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) Following the suggestions of the referee, we could proceed as follows:

- We consider the application reported in the paper (Section 7).
- We choose n to match the value in the application, i.e., we set $n = 8$, (or $n = 10$ to get a round number) and choose different values for T (one of them very close to the time series length in the application, e.g., $T = 100$).
- We assume that each covariate process $\{X_{it,j}\}$ $(j = 1, \ldots, 4)$ is an AR(1) process of the form $X_{it,j} = a_j X_{i(t-1),j} + \nu_{it,j}$ and estimate the AR parameters a_j as well as the innovation variances $\sigma_j^2 = \mathbb{E}[\nu_{it,j}^2]$ from the data. (Note: As the estimated a_j and σ_j^2 may differ across time series i, we simply take their average over i: $a_j = (a_{1j} + a_{2j} + \ldots + a_{8j})/8$ and $\sigma_j^2 = (\sigma_{1j}^2 + \sigma_{2j}^2 + \ldots + \sigma_{8j}^2)/8$. The estimated values are presented in Tables 1 and 2.) Assuming that the innovations $\nu_{it,j}$ are normally distributed and independent across j, we can then easily simulate sample paths of the covariates.
- We assume that the errors ε_{it} follow the AR(1) model $\varepsilon_{it} = a\varepsilon_{i,t-1} + \eta_{it}$, where we set a and the innovation variance $\mathbb{E}[\eta_{it}^2]$ equal to the estimated values (averaged over i) from the application: see the last column in Tables 1 and 2.
- We let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a normally distributed random vector. In particular, $\alpha \sim N(0, c\Sigma)$ with

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

Country	GDP	Population	Long-term rate	Inflation	Error process
AUS	0.971	0.985	0.932	0.746	0.786
BEL	0.939	0.966	0.927	0.666	0.920
DNK	0.978	0.992	0.964	0.865	0.915
FRA	0.971	0.975	0.919	0.948	0.943
NLD	0.933	0.993	0.892	0.735	0.927
NOR	0.988	0.983	0.953	0.754	0.930
SWE	0.991	0.988	0.966	0.726	0.914
USA	0.959	0.989	0.931	0.670	0.845
mean	0.966	0.984	0.935	0.764	0.897

Table 1: Estimated a_{ij} for the house prices application.

Country	GDP	Population	Long-term rate	Inflation	Error process
AUS	0.034	0.003	0.598	0.010	0.080
BEL	0.064	0.002	0.528	0.013	0.087
DNK	0.037	0.001	0.889	0.008	0.049
FRA	0.063	0.004	0.632	0.007	0.065
NLD	0.060	0.001	0.482	0.008	0.061
NOR	0.035	0.001	0.561	0.009	0.070
SWE	0.035	0.001	0.653	0.012	0.071
USA	0.043	0.001	0.576	0.010	0.076
mean	0.048	0.002	0.626	0.010	0.071

Table 2: Estimated σ_{ij} (square root of the variance) for the house prices application.

	nominal size α		
T	0.01	0.05	0.1
100	0.528	0.619	0.678
250	0.999	1.000	1.000

Table 3: Size of the multiscale test for different sample sizes T and nominal sizes α.

where $\rho = 0.1$ (or a bit larger, e.g., $\rho = 0.25$) gives the correlation across time series i. To start with, we can set $c = 1$ and then adjust it if needed. I have run the size computations with α drawn from a multivariate normal distribution with $\rho = 0.1$, but with more "reasonable" value of a_j, i.e. $a_j = 0.25$ for all j, and $\mathbb{E}[\nu_{it,j}^2] = 1$ and the actual size is fairly close to the nominal target 0.05. I do not include the results here since we first need to decide what to do with the covariates.

