Empir Econ (2017) 53:329–350
DOI 10.1007/s00181-017-1230-9

Combination of "combinations of p values"

Lan Cheng[1] · Xuguang Simon Sheng[2]

Received: 11 November 2015 / Accepted: 17 October 2016 / Published online: 11 February 2017

Abstract We investigate the impact of an uncertain number of false individual null hypotheses on commonly used p value combination methods. Under such uncertainty, these methods perform quite differently and often yield conflicting results. Consequently, we develop a combination of "combinations of p values" (CCP) test aimed at maintaining good power properties across such uncertainty. The CCP test is based on a simple union–intersection principle that exploits the weak correspondence between two underlying p value combination methods. Monte Carlo simulations show that the CCP test controls size and closely tracks the power of the best individual methods. We empirically apply the CCP test to explore the stationarity in real exchange rates and the information rigidity in inflation and output growth forecasts.

Keywords Hypothesis testing · Information rigidity · p value · Panel unit root

This paper was presented at the 2nd IAAE Conference, 22nd SNDE Conference, North American Summer Meeting of the Econometric Society, AEA Annual Meeting, and the Seaway Section of the Mathematical Association of America. We thank Badi Baltagi, Nancy Boynton, Raffaella Giacomini, Bruce Hansen, Kajal Lahiri, Ulrich Müller, Eric Renault, Aman Ullah, Herman van Dijk, Jonathan Wright, and conference participants for helpful comments and suggestions. We also thank three anonymous referees for the comments that have significantly improved the paper.

Electronic supplementary material The online version of this article (doi:10.1007/s00181-017-1230-9) contains supplementary material, which is available to authorized users.

✉ Xuguang Simon Sheng
sheng@american.edu

[1] Department of Mathematical Sciences, SUNY Fredonia, Fredonia, NY, USA

[2] Department of Economics, American University, 4400 Massachusetts Ave., N.W.,
Washington, DC 20016, USA

1 Introduction

P value combinations from a set of hypothesis tests are a common tool in meta-analysis. The combinations resolve data and methodological problems: Full datasets may not be available in many published studies, and individual methodologies are too dissimilar for the underlying data to be combined. Given n individual hypotheses H_{0i}, $i = 1, 2, \ldots, n$, 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, Westberg (1985), Neuhäuser (2003), Loughin (2004), Sheng and Yang (2013a), and Hanck (2013). 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, other tests, such as Fisher's and Stouffer's methods, are powerful when there are many false individual null hypotheses. Since the number of false individual null hypotheses is *a priori* unknown in practice, researchers have difficulty selecting the optimal method to use. This problem is most severe when results from differing p value combination methods conflict. In response, Loughin (2004, p. 484) suggested that "Meta-combinations, or combinations of combinations, could be considered rather than relying on a single function for all purposes." We follow Loughin's suggestion by proposing a combination of "combinations of p values" (CCP) test that maintains good power properties across an uncertain number of false individual hypotheses.

The CCP test is based on a simple union–intersection principle, originally proposed by Roy (1953) as a heuristic method to test an intersection of some component hypotheses. The union–intersection method allows us to bound the probability of the union of a finite set of p value combination methods, without knowing their dependence structure. Recent studies have applied a similar approach to other statistical problems, but the application to combinations of p values is novel.[1] More specifically, the CCP test rejects the joint null hypothesis at the given significance level α when at least one of the two p value combination methods yields a rejection at the designed individual level γ. The value of γ is selected such that the CCP test (i) controls the overall size at α and (ii) has good power, robust to the uncertain number of false individual null hypotheses. Monte Carlo simulations show that the power of the CCP test closely tracks that of the best individual method, irrespective of the number of false individual null hypotheses.

We demonstrate the usefulness of the CCP test in two applications. In the first empirical study, we investigate whether real exchange rates are stationary among a group of OECD countries. Our results from the CCP test suggest that the underlying source of non-stationarity in the observed real exchange rates is likely the common factor. This source of non-stationarity explains why evidence against purchasing power

[1] Harvey et al. (2009) used the union–intersection method to combine two unit root tests, and their combination approach comes close to exploiting the available information efficiently as shown by Müller (2009). Dumitru and Urga (2012) used the similar strategy across nine procedures and across sampling frequencies to minimize spurious jump detection in financial assets. Bayer and Hanck (2013) applied a similar method to combine non-cointegration tests.

parity tends to accumulate. In the second empirical application, we assess which forms of information constraints are most relevant to the expectation formation process of professional forecasters. While individual p value combination methods often give conflicting results, the CCP test clearly indicates the presence of substantial heterogeneity in estimated levels of information rigidity. This heterogeneity is across forecasting horizons and countries and in line with the noisy information model *à la* Sims (2003).

The rest of the paper is organized as follows: Sect. 2 develops the CCP test. Section 3 uses Monte Carlo simulations to explore the power of the test. Section 4 illustrates the use of the test in two empirical examples, and Sect. 5 concludes. The online appendix provides additional simulation results.

2 The combination of "combinations of *p* values" test

Let p_i be the p value for testing an individual hypothesis H_{0i}, $i = 1, 2, \ldots, n$. We consider the problem of testing the joint null hypothesis $H_0 = \bigcap_{i=1}^n H_{0i}$ when the underlying p values are independent. We reject H_0 at an overall significance level α if at least one of H_{0i}, $i = 1, 2, \ldots, n$, is false. Various methods have been proposed in the literature to combine p values. Broadly speaking, these methods can be classified either as quantile combination methods or as ordered statistic methods.

The quantile combination methods transform the p values into distributional quantiles. Two well-known examples include Fisher's (1932) method, $t = -2\sum_{i=1}^n \ln(p_i) \sim \chi^2_{2n}$, and Stouffer's method, attributed to Stouffer et al. (1949), defined as $z = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Phi^{-1}(p_i) \sim N(0, 1)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (c.d.f.).[2]

Let $p_{(i)}$ be the ordered p values such that $p_{(i)} \le p_{(i+1)}$. The ordered statistic methods take advantage of the fact that, under H_0, $p_{(1)} \le p_{(2)} \le \cdots \le p_{(n)}$ are ordered statistics from a uniform distribution from zero to one. The underlying assumptions are that under H_0, these statistics follow a continuous distribution that is free of nuisance parameters. The typical methods include Tippett's (1931) method: $p_{\min} = \min_{i=1,\ldots,n}\{p_i\}$, and Simes' (1986) method: $p_{\min} = \min_{i=1,\ldots,n}\{np_{(i)}/i\}$, where for both methods H_0 is rejected if $p_{\min} \le \alpha$.[3]

Choosing an optimal method *ex ante* is complicated because H_0 may be false in numerous ways. As shown by the recent simulation studies in Neuhäuser (2003), Loughin (2004), Sheng and Yang (2013a), and Hanck (2013), Tippett's and Simes' methods are powerful when the total evidence against H_0 is concentrated in very few of the combined p values. However, Fisher's and Stouffer's methods perform well when evidence against H_0 is broadly spread among the combined p values. Therefore, the quantile combination methods and the order statistic methods frequently give conflicting test decisions. We cannot generally expect the methods from different groups to be sensitive to all possible alternatives. One of the contributions of this

[2] These two methods were introduced to panel unit root literature independently by Maddala and Wu (1999) and Choi (2001). See Costantini and Lupi (2013) and Sheng and Yang (2013a) for recent applications.

