Combination of “Combinations of P-values”^{*}

Lan Cheng

Department of Mathematical Sciences, SUNY Fredonia

lan.cheng@fredonia.edu

and

Xuguang Sheng [†]

Department of Economics, American University

sheng@american.edu

Abstract

We investigate the impact of an uncertain number of false individual null hypotheses on commonly used p -value combination methods and find that the performance of these methods varies substantially under such uncertainty. These variations yield conflicting results in meta-analysis, motivating the development of a new, reconciling test. We consequently develop a combination of “combinations of p -values” (CCP) test that maintains good power properties across such uncertainty. We base the CCP test on a simple union of rejections decision rule that exploits the similarity between any two p -value combination methods. Monte Carlo simulations show that our test controls size and closely tracks the power of the best individual methods.

JEL Classification: C12; C33

Key words: Combination methods; Hypothesis testing; p value; Union of rejections.

^{*}This paper was presented at the Seaway Section of the Mathematical Association of America, North American Summer Meeting of the Econometric Society and American Economic Association Annual Meeting. We thank Badi Baltagi, Nancy Boynton, Bruce Hansen, Ulrich Müller, Eric Renault, Aman Ullah, Jonathan Wallen, Jonathan Wright and conference participants for helpful comments and suggestions. The usual disclaimer applies.

[†]Corresponding author. Contact information: 4400 Massachusetts Ave., N.W., Washington, DC 20016; Tel: (202) 885-3782; Fax: (202) 885-3790.

1 Introduction

Combining independent p -values from a set of hypothesis tests has long been used in meta-analysis. Given n individual hypotheses H_{0i} , $i = 1, 2, \dots, n$, consider a test for the joint null hypothesis, $H_0 = \bigcap_{i=1}^n H_{0i}$. H_0 is true if all the individual null hypotheses are true, but false if at least one of its components is false. Many attempts have been made to compare the size and power of different combination methods. See, for example, Hedges and Olkin (1985), Zaykin, et al. (2002), and Sheng and Yang (2012), among others. A key result of this literature is that no uniformly most powerful test exists. Some tests, such as Tippett's and Simes' methods, are powerful when there are very few false individual null hypotheses. In contrast, others, such as Fisher's and Stouffer's methods, are powerful when there are many false individual null hypotheses. Since whether few or many individual null hypotheses are false is *a priori* unknown in practice, researchers have difficulty selecting the optimal method to use, especially if the results from differing p -value combination methods conflict. To address this problem, we propose a combination of "combinations of p -values" (CCP) test that maintains good power properties across such uncertainty.

The CCP test is based on a simple union of rejections decision rule. According to the CCP test, if at least one of the two p -value combination methods yields a rejection at the significance level γ , then the joint null hypothesis is rejected at the significance level α . The value of γ is selected such that the CCP test (i) maintains the size of α , (ii) has robust, good power across the uncertain number of false individual null hypotheses. The union of rejections decision rule has been widely used in hypothesis testing across the fields of statistics and econometrics. Recent contributions to the decision rule include Neuhäuser (2003), Harvey, et al. (2009, 2011), Müller (2009) and Dumitru and Urga (2012). Neuhäuser (2003) introduced a combination of Simes' method and the truncated product method. Harvey, et al. (2009, 2011) used a similar strategy to combine two unit root tests. Müller (2009) further showed

that Harvey, et al. (2009)’s test comes close to exploiting the available information efficiently because the test has a power similar to that of the weighted average power maximizing test under uncertainty over the trend. Dumitru and Urga (2012) used the union of rejections across nine procedures and across sampling frequencies to minimize spurious jump detection in financial assets. However, as aptly noted by these authors, this simple decision rule often leads to size distortions. Our CCP test avoids the size distortion from separately applying p -value combination methods to the same data set, thus circumventing the problem of “conservative combination.” Our Monte Carlo simulations show that the power of the CCP test closely tracks that of the better individual method, irrespective of the number of false individual null hypotheses. As illustrated by one empirical example of testing purchasing power parity, the proposed test avoids arbitrarily choosing a method when the results from individual p -value combination methods conflict.

The rest of the paper is organized as follows. Section 2 develops the CCP test. Section 3 uses Monte Carlo simulations to explore the size and power of the CCP test. Section 4 applies the CCP test to an empirical example. Section 5 concludes. For further technical details, see the Appendix.

2 The combination of “combinations of p -values” test

Suppose that tests were conducted for n individual hypotheses H_{0i} , $i = 1, 2, \dots, n$. For each test the p -value p_i , $i = 1, 2, \dots, n$, was calculated. We consider the problem of testing the joint null hypothesis $H_0 = \bigcap_{i=1}^n H_{0i}$ at a specific significance level α when the p -values are independent. The combined alternative, denoted by H_A , is true if at least one of H_{0i} , $i = 1, 2, \dots, n$, is false. Various methods have been proposed in the literature to combine p -values. Here we briefly summarize four popular methods.

- Tippett (1931) suggested rejecting H_0 if $p_i \leq 1 - (1 - \alpha)^{1/n}$ for at least one $i =$

$1, 2, \dots, n$.

- Simes (1986) proposed that H_0 be rejected if $p_{(i)} \leq i\alpha/n$ for at least one $i = 1, 2, \dots, n$, where $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$ are the ordered p -values of $p_1, p_2, \dots, p_n$.
- Fisher (1932) noted that the test statistic

$$t = -2 \sum_{i=1}^n \ln(p_i) = -2 \ln \prod_{i=1}^n p_i$$

has a χ^2 distribution with $2n$ degrees of freedom. Thus, the p -value for testing H_0 is

$$p = 1 - F_{\chi^2}(t; 2n),$$

where $F_{\chi^2}(x; k)$ is the cumulative distribution function (c.d.f.) for a χ^2 distribution with k degrees of freedom.

