for their potential joining. The EU currencies currently included are the Danish krone, the Lithuanian litas and the Latvian lats; other countries such as Poland and Romania are expected to join the mechanism in the near future. However, Sweden and the United Kingdom (which left the EMS in 1992) have preferred not to participate in this mechanism. In our sample, this leaves the Danish krone as the only currency linked to the Euro by exchange rate agreements. Thus, the outcome on German or Eurozone linkage may, if at all, only slightly be affected by the ERM.

Appendix

A. Variables and data sources

In the study at hand, monthly data for the period 1990.01 to 2012.10 is used. Long and short-term interest rates are obtained from OECD.StatExtracts. Medium-term interest rates are taken from Global Financial Data (Series IGAUS5D, IGCAN5D, IGCHE5D, IGDEU5D, IGDNK5D, IGJPN5D, IGNOR5D, IGNZL5D, IGSWE5D, IGGBR5D, IGUSA5D). Monthly values are mostly arithmetic averages "relating to all days or specified days in the month". The rates labeled "10 year" are "long-term (in most cases 10 year) government bonds whose yield is is used as the representative 'interest rate' for the respective country". For the rates named "5 year" benchmark bonds are used. Global Financial Data states that "the benchmark bond is the bond that is closest to the stated maturity without exceeding it. When the government issues a new bond of the stated maturity, it replaces the bond used for the index to keep the maturity as close to the stated time period as possible.". The rates denoted as "3 month" are short-term rates and "usually either the three month interbank offer rate [. . .] or the rate associated with Treasury bills, Certificates of Deposit or comparable instruments, each of three month maturity". For more detailed information on the data, see the webpage stats.oecd.org or the GFD database, respectively.

B. The sieve bootstrap algorithm

The sieve bootstrap is employed for all procedures warranting a resampling procedure in order to approximate the dependence structure across the units in the panel. An equally adequate alternative would be the (overlapping) block bootstrap approach as used by Moon and Perron (2012). Steps (1) to (5) included of the sieve bootstrap are the same for the Fisher-type test as well as for the two joint resampling-based procedures by Romano and Wolf (2005) and Romano et al. (2008), respectively. From (6) on, the steps differ according to the testing procedure used.

According to Hanck (2009a) and Deckers and Hanck (2013), the bootstrap algorithm takes the following steps:

(1) Fit an AR process to $\Delta y_{i,t}$ where $y_{i,t}$ denotes the interest rate differential between country i and the US or Germany at month t, respectively. It is preferable to use the Yule-Walker method since it always yields an autoregression which is invertible. With $\overline{\Delta y_i} := (T-1)^{-1} \sum_{t=2}^{T} \Delta y_{i,t}$, compute the empirical autocovariances of $\Delta y_{i,t}$ up to order q which is chosen via the Schwert criterion $[4 \cdot (T/100)^{0.25}]$

$$\widehat{\gamma_i}(j) := \frac{1}{T-1-j} \sum_{t=2}^{T-j} (\Delta y_{i,t} - \overline{\Delta y_i})(\Delta y_{i,t+j} - \overline{\Delta y_i}), \quad i = 1, \ldots, N, \; j = 1, \ldots, q.$$

Let

$$\widehat{\Gamma}_i := \begin{pmatrix} \widehat{\gamma}_i(0) & \cdots & \widehat{\gamma}_i(q-1) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}_i(q-1) & \cdots & \widehat{\gamma}_i(0) \end{pmatrix}$$

and define $\widehat{\gamma}_i := (\widehat{\gamma}_i(1), \ldots, \widehat{\gamma}_i(q))'$ such that the autoregressive coefficient vector is obtained as

$$(\widehat{\phi}_{i,1}, \ldots, \widehat{\phi}_{i,q})' := \widehat{\Gamma}_{i,q}^{-1} \widehat{\gamma}_i \,, \quad i = 1, \ldots, N\,.$$

(2) The residuals are then defined as

$$\widehat{\varepsilon}_{i,t} := \Delta y_{i,t} - \sum_{l=1}^{q} \widehat{\phi}_{i,l} \Delta y_{i,t-l}\,.$$

Center the residuals

$$\widetilde{\varepsilon}_{i,t} := \widehat{\varepsilon}_{i,t} - \frac{1}{T-q-1} \sum_{g=q+2}^{T} \widehat{\varepsilon}_{i,g}\,.$$