[3] Hanck (2013) advocated Simes' method in testing panel unit roots.

paper is to provide a robust test that is relatively powerful for all situations under the alternative hypothesis.

To this end, we propose a further combination of a pair of p value combination methods coming from different groups. Let T_j, $j = 1, 2$, denote the test statistics of two p value combination methods (one from each group), c_j the critical value, and γ_j the corresponding significance level. Then, we have

$$\gamma_j = \Pr(T_j \geq c_j | H_0 \text{ is true}),\ j = 1, 2.$$

Clearly, c_j is uniquely and implicitly determined by γ_j. We define the test statistic of the CCP test as

$$T_c = 1_{\{\bigcup_{j=1}^{2} T_j \geq c_j(\gamma_j)\}}, \tag{1}$$

where $1_{\{A\}}$ is the indicator function of event A. To our knowledge, there is no optimal criterion to choose individual levels. 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. In the absence of any prior information about the number of false null hypotheses, we choose γ_1 and γ_2 to minimize the number of instances where both methods reject H_0, while still controlling the overall size at the level of α:

$$\Pr(T_c = 1 | H_0 \text{ is true}) = \alpha. \tag{2}$$

Theorem 2.1 shows that this minimization can be achieved at $\gamma_1 = \gamma_2 = \gamma$.

Theorem 2.1 *Let T_j, $j = 1, 2$, denote the test statistics of two p value combination methods, c_j the critical value, and γ_j the corresponding significance level. In order to minimize the number of instances where both tests reject H_0, while still maintaining the size α, we select $\gamma_1 = \gamma_2$.*

Proof It is sufficient to solve for γ_1, since γ_2 can be determined by equation (2). In order for the CCP test to have good power properties across the uncertain number of false individual null hypotheses, we minimize the probability that both methods reject H_0:

$$\min_{\gamma_1 \in (0,1)} \frac{\Pr\left(T_1 \geq c_1(\gamma_1) \bigcap T_2 \geq c_2(\gamma_2)\right)}{\min\{\Pr(T_1 \geq c_1(\gamma_1)), \Pr(T_2 \geq c_2(\gamma_2))\}}, \tag{3}$$

subject to equation(2). In (3), the numerator can be expressed as

$$\begin{aligned}\Pr\left(T_1 \geq c_1(\gamma_1) \bigcap T_2 \geq c_2(\gamma_2)\right) &= \Pr(T_1 \geq c_1(\gamma_1)) + \Pr(T_2 \geq c_2(\gamma_2)) \\ &\quad - \Pr\left(\bigcup_{j=1}^{2} T_j \geq c_j(\gamma_j)\right) \\ &= \Pr(T_1 \geq c_1(\gamma_1)) + \Pr(T_2 \geq c_2(\gamma_2)) - \alpha.\end{aligned}$$

Without loss of generality, assume $\Pr(T_1 \geq c_1(\gamma_1)) \leq \Pr(T_2 \geq c_2(\gamma_2))$, then (3) becomes

$$\min_{\gamma_1 \in (0,1)} 1 + \frac{\Pr(T_2 \geq c_2(\gamma_2)) - \alpha}{\Pr(T_1 \geq c_1(\gamma_1))}. \tag{4}$$

Taking the derivative with respect to $\Pr(T_1 \geq c_1(\gamma_1))$ yields

$$\frac{\frac{\partial \Pr(T_2 \geq c_2(\gamma_2))}{\partial \Pr(T_1 \geq c_1(\gamma_1))} \Pr(T_1 \geq c_1(\gamma_1)) - [\Pr(T_2 \geq c_2(\gamma_2)) - \alpha]}{(\Pr(T_1 \geq c_1(\gamma_1)))^2}. \tag{5}$$

If the minimization problem in (4) has an interior solution, that is, $\Pr(T_1 \geq c_1(\gamma_1)) < \Pr(T_2 \geq c_2(\gamma_2))$, then (5) equals zero, which implies that $\frac{\partial \Pr(T_2 \geq c_2(\gamma_2))}{\partial \Pr(T_1 \geq c_1(\gamma_1))} < 0$. However, this inequality will eventually lead to $\Pr(T_1 \geq c_1(\gamma_1)) > \Pr(T_2 \geq c_2(\gamma_2))$, which contradicts the assumption $\Pr(T_1 \geq c_1(\gamma_1)) \leq \Pr(T_2 \geq c_2(\gamma_2))$. Hence, the minimum in (4) is obtained along the boundary by taking $\Pr(T_1 \geq c_1(\gamma_1)) = \Pr(T_2 \geq c_2(\gamma_2))$, that is, $\gamma_1 = \gamma_2$.

Remark 2.1 In Theorem 2.1, we show that $\gamma_1 = \gamma_2$ can be achieved by minimizing the probability that both p value combination methods reject H_0. Note that the condition $\gamma_1 = \gamma_2$ is similar to Dufour and Torres (1998) and Bayer and Hanck (2013) where they choose the weights to ensure the same null rejection probabilities for both methods.

Remark 2.2 By taking $\gamma_1 = \gamma_2 = \gamma$, the CCP test can be equivalently stated as: H_0 is rejected at the overall significance level α if $min(\Theta_j) \leq \gamma$ for $j = 1, 2$, where Θ_j is the p value from the test statistic T_j in testing the joint null hypothesis H_0. This alternative form proves particularly useful when demonstrating admissibility properties for the CCP test. According to Birnbaum (1954), every monotone combined test procedure is admissible in the class of all combined test procedures. A combined test procedure Y is monotone if Y is a non-decreasing function, that is, if $p_i^* \leq p_i$, $i = 1, 2, \ldots, n$, then $Y(p_1^*, p_2^*, \ldots, p_n^*) \leq Y(p_1, p_2, \ldots, p_n)$. In our case, $Y = min(\Theta_j)$, $j = 1, 2$, is clearly non-decreasing and thus the CCP test is admissible.

Remark 2.3 Although the CCP test is similar to Bonferroni procedure, these two methods are different. According to Bonferroni procedure, H_0 is rejected at the overall significance level α if $min(\Theta_j) \leq \alpha/2$ for $j = 1, 2$. Bonferroni procedure controls the family-wise error rate but at the cost of low false null hypothesis detection, because Θ_j is compared to the level $\alpha/2$, which is smaller than γ as in the CCP test—as shown in equation (6). Thus, the CCP test has a higher ability to detect false null hypothesis and is more powerful than Bonferroni procedure.[4]

The CCP test is fully specified after γ is determined. For convenience, let

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

[4] See Lehmann and Romano (2005, p.350) for further discussions on the power of Bonferroni procedure.

denote the size of the CCP test. Given the overall significance level α, γ is determined such that $g(\gamma; n) = \alpha$. In most cases, $g(\gamma; n)$ does not have a closed-form formula because the underlying multiple testing methods are not independent, with possibly a complex dependence structure. To address such problems, we use Monte Carlo simulation to generate $g(\gamma; n)$ for $0 < \gamma < 0.2$. 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 $k = 1, 2, \ldots, K$. 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 $p_1, p_2, \ldots p_n$ 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 $T_{c,k}^1, T_{c,k}^2, \ldots, T_{c,k}^M$. Then, $y_k = g(\gamma_k; n) \approx \overline{T_{c,k}} = \dfrac{1}{M} \displaystyle\sum_{m=1}^{M} T_{c,k}^m$.

