Revision of the paper
"Multiscale inference and long-run variance estimation in nonparametric regression with time series errors"

First of all, we would like to thank the editor, the associate editor and the reviewers for their many comments and suggestions which were very helpful in improving the paper. In the revision, we have addressed all comments and have rewritten the paper accordingly. Please find our point-by-point responses below. Before we reply to the specific comments of the referees, we summarize the major changes in the revision.

Generalization of the theoretical results. We have extended the theoretical results as suggested by Referee 1:

(i) We have derived the following consistency result in addition to Proposition 3.3: Let the significance level $\alpha = \alpha_T \in (0,1)$ depend on the sample size T. If $\alpha_T \to 0$, then $\mathbb{P}(E_T^{\ell}) \to 1$.

(ii) We have generalized our estimator of the long-run error variance. The estimation procedure is shown to be valid not only for AR(p) processes of known finite order p but for any stationary error process $\{\varepsilon_t\}$ with an AR(∞) representation. This greatly extends the applicability of the estimator.

Comparison to SiZer. As requested by the Associate Editor, we give a clear account of the main contributions and innovations of our paper relative to the SiZer approach in the revision. Please see the new Section ?? for the details. In what follows, we give a slightly rephrased and condensed version of the new Section ??.
Informally speaking, both our approach and SiZer for dependent data (dependent SiZer for short) are methods to test for local increases/decreases of a nonparametric trend function m. The formal problem is to test the hypothesis

$$H_0(u,h) : \text{The trend } m \text{ is constant on the time interval } [u-h, u+h]$$

simultaneously for a large number of different time intervals $[u-h, u+h]$, in particular, for all intervals with $u \in I$ and $h \in H$, where I is the set of locations and H the set of bandwidths or scales H under consideration.
Let $s_T(u,h)$ be the SiZer statistic to test $H_0(u,h)$, which corresponds to the statistic $\widehat{\psi}_T(u,h)$ in our approach and which is properly defined in Section ??. There are two versions of dependent SiZer:

(a) The global version aggegrates the individual statistics $s_T(u,h)$ into the overall statistic $S_T = \max_{h \in H} S_T(h)$, where $S_T(h) = \max_{u \in I} |s_T(u,h)|$. The statistic S_T is the counterpart to the multiscale statistic $\widehat{\Psi}_T$ in our approach.

(b) The row-wise SiZer version considers each scale $h \in H$ separately. In particular, for each bandwidth $h \in H$, a test is carried out based on the statistic $S_T(h)$.

In practice, SiZer is commonly implemented in its row-wise form. The main reason is that it has more power than the global version by construction. However, this gain of power comes at a cost: Row-wise SiZer carries out a test *separately* for each scale $h \in H$, thus ignoring the simultaneous test problem across scales h. Hence, it is not a rigorous level-α-test of the overall null hypothesis H_0. For this reason, we focus on global SiZer in what follows.

Even though related, our methods and theory are markedly different from those of the SiZer approach:

(i) Theory for SiZer is derived under the assumption that the set of bandwidths H is a compact subset of $(0, 1)$. As already pointed out in Chaudhuri and Marron (2000), this is a quite substantial/severe restriction: Only bandwidths h are taken into account that remain bounded away from zero as the sample size T grows. Bandwidths h that converge to zero as T increases are excluded. As Chaudhuri and Marron (2000) put it (on p.420):

> *Note that all the weak convergence results in this section have been established under the assumption that both H and I are fixed compact subintervals of $(0, \infty)$ and $(-\infty, \infty)$ respectively. Compactness of the set $H \times I$ enables us to exploit standard results on weak convergence of a sequence of probability measures on a space of continuous functions defined on a common compact metric space. However, conventional asymptotics for nonparametric curve estimates allows the smoothing parameter h to shrink with growing sample size. There frequently one assumes that h_n is of the order $n^{-\gamma}$ for some appropriate choice of $0 < \gamma < 1$ so that the estimate $\hat{f}_{h_n}(x)$ converges to the "true function" $f(x)$ at an "optimal rate". This makes one wonder about the asymptotic behaviour of the empirical scale space surface when h varies in $H_n = [an^{-\gamma}, b]$, where $a, b > 0$ are fixed constants. Extension of our weak convergence results along that direction will be quite interesting, and we leave it as a challenging open problem here.*

