

Reply to reports on article JBES-P-2022-0567

Reply to the editor

1. *The reviewers have diversified opinions, and they raise many issues. Although I think you can address many of them, there is one main issue I don't see how to handle: As the AE points out, you do not have a compelling case for the null hypothesis that time trends are identical across multiple time series. Like the AE, I can't think of different economic time series sharing the same trend. I appreciate your discussions on clustering much more, but the same issue arises within a group.*

There is a small econometric literature on co-trending and co-breaking where common time trends and breaks are used to capture comovements of multiple time series. Rather than time trends being identical, they are proportional to each other. That might be more plausible for modeling multiple economic time series. Because this may not be difficult in your framework, I do not mean to impose it. Instead, I am pointing out that such literature may or may not be helpful.

Although I doubt the plausibility and usefulness of the null hypothesis, I would like to give the benefit of the doubt. I am willing to consider a revision of your paper, though with no guarantee that it will ultimately be accepted. I ask you to address all of the reviewers' concerns. Most importantly, please make a much more compelling and convincing case for the null hypothesis, which may be very difficult. I should note that this is not a typical R&R and is a rather weak R&R.

Reply to the associate editor

1. *The less enthusiastic referee wrote: "Perhaps it is my unfamiliarity with the problem (or my tendencies toward Bayesian methodologies...), but I do not find it to be a particularly compelling research question. The goal is to test whether a time trend—after adjusting for covariates—is identical across multiple time series. This does not seem to be a high priority for multiple time series and dynamic regression analysis, and it's not clear whether a hypothesis test generates much useful information in this context." This is a comment I broadly agree with: would one really want/need to test for the exact "sameness" of time series trends? Or is it the case that the null hypothesis is uninteresting, but the alternative is, especially if I am able to see which trends are different and where? I am thinking aloud here, but overall I don't think the testing problem, as stated, is interesting enough for JBES readers.*

2. *There are two prior papers by the submitting author, which consider a similar problem but in the absence of external covariates. I don't think the current paper makes it clear early enough what is different between the current work and those earlier papers.*
3. *Due to the various approximations, the size control is only approximate. I don't see it as a "state of the art" way of thinking in these types of FWER control problems; please see e.g. <https://arxiv.org/abs/2009.05431>, where size control, in a different but related multiscale testing problem, is exact.*
4. *I suspect the procedure must be really difficult to use in practice with confidence, as it depends on so many tuning parameters including the bandwidth. The authors say their software is at https://github.com/marina-khi/multiscale_inference, but the link is broken.*
5. *Both referees, including the more enthusiastic one, mention several further issues with the paper, including issues related to the practicalities of the method, the simulation study and the asymptotic nature of the method.*

Reply to referee 1

Thank you very much for the careful reading of our manuscript and the interesting suggestions. In our revision, we have addressed all your comments. Please see our replies to them below.

1. *The assumptions and requirements for the variance σ^2 deserve further consideration. First, it is claimed that the variances are assumed to be constant across series, but that a different estimator is used for each series. Which is the correct assumption for practice and theory? Second, given the economic and potential financial applications, how might volatility (or time-varying variance) be incorporated into the testing procedure? Is this plausible within the proposed framework, even if additional assumptions are required? If it is not plausible to account for volatility explicitly, then is the procedure robust in the presence of volatility?*
2. *There are several issues with the simulation study.*
 - (i) *Setting the fixed effect to zero and including a single covariate both make for a much simpler design than considered in the theory. More challenging scenarios, including nonzero fixed effects and multiple predictors (e.g., using the estimated values and/or covariates from the application) would better demonstrate the capabilities of this approach.*
 - (ii) *The data from the null fix $m_i = 0$ and claim this is WLOG. However, this*

is also quite a simple case: the shared $m_i()$ curve could be quite complex under the null, which only maintains that the trends are shared among the series.

(iii) There are no competing methods considered; some alternative approach or benchmark must be added. A reasonable alternative might consider an additive model and compute confidence intervals (or bands) for the trends, with a simple heuristic to determine whether the functions are identical. The proposed approach should do better, but demonstrating improvements over a reasonable alternative is important.

(iv) only a small number of series is considered. How does the approach perform when n is large?

In the simulation study, we need to take into account (i), (iii) and (iv). The point (ii) is not correct and can be taken care of by verbal argument.