- Stouffer's method, attributed to Stouffer et al. (1949), transforms the p values via the standard normal distribution. The test statistic is defined as

$$z = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi^{-1}(p_i),$$

where $\Phi(\cdot)$ is the standard normal c.d.f. When p_i , $i = 1, 2, \dots, n$, are independent, $z \sim N(0, 1)$. Therefore, the p -value for testing H_0 is

$$p = \Phi(z) = \Phi \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi^{-1}(p_i) \right).$$

Since H_0 may be false in numerous ways, the selection of a method becomes increasingly complicated. Consequently, we cannot generally expect the methods to be sensitive to all possible alternatives. As shown in the simulation results from Neuhausser (2003), Loughin

(2004) and Sheng and Yang (2012), Tippett's and Simes' methods are powerful when the total evidence against the joint null hypothesis is concentrated in very few of the combined p -values, but Fisher's and Stouffer's methods perform well when evidence against the joint null is more broadly spread among the combined p -values. Often one method rejects H_0 whereas the other does not, complicating the test decision.

2.1 Construction of the CCP test

To construct a robust test that is relatively powerful for all situations under the alternative hypothesis, we propose a further combination of different p -value combination methods. Let T_i , $i = 1, 2$, denote the test statistics of two p -value combination methods. Given the significance level γ_i , the critical values to reject H_0 , denoted by $c_i(\gamma_i)$, can be uniquely and implicitly determined by

$$\gamma_i = \Pr(T_i \geq c_i | H_0 \text{ is true}), i = 1, 2.$$

Given the significance level α , we define the test statistic of the CCP test as

$$T_c = 1_{\{\cup_{i=1}^2 T_i \geq c_i(\gamma_i)\}},$$

where $1_{\{A\}}$ is the indicator function of event A , and γ_1 and γ_2 are selected such that

$$\Pr(T_c = 1 | H_0 \text{ is true}) = \alpha. \tag{1}$$

Let w denote the probability that both p -value combination methods incorrectly reject H_0

when it is true, that is

$$w(\gamma_1, \gamma_2) = \Pr \left(\bigcap_{i=1}^2 T_i \geq c_i(\gamma_i) \mid H_0 \text{ is true} \right).$$

Since

$$\Pr(T_c = 1 \mid H_0 \text{ is true}) = \Pr \left(\bigcup_{i=1}^2 T_i \geq c_i(\gamma_i) \mid H_0 \text{ is true} \right) = \gamma_1 + \gamma_2 - w(\gamma_1, \gamma_2), \quad (2)$$

equation (1) can also be written as

$$\gamma_1 + \gamma_2 - w(\gamma_1, \gamma_2) = \alpha. \quad (3)$$

Equation (3) defines an entire family of CCP tests that control the size at the level of α . With different pairs of γ_1 and γ_2 , the CCP test assigns different weights to each p -value combination method. If we have additional information about the proportion of false individual hypotheses, we could select the specific weights to achieve a more powerful CCP test. However, such a situation without uncertainty is beyond the scope of the present analysis. Instead, we focus our analysis on the situation without information on the number of false null hypotheses.

Because no single p -value combination method is among the best across the uncertain number of false individual null hypotheses, we want to avoid the extreme cases when one method dominates. For example, if $\gamma_1 \approx 0$ and $\gamma_2 \approx \alpha$, then the CCP test almost reduces to the second p -value combination method. To avoid such extreme cases, we choose γ_1 and γ_2 to minimize $\left| \frac{\gamma_1}{\gamma_2} - 1 \right|$.

Theorem 2.1. $\left| \frac{\gamma_1}{\gamma_2} - 1 \right|$ is minimized when $\gamma_1 = \gamma_2$.

Proof. Given γ_2 , the value of γ_1 is uniquely determined by equation (3), and vice versa. Without loss of generality, we assume that γ_1 depends on γ_2 , that is, $\gamma_1 = f(\gamma_2)$, where f is

a continuous function. We may easily see that $w \geq 0$ and

$$\gamma_2 - w = \Pr(T_2 \geq c_2(\gamma_2) | H_0 \text{ is true}) - \Pr\left(\bigcap_{i=1}^2 T_i \geq c_i(\gamma_i) | H_0 \text{ is true}\right) \geq 0.$$

If $\gamma_2 \in [0, \alpha]$, then $0 \leq \gamma_2 - w \leq \gamma_2 \leq \alpha$. From this, $f(\gamma_2) = \gamma_1 = \alpha - \gamma_2 + w \in [0, \alpha]$ follows. Thus, f has at least one fixed point in $[0, \alpha]$; that is, there exists at least one γ_2 such that $\gamma_2 = f(\gamma_2) = \gamma_1$. Apparently, $\left|\frac{\gamma_1}{\gamma_2} - 1\right|$ achieves its minimum value of 0 at the fixed point of $f(\gamma_2)$, on which $\gamma_1 = \gamma_2$. $\square$

Consequently, we assume that $\gamma_1 = \gamma_2 = \gamma$. Then the test statistic of the CCP test can be rewritten as

$$T_c = 1_{\{\bigcup_{i=1}^2 T_i \geq c_i(\gamma)\}}.$$

2.2 Size of the CCP test

Given any two p -value combination methods, the size of the CCP test, $\Pr(T_c = 1 | H_0 \text{ is true})$, depends on both γ and n . We denote it as

$$g(\gamma; n) = \Pr(T_c = 1 | H_0 \text{ is true}).$$

From equation (2), we have $g(\gamma; n) = 2\gamma - w$. Given the significance level for the CCP test $\alpha \in (0, 1)$, γ is to be determined such that

$$g(\gamma; n) = \alpha.$$

Since $g(\gamma; n) = 2\gamma - w$ and $0 \leq w \leq \gamma$, it follows that

$$\frac{\alpha}{2} \leq \gamma \leq \alpha.$$

In order to determine γ , we need to find $g(\gamma; n)$, which can be obtained analytically or numerically.