(3) To obtain $\varepsilon^*_{i,t}$, resample nonparametrically from $\widetilde{\varepsilon}_{i,t}$. In order to maintain the empirical dependence structure across the units $i = 1, \ldots, N$, resample complete vectors

$$\widetilde{\varepsilon}_{\bullet,t} := (\widetilde{\varepsilon}_{1,t}, \ldots, \widetilde{\varepsilon}_{N,t})'\,.$$

(4) Then, construct the bootstrap samples recursively as

$$\Delta y^*_{i,t} = \sum_{l=1}^{q} \widehat{\phi}_{i,l} \Delta y^*_{i,t-l} + \varepsilon^*_{i,t}\,.$$

(5) Impose the null of a unit root by integrating $\Delta y^*_{i,t}$ to obtain $y^*_{i,t}$.

(6) Compute the individual test statistics $t^*_{b,i}$ for each bootstrap sample.

a) Fisher-type test

Then compute the corresponding p-values and the joint test statistic F^*_b.

b) Romano and Wolf (2005)

Compute

$$\min{}^*_{b,k} := \min_{R_{k-1}+1 \leq i \leq N} t^*_{b,r_i}\,.$$

(7) Repeat steps (3) to (6) B times.

(8) a) Romano and Wolf (2005)

Compute the critical value $\widehat{d}_k$ as the α - quantile of the values $\min^*_{1,k}, \ldots, \min^*_{B,k}$.

b) Romano et al. (2008)

Compute the critical values $\widehat{c}_1, \ldots, \widehat{c}_N$ by recursively solving the left-tailed analogue of equation (7) in the paper by Romano et al. (2008) for the case that $N_0 = N$.

(9) a) Fisher-type test

Register a rejection of the joint unit root null if

$$B^{-1} \sum_{b=1}^{B} 1_{\{F_b^* > F\}} < \alpha \,,$$

where $1_{\{\}}$ denotes the indicator function.

b) Romano and Wolf (2005)

Compare the test statistics t_j to the critical value $\widehat{d}_k$ and apply the RW step-down decision rule described in Section 2.2. If no null can be rejected, stop. Otherwise, follow the steps of the RW algorithm described in Section 2.2 and compute further critical values.

c) Romano et al. (2008)

Compare the test statistics t_j to the critical values $\widehat{c}_j$ and apply the RSW step-down decision rule described in Section 2.2.

References

Alexius, A. (2001), Uncovered Interest Parity Revisited. Review of International Economics 9: 505–517.

Benjamini, Y., Y. 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.

Breitung, J., S. Das (2005), Panel Unit Root Tests under Cross-Sectional Dependence. Statistica Neerlandica 59: 414–433.

Breitung, J., M.H. Pesaran (2008), Unit Roots and Cointegration in Panels. Pp. 279–322 in: L. Matyas, P. Sevestre (eds.), The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice. Dordrecht: Kluwer Academic Publishers, 3 ed.

Brüggemann, R., H. Lütkepohl (2005), Uncovered Interest Rate Parity and the Expectations Hypothesis of the Term Structure: Empirical Results for the U.S. and Europe. Applied Economics Quarterly 51: 143–154.

Chang, Y., J.Y. Park (2003), A Sieve Bootstrap for the Test of a Unit Root. Journal of Time Series Analysis 24: 379–400.

Chinn, M.D., G. Meredith (2004), Monetary Policy and Long-Horizon Uncovered. Interest Parity, IMF Staff Papers 51: 409–430.

Choi, I. (2001), Unit Root Tests for Panel Data. Journal of International Money and Finance 20: 249–272.

Cumby, R.E., F.S. Mishkin (1986), The International Linkage of Real Interest Rates: The European – U.S. Connection. Journal of International Money and Finance 5: 5–23.

Deckers, T., C. Hanck (2013), Multiple Testing for Output Convergence. Macroeconomic Dynamics, forthcoming, doi: 10.1017/S1365100512000338.

Demetrescu, M., U. Hassler, A.I. Tarcolea (2006), Combining Significance of Correlated Statistics with Application to Panel Data. Oxford Bulletin of Economics and Statistics 68: 647–663.

Dmitrienko, A., A.C. Tamhane, F. Bretz (eds.) (2010), Multiple Testing Problems in Pharmaceutical Statistics. Boca Raton: Chapman & Hall/CRC. Taylor and Francis Group.

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

Fisher, R.A. (1954), Statistical Methods for Research Workers. Berlin: Oliver & Bond, 12th ed.