(i) As we cannot use the estimated AR(1) values from the application, let's take the old simulation setup and modify it a bit (for the 'robustness' check we also consider a case without any covariates, see the end of this section):

- As before, we choose $n = 15$ and $T = 100, 250, 500$ (for an easy DGP without any covariates we use $T = 100, 250, 500, 750, 1000$. We also use this for comparison with SiZer in (iii).).
- We include 3 covariates and model them by a very simple VAR(3) process:

$$\underbrace{\begin{pmatrix} X_{it,1} \\ X_{it,2} \\ X_{it,3} \end{pmatrix}}_{=: \mathbf{X}_{it}} = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{pmatrix} \begin{pmatrix} X_{it-1,1} \\ X_{it-1,2} \\ X_{it-1,3} \end{pmatrix} + \underbrace{\begin{pmatrix} \nu_{it,1} \\ \nu_{it,2} \\ \nu_{it,3} \end{pmatrix}}_{=: \boldsymbol{\nu}_{it}}.$$

We choose $a_1 = a_2 = a_3 = 0.25$. The innovations $\boldsymbol{\nu}_{it}$ are drawn i.i.d. from a multivariate normal $N(0, \Phi)$ with

$$\Phi = \begin{pmatrix} 1 & \varphi & \varphi \\ \varphi & 1 & \varphi \\ \varphi & \varphi & 1 \end{pmatrix},$$

where $\varphi = 0, 0.1, 0.25$ and setting $\varphi = 0$ results in simulating 3 independent covariate processes (case $\varphi = 0$ is not reported).

- We set $\boldsymbol{\beta}_i = (\beta_{i,1}, \beta_{i,2}, \beta_{i,3}) = (1, 1, 1)$ for all i .
- We calculate the Gaussian quantiles based on 5000 samples, and the size and power calculations are done based on 5000 repetitions.
- As before, we assume that the errors ε_{it} follow the AR(1) model $\varepsilon_{it} = a\varepsilon_{i(t-1)} + \eta_{it}$, where $a = 0.25$ and the innovations η_{it} are i.i.d. normal with zero mean $E[\eta_{it}] = 0$ and variance $E[\eta_{it}^2] = 0.25^2$ (setting higher variance results in lower signal-to-noise ratio and we are not hitting the size level).

- We let $\alpha = (\alpha_1, \dots, \alpha_n)$ be a normally distributed random vector. In particular, $\alpha \sim N(0, c\Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix},$$

where $\rho = 0.1, 0.25$ (larger values do not seem to affect the size that much) gives the correlation across time series i . To start with, we set $c = 1$ and then adjust it if needed.

- To generate data under the null $H_0 : m_1 = \dots = m_n$, we let $m_i = 0$ for all i as before. To produce data under the alternative, we use the bump functions $m_1(u) = b \cdot \mathbf{1}(u \in [0.3, 0.7]) \cdot (1 - \{\frac{u-0.5}{0.2}\}^2)^2$ for $b = 0.25, 0.5, 1$ (depicted in Figure 1) and $m_i = 0$ for $i \neq 1$. For the cluster analysis, it is presumably also better to use bump functions rather than linear maps. For example, we could use 3 groups with (scaled versions of) the functions g_1 , g_2 and g_4 from Figure 2 in Vogt & Linton (2020, Multiscale clustering of nonparametric regression curves).
- We take the grid \mathcal{G}_T to be the same as before: $\mathcal{G}_T = U_T \times H_T$, where $U_T = \{u \in [0, 1] : u = \frac{5t}{T} \text{ for some } t \in \mathbb{N}\}$ and $H_T = \{h \in [\frac{\log T}{T}, \frac{1}{4}] : h = \frac{5t-3}{T} \text{ for some } t \in \mathbb{N}\}$ (the grid is sparser for the DGP without any covariates: $\mathcal{G}_T = U_T \times H_T$, where $U_T = \{u \in [0, 1] : u = \frac{10t}{T} \text{ for some } t \in \mathbb{N}\}$ and $H_T = \{h \in [\frac{\log T}{T}, \frac{1}{4}] : h = \frac{5t-3}{100} \text{ for some } t \in \mathbb{N}\}$). Hence, the grid is growing only a bit with the sample size, because with the usual grid the actual size is still a bit higher than the nominal size even though the problem should be “easy”. I guess this is because of the fact that with the growing sample size, the multiple testing problem becomes much more complicated... Moreover, the computational time is not feasible anymore for large values of T .)
- In order to estimate the long-run variance σ_i^2 , we follow the procedure described in Khismatullina and Vogt (2020) with the following tuning parameters: $q = 25$ and $r = 10$.