2.2.1 Combining Tippett's and Simes' methods: closed-form solution

Let T_1 =Tippett's method and T_2 =Simes' method. Now we show that $g(\gamma; n)$ has a closed-form formula.

Theorem 2.2. *Let $P_{(1)}, \dots, P_{(n)}$ be the ordered p -values for testing individual hypotheses H_{0i} , $i = 1, \dots, n$. Let $0 < \gamma < 0.2$ and $\zeta = 1 - (1 - \gamma)^{1/n}$. Then*

$$\begin{aligned} g(\gamma; n) &= Pr \left[\left(\bigcup_{i=1}^n \left(P_{(i)} \leq \frac{i\gamma}{n} \right) \right) \cup \left(\bigcup_{i=1}^n (P_i \leq \zeta) \right) \right] \\ &= \gamma + n(\zeta - \frac{\gamma}{n})(1 - \gamma)(1 - \frac{\gamma}{n})^{n-2}. \end{aligned} \quad (4)$$

Proof. See the appendix. □

Once $g(\gamma; n)$ is obtained by equation (4), γ can be determined by solving

$$g(\gamma; n) = \gamma + n(\zeta - \frac{\gamma}{n})(1 - \gamma)(1 - \frac{\gamma}{n})^{n-2} = \alpha. \quad (5)$$

When $n = 1$, equation (5) is reduced to $\gamma = \alpha$. When $n \geq 2$, equation (5) is nonlinear and Newton-Raphson's iteration can be applied. Since $\gamma > 0$ is usually small, we may alternatively use a second order Taylor polynomial at $\gamma = 0$ to approximate $g(\gamma; n)$. Then

equation (5) becomes

$$\alpha = g(\gamma; n) \approx \gamma + \frac{n-1}{2n}\gamma^2,$$

which has only one positive solution

$$\gamma \approx \frac{-n + \sqrt{n^2 + 2n^2\alpha - 2n\alpha}}{n-1}.$$

Remark 2.1. In Theorem 2.2, we only consider the case where $0 < \gamma < 0.2$, since $\gamma < \alpha$ and conventional significance levels include 0.01, 0.05 or 0.10. Since γ is small, by equation (5), we have

$$\alpha = g(\gamma; n) = \gamma + n(\zeta - \frac{\gamma}{n})(1 - \gamma)(1 - \frac{\gamma}{n})^{n-2} = \gamma + o(\gamma^2) \approx \gamma.$$

2.2.2 Combining other tests: numerical approach

In most cases, $g(\gamma; n)$ does not have a closed-form formula. In such cases, Monte Carlo simulation can be used to generate $g(\gamma; n)$.

Let K be a positive integer and simulate K points for $g(\gamma; n)$. Define the step size to be $h = \frac{0.2}{K}$ and $\gamma_k = hk$, for $k = 1, 2, \dots, K$. Each $y_k = g(\gamma_k; n)$ is generated by the Monte Carlo simulation as follows.

1. Generate n i.i.d. uniformly random numbers $p_1, p_2, \dots, p_n$ on $[0, 1]$ as a set of p -values.
2. Compute the value of test statistics $T_{c,k}$ when the significance level for the individual p -value combination method is γ_k .
3. Repeat the steps above M times and get a random sample $T_{c,k}^1, T_{c,k}^2, \dots, T_{c,k}^M$. Then

$$y_k = g(\gamma_k; n) \approx \overline{T_{c,k}} = \frac{1}{M} \sum_{m=1}^M T_{c,k}^m.$$

Since we do not expect to see many oscillations over a small interval, we use a second order polynomial to fit the simulated data:

$$y_k = a_0^n + a_1^n \gamma_k + a_2^n (\gamma_k)^2, \quad k = 1, 2, \dots, K. \quad (6)$$

We estimate a_0 , a_1 and a_2 by the least square approximation. This system of equations can be expressed in matrix form as

$$y = BA,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}, B = \begin{pmatrix} 1 & \gamma_1 & \gamma_1^2 \\ 1 & \gamma_2 & \gamma_2^2 \\ \vdots & \vdots & \vdots \\ 1 & \gamma_K & \gamma_K^2 \end{pmatrix}, A = \begin{pmatrix} a_0^n \\ a_1^n \\ a_2^n \end{pmatrix}.$$

The least square estimator for A is

$$\hat{A} = (B'B)^{-1}B'y.$$

After the coefficients in equation (6) are estimated, γ is then obtained by solving

$$g(\gamma; n) \approx a_0^n + a_1^n \gamma + a_2^n \gamma^2 = \alpha.$$

2.3 Power of the CCP test

When the joint null hypothesis H_0 is false, whether one, few or almost all individual null hypotheses H_{0i} , $i = 1, 2, \dots, n$ are false is *a priori* unknown. Let $L \in \{1, 2, \dots, n\}$ denote

the number of false individual hypotheses. The goal of this subsection is to examine how changes in L affect the power of the CCP test.

A model is needed to simulate the p -value of each individual hypothesis, H_{0i} . According to Loughin (2004), the p -value is a random variable, denoted by P , with the following properties: (i) under the individual null hypothesis H_{0i} , $P \sim U(0, 1)$; (ii) under the alternative hypothesis, denoted by H_{Ai} , the density of P should be non-increasing; (iii) one or more parameters that can stochastically order the densities must be definable, thereby mimicking the effects of increasing sample sizes under H_{Ai} , with H_{0i} as a limiting case. Although there are many functions that can serve as the density function of P , we follow Loughin (2004) to use beta function because of its non-increasing density and easiness to work with computationally. The beta density function is

$$f_{\beta}(p) = \beta(a, b)p^{a-1}(1-p)^{b-1},$$

where $\beta(a, b) = \left(\int_0^1 t^{a-1}(1-t)^{b-1} dt \right)^{-1}$. When $a = 1$ and $b \geq 1$, we have