Since we do not expect to see many oscillations over a small interval, we use a second-order polynomial to fit the simulated data[5]

$$y_k = a_0^n + a_1^n \gamma_k + a_2^n (\gamma_k)^2, \qquad k = 1, 2, \ldots, K.$$

We estimate the coefficients in a regression of y_k on $(1, \gamma_k, \gamma_k^2)$ and then obtain γ by solving

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

Table 1 provides the values of γ for combining two of the four methods considered in this paper, for $\alpha = 0.01, 0.05, 0.10$ and $n = 2, 5, 10, 20, 40, 80, 160, 500$.[6]

The values of γ also provide information about the correspondence between two p value combination methods. To illustrate this, we define the correspondence between two combination methods under H_0 as

$$\rho = \frac{\Pr\left(\bigcap_{j=1}^{2} T_j \geq c_j(\gamma_j) \middle| H_0 \text{ is true}\right)}{\sqrt{\Pr\left(T_1 \geq c_1(\gamma_1) \middle| H_0 \text{ is true}\right) \Pr\left(T_2 \geq c_2(\gamma_2) \middle| H_0 \text{ is true}\right)}} = 2 - \frac{\alpha}{\gamma} \in [0, 1], \tag{6}$$

where the last equality holds as we set $\gamma_j = \gamma$ for $j = 1, 2$.

[5] In cases where $g(\gamma; n)$ has a closed-form formula, e.g., combining Tippett's and Simes' methods, we find that a second-order polynomial is sufficient in approximating the function $g(\gamma; n)$. We also conducted the simulation study by combining Tippett's and Simes' methods. The simulated values of γ are virtually the same as the ones obtained analytically by solving the equation $g(\gamma; n)$, implying that the Monte Carlo simulation methods are reliable.

[6] We experimented with different values of K and step size h. Our results show that the values of γ remain almost the same across different values of $K = 10, 20$, and 40. For the detailed results, see the online appendix. In our simulation, we used $K = 20$, $h = 0.01$, and $M = 10,000$.

Table 1 Estimation of individual significance level γ

	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

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

Remark 2.4 According to equation (6), given the overall significance level α, ρ is uniquely determined by γ, and vice versa. We consider three different cases regarding the correspondence between two p value combination methods under H_0.

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

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

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

The findings reported in Table 1 imply that the correspondence between the methods from different groups is relatively low. More specifically, the correspondence between Fisher's and Simes' (or Tippett's) methods ranges from 0.10 to 0.46, and between Stouffer's and Simes' (or Tippett's) methods, from 0.03 to 0.25 for $n \geq 10$. These findings, in turn, justify our practice of combining a pair of methods from different groups.[7]

At the end of this section, we need to point out that there may exist situations where we assign different weights to different p value combination methods. Although the number of false individual hypotheses is a priori unknown in practice, the set of p values will provide some information. To fix ideas, let τ be the percentage of the p values that are less than a given threshold, ϵ. Recall that when τ is larger, Tippett's and

[7] In separate simulation studies not shown here, we combined the two methods from the same group. We find that Tippett's and Simes' methods are almost perfectly corresponded, and Fisher's and Stouffer's methods are highly corresponded. To save space, these results are not reported here. They are available upon request.

Simes' methods become less powerful but Fisher's and Stouffer's methods become more powerful. Thus, it is possible to construct the weighted CCP test with the individual significance level γ_1 (for Tippett's or Simes' method) and γ_2 (for Fisher's or Stouffer's method) determined by equations (2) and (7):

$$\frac{\gamma_1}{\gamma_2} = \frac{1-\tau}{\tau}. \tag{7}$$

In our simulations, we take different threshold values $\epsilon = 0.1, 0.2, 0.4, 0.6, 0.8$, or 0.9 and $\tau = k \cdot 10\%$, $k = 0, 1, \cdots, 10$.

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}, $i = 1, 2, \ldots, n$ are false is *a priori* unknown. Let $L \in \{1, 2, \ldots, n\}$ denote the number of false individual hypotheses. The goal of this section is to examine how changes in L affect the power of the CCP test via Monte Carlo simulations.

3.1 A simulation study using exact p values

A model is needed to simulate the p value of each individual hypothesis, H_{0i}. According to Loughin (2004), the p value (P) is a random variable 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 is non-increasing; and (iii) one or more parameters that can stochastically order the densities must be definable. The last property mimics the effects of increasing sample sizes under H_{Ai}, with H_{0i} as a limiting case. Although there are many potential density functions for P, we follow Loughin (2004) in using the 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 $\beta(a, b)$ is the beta coefficient. When $a = 1$ and $b \geq 1$, we have the density $f_\beta(p) = b(1-p)^{b-1}$ and the c.d.f. $F_\beta(p) = 1 - (1-p)^b$ that satisfy all of the properties above. Let S be the probability integral transformation of P, that is, $S = F_\beta(P) = 1 - (1-P)^b \sim U(0, 1)$. We simulate p values using $P = F_\beta^{-1}(S) = 1 - (1-S)^{1/b}$. When $b = 1$, we have p values generated under the null hypothesis, where $P \sim U(0, 1)$. When $b > 1$, we generate p values under the alternative hypothesis, where b measures the strength of evidence against the individual null hypothesis. Without loss of generality, let $b_1 = \cdots = b_L > 1$ and $b_{L+1} = \cdots = b_n = 1$. In detail, the simulation follows:

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

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

3. Apply the CCP test to the p values generated in Step 2.

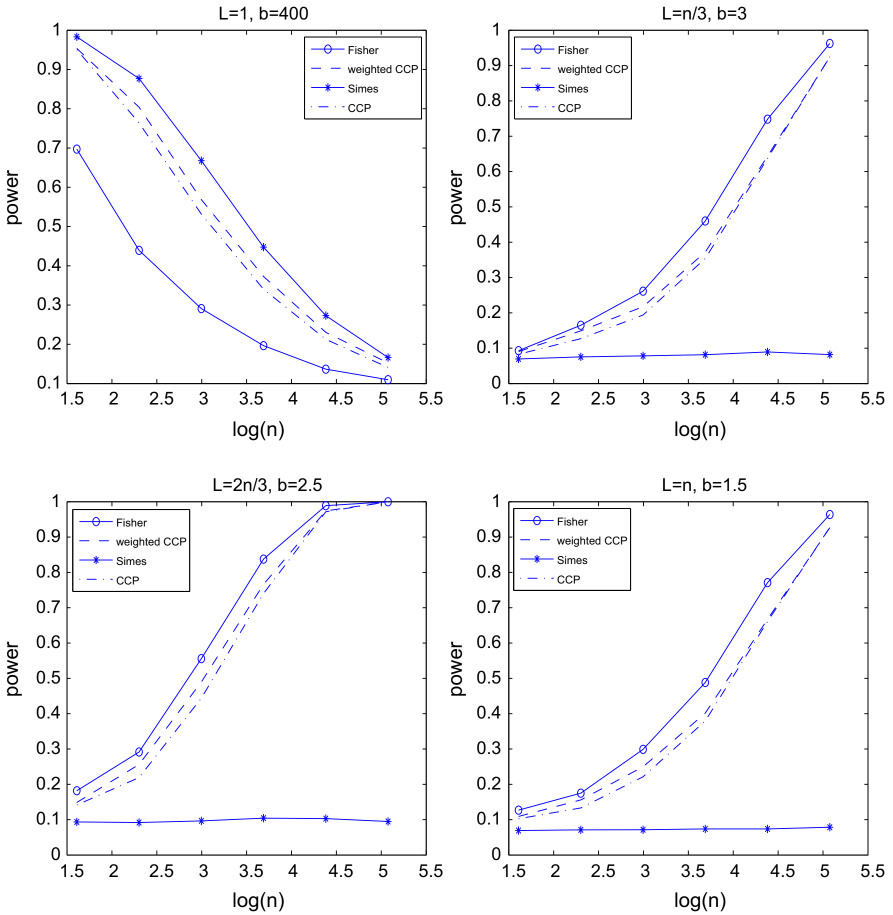

Fig. 1 Power of the test when combining Fisher’s and Simes’ methods

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 N/M.

Figures 1, 2, 3, 4 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 values of 1, $n/3$, $2n/3$, and n, respectively. The results are obtained based on $M = 10,000$ simulations with the nominal size set at 5%. Three points are worth noting.

Consistent with the literature, 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 optimal method is Simes’ or Tippett’s. In contrast, the optimal method is Fisher’s or Stouffer’s 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

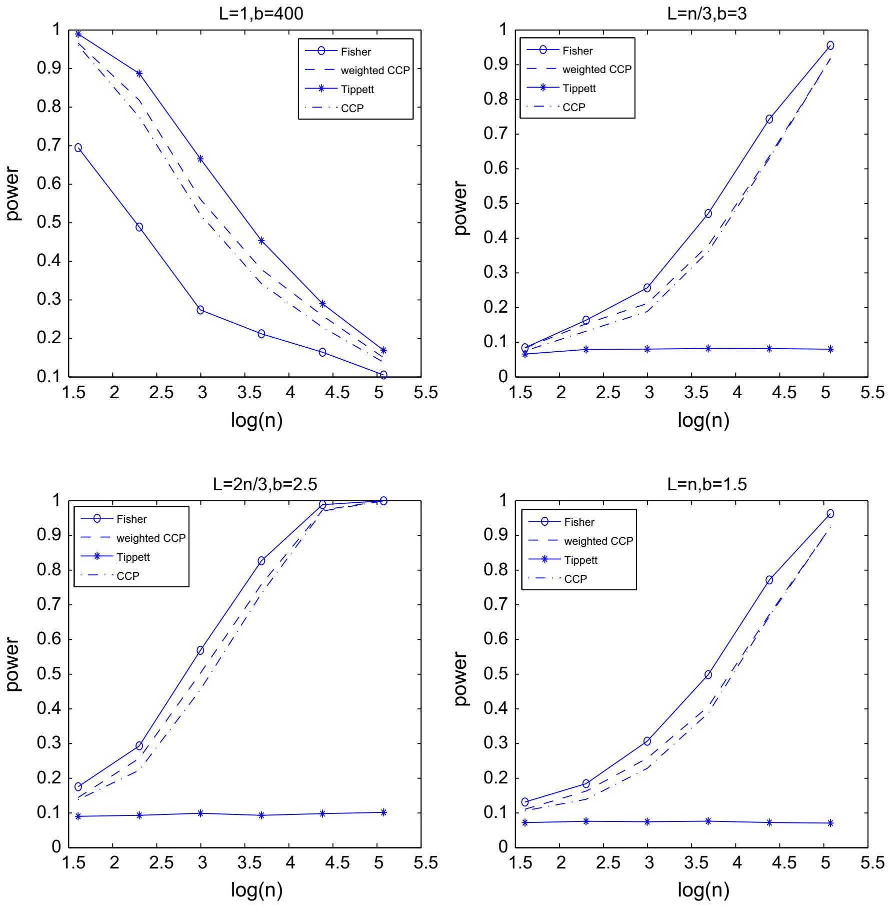

Fig. 2 Power of the test when combining Fisher's and Tippett's methods

practice. Therefore, a risk-averse strategy is to further combine Simes' (or Tippett's) method with Fisher's (or Stouffer's) method.

Second, even though the CCP test is not the most powerful method in all cases, it performs very well, closely tracking the best method. The CCP test utilizes the weak correspondence 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 takes other values. The simulation results confirm our expectation that the CCP test insures against selecting an inferior individual method without sacrificing much power.[8]

[8] In separate simulation studies, we also combined the two methods within the same group. We find that the power of the individual methods is virtually indistinguishable from that of the CCP test. Thus, the gains from combining two highly correlated methods are very small. For this reason, we do not recommend combining the two methods coming from the same group.