The results of the simulation study with the covariates are presented in Tables 1 and 2 for $\phi = 0.1$ and $\rho = 0.1$ and in Tables 3 and 4 for $\phi = 0.25$ and $\rho = 0.25$. We are a bit more liberal with higher values of ϕ and ρ , but not overly so. The results of the size simulation study without any covariates and for $\rho = 0.25$ are presented in Table 5 (no power computations since this is more of a robustness check). There is a “bump” in the actual

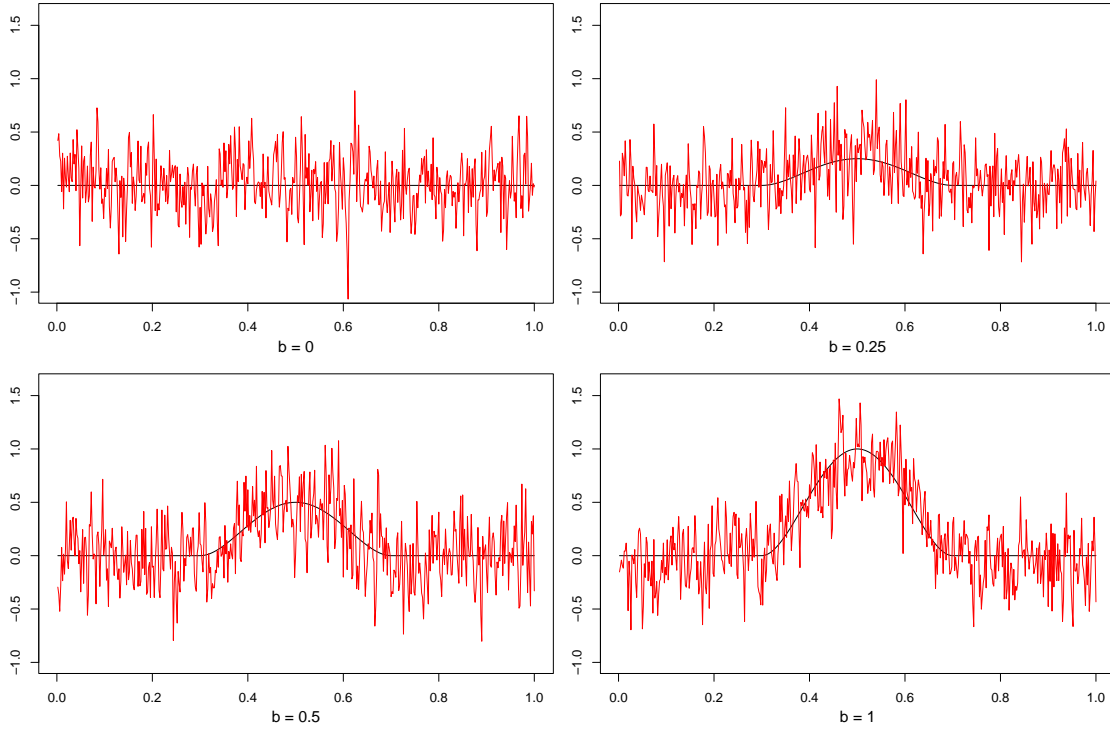


Figure 1: In black, the bump function $m_1(t/T)$ is plotted for different heights of the bump: $b = 0, 0.25, 0.5, 1$ ($b = 0$ corresponds to the data under the null H_0). In red, we depict the bump function plus the error term, $m_1(t/T) + \varepsilon_{it}$.

size for moderate sample sizes ($T = 250, 500$) which is present under a lot specifications (even without dependence in the fixed effect), I guess this is due to the grid specifications: for small sample size ($T = 100$) the grid is a bit too sparse to detect anything spurious, whereas starting from $T = 250$ the grid becomes more sensitive to the noise which stabilizes with larger values of the sample size. I have checked with different grid specification and this “bump” is always present but occurs at different sample sizes.