$$f_{\beta}(p) = b(1-p)^{b-1},$$

where b measures the strength of evidence against the individual null hypothesis. The c.d.f. of P is

$$F_{\beta}(p) = 1 - (1-p)^b.$$

Let S be the probability integral transformation of P , that is

$$S = F_{\beta}(P) = 1 - (1-P)^b.$$

It can be shown that $S \sim U(0, 1)$. Here

$$P = F_{\beta}^{-1}(S) = 1 - (1 - S)^{1/b}$$

is used to simulate P , where $b = 1$ for a true individual hypothesis and $b > 1$ for a false hypothesis. Then a set of p -values can be simulated for the case of L false individual hypotheses. Without loss of generality, let $b_1 = \dots = b_L > 1$ and $b_{L+1} = \dots = b_n = 1$. The detailed procedure is described as follows.

1. Generate n i.i.d. uniformly distributed random numbers, $s_1, s_2, \dots, s_n$ on $[0, 1]$.
2. Convert s_i to p_i by

$$p_i = F_{\beta}^{-1}(s_i) = 1 - (1 - s_i)^{1/b}, i = 1, 2, \dots, n.$$

3. Apply the CCP test to the p -values generated in step 2.
4. Repeat steps 1 – 3 M times. Count the number of times, denoted by N , that H_0 is rejected. The power of the CCP test is then given by N/M .

3 Monte Carlo simulation

Monte Carlo simulations are conducted to explore the size and power of the CCP test. We combine any two from the following four methods: Fisher's, Simes', Stouffer's and Tippett's methods. Our simulations are performed in Matlab.

3.1 Size consideration

The CCP test is fully specified after γ is determined. The Monte Carlo simulation described in Section 2.2.2 is used to obtain γ , for $\alpha = 0.01, 0.05, 0.10$ and $n = 2, 5, 10, 20, 40, 80, 160$,

500 (Tables 1 and 2).

Recall that, when combining Tippett's and Simes' methods, the size of the CCP test can also be determined analytically. In Table 1, we include the exact values of γ obtained by solving equation (5). Reassuringly, the simulation results are virtually indistinguishable from the analytical results, implying that the Monte Carlo simulation results are reliable and accurate.

To measure the similarity between two p -value combination methods, we define the degree of similarity as

$$\rho = \frac{w}{\gamma} = \frac{2\gamma - \alpha}{\gamma} = 2 - \frac{\alpha}{\gamma} \in [0, 1]. \quad (7)$$

Remark 3.1. *According to equation (7), ρ is uniquely determined by γ , and vice versa. We consider three different cases regarding the similarity between two p -value combination methods under the joint null hypothesis.*

Case (i): If $\gamma = \frac{\alpha}{2}$, then $\rho = 0$ and the two p -value combination methods make opposite decisions;

Case (ii): If $\gamma = \alpha$, then $\rho = 1$ and the two p -value combination methods make the same decisions;

Case (iii): If $\frac{\alpha}{2} < \gamma < \alpha$, then $0 < \rho < 1$. The degree of similarity between the two methods increases as ρ increases.

Since γ measures the similarity between two p -value combination methods under the joint null hypothesis, we also observe the following from Tables 1 and 2:

- Tippett's and Simes' methods are almost identical, echoing earlier results in Remark 2.1;
- Fisher's and Stouffer's methods are very similar;

- The similarity between Simes' (Tippett's) and Fisher's (Stouffer's) methods is very low.

Remark 3.2. *The previous observation justifies Neuhäuser (2003)'s combination of Simes' method and the truncated product method (TPM) at the individual significance level $\alpha/2$. Since the TPM is similar to Fisher's method when the truncation point is relatively large ($\tau = 0.7$), we conjecture that the similarity between Simes' method and the TPM with large τ is very low, i.e., $\rho \approx 0$. According to equation (7), $\gamma \approx \alpha/2$, that is, H_0 is rejected at the significance level α if either Simes' method or the TPM yields a rejection at the level $\alpha/2$.*

Remark 3.3. *The CCP test is easy to carry out. When combining any two of the four well-known p -value combination methods considered in this paper, simply refer to Tables 1 and 2 for the values of γ . Otherwise, the procedures described in Section 2.2.2 outline how to find γ for a given significance level and number of combined hypotheses.*

3.2 Power consideration

To explore the power of the tests, we conducted simulations in accordance with the procedures described in Section 2.3. Figures 1-6 plot the power of the CCP test and the underlying two p -value combination methods, when the number of false individual null hypotheses, L , takes the value of 1, $n/3$, $2n/3$ and n , respectively. Several points are worth noting.

First, as the similarity between two p -value combination methods increases, the gains from combining them decrease. For example, when combining Tippett's and Simes' methods (Figure 6), their power is virtually indistinguishable from that of the CCP test; when combining Fisher's and Stouffer's methods (Figure 5), their power is very close to that of the CCP test. Therefore, we do not recommend combining two highly similar p -value combination methods.

Second, there is no dominant test for various levels of L . When there is only one false

individual null hypothesis (i.e. $L = 1$) and the evidence against the individual null is very strong (i.e. $b = 400$), the practitioner should employ Simes' or Tippett's method. In contrast, one would most certainly want to employ Fisher's or Stouffer's method when many individual null hypotheses are false (i.e. $L = n/3, 2n/3$ and n) and the overall amount of evidence against the individual null is moderate (i.e. $b = 3, 2.5$ and 1.5). However, we are uncertain whether the number of false hypotheses is small or large in practice. Therefore, a risk-averse strategy is to further combine Simes' or Tippett's method with Fisher's or Stouffer's method.