- To generate data under the null $H_0 : m_1 = \ldots = m_n$, we let $m_i = 0$ for all i as before. To produce data under the alternative, we define $m_1(u) = b\,(u - 0.5)$ with different values of b as before. (Maybe we can fit the green curve in Figure 8 by a linear function and take the estimated slope value as one of the b-values.)
- We take the grid $\mathcal{G}_T$ to be the same as in the application: $\mathcal{G}_T = U_T \times H_T$, where $U_T = \left\{u \in [0,1] : u = \frac{t}{T} \text{ for some } t \in \mathbb{N}\right\}$ and $H_T = \left\{h \in \left[\frac{\log T}{T}, \frac{1}{4}\right] : h = \frac{5t-3}{T} \text{ for some } t \in \mathbb{N}\right\}$. We thus take into account all locations u on an equidistant grid U_T with step length $1/T$ and all scales $h = 2/T, 7/T, 12/T, \ldots$ with $\log T/T \leq h \leq 1/4$. This implies that each interval $\mathcal{I} = [u - h, u + h]$ with $(u, h) \in \mathcal{G}_T$ spans $5, 15, 25, \ldots$ years.

The results of the size computations are presented in Table 3. The actual size is not even close to the target nominal size, and the main reason for that is very poor estimates of the parameters $\boldsymbol{\beta}_i$ (especially $\beta_{i,2}$ and $\beta_{i,4}$ responsible for the population and for the inflation) that get marginally better with increase of the sample size T: see Figures 1 - 8.

(iii) Maybe, we can use the test from Degras et al. (2012, Testing for parallelism among trends in multiple time series) as a competitor because it should be fairly straightforward to implement the test. (One needs to compute kernel estimates and from these the test statistic $\hat{\Delta}_{n,T}$ in (10), then one needs to compare this statistic with a normal distribution (see (20)) or with a Gaussian version of the statistic (similar as in our approach, see Section IV, C).)

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 Degras et

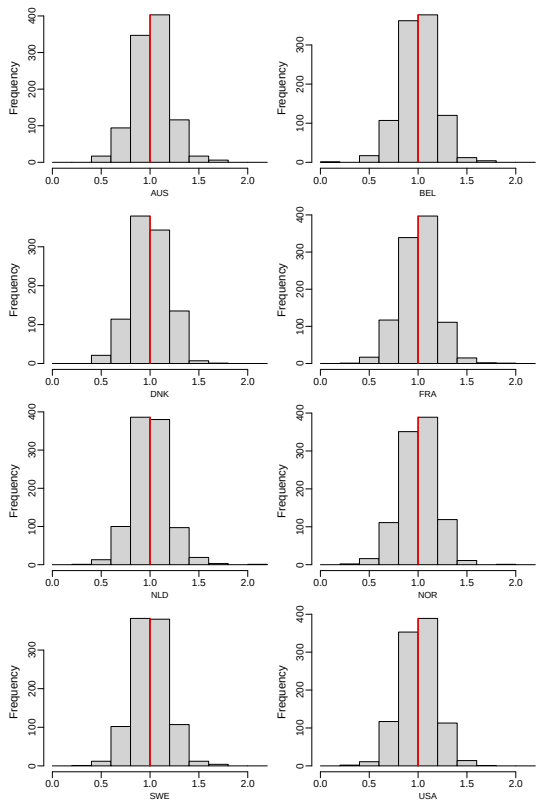

Figure 1: Histogram of the estimated value of $\beta_{i,1}$ (GDP) for various countries for $T = 100$.

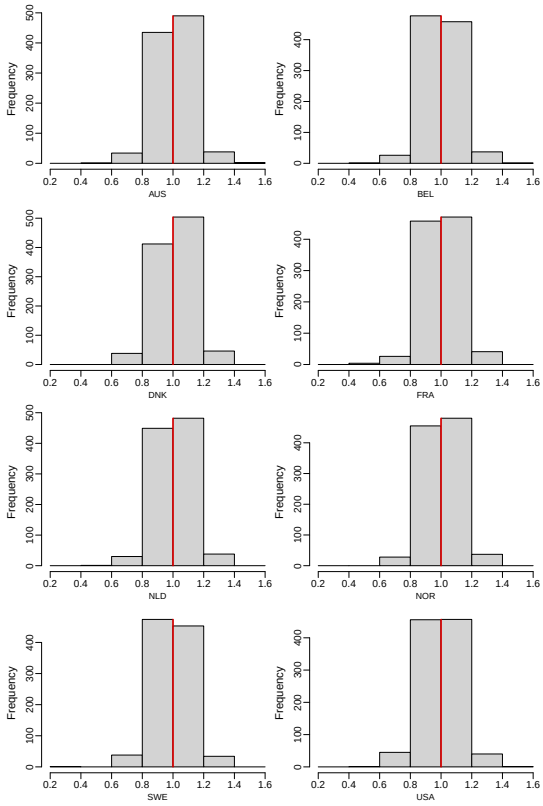

Figure 2: Histogram of the estimated value of $\beta_{i,1}$ (GDP) for various countries for $T = 250$.