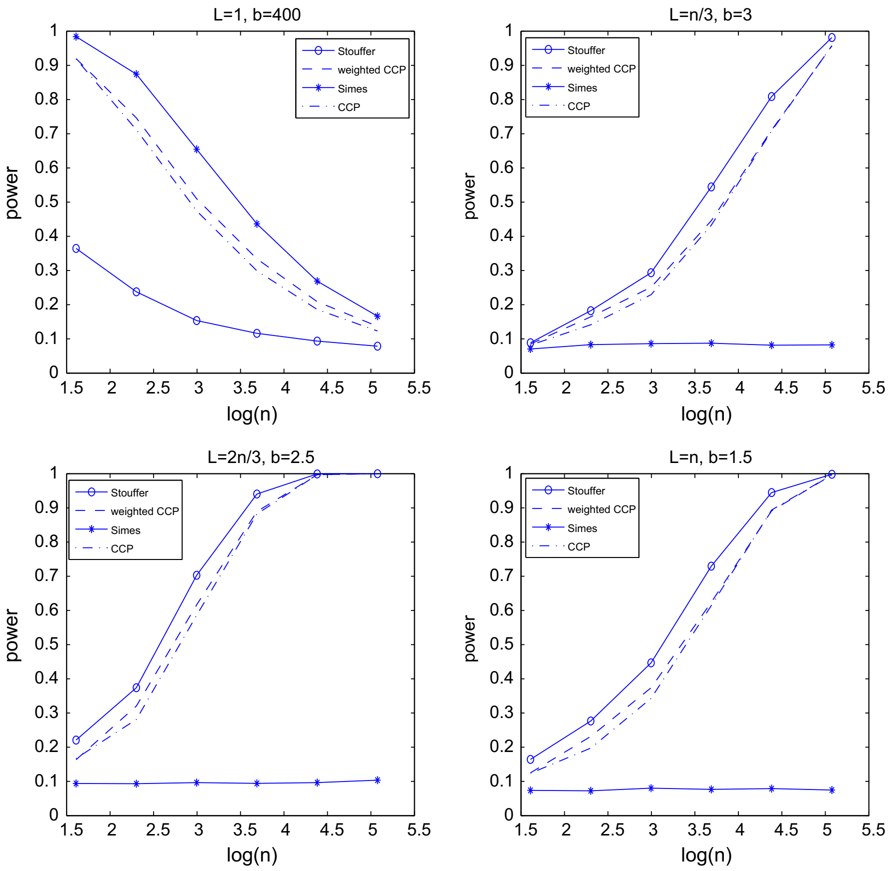

Fig. 3 Power of the test when combining Stouffer’s and Simes’ methods

Third, the weighted test outperforms the CCP test for selected parameter values, consistent with our argument that there are power gains by employing additional information. However, the power gains depend on the number of false individual hypotheses (L), the strength of evidence against the individual hypotheses (b), and the threshold (ϵ). Indeed, in situations where the weights are inappropriately chosen, the weighted test is less powerful than the CCP test, as shown by the figures in the online appendix. In light of the marginal power improvement to such a weighting mechanism, we generally recommend the use of the CCP test and restrict our discussion on the equally weighted CCP test below.

3.2 Power of panel unit root tests

We assess the ability of our measure to supplement a common empirical test, the panel unit root test. More specifically, we explore the power of panel unit root test using the following data generating process:

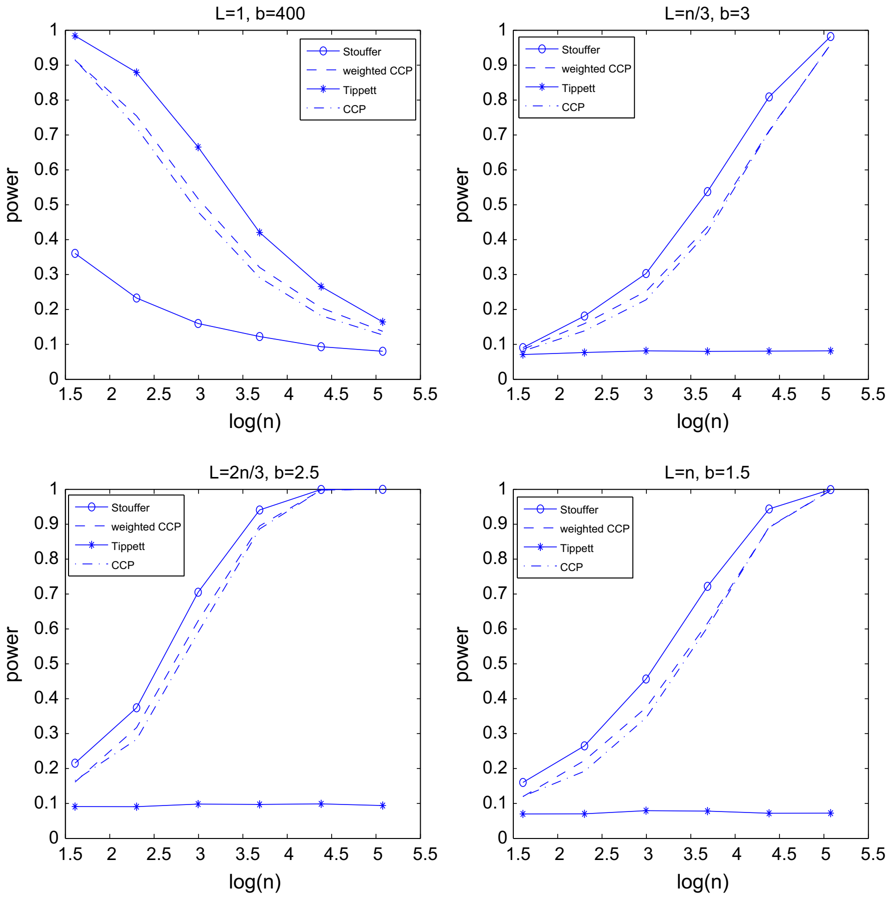

Fig. 4 Power of the test when combining Stouffer's and Tippett's methods

$$y_{it} = \mu_i + \gamma_i f_t + e_{it}, \tag{8}$$

$$f_t = \phi f_{t-1} + \eta_t, \tag{9}$$

$$e_{it} = \rho_i e_{i,t-1} + w_{it}, \tag{10}$$

for $i = 1, \ldots, N$, $t = 1, \ldots, T$. The individual fixed effect μ_i and the factor loading γ_i in equation (8) are drawn independently of each other as $\mu_i \sim$ i.i.d. $U[0, 0.2]$ and $\gamma_i \sim$ i.i.d. $U[-1, 3]$. The error terms η_t in equation (9) and w_{it} in equation (10) are simply drawn as $\eta_t \sim$ i.i.d. $N(0, 1)$ and $w_{it} \sim$ i.i.d. $N(0, 1)$. The strong dependence across panel units is driven by the common factor f_t.

We explore size of the tests under $H_0 : \phi = 1$ and $\rho_i = 1$ for all i. Note that the null hypothesis allows for non-stationarity in both common factor and idiosyncratic errors. We explore power of the tests when $\phi = 0.5$ and

$$\rho_i = \begin{cases} 1 & \text{for } x_i \geq 1 \\ x_i & \text{for } 0 < x_i < 1, \end{cases}$$

where $f(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(\frac{x}{\lambda})^k}$ for $x > 0$ and 0 otherwise. The parameters k and λ of the Weibull distribution are selected such that

$$P(x_i \geq 1) = P(\rho_i = 1) = 1 - L,$$
$$E(\rho_i) = 1 \times (1 - \delta) + \int_0^1 xf(x)dx = 1 - b,$$

where $L \in (0, 1)$ and $b \in (0, 1)$ are pre-specified constants. L indicates the fraction of stationary series (i.e., false individual null) in the panel, and b captures the deviation of the autoregressive parameters from the unit root null on average. Taken together, L measures the relative amount of evidence against H_0 or "patterns," and b measures the overall amount of evidence against H_0 or "strength." Thus, our Monte Carlo designs allow us to study the impact of changes in either "patterns" or "strength" on power of the tests. The tests are one-sided with the nominal size set at 5% and conducted for all combinations of $N \in \{20, 40\}$ and $T \in \{50, 100\}$ using M=10,000 simulations. To control for cross section dependence, we use Bai and Ng's (2004) panel analysis of non-stationarity 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. We conduct the augmented Dickey–Fuller (ADF) test on the defactored residuals and obtain p values of the ADF test using response surface regressions.