- (iii) As suggested by you, we compare with the Sizer method of Park et al. (2009). For the comparison, we could consider the design proposed in (i) without covariates and fixed effects (as these are not part of the model in Park et al.). I’d suggest to add the comparison to the supplement rather than to the main body of the text.

Currently running. Do I understand correctly that we are comparing both in terms of size and power?

- (iv) We rerun the simulations from (i) with a much smaller grid. This allows us to deal not only with $n = 15$ but also with larger values of n : $n = 25, 50, 100$. Specifically, we take the grid \mathcal{G}_T to be a dyadic scheme (as in Wavelet

Table 1: Size of the multiscale test for $\phi = 0.1$ and $\rho = 0.1$ for different sample sizes T and nominal sizes α .

T	nominal size α		
	0.01	0.05	0.1
100	0.009	0.038	0.067
250	0.018	0.068	0.124
500	0.015	0.057	0.115

Table 2: Power of the multiscale test for $\phi = 0.1$ and $\rho = 0.1$ for different sample sizes T and nominal sizes α . Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$				(b) $b = 0.50$				(c) $b = 1.00$			
nominal size α				nominal size α				nominal size α			
T	0.01	0.05	0.1	T	0.01	0.05	0.1	T	0.01	0.05	0.1
100	0.019	0.073	0.128	100	0.058	0.198	0.299	100	0.012	0.060	0.114
250	0.242	0.440	0.565	250	0.981	0.996	0.999	250	1.000	1.000	1.000
500	0.731	0.864	0.914	500	1.000	1.000	1.000	500	1.000	1.000	1.000

analysis) with scales in the set

$$\mathcal{H}_T = \{h = 2^k h_{\min} \text{ for } k = 0, \dots, K\},$$

where

$$h_{\min} = \frac{\lceil \log T \rceil}{T}$$

and K is such that $2^K h_{\min} \leq \frac{1}{4}$, i.e.,

$$K \leq \left\lfloor \log \left(\frac{T}{4 \lceil \log T \rceil} \right) \right\rfloor \frac{1}{\log(2)},$$

and

$$\mathcal{G}_T = \{(u, h) \subseteq [0, 1] : (u, h) = ((2s + 1)h, h) \text{ for } s = 0, \dots, \left\lfloor \frac{h^{-1} - 1}{2} \right\rfloor \text{ and } h \in \mathcal{H}_T\}.$$

We rerun the simulations as specified in (i) (with the exception of the grid) for the case $\phi = 0.1$ and $\rho = 0.1$ with $T = 100, 250, 500$ and different values of n : $n = 15, 25, 50, 100$. (This is very feasible, even with $n = 100$ and $T = 500$ the running time is around 45 minutes for all size and power calculations). The results are presented in Tables 6 - 13. The results are not very different from using the full grid: we are not even much more liberal.

Table 3: Size of the multiscale test for $\phi = 0.25$ and $\rho = 0.25$ for different sample sizes T and nominal sizes α .

T	nominal size α		
	0.01	0.05	0.1
100	0.009	0.038	0.079
250	0.014	0.068	0.131
500	0.015	0.065	0.121

Table 4: Power of the multiscale test for $\phi = 0.25$ and $\rho = 0.25$ for different sample sizes T and nominal sizes α . Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$				(b) $b = 0.50$			
T	nominal size α			T	nominal size α		
	0.01	0.05	0.1		0.01	0.05	0.1
100	0.017	0.069	0.139	100	0.062	0.196	0.318
250	0.235	0.459	0.583	250	0.982	0.997	0.999
500	0.704	0.860	0.914	500	1.000	1.000	1.000

I attribute it to the “simplicity” of the problem: not that many multiple comparisons.

3. *Similarly, there are many related clustering methods, including (Bayesian and non-Bayesian) methods for clustering functional data. The proposed approach is reasonable, yet should be placed in a broader context and evaluated against appropriate competitors.*

I think we could use the following clustering procedure as a benchmark:

- Estimate the trends m_i by a local linear estimator \hat{m}_i with a fixed bandwidth (chosen adhoc).
- Compute a simple distance measure d_{ij} between \hat{m}_i and \hat{m}_j , e.g.

$$d_{ij} = \int_0^1 (\hat{m}_i(w) - \hat{m}_j(w))^2 dw.$$

- Construct the following dissimilarity measure from these distances:

$$\hat{\Delta}(S, S') = \max_{i \in S, j \in S'} d_{ij}.$$

- Run a HAC algorithm with the computed dissimilarities.