Third, even though the CCP test is not the most powerful method in all cases, it performs very well, closely tracking the better method. The CCP test utilizes the weak similarity between the two underlying methods, thereby deriving its power from Simes' or Tippett's method when $L = 1$ and from Fisher's or Stouffer's method when L take other values (Figures 1-4). The simulation results confirm our expectation that the CCP test insures against selecting an inferior individual method without sacrificing much power.

4 Empirical example

Purchasing Power Parity (PPP) is a key assumption in many theoretical models of international economics. A common way to test for evidence of long-run PPP is to test for real exchange rate stationarity. However, empirical evidence of PPP for the floating regime period (1973-1998) is mixed. While several authors, such as Wu and Wu (2001) and Lopez (2008), found supporting evidence, others (Choi and Chue, 2007 and Pesaran, 2007) questioned the validity of PPP for this period. In this section, we use the CCP test to investigate whether real exchange rates are stationary among a group of countries within the Organization for Economic Cooperation and Development (OECD).

The log real exchange rate between country i and the U.S. is given by

$$q_{it} = s_{it} + p_{us,t} - p_{it}, i = 1, \dots, n; t = 1, \dots, T, \quad (8)$$

where s_{it} is the logarithm of the nominal exchange rate of the i th country's currency in terms of U.S. dollars, $p_{us,t}$ and p_{it} are the logarithm of consumer price indices in the U.S. and country i , respectively. We use quarterly data from the first quarter of 1973 to the second quarter of 1998 for 23 OECD countries ($n = 23, T = 102$), as listed in Table 3. All data are obtained from the IMF's International Financial Statistics.¹

Although tests that combine the data generally provide superior power to tests that combine p -values, combining dependent data often leads to excessive size distortions. By construction, real exchange rates are dependent induced by the numeraire country. Therefore, when assessing PPP, tests that combine p -values are preferred over those that combine the data. The literature verifies our decision, as seen in O'Connell (1998) who showed that when the independence assumption is violated, panel unit root tests by directly pooling the data will over-reject the null hypothesis. Similarly, Banerjee, et al. (2005) argued against the use of panel unit root tests because of this potential problem. In this paper, we use Bai and Ng (2004)'s panel analysis of nonstationarity in idiosyncratic and common components (PANIC) approach to remove the common factors from the data. As a result, the defactored residuals are independent across panel units, which then allows us to construct valid pooled test statistics. As originally pointed out by Maddala and Wu (1999) and Choi (2001), combining p -values has the added advantages that it allows different specifications for each panel unit and that panel tests can be constructed whether stationarity or nonstationarity is taken as the null hypothesis.

The PANIC approach decomposes q_{it} in the following way:

¹Note that, for Iceland, the consumer price indices are missing during 1982:Q1-1982:Q4 in the original data. We filled out this gap by calculating the level of CPI from its percentage changes in the IMF database.

$$q_{it} = c_i + \lambda_i' F_t + e_{it}, \quad (9)$$

where F_t is an $r \times 1$ vector of common factors that induce correlation across panel units, λ_i is an $r \times 1$ vector of factor loadings and e_{it} is an idiosyncratic error. Compared to the existing tests that applied to the observed data, the PANIC approach tests the common and idiosyncratic components separately. Therefore, the PANIC approach is more precise than testing the sum of two components, especially when the two pieces are integrated of different order. Furthermore, the tests on F_t can be performed without knowing if e_{it} are stationary or nonstationary, and the tests on e_{it} do not depend on whether F_t is stationary or not. For real exchange rates to be stationary, F_t and e_{it} must be stationary as well. Nonstationarity, on the other hand, could arise because of a unit root in any of the r factors or in e_{it} .

We begin with estimating the factors and the loadings using the method of principal components. Bai and Ng (2004) proposed an information based procedure, IC_1 , that can consistently estimate the number of factors r . The IC_1 selects one factor in our sample. We proceed with estimation assuming there is one common factor. The augmented Dickey-Fuller (ADF) test on the common factor yields -1.793 with a p -value of 0.385. We cannot reject the null hypothesis that the common component is nonstationary.

We consequently assume that F_t is nonstationary and test the idiosyncratic errors. To conduct the unit root test, we replace e_{it} in equation (9) by $\hat{e}_{it}$, the accumulated principal components estimator of Δe_{it} , and then run the following regression

$$\Delta \hat{e}_{it} = \phi_i \hat{e}_{i,t-1} + \sum_{j=1}^{k_i} \varphi_{ij} \Delta \hat{e}_{i,t-j} + u_{it}. \quad (10)$$

We conduct the ADF test, which is the individual t -test for testing $\phi_i = 0$ in equation (10). By removing the common component from the observed data, the PANIC residual-based ADF test statistics and the associated p -values are independent. Thus, the corresponding

pooled test statistic is valid. For a more rigorous justification of the pooled PANIC test by Bai and Ng (2004), see the recent paper by Westerlund and Larsson (2009). Table 3 shows the estimation results. The null hypothesis that the idiosyncratic components are nonstationary cannot be rejected at $\alpha = 5\%$ according to Simes' and Tippett's methods, since the smallest p -value exceeds the threshold level of either method. Specifically, $p_{(1)} = 0.01 > \frac{\alpha}{N} = 0.002$ for Simes' method and $p_{(1)} = 0.01 > 1 - (1 - \alpha)^{1/N} = 0.002$ for Tippett's method. However, the null hypothesis is rejected by Fisher's and Stouffer's methods, since they result in very small p -values. Now turning to the CCP test, we first consider the combination of Simes' and Fisher's methods. If either of the methods individually yields a rejection at the significance level γ , then we reject the null hypothesis that the idiosyncratic components are nonstationary. According to Table 2, we use $\gamma = 0.0287$. Since the p -value of Fisher's method equals 0.0104, which is less than the significance level 0.0287, Fisher's method and thus the CCP test reject the null hypothesis. When combining other pairs of methods that give conflicting results, e.g. Simes and Stouffer, Tippett and Fisher, Tippett and Stouffer, we reach the same rejection decision. In summary, the results from the CCP tests unanimously reject the null hypothesis that the idiosyncratic components are nonstationary.