In the presence of strong cross section dependence, the CCP test yields good empirical size. For all four combinations of individual methods considered in this paper, the size of the CCP test ranges from 0.042 to 0.054. The power of the CCP test increases when T increases and when N increases as long as L is fixed, which justifies the use of panel data in unit root tests. Figures 5 and 6 plot the power of the CCP test when combining Stouffer's and Simes' methods with $b = 0.10$ and 0.06, respectively.[9] Consistent with the results from previous simulations, the CCP test closely tracks the most powerful method, deriving its power from Simes' method when $L < 0.30$ and from Stouffer's method when $L > 0.30$. The CCP test even outperforms the best single method when the constituent methods have very similar power. Intuitively, this is because each constituent method rejects the null hypothesis only marginally, but taken together, the CCP test provides sufficient evidence to reject H_0. The CCP test takes advantage of the imperfect correlation of the underlying methods. Finally, the power increases as the overall amount of evidence against H_0 accumulates with b increasing from 0.06 in Fig. 5 to 0.10 in Fig. 6.

[9] The power of the CCP test when combining other methods, that is, Fisher's and Simes', Fishers' and Tippett's, and Stouffer's and Tippett's, is qualitatively similar to that of combining Stouffer's and Simes' and is thus not reported here for the sake of brevity.

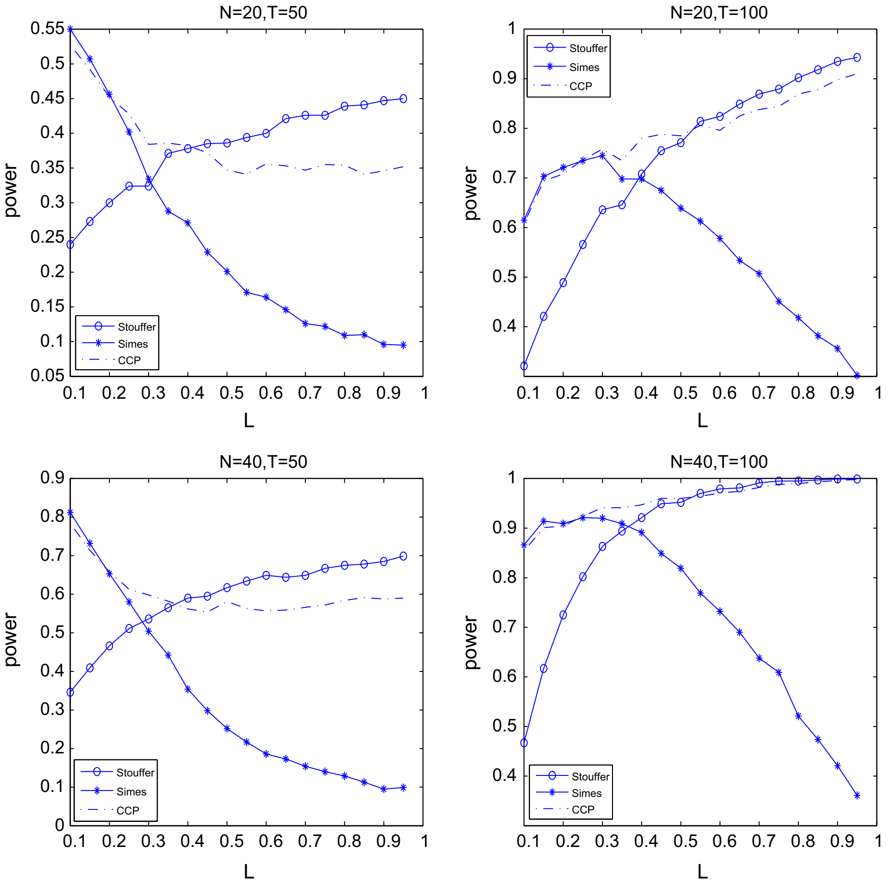

Fig. 5 Power of panel unit root test with a small amount of evidence against H_0

4 Empirical examples

We demonstrate the usefulness of the CCP test in two applications: (i) testing the stationarity in real exchange rates and (ii) testing heterogeneity in information rigidity in macro-forecasts.

4.1 Testing the purchasing power parity hypothesis

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 Lopez (2008) and Costantini and Lupi (2013), found supporting evidence, others (Choi and Chue 2007; Pesaran 2007) questioned the validity of PPP for this period. In this section, we use

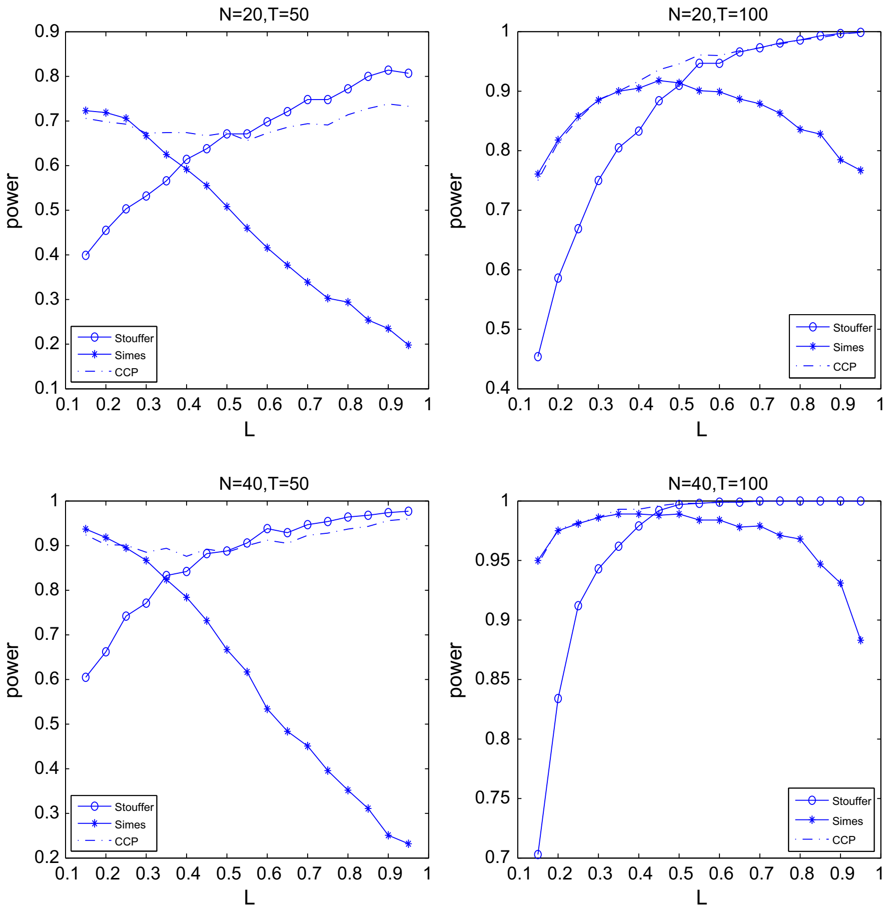

Fig. 6 Power of panel unit root test with a large amount of evidence against H_0

the CCP test to investigate whether real exchange rates are stationary among a group of OECD countries.