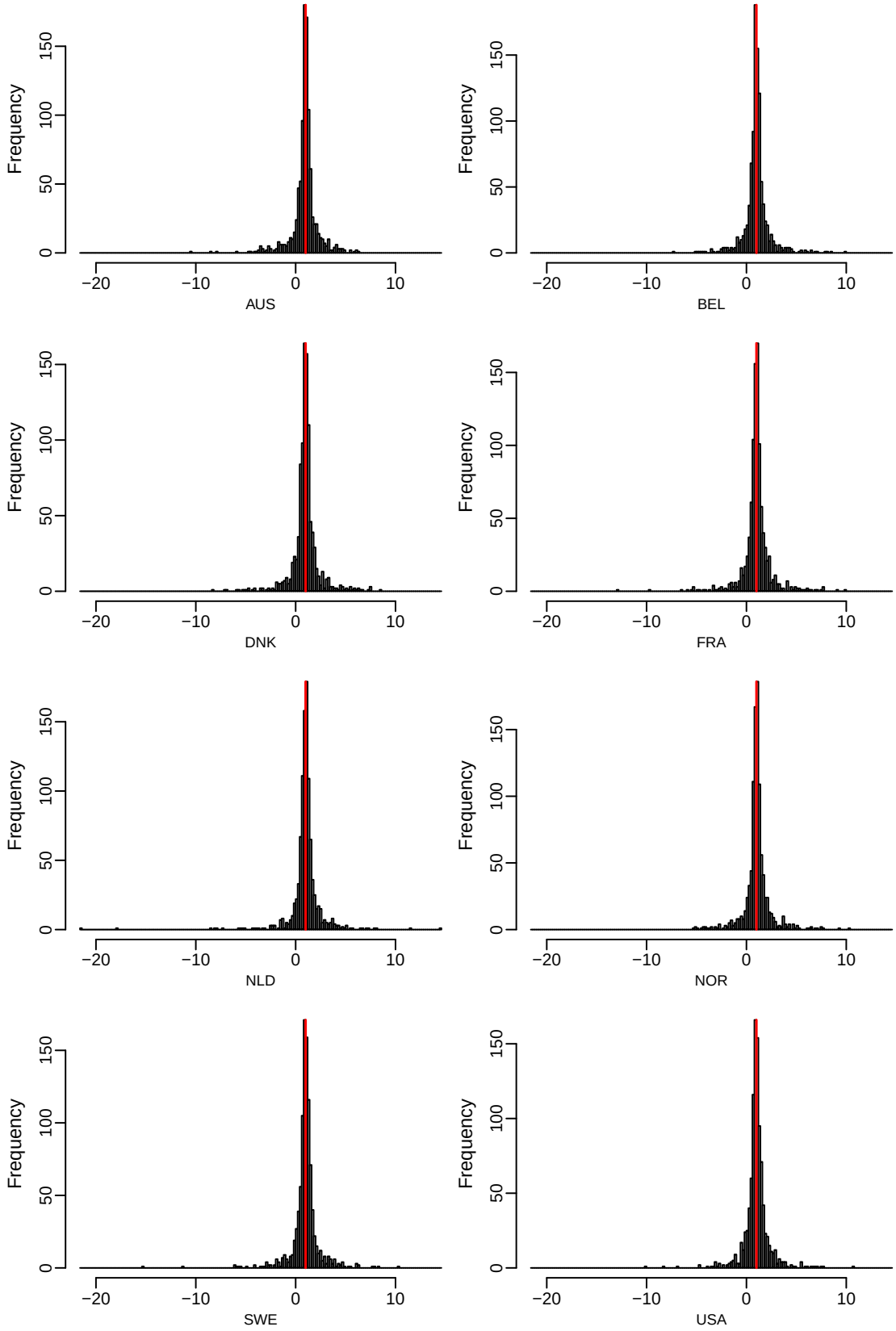

Figure 3: Histogram of the estimated value of $\beta_{i,2}$ (population) for various countries for $T = 100$.