The theory of our paper allows to deal with this problem. In particular, it allows to simultaneously consider scales h that remain bounded away from zero and scales $h = h_T$ that converge to zero at various different rates $T^{-\gamma}$. To achieve this, we come up with a proof strategy which is very different from that in the SiZer literature: As proven in Chaudhuri and Marron (2000) for the i.i.d. case and in Park et al. (2009) for the dependent data case, S_T weakly converges to some limit process S under the overall null hypothesis H_0. This is the central technical result on which the theoretical properties of SiZer are based. In contrast to this, our proof strategy does not even require the statistic $\widehat{\Psi}_T$ to have a weak limit and is thus not restricted by the limitations of classic weak convergence theory.

(ii) There are different ways to combine the test statistics $S_T(h) = \max_{u \in I} |s_T(u, h)|$ for different scales $h \in H$. One way is to take their maximum, which leads to the

SiZer statistic $S_T = \max_{h \in H} S_T(h)$. We could proceed analogously and consider the multiscale statistic $\widehat{\Psi}_{T,\text{uncorrected}} = \max_{h \in H} \widehat{\Psi}_T(h) = \max_{(u,h) \in I \times H} |\widehat{\psi}_T(u,h)/\widehat{\sigma}|$. However, as argued in Dümbgen and Spokoiny (2001) and as discussed in Section ?? of our paper, this aggregation scheme is not optimal when the set $H = H_T$ contains scales h of many different rates. Following the lead of Dümbgen and Spokoiny (2001), we consider the test statistic $\widehat{\Psi}_T = \max_{(u,h) \in I \times H} \{|\widehat{\psi}_T(u,h)/\widehat{\sigma}| - \lambda(h)\}$ with the additive correction terms $\lambda(h)$. Hence, even though related, our multiscale test statistic $\widehat{\Psi}_T$ differs from the SiZer statistic S_T in important ways.

(iii) The main complication in carrying out both our multiscale test and SiZer is to determine the critical values, that is, the quantiles of the test statistics $\widehat{\Psi}_T$ and S_T under H_0. In order to approximate the quantiles, we proceed quite differently than in the SiZer literature. The quantiles of the SiZer statistic S_T can be approximated by those of the weak limit process S. Usually, however, the quantiles of S cannot be determined analytically but have to be approximated themselves (e.g. by the bootstrap procedures of Chaudhuri and Marron (1999, 2000)). Alternatively, the quantiles of S_T can be approximated by procedures based on extreme value theory (as proposed in Hannig and Marron (2006) and Park et al. (2009)). In our approach, the quantiles of $\widehat{\Psi}_T$ under H_0 are approximated by those of a suitably constructed Gaussian analogue of $\widehat{\Psi}_T$. It is far from obvious that this Gaussian approximation is valid when the data are dependent. To see this, deep strong approximation theory for dependent data (as derived in Berkes et al. (2014)) is needed. It is important to note that our Gaussian approximation procedure is not the same as the bootstrap procedures proposed in Chaudhuri and Marron (1999, 2000). Both procedures can of course be regarded as resampling methods. However, the resampling is done in a quite different way in our case.

We hope the above points make clear that the methodological and theoretical contributions of our paper are quite substantial relative to the SiZer methodology.

Simulations and application examples. We have thoroughly revised the simulation study in Section 5. In order to take into account the many suggestions of Referees 1 and 2, we have completely re-designed the size and power simulations for our multiscale test (formerly Section 5.1) and the comparison with SiZer (formerly Section 5.2), which are combined in the new Section 5.1. Among other things, we consider different trend signals for our power simulations as suggested by Referee 2 and AR error terms with strong autocorrelation (AR(1) errors with parameter ± 0.9) as suggested by Referee 1. Note that we have removed the simulation setup with AR(2) errors which mimics the situation in the application example in order to keep the simulation study to a reasonable length. Finally, we have added a second application example to global temperature data as requested by Referee 1.