The log real exchange rate between country i and the USA is given by

$$q_{it} = s_{it} + p_{us,t} - p_{it}, i = 1, \ldots, n; t = 1, \ldots, T, \tag{11}$$

where s_{it} is the logarithm of the nominal exchange rate of the ith country’s currency in terms of US dollars; $p_{us,t}$ and p_{it} are the logarithm of consumer price indices in the USA 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 2. All data are obtained from the IMF’s International Financial Statistics.

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 because of the common numeraire.

Table 2 Augmented Dickey–Fuller test for stationarity in real exchange rates

Country	Lag	p 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
UK	1	0.120

Therefore, when assessing PPP, tests that combine p values are preferred over those that combine the data. The literature verifies our decision. O'Connell (1998) showed that when the independence assumption is violated, panel unit root tests by directly pooling the data will over-reject the null hypothesis. As also pointed out by Maddala and Wu (1999) and Choi (2001), combining p values has the added advantages of allowing different specifications for each panel unit and for the null hypothesis. See Baltagi (2008, chapter 12) for a recent review on non-stationary panels and p values.

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

$$q_{it} = c_i + \lambda_i' F_t + e_{it}, \tag{12}$$

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. For real exchange rates to be stationary, F_t and e_{it} must be stationary as well. Non-stationarity, 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 principal component analysis. 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 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 non-stationary.

We consequently assume that F_t is non-stationary and test the idiosyncratic errors. To conduct the unit root test, we replace e_{it} in equation (12) by $\hat{e}_{it}$, the accumulated principal components estimator of Δe_{it}, and then run the following regression

$$\Delta \hat{e}_{it} = \phi_i \hat{e}_{i,t-1} + \sum_{j=1}^{k_i} \varphi_{ij} \Delta \hat{e}_{i,t-j} + u_{it}. \tag{13}$$

We conduct the ADF test, which is the individual t-test for testing $\phi_i = 0$ in equation (13). Table 2 shows the estimation results. The null hypothesis that the idiosyncratic components are non-stationary cannot be rejected at $\alpha = 5\%$ according to Simes' and Tippett's methods. The smallest p value exceeds the threshold level of either method: $p_{(1)} = 0.01 > \frac{\alpha}{n} = 0.002$ for Simes' method and $p_{(1)} = 0.01 > 1 - (1 - \alpha)^{1/n} = 0.002$ for Tippett's method. However, the null hypothesis is strongly rejected by Fisher's and Stouffer's methods with p values of 0.0104 and 0.0046, respectively. The different methods yield conflicting results or "mixed signals." 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 non-stationary. According to Table 1, 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 non-stationary.

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

4.2 Testing information rigidity in macro-forecasts

Imperfect information models have recently regained interest in the macroeconomics literature: See Coibion and Gorodnichenko (2012), and Andrade and Le Bihan (2013), among others. In this section, we focus on two types of rational expectation models with information rigidities: the sticky information model *à la* Mankiw and Reis (2002) and the noisy information model *à la* Sims (2003). The goal of this section is to assess these two models by comparing the estimated degrees of information rigidity across forecasting horizons and countries.

In the sticky information model, due to limited resources and the cost of updating information sets, forecasters update their information infrequently. In contrast, in the noisy information model, forecasters continuously update their information sets, but form expectations via Kalman filter because they cannot fully observe the true state. Despite their differences, Coibion and Gorodnichenko (2015) show that both models predict the same relationship between the average forecast error and the average forecast revision as specified in the following regression:

$$x_t - F_{\text{th}} = \alpha + \beta R_{\text{th}} + e_t, \tag{14}$$

where x_t is the actual inflation (or output growth), F_{th} is the h-quarter ahead mean forecast, $R_{\text{th}} = F_{\text{th}} - F_{t,h+1}$ is the revision in mean forecasts, and e_t is the full-information rational expectations error and thus uncorrelated with information dated t or earlier. In equation (14), the coefficient on forecast revisions, β, maps one to one into the underlying degree of information rigidity. In the sticky information model, $\beta = \frac{\lambda}{1-\lambda}$, where λ is the proportion of forecasters not updating information at each period. Since λ is defined as a fixed parameter common across countries and forecasting horizons, a testable implication of the canonical sticky information model is that the estimated degree of information rigidity is invariant across countries and forecasting horizons. In the noisy information model, $\beta = \frac{1-G}{G}$, where G is the Kalman gain, measuring the weight given to new information. Since the Kalman gain depends on the persistence of the target variable and the signal-to-noise ratio of information for each country at each forecasting horizon, the noisy information model implies that the estimated degree of information rigidity may vary across countries and forecasting horizons. Hence, we test the sticky vs. noisy information model with the following hypotheses

H_0^a: Information rigidity is the same across *forecast horizons*.
H_0^b: Information rigidity is the same across *countries*.

Evidence for both H_0^a and H_0^b would favor the sticky information model. On the other hand, rejection of either H_0^a or H_0^b would support the noisy information model.

To estimate information rigidities, we use professional forecasts of inflation and output growth from *Consensus Forecasts*. Relative to other economic agents, professional forecasters have access to a wider range of news and have a comparative advantage in processing news. For these reasons, the extent of information rigidity among professional forecasters can be seen as a lower bound for other agents' inattention to news. Our dataset covers 22 target years (1990–2012), 7 horizons from one-quarter to seven-quarter ahead, and 7 major industrialized countries: Canada, France, Germany, Italy, Japan, the UK and the USA.

Tables 3 and 4 show the results of testing the homogeneity of information rigidities across forecasting horizons and countries, respectively, at the 5% significance level. When all individual p value combination methods do or do not reject, the CCP test does so too. However, agreeing methods are consistent only in 57–86% of cases. For the remaining cases, the two individual methods give conflicting results. The mixed signals arise either because the order statistic methods reject the null but the quantile combination methods do not reject, or vice versa. Among those cases of mixed signals from individual methods, the CCP test yields a clear decision: In 82% of the mixed cases, the CCP test rejects the null hypotheses, and in the remaining cases, the

Table 3 Test of the equality in information rigidity across forecasting horizons

	Percentage of cases in which two methods			In case of disagreement	
	Agree		Disagree	Percentage in which CCP test	
	Reject	Not reject		Reject	Not reject
(a) Inflation forecast					
A (%)	71.4	14.3	14.3	100.0	0.0
B (%)	71.4	14.3	14.3	100.0	0.0
C (%)	71.4	14.3	14.3	100.0	0.0
D (%)	71.4	14.3	14.3	100.0	0.0
(b) GDP forecast					
A (%)	85.7	0.0	14.3	100.0	0.0
B (%)	85.7	0.0	14.3	100.0	0.0
C (%)	85.7	0.0	14.3	100.0	0.0
D (%)	85.7	0.0	14.3	100.0	0.0

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

CCP test does not reject.[10] Overall, the null hypothesis that the estimated degrees of information rigidity are the same across *horizons* is rejected in about 86% of inflation forecasts and 100% of GDP forecasts (Table 3). Specifically, we find that information rigidities tend to increase with forecast horizons. At very long horizons, signals are noisy and the target variable predictability is low, cf. Lahiri and Sheng (2008). The null hypothesis that information rigidity is the same across *countries* is rejected in about 71% of inflation forecasts and 86% of GDP forecasts (Table 4). The presence of heterogeneity in estimated levels of information rigidity across forecasting horizons and countries is in line with the noisy information model. However, this substantial heterogeneity contrasts sharply with the canonical sticky information model, suggesting that future work on models of sticky information should allow for differential information updating rates across forecasting characteristics.