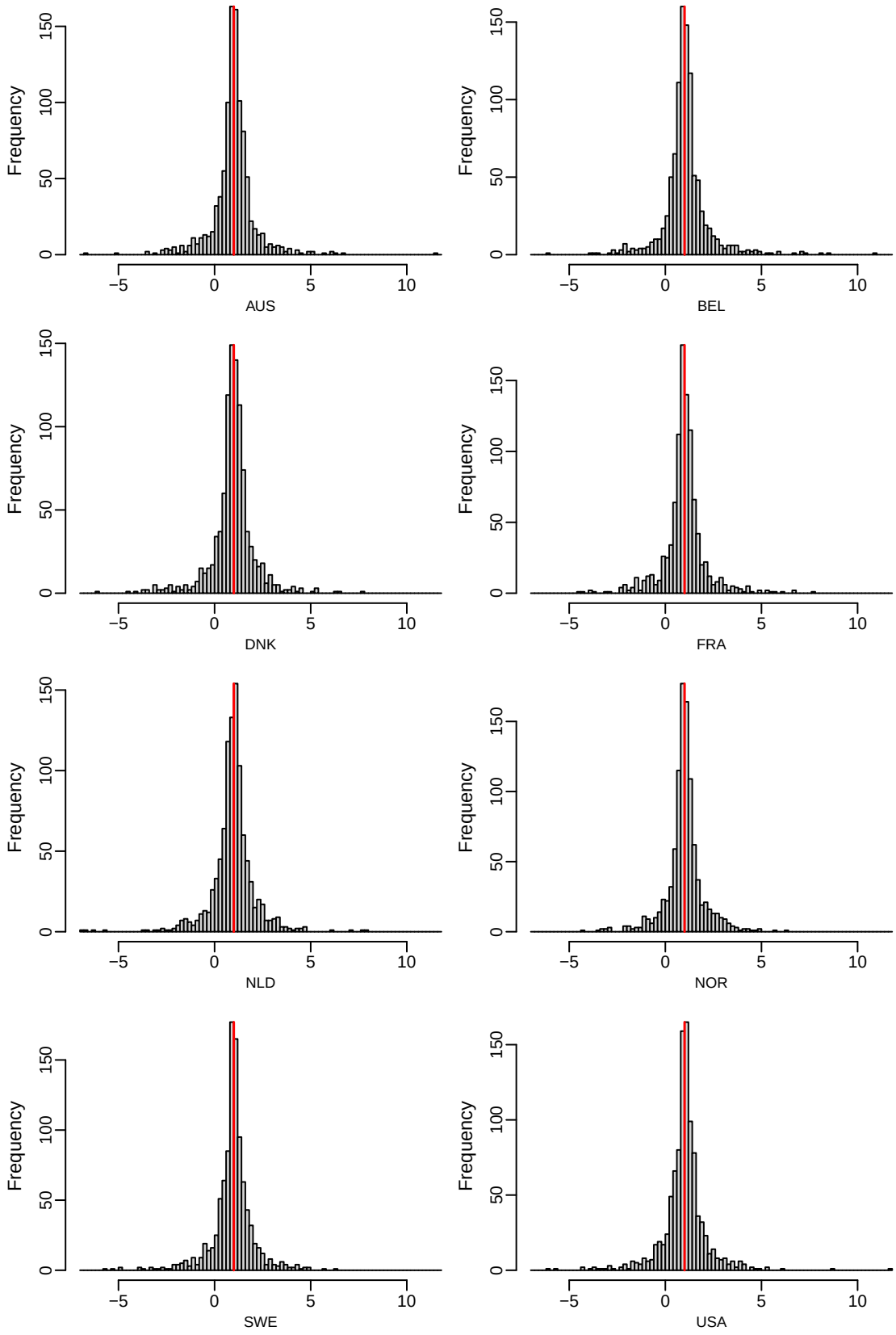

Figure 4: Histogram of the estimated value of $\beta_{i,2}$ (population) for various countries for $T = 250$.