Table 5: Size of the multiscale test for the case without any covariates and $\rho = 0.25$ for different sample sizes T and nominal sizes α .

T	nominal size α		
	0.01	0.05	0.1
100	0.004	0.030	0.069
250	0.015	0.072	0.131
500	0.013	0.065	0.122
750	0.017	0.063	0.122
1000	0.012	0.072	0.123

Our procedure can be regarded as a further development of this very simple and natural benchmark procedure. In particular: our procedure replaces the simple distance measure d_{ij} by a more advanced multiscale distance measure and provides a way to estimate the number of clusters, which is not part of the simple benchmark procedure.

As the benchmark procedure does not provide an estimate of the number of clusters K , it is presumably best to compare our procedure with the benchmark for known K .

4. *A related Bayesian strategy is to use simultaneous band scores (simBaS) to assess whether a function differs from zero. This could be applied pairwise to the differences between functions to establish a Bayesian competitor to the proposed approach, and simply requires posterior draws from an analogous Bayesian model.*

We may argue here that we use the benchmark discussed in the previous comment instead of the proposed Bayesian strategy because it is naturally linked to our approach.

5. *The application includes numerous tuning parameters (including kernels, intervals, etc.). Are the results robust to these choices? Further details are needed.*

The tuning parameters are:

- (a) the grid \mathcal{G}_T
- (b) tuning parameters to estimate the error variances σ_i^2
- (c) the number of bootstrap samples L to compute the Gaussian quantile
- (d) the kernel K .

(Have I forgotten anything here?)

We should run robustness checks:

Table 6: Size of the multiscale test for $n = 15$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

T	nominal size α		
	0.01	0.05	0.1
100	0.007	0.030	0.064
250	0.012	0.054	0.103
500	0.009	0.050	0.091
750	0.011	0.057	0.107
1000	0.009	0.053	0.103

Table 7: Power of the multiscale test for $n = 15$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$				(b) $b = 0.50$				(c) $b = 1.00$			
nominal size α				nominal size α				nominal size α			
T	0.01	0.05	0.1	T	0.01	0.05	0.1	T	0.01	0.05	0.1
100	0.013	0.059	0.106	100	0.051	0.175	0.274	100	0.012	0.059	0.116
250	0.108	0.229	0.326	250	0.770	0.893	0.935	250	1.000	1.000	1.000
500	0.509	0.721	0.801	500	1.000	1.000	1.000	500	1.000	1.000	1.000

- (a) We should consider different grids \mathcal{G}_T . In the simulation study, we use $\mathcal{G}_T = U_T \times H_T$ with

$$U_T = \left\{ u \in [0, 1] : u = \frac{5t}{T} \text{ for some } t \in \mathbb{N} \right\}$$

$$H_T = \left\{ h \in \left[\frac{\log T}{T}, \frac{1}{4} \right] : h = \frac{5t-3}{T} \text{ for some } t \in \mathbb{N} \right\}.$$

We could additionally consider a finer grid with, e.g.,

$$U_T = \left\{ u \in [0, 1] : u = \frac{t}{T} \text{ for some } t \in \mathbb{N} \right\}$$

$$H_T = \left\{ h \in \left[\frac{\log T}{T}, \frac{1}{4} \right] : h = \frac{t}{T} \text{ for some } t \in \mathbb{N} \right\}$$

and a sparser one with, e.g.,

$$U_T = \left\{ u \in [0, 1] : u = \frac{10t}{T} \text{ for some } t \in \mathbb{N} \right\}$$

$$H_T = \left\{ h \in \left[\frac{\log T}{T}, \frac{1}{4} \right] : h = \frac{10t-3}{T} \text{ for some } t \in \mathbb{N} \right\}.$$

I guess it is enough to consider only $n = 15$ and $T = 100$ for the robustness check. But I guess we should consider both the null and the alternative(s).

