Review report of JBES manuscript P-2022-0567 Multiscale Comparison of Nonparametric Trend Curves

Summary: This paper is a very well written detailed theoretical development of a multiscale test and subsequent clustering method for nonparametrically comparing trends of multiple time series. The method is designed to provide the intervals in time along which the curves can be considered to have significantly differences in their trends over time. These can then be used to partitioning the collection of observed time series into clusters of similar trends. The sound theoretical asymptotic results on control of the level, power (against a class of local alternatives), and clustering error are completed by a simulation study and applications to two real data sets (on to be found in the Supplement, as well as all the proofs).

Evaluation:

The paper is really quite nicely written, both concerning the economical and statistical motivation and the presentation of the rather technical asymptotic results. If I give in the following a list of questions/comments that I would like to be addressed, then this is mainly in order to increase the amount of insights for the general JBES reader:

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)?

Minor questions and comments:

1. 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.

2. 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?

3. 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 $\mathcal{G}_{\mathcal{T}}$ 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.

4. Section 7, page 41, lines 45-50 : can you develop this conjecture a bit?

5. 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.

6. 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).

7. 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?

8. 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.