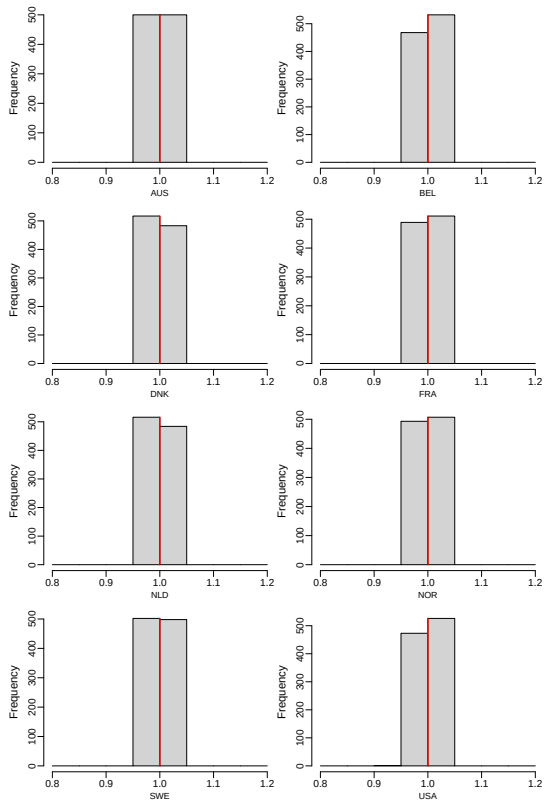

Figure 5: Histogram of the estimated value of $\beta_{i,3}$ (long-term interest rate) for various countries for $T = 100$.

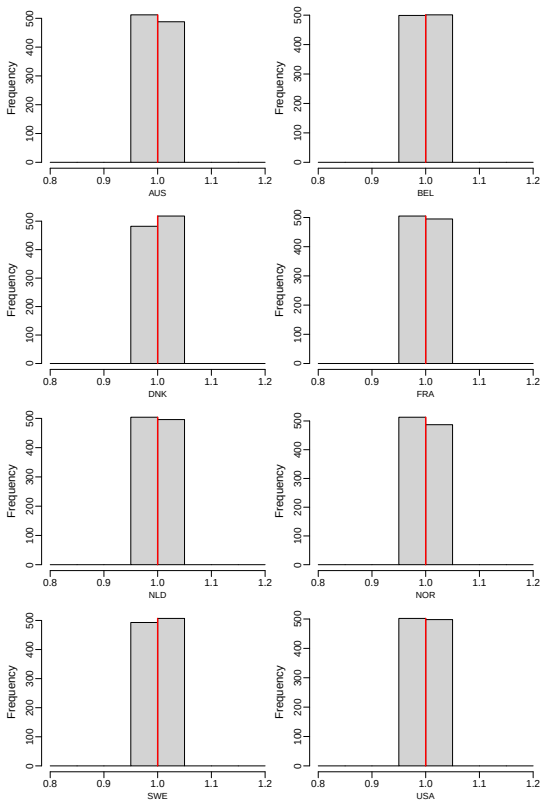

Figure 6: Histogram of the estimated value of $\beta_{i,3}$ (long-term interest rate) for various countries for $T = 250$.