Hanck, C. (2009a), Cross-sectional Correlation Robust Tests for Panel Cointegration. Journal of Applied Statistics 36: 817–833.

Hanck, C. (2013), An Intersection Test for Panel Unit Roots. Econometric Reviews 32: 183–203.

Hartung, J. (1999), A Note on Combining Dependent Tests of Significance. Biometrical Journal 41: 849–855.

Hassler, U., V. Werkmann (2012), New Panel Evidence on International Interest Rate Linkage. Available at SSRN: http://ssrn.com/abstract=2194831 or http://dx.doi.org/10.2139/ssrn.2194831.

Hochberg, Y. (1988), A Sharper Bonferroni Procedure for Multiple Tests of Significance Biometrika 75: 800–802.

Holm, S. (1979), A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6: 65–70.

Hommel, G. (1988), A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. Biometrika 75: 383–386.

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

Kirchgässner, G., J. Wolters (1995), Interest Rate Linkages in Europe Before and After the Introduction of the European Monetary System – Some Empirical Results. Empirical Economics 20: 435–454.

Kirchgässner, G., J. Wolters, U. Hassler (2013): Introduction to Modern Time Series Analysis. Berlin: Springer, 2nd edition.

Levin, A., C.-F. Lin, C.-S. J. Chu (2002), Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. Journal of Econometrics 108: 1–24.

Littell, R.C., J.L. Folks (1971), Asymptotic Optimality of Fisher's Method of Combining Independent Tests. Journal of the American Statistical Association 66: 802–806.

MacKinnon, J.G. (1996), Numerical Distribution Functions for Unit Root and Cointegration Tests Journal of Applied Econometrics 11: 601–618.

Maddala, G.S., S. Wu (1999), A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. Oxford Bulletin of Economics and Statistics 61: 631–652.

Moon, H.R., B. Perron (2004), Testing for a Unit Root in Panels with Dynamic Factors. Journal of Econometrics 122: 81–126.

Moon, H.R., B. Perron (2012), Beyond Panel Unit Root Tests: Using Multiple Testing to Determine the Nonstationarity Properties of Individual Series in a Panel. Journal of Econometrics 169: 29–33.

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

Romano, J.P., A.M. Shaikh, M. Wolf (2008), Control of the False Discovery Rate under Dependence Using the Bootstrap and Subsampling. Test 17: 417–442.

Romano, J.P., M. Wolf (2005), Stepwise Multiple Testing as Formalized Data Snooping. Econometrica 73: 1237–1282.

Romano, J.P., M. Wolf 2007), Control of Generalized Error Rates in Multiple Testing. The Annals of Statistics 35: 1378–1408.

Roquain, E. (2011), Type I Error Rate Control for Testing Many Hypotheses: A Survey with Proofs. Journal de la Société Française de Statistique 152: 3–38.

Sarkar, S.K. (1998), Some Probability Inequalities for Ordered MTP2 Random Variables: A Proof of the Simes Conjecture. The Annals of Statistics 26: 494–504.

Sarkar, S.K., C.-K. Chang (1997), The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics. Journal of the American Statistical Association 92: 1601–1608.

Simes, R.J. (1986), An Improved Bonferroni Procedure for Multiple Tests of Significance Biometrika 73: 751–754.

Swensen, A.R. (2003), Bootstrapping Unit Root Tests for Integrated Processes. METEOR Research Memorandum RM/11/003, working Paper.

Werkmann, V. (2013), Performance of Unit Root Tests in Unbalanced Panels: Experimental Evidence. Advances in Statistical Analysis 97: 271–285.

Wolters, J. (2002), Uncovered Interest Parity and the Expectation Hypothesis of the Term Structure: Empirical Results for the U.S. and Europe. Pp. 271–282 in: I. Klein, S. Mittnik (eds), Contributions to Modern Econometrics: From Data Analysis to Economic Policy. Honour of Gerd Hansen: Kluwer Academic Publishers.

Prof. Dr. Uwe Hassler, Statistik und Methoden der Ökonometrie, Goethe-Universität Frankfurt, RuW-Gebäude, Grüneburgplatz 1, 60323 Frankfurt, Germany.
hassler@wiwi.uni-frankfurt.de; http://www.wiwi.uni-frankfurt.de/~hassler

Corresponding author: Dr. Verena Werkmann, Goethe-Universität Frankfurt, RuW-Gebäude, Grüneburgplatz 1, 60323 Frankfurt, Germany.
werkmann@wiwi.uni-frankfurt.de