5 Concluding remarks

This paper makes two contributions to the literature on meta-analysis. First, we address the problem of substantially differing performance in commonly used *p* value combination methods across an uncertain number of false individual hypotheses. We propose a new procedure, the CCP test, that retains good power properties despite such uncertainty. Based on a simple union–intersection principle, the proposed test uses the weak correspondence between the two individual methods to extract the superior

[10] The 82% is calculated as the ratio of the number of rejections of the null by the CCP test (18) over the number of cases where two constituent methods give conflicting results (22).

Table 4 Test of the equality in information rigidity across countries

	Percentage of cases in which two methods			In case of disagreement	
	Agree		Disagree	Percentage in which CCP test	
	Reject	Not reject		Reject	Not reject
(a) Inflation forecast					
A (%)	42.9	28.6	28.6	100.0	0.0
B (%)	42.9	28.6	28.6	100.0	0.0
C (%)	42.9	14.3	42.9	66.7	33.3
D (%)	42.9	14.3	42.9	66.7	33.3
(b) GDP forecast					
A (%)	71.4	14.3	14.3	0.0	100.0
B (%)	71.4	14.3	14.3	0.0	100.0
C (%)	71.4	14.3	14.3	100.0	0.0
D (%)	71.4	14.3	14.3	100.0	0.0

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

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 close to the nominal level.

Second, we present two empirical examples where the results from individual p value combination methods contradict. In one example, we investigate whether real exchange rates are stationary among a group of OECD countries. In another, we assess which forms of information constraints are most relevant to the expectation formation process of professional forecasters. While the purchasing power parity hypothesis has been well studied in the literature, to our knowledge, this is the first work to apply meta-analysis in testing the heterogeneity of information frictions across forecasting horizons and countries. As illustrated in both empirical examples, the CCP test avoids arbitrarily choosing a method when the constituent methods give "mixed signals."

Since our test applies to any situation with multiple competing tests, the CCP test has the potential to serve a wide variety of empirical applications. Examples include testing for panel cointegration, jumps, and structural breaks-to name just a few. By applying the CCP test to empirical studies, practitioners would take a significant step toward robustly using all available information in hypothesis testing. Other worthwhile extensions include (i) comparing the CCP test to the resampling-based combined test in Dufour et al. (2015) and (ii) comparing the weighted CCP test to the truncated product method (TPM) proposed by Zaykin et al. (2002) and the adaptive TPM in Sheng and Yang (2013b). We leave these for future research.

References

Andrade P, Le Bihan H (2013) Inattentive professional forecasters. J Monet Econ 60:967–982

Bai J, Ng S (2004) A PANIC attack on unit roots and cointegration. Econometrica 72:1127–1177

Baltagi B (2008) Econometric analysis of panel data, 4th edn. Wiley, Hoboken

Bayer C, Hanck C (2013) Combining non-cointegration tests. J Time Ser Anal 34:83–95

Birnbaum A (1954) Combining independent tests of significance. J Am Stat Assoc 49:559–574

Choi I (2001) Unit root tests for panel data. J Int Money Financ 20:249–272

Choi I, Chue TK (2007) Subsampling hypothesis tests for nonstationary panels with applications to exchange rates and stock prices. J Appl Econom 22:233–264

Coibion O, Gorodnichenko Y (2012) What can survey forecasts tell us about informational rigidities? J Polit Econ 120:116–159

Coibion O, Gorodnichenko Y (2015) Information rigidity and the expectations formation process: a simple framework and new facts. Am Econ Rev 105:2644–2678

Costantini M, Lupi C (2013) A simple panel-CADF test for unit roots. Oxford Bull Econ Stat 75:276–296

Dufour J-M, Khalaf L, Voia M (2015) Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability. Commun Stat—Simul Comput 44:2329–2347

Dufour J-M, Torres O (1998) Union-intersection and sample-split methods in econometrics with applications to MA and SURE Models. In: Giles D, Ullah A (eds) Handbook of applied economic statistics, vol 14. Marcel Dekker Inc., New York, pp 465–505

Dumitru A-M, Urga G (2012) Identifying jumps in financial assets: a comparison between nonparametric jump tests. J Bus Econ Stat 30:242–255

Fisher RA (1932) Statistical methods for research workers, 4th edn. Oliver and Boyd, London

Hanck C (2013) An intersection test for panel unit roots. Econom Rev 32:183–203

Harvey DI, Leybourne SJ, Robert Taylor AM (2009) Unit root testing in practice: dealing with uncertainty over the trend and initial condition. Econom Theory 25:587–636

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press, San Diego

Lahiri K, Sheng X (2008) Evolution of forecast disagreement in a Bayesian learning model. J Econom 144:325–340

Lehmann EL, Romano JP (2005) Testing statistical hypotheses, 3rd edn. Springer, New York

Lopez C (2008) Evidence of purchasing power parity for the floating regime period. J Int Money Financ 27:156–164

Loughin TM (2004) A systematic comparison of methods for combining *p* values from independent tests. Comput Stat Data Anal 47:467–485

Maddala GS, Wu S (1999) A comparative study of unit root tests with panel data and a new simple test. Oxford Bull Econ Stat 61:631–652

Mankiw G, Reis R (2002) Sticky information versus sticky prices: a proposal to replace the New Keynesian Phillips curve. Quart J Econ 117:1295–1328

Müller UK (2009) Comment on unit root testing in practice: dealing with uncertainty over the trend and initial condition. Econom Theory 25:643–648

Neuhäuser M (2003) Tests for genetic differentiation. Biom J 8:974–984

O'Connell PGJ (1998) The overvaluation of purchasing power parity. J Int Econ 44:1–19

Pesaran MH (2007) A simple panel unit root test in the presence of cross-section dependence. J Appl Econom 22:265–312

Roy SN (1953) On a heuristic method of test construction and its use in multivariate analysis. Ann Math Stat 24:220–238

Sheng X, Yang J (2013a) Truncated product methods for panel unit root tests. Oxford Bull Econ Stat 75:624–636

Sheng X, Yang J (2013b) An adaptive truncated product method for combining dependent *p* values. Econ Lett 119:180–182

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Sims CA (2003) Implications of rational inattention. J Monet Econ 50:665–690

Stouffer Samuel A, Suchman Edward A, DeVinney Leland C, Star Shirley A, Williams Robin M Jr (1949) Studies in social psychology in world war II: The American soldier, Adjustment during Army Life. vol 1. Princeton University Press, Princeton

Tippett LHC (1931) The method of statistics. Williams & Norgate, London

Westberg M (1985) Combining independent statistical tests. The Statistician 34:287–296

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002) Truncated product method for combining p-values. Genet Epidemiol 22:170–185