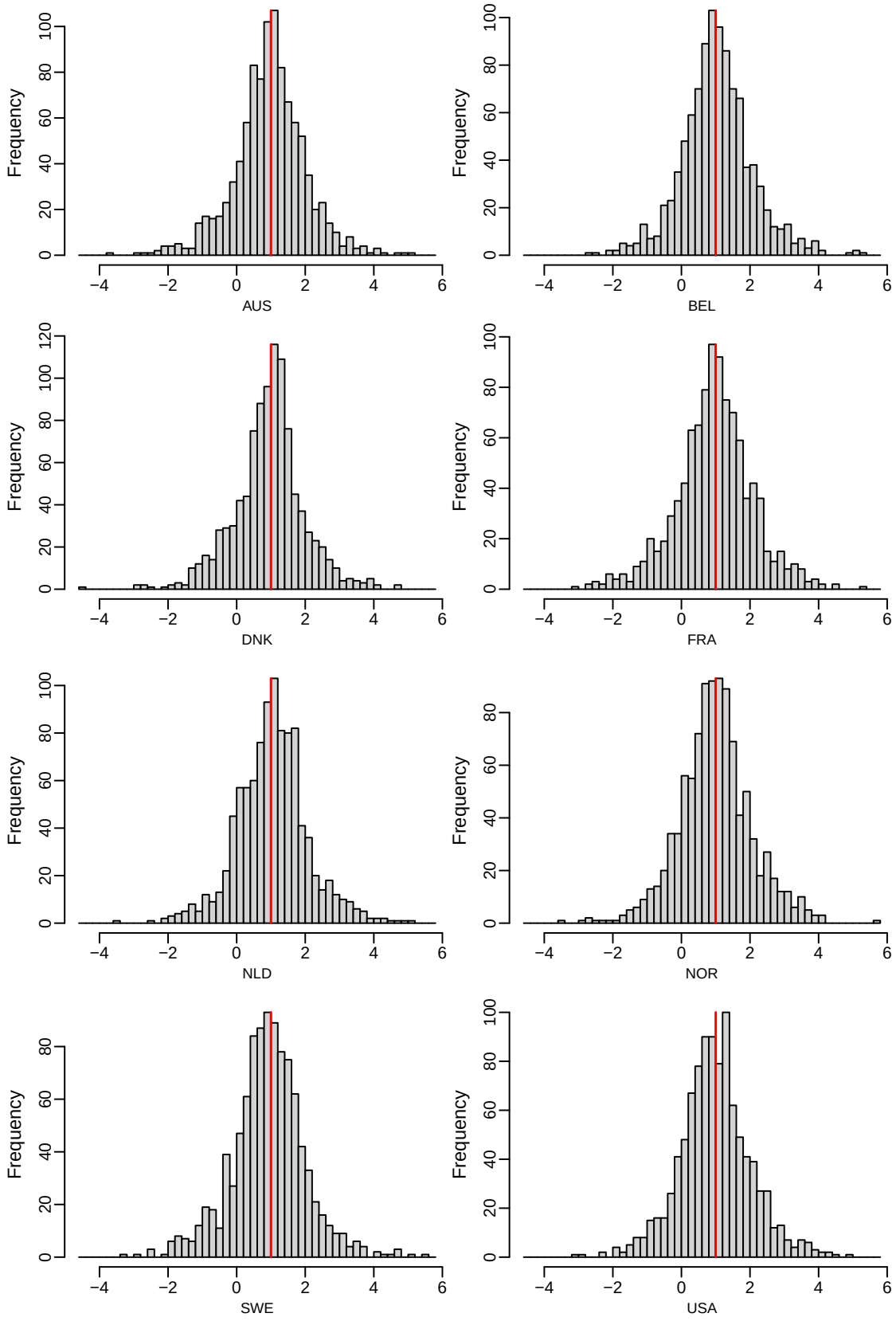

Figure 7: Histogram of the estimated value of $\beta_{i,4}$ (inflation) for various countries for $T = 100$.