Taken as a whole, evidence from testing the common component suggests the presence of one nonstationary factor. However, the results from the CCP test strongly reject the nonstationarity in the idiosyncratic components. Therefore, our tests suggest that the underlying source of nonstationarity in the observed real exchange rates is likely the common component. Understanding the structural source of this nonstationarity provides insight to why evidence against PPP tends to accumulate.

5 Concluding remarks

In this paper, we addressed the problem of substantially differing performance in commonly used p -value combination methods across an uncertain number of false individual hypotheses. We proposed a new procedure that retains good power properties despite such uncertainty. Based on a simple union of rejections decision rule, the proposed test controls for the weak similarity between the two individual methods and simultaneously extracts the superior power of the two. Our Monte Carlo simulations show that the power of the CCP test closely tracks the best achievable power, while keeping the size at the level of α .

As illustrated by our empirical example, the CCP test avoids arbitrarily choosing a method when the results from individual p -value combination methods contradict. Since our test can be applied to any problem for which several tests have been proposed, the CCP test has the potential to serve a wide variety of empirical applications. Examples include testing for panel cointegration, testing for jumps, testing for structural breaks - to name just a few. By applying the CCP test to empirical studies, practitioners would take a significant step towards objectively integrating and synthesizing results.

The CCP test can be generalized to combine more than two p -value combination methods or test statistics. This procedure would be similar to that of Harvey, et al. (2011), where they combined three unit root tests to deal with uncertainty over the trend. Furthermore, given information about the proportion of false individual hypotheses, the power of the CCP test might be further improved. This issue is currently under investigation by the authors.

6 Appendix: Proof of theorem 2.2

For each $n \geq 1$ note that $\zeta = 1 - (1 - \gamma)^{1/n} > \frac{\gamma}{n}$. Now we claim that $\zeta < \frac{2\gamma}{n}$ when $0 < \gamma < 0.2$ for each $n \geq 1$.

For $n = 1$, it is clear that $\zeta = \gamma < 2\gamma$. For each $n \geq 2$, we define

$$f_n(\gamma) := \zeta - \frac{2\gamma}{n} = 1 - (1 - \gamma)^{1/n} - \frac{2\gamma}{n}.$$

When $\gamma < 1 - \frac{1}{2}(\frac{1}{2})^{\frac{1}{n-1}}$, $f'_n(\gamma) = \frac{1}{n}(1 - \gamma)^{\frac{1}{n}-1} - \frac{2}{n} < 0$. Hence $f_n(\gamma)$ strictly decreases on $(0, 0.2)$ since $0 < \gamma < 0.2 < 1 - \frac{1}{2}(\frac{1}{2})^{\frac{1}{n-1}}$. Then $\zeta - \frac{2\gamma}{n} = f_n(\gamma) < f_n(0) = 0$. Therefore $\zeta < \frac{2\gamma}{n}$. Thus

$$\frac{\gamma}{n} < \zeta < \frac{2\gamma}{n}, \quad \forall n \geq 1 \text{ and } 0 < \gamma < 0.2.$$

Now

$$\begin{aligned} g(\gamma; n) &= Pr \left[\left(\bigcup_{i=1}^n \left(P_{(i)} \leq \frac{i\gamma}{n} \right) \right) \cup \left(\bigcup_{i=1}^n (P_i \leq \zeta) \right) \right] \\ &= Pr \left[\left(\bigcup_{i=1}^n \left(P_{(i)} \leq \frac{i\gamma}{n} \right) \right) \cup (P_{(1)} \leq \zeta) \right] \\ &= Pr \left[(P_{(1)} \leq \zeta) \cup \left(\bigcup_{i=2}^n \left(P_{(i)} \leq \frac{i\gamma}{n} \right) \right) \right] \\ &= 1 - Pr \left[(P_{(1)} > \zeta) \cap \left(\bigcap_{i=2}^n \left(P_i > \frac{i\gamma}{n} \right) \right) \right]. \end{aligned}$$

Since the joint density function of the ordered p -values $P_{(1)}, \dots, P_{(n)}$ is $n!$,

$$\begin{aligned}
g(\gamma; n) &= 1 - \int_{\gamma}^1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} \int_{\gamma}^{p_{(2)}} n! dp_{(1)} dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} \\
&= 1 - \left(\int_{\gamma}^1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} \int_{\frac{\gamma}{n}}^{p_{(2)}} n! dp_{(1)} dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} \right. \\
&\quad \left. - \int_{\gamma}^1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} \int_{\frac{\gamma}{n}}^{\zeta} n! dp_{(1)} dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} \right) \\
&= 1 - Pr \left[\bigcap_{i=1}^n P_{(i)} > \frac{i\gamma}{n} \right] + \int_{\gamma}^1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} \int_{\frac{\gamma}{n}}^{\zeta} n! dp_{(1)} dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} \\
&= Pr \left[\bigcup_{i=1}^n P_{(i)} \leq \frac{i\gamma}{n} \right] \\
&\quad + \int_{\gamma}^1 n dp_1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} \int_{\frac{\gamma}{n}}^{\zeta} (n-1)! dp_{(1)} dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} \\
&= \gamma + n \left(\zeta - \frac{\gamma}{n} \right) \int_{\gamma}^1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} (n-1)! dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} \\
&= \gamma + n \left(\zeta - \frac{\gamma}{n} \right) I(\gamma; n),
\end{aligned}$$

where

$$I(\gamma; n) = \begin{cases} 1 & n = 1, \\ \int_{\gamma}^1 \int_{\frac{(n-1)\gamma}{n}}^{p_{(n)}} \cdots \int_{\frac{2\gamma}{n}}^{p_{(3)}} (n-1)! dp_{(2)} \cdots dp_{(n-1)} dp_{(n)} & n \geq 2. \end{cases}$$