- (b) I don't know whether it is necessary to consider different tuning parameters for the estimation of the error variance. Maybe we could run the procedure

Table 8: Size of the multiscale test for $n = 25$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

T	nominal size α		
	0.01	0.05	0.1
100	0.006	0.031	0.065
250	0.013	0.053	0.102
500	0.011	0.063	0.114

Table 9: Power of the multiscale test for $n = 25$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$				(b) $b = 0.50$				(c) $b = 1.00$			
nominal size α				nominal size α				nominal size α			
T	0.01	0.05	0.1	T	0.01	0.05	0.1	T	0.01	0.05	0.1
100	0.010	0.052	0.106	100	0.035	0.137	0.246	100	0.009	0.043	0.094
250	0.082	0.193	0.285	250	0.725	0.873	0.920	250	1.000	1.000	1.000
500	0.448	0.675	0.768	500	0.999	1.000	1.000	500	1.000	1.000	1.000

with the true $\sigma^2 = \sigma_i^2$ as a benchmark. Ideally, we can then report that the results produced by the benchmark are very similar to those produced by the feasible algorithm with estimated σ_i^2 . I guess this should be enough.

- (c) One may compute the Gaussian quantile with different values of L . I guess the results should be very similar.
- (d) I think there is no need to try out different kernels K as from nonparametrics it is well known that the kernel is not so important.

6. *The multiscale tests are designed to control the FWER. Why is that the right criterion for the types of applications in mind (compared to e.g., FDR)? Given that other reasonable choices exist, additional motivation for this objective is warranted.*
7. *I'm wondering if there might be some clarification about the independence of ε_{it} across series i . In particular, suppose the intercepts α_i were instead considered random, like in mixed modeling (or Bayesian inference). Then marginally, the "new" errors ($\alpha_i + \varepsilon_{it}$) would be dependent across series i . Similar reasoning might apply to the covariates. From this perspective, the class of models might be considered more general.*
8. *It is claimed on p.6 that the mean function integrating to zero is "required" for*

Table 10: Size of the multiscale test for $n = 50$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

T	nominal size α		
	0.01	0.05	0.1
100	0.005	0.026	0.062
250	0.013	0.053	0.098
500	0.015	0.059	0.113

Table 11: Power of the multiscale test for $n = 50$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$				(b) $b = 0.50$				(c) $b = 1.00$			
nominal size α				nominal size α				nominal size α			
T	0.01	0.05	0.1	T	0.01	0.05	0.1	T	0.01	0.05	0.1
100	0.007	0.038	0.081	100	0.022	0.098	0.168	100	0.006	0.036	0.071
250	0.078	0.170	0.248	250	0.699	0.841	0.894	250	1.000	1.000	1.000
500	0.487	0.651	0.737	500	1.000	1.000	1.000	500	1.000	1.000	1.000

Table 12: Size of the multiscale test for $n = 100$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$).

T	nominal size α		
	0.01	0.05	0.1
100	0.005	0.029	0.065
250	0.017	0.068	0.126
500	0.019	0.068	0.132

Table 13: Power of the multiscale test for $n = 100$ for different sample sizes T and nominal sizes α for the dyadic grid ($\phi = 0.1$ and $\rho = 0.1$). Each panel corresponds to a different height parameter b of the bump function.

(a) $b = 0.25$				(b) $b = 0.50$				(c) $b = 1.00$			
nominal size α				nominal size α				nominal size α			
T	0.01	0.05	0.1	T	0.01	0.05	0.1	T	0.01	0.05	0.1
100	0.008	0.044	0.090	100	0.021	0.079	0.148	100	0.006	0.038	0.085
250	0.071	0.168	0.253	250	0.677	0.812	0.875	250	1.000	1.000	1.000
500	0.426	0.609	0.700	500	1.000	1.000	1.000	500	1.000	1.000	1.000

identification of the intercept. I think this is a sufficient, not necessary, condition, since others might suffice.

Reply to referee 2

Thank you very much for the careful reading of our manuscript and the interesting suggestions. We have addressed all your comments in the revision. Please see our replies to them below.