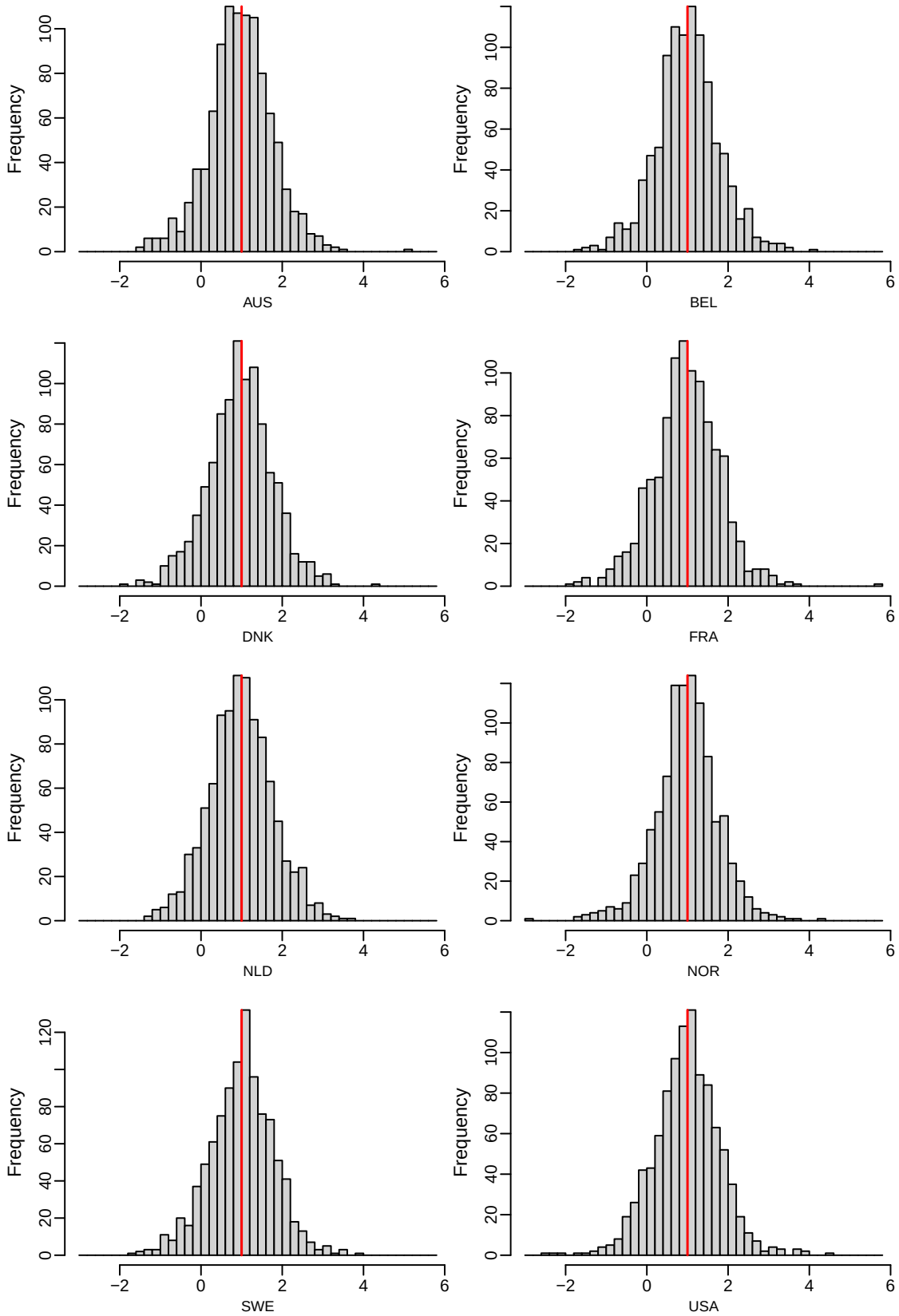

Figure 8: Histogram of the estimated value of $\beta_{i,4}$ (inflation) for various countries for $T = 250$.

al.). I'd suggest to add the comparison to the supplement rather than to the main body of the text.

(iv) We could re-run the simulations from (i) with a much smaller grid than the one from the application. This hopefully allows us to deal not only with $n = 8$ / $n = 10$ but also with larger values of n (say $n = 25$, $n = 50$ or even $n = 100$). As a suggestion, we may take the grid $\mathcal{G}_T$ to be a dyadic scheme (as in Wavelet analysis) with scales in the set

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

where

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

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

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

and

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

I think it would be enough to re-run the simulations with $T = 150$ (which approximates the time series length in the application) and different values of n, say, $n = 10$ (or $n = 8$), $n = 25$, $n = 50$ (and maybe even $n = 100$). With $T = 150$ and $n = 50$, we would get

$$\#\Big(\mathcal{G}_T \cup \{(i,j) : 1 \leq i < j \leq n\}\Big) \approx 27500$$

(provided that I did not make any errors in my calculation ...). Is this still feasible for the simulations???

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.

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:

   (a) We should consider different grids $\mathcal{G}_T$. I think using the grid from the application + the dyadic scheme discussed above should be enough.
   (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 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 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.*

Bibliography