Now we claim that

$$I(\gamma; n) = (1 - \gamma) \left(1 - \frac{\gamma}{n} \right)^{n-2}. \tag{11}$$

It is easy to see that equation (11) holds when $n = 1$ and $n = 2$. By induction, we assume

that equation (11) holds for $n = k$, that is $I(\gamma; k) = (1 - \gamma)(1 - \frac{\gamma}{k})^{k-2}$. When $n = k + 1$,

$$I(\gamma; k + 1) = \int_{\gamma}^1 \int_{\frac{k\gamma}{k+1}}^{p(k+1)} \cdots \int_{\frac{2\gamma}{k+1}}^{p(3)} k! dp_{(2)} \cdots dp_{(k)} dp_{(k+1)}.$$

By using the change of variables, let $\widetilde{p}_{(i)} = \frac{p_{(i)}}{p_{(k+1)}}$, $i = 2, \dots, k$. Then

$$\begin{aligned} I(\gamma; k + 1) &= \int_{\gamma}^1 \int_{\frac{k\gamma}{k+1}}^{p(k+1)} \cdots \int_{\frac{2\gamma}{k+1}}^{p(3)} k! dp_{(2)} \cdots dp_{(k)} dp_{(k+1)} \\ &= \int_{\gamma}^1 \left(\int_{\frac{k\gamma}{p(k+1)}}^1 \int_{\frac{(k-1)\gamma}{p(k+1)}}^{\widetilde{p}_{(k)}} \cdots \int_{\frac{2\gamma}{p(k+1)}}^{\widetilde{p}_{(3)}} k! p_{(k+1)}^{k-1} d\widetilde{p}_{(2)} \cdots d\widetilde{p}_{(k-1)} d\widetilde{p}_{(k)} \right) dp_{(k+1)} \\ &= \int_{\gamma}^1 k p_{(k+1)}^{k-1} \left(\int_{\frac{k\gamma}{p(k+1)}}^1 \int_{\frac{(k-1)\gamma}{p(k+1)}}^{\widetilde{p}_{(k)}} \cdots \int_{\frac{2\gamma}{p(k+1)}}^{\widetilde{p}_{(3)}} (k-1)! d\widetilde{p}_{(2)} \cdots d\widetilde{p}_{(k-1)} d\widetilde{p}_{(k)} \right) dp_{(k+1)} \\ &= \int_{\gamma}^1 k p_{(k+1)}^{k-1} I\left(\frac{k\gamma}{(k+1)p_{(k+1)}}; k\right) dp_{(k+1)} \\ &= \int_{\gamma}^1 k p_{(k+1)}^{k-1} \left(1 - \frac{k\gamma}{(k+1)p_{(k+1)}}\right) \left(1 - \frac{\gamma}{(k+1)p_{(k+1)}}\right)^{k-2} dp_{(k+1)} \\ &= (1 - \gamma) \left(1 - \frac{\gamma}{k+1}\right)^{k-1}. \end{aligned}$$

By induction, equation (11) holds for all $n \geq 1$. Hence

$$g(\gamma; n) = \gamma + n \left(\zeta - \frac{\gamma}{n} \right) I(\gamma; n) = \gamma + n \left(\zeta - \frac{\gamma}{n} \right) (1 - \gamma) \left(1 - \frac{\gamma}{n}\right)^{n-2}.$$

References

- Bai, Jushan and Serena Ng (2004). “A PANIC Attack on Unit Roots and Cointegration,” *Econometrica*, 72, 1127-1177.
- Banerjee, Anindya, Massimiliano Marcellino and Chiara Osbat (2005). “Testing for PPP: Should We Use Panel Methods?” *Empirical Economics*, 30, 77-91.
- Choi, In (2001). “Unit Root Tests for Panel Data,” *Journal of International Money and Finance*, 20, 249-272.
- Choi, In and Timothy K. Chue (2007). “Subsampling Hypothesis Tests for Nonstationary Panels with Applications to Exchange Rates and Stock Prices,” *Journal of Applied Econometrics*, 22, 233-264.
- Dumitru, Ana-Maria and Giovanni Urga (2012). “Identifying Jumps in Financial Assets: a Comparison between Nonparametric Jump Tests,” *Journal of Business and Economic Statistics*, DOI: 10.1080/07350015.2012.663250.
- Fisher, Ronald A. (1932). *Statistical Methods for Research Workers*, 4th edition, Oliver and Boyd, London.
- Harvey, David I., Stephen J. Leybourne and A.M. Robert Taylor (2009). “Unit Root Testing in Practice: Dealing with Uncertainty over the Trend and Initial Condition,” *Econometric Theory*, 25, 587-636.
- Harvey, David I., Stephen J. Leybourne and A.M. Robert Taylor (2011). “Testing for Unit Roots and the Impact of Quadratic Trends, with an Application to Relative Primary Commodity Prices,” *Econometric Reviews*, 30, 514-547.
- Hedges, Larry V. and Ingram Olkin (1985). *Statistical Methods for Meta-Analysis*, Academic Press, San Diego.