1. *Although you correctly cite Khismatullina and Vogt (2020, 2021) on which quite a bit of this new work seems to be based can you please summarize more in particular about the test proposed in Khismatullina and Vogt (2021, Journal of Econometrics), and explain where and how your proofs differ from that (e.g. by the complexity in needing to treat the covariates).*
2. *As your results are of asymptotic nature, it would be good to discuss limitations – even give an example where the procedure would cease to work.*
3. *Moreover, can you at least sketch out if by something like a Bootstrap procedure (cf. Zhang et al, 2012) more of the “asymptotic flavour” of your test/cluster procedure could be remedied?*
4. *I am having a slight (finite sample) identification concern with (not only your) model(s) mixing deterministic (nonparametric) trends with covariate (and also error) structure which is allowed to be positively serially dependent, e.g. autoregressive (as in your examples): I think that for “any” fixed sample size it might always occur that the trajectory of a stochastic trend, an autoregressive process with roots relatively close to the unit circle, say, cannot be distinguished from the deterministic trend. Wouldn’t that be potentially a problem for your (and any related) test procedure? As a follow-up on this, wouldn’t you need (or to say it differently, wouldn’t it be perhaps beneficial to add) some extra conditions on the nature of your covariates (and potentially also your errors ε ?) to avoid this problem?*
5. *What about a naive competitor that is just based on the second derivative (= change of the slope parameters) rather than the distance based on the curves and the first derivative (as in your local linear estimator)? I believe that this could also work rather well on your economic example data in Figures 3–6? Maybe you can “benchmark” your procedure against such a simple competitor (as such a comparison is somewhat missing explicitly – although you orally compare sometimes with Zhang et al (2012)).*

6. *Can your proposed test procedure be considered somehow to be equivalent to constructing a uniform confidence region where you would need to control if two (or more curves) are within the same tube (not just pointwise)? If so, it would be perhaps interesting to explain the link, and why for your test procedure it is sufficient/adequate to control the “familywise” error (does this correspond to what one does for a “uniform” region?).*
7. *How in all of this does the number of curves (larger than two) play a role, in practice, for correctly calibrating your test (as least asymptotically as possible)? On the other hand, do your results reflect the fact that obviously they depend on the number of time series (or rather the number of series where trends are different, a number that you would have access to in an oracle situation)?*
8. *Here is a small series of remarks towards needing to choose (u, h) – an example for a practical choice is given in Section 7, only (a bit late): Your localised multiscale method requires to discretize the continuous (u, h) . I am wondering if the way to do this plays a role for the properties of the resulting practical procedure. Can you please also compare with wavelet-based multiscale methods which are based somehow on a “built-in” way of choosing the location-scale parameters (u, h) ?*
9. *page 12, around equation (3.6): it took me a moment to understand that you are talking about the standard local linear estimator (of Fan and Gijbels) here, you might want to make this clearer.*
10. *I understand the heuristics behind using the Gaussian version (3.12) of the test statistics in the “idealised” situation but what about the “non-idealised” situation of unknown variances σ^2 and unknown parameters β ? Is the Gaussian-based MC simulation method still valid when you need to estimate those parameters?*
11. *Again about the choice of (u, h) : what happens with expressions (such as in equation (4.1)) which depend on $\max_{(u, h)}$ in practice where you have to discretize this (u, h) ? I do not think that a maximum over a continuous location-scale parameter can be treated the same way as one over a discrete one? Does the choice of the grid \mathcal{G}_T influence the results here, don’t you need some (additional) conditions on the grid (its spacing etc)? This refers, e.g. to the simulation section 6, page 24 where – in passing – you might want to change the strange wording there where you say “for some t in N ” and rather detail the specification of the grid in t here as you do later in Section 7.*
12. *Section 7, page 41, lines 45-50: can you develop this conjecture a bit?*

13. *You might want to add a Conclusion Section which could both serve to recall the difficulties encountered in treating the more general situation of more than two curves and the presence of covariates, and also discuss some of the aforementioned points on Bootstrap alternatives or on potential competitors.*
14. *Develop more to which extent the second data application (in the Supplement) brings insights beyond the one of the first (and why you chose to present the first and not the second in the main body of the text).*
15. *Supplement section page 15, line 49 – a notational detail: should the first o_p be O_p if the $\rho_T = o(1/\log(T))$ or vice versa?*
16. *It would be good to explain somewhere in the main body (Section 3 or 4?) the additional difficulties in proving the results in the presence of the covariates.*

References

- KHISMATULLINA, M. and VOGT, M. (2020). Multiscale inference and long-run variance estimation in non-parametric regression with time series errors. *Journal of the Royal Statistical Society: Series B*, **82** 5–37.