- Lopez, Claude (2008). "Evidence of Purchasing Power Parity for the Floating Regime Period," *Journal of International Money and Finance*, 27, 156-164.
- Loughin, Thomas M. (2004). "A Systematic Comparison of Methods for Combining p -values from Independent Tests," *Computational Statistics and Data Analysis*, 47, 467-485.
- Maddala, G.S. and Shaowen 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.
- Müller, Ulrich K. (2009). "Comment on Unit Root Testing in Practice: Dealing with Uncertainty over the Trend and Initial Condition," *Econometric Theory*, 25, 643-648.
- Neuhäuser, Markus (2003). "Tests for Genetic Differentiation," *Biometrical Journal*, 8, 974-984.
- O'Connell, Paul G.J. (1998). "The Overvaluation of Purchasing Power Parity," *Journal of International Economics*, 44, 1-19.
- Pesaran, M. Hashem (2007). "A Simple Panel Unit Root Test in the Presence of Cross-section Dependence," *Journal of Applied Econometrics*, 22, 265-312.
- Sheng, Xuguang and Jingyun Yang (2012). "Truncated Product Methods for Panel Unit Root Tests," *Oxford Bulletin of Economics and Statistics*, DOI: 10.1111/j.1468-0084.2012.00705.x.
- Simes, R.J. (1986). "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751-754.
- Stouffer, Samuel A., Edward A. Suchman, Leland C. DeVinney, Shirley A Star. and Robin M. Williams Jr. (1949). *Studies in Social Psychology in World War II: The American Soldier. Vol. 1, Adjustment during Army Life*, Princeton, Princeton University Press.

- Tippett, Leonard H.C. (1931). *The Method of Statistics*, Williams & Norgate, London.
- Westerlund, Joakim and Rolf Larsson (2009). "A Note on the Pooling of Individual PANIC Unit Root Tests," *Econometric Theory*, 25, 1851-1868.
- Wu, Jyh-Lin and Shaowen Wu (2001). "Is Purchasing Power Parity Overvalued?" *Journal of Money, Credit and Banking*, 33, 804-812.
- Zaykin, Dmitri V., Lev A. Zhivotovsky, Peter H. Westfall, and Bruce S. Weir (2002). "Truncated Product Method for Combining p -Values," *Genetic Epidemiology* 22, 170-185.

Table 1: Selection of γ : combination of Tippett's and Simes' methods

	α	$n = 2$	5	10	20	40	80	160	500
I	0.01	0.0100	0.0100	0.0100	0.0100	0.0099	0.0099	0.0100	0.0100
	0.05	0.0494	0.0490	0.0489	0.0489	0.0488	0.0488	0.0489	0.0489
	0.10	0.0977	0.0964	0.0960	0.0956	0.0956	0.0954	0.0955	0.0956
II	0.01	0.0100	0.0100	0.0100	0.0100	0.0100	0.0100	0.0100	0.0100
	0.05	0.0494	0.0491	0.0490	0.0489	0.0489	0.0489	0.0488	0.0489
	0.10	0.0977	0.0966	0.0963	0.0961	0.0960	0.0960	0.0960	0.0960

I: Simulated values; II: Exact values

Table 2: Selection of γ : combination of other methods

	α	$n = 2$	5	10	20	40	80	160	500
A	0.01	0.0071	0.0060	0.0055	0.0053	0.0048	0.0046	0.0046	0.0047
	0.05	0.0385	0.0330	0.0304	0.0287	0.0270	0.0261	0.0257	0.0254
	0.10	0.0817	0.0708	0.0648	0.0606	0.0573	0.0551	0.0538	0.0527
B	0.01	0.0071	0.0059	0.0055	0.0053	0.0048	0.0046	0.0046	0.0047
	0.05	0.0383	0.0328	0.0303	0.0286	0.0270	0.0260	0.0256	0.0254
	0.10	0.0808	0.0697	0.0641	0.0601	0.0570	0.0549	0.0537	0.0526
C	0.01	0.0058	0.0052	0.0051	0.0051	0.0047	0.0047	0.0047	0.0048
	0.05	0.0325	0.0286	0.0273	0.0265	0.0257	0.0253	0.0252	0.0252
	0.10	0.0696	0.0605	0.0570	0.0550	0.0534	0.0525	0.0520	0.0517
D	0.01	0.0058	0.0052	0.0051	0.0051	0.0047	0.0047	0.0047	0.0048
	0.05	0.0324	0.0285	0.0272	0.0265	0.0256	0.0252	0.0252	0.0252
	0.10	0.0690	0.0601	0.0568	0.0549	0.0533	0.0524	0.0520	0.0517
E	0.01	0.0079	0.0072	0.0070	0.0069	0.0062	0.0060	0.0061	0.0063
	0.05	0.0410	0.0382	0.0374	0.0370	0.0361	0.0357	0.0357	0.0359
	0.10	0.0846	0.0793	0.0778	0.0771	0.0771	0.0768	0.0765	0.0763

A: Combination of Fisher's and Simes' methods

B: Combination of Fisher's and Tippett's methods

C: Combination of Stouffer's and Simes' methods

D: Combination of Stouffer's and Tippett's methods

E: Combination of Fisher's and Stouffer's methods

Table 3: PANIC residual-based unit root tests for 23 OECD real exchange rates

Panel (a): augmented Dickey-Fuller test		
Country	Lag	<i>p</i> -value
Australia	0	0.585
Austria	6	0.760
Belgium	0	0.365
Canada	0	0.905
Denmark	0	0.080
Finland	0	0.265
France	0	0.405
Germany	0	0.760
Greece	2	0.100
Iceland	5	0.250
Ireland	1	0.185
Italy	0	0.115
Japan	0	0.525
Korea	0	0.035
Luxembourg	0	0.650
Mexico	0	0.035
Netherlands	2	0.075
Norway	0	0.010
Portugal	0	0.205
Sweden	0	0.430
Switzerland	0	0.520
Turkey	0	0.435
United Kingdom	1	0.120
Panel (b): <i>p</i> -values for panel unit root tests		
	Test	<i>p</i> -value
	Fisher's method	0.0104
	Stouffer's method	0.0046

Figure 1: Power of the CCP test: combination of Fisher's and Simes' methods

Figure 2: Power of the CCP test: combination of Fisher's and Tippett's methods

Figure 3: Power of the CCP test: combination of Stouffer's and Simes' methods

Figure 4: Power of the CCP test: combination of Stouffer's and Tippett's methods

Figure 5: Power of the CCP test: combination of Fisher's and Stouffer's methods

Figure 6: Power of the CCP test: combination of Tippett's and Simes' methods