Reply to Referee 1

Thank you very much for the constructive and helpful comments. In our revision, we have addressed all of them. Please see our replies to your comments below.

(1) *A consistency result of Proposition 3.3.*

I believe that the following type of result can be obtained: $\mathbb{P}(E_T^\ell) \to 1$. *Theorem 3.1 is for testing purpose. In certain application one might be interested in such consistency result. Basically one needs to study the behavior of* $q_T(\alpha)$ *when* $\alpha \to 0$.

Under our regularity conditions, it can indeed be proven that $\mathbb{P}(E_T^\ell) \to 1$ as $\alpha = \alpha_T \to 1$. We have added this consistency result as Corollary ?? to the paper. The proof is provided in the Supplementary Material.

(2) *Estimation of long run variance using autoregressive processes.*

The authors considered estimating σ^2 *using AR processes. A limitation is that the order p is fixed and finite. It appears that the latter limitation can be relaxed. For a stationary process* ε_t *(not necessarily linear), one can fit an AR process with large p*

$$\varepsilon_t = \sum_{j=1}^{p} a_j \varepsilon_{t-j} + \eta_t,$$

properties of fitted $\widehat{a}_1, \dots, \widehat{a}_p$ *can be obtained from the results in the following papers: Wu and Pourahmadi (2009) and Xiao and Wu (2012). A similar version of the authors' estimate (4.14) can be used. Rate of convergence (cf. Proposition 4.1) can be derived with rate* $T^{-1/2}$ *therein possibly replaced by a larger term of the form* T^{-c} *with* $c < 1/2$.

Many thanks for this interesting suggestion. We have generalized our procedure for estimating the long-run variance σ^2 along the lines suggested by you: Rather than considering AR(p) processes of known finite order p, we consider the much more general case of AR(∞) processes. We thus assume the error process $\{\varepsilon_t\}$ to have the form

$$\varepsilon_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j} + \eta_t, \qquad (*)$$

where $a_1, a_2, a_3, \dots$ are unknown parameters and η_t are i.i.d. innovations (which fulfill certain regularity conditions as detailed in the revised Section 4). Notably, the model $(*)$ for the error terms nests AR(p^*) processes of any finite (unknown) order p^* as a special case: If $a_{p^*} = 0$ and $a_j = 0$ for all $j > p^*$, then $\{\varepsilon_t\}$ is an AR process of order p^*.

In order to generalize our theory to the case that $\{\varepsilon_t\}$ has an AR(∞) representation of the form $(*)$, we proceed as follows: We fit AR(p) processes to the data whose order $p = p_T$ grows with the sample size T. We thus approximate the AR(∞) process $\{\varepsilon_t\}$ by a sequence of AR(p) processes whose order $p = p_T$ goes to infinity. This is somewhat different but related to the banding techniques developed in Wu and Pourahmadi (2009) and Xiao and Wu (2012). As you mention yourself in your comment, the resulting estimators are not $\sqrt{T}$-consistent any more. Their convergence rate is somewhat slower than $1/\sqrt{T}$, in particular, it is of the form $p^c/\sqrt{T}$ with some $c > 0$. Please see the revised Proposition 4.1 and its proof in the Supplementary Material for the details.

(3) *Real data application.*

The authors analyzed the yearly mean Central England temperature data. It will be interesting to apply their approach to the global temperature data. In the paper Wu et al. (2001), an increasing trend function is fitted. It will be important to know which period the sequence in increasing/decreasing.

(4) *Simulation Study.*

In the simulation study, the authors considered AR(1) processes with relatively weaker dependence: $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$. *One should consider the stronger positive/negative dependence case with* $a = \pm 0.9$ *(say). How does the strength of dependence affect the performance of the procedure?*

We have added the AR(1) case with $a = \pm 0.9$ to our revised simulation study in Section ??. In particular, we have carried out additional size simulations for the case that $a = \pm 0.9$. The results are reported in Table 2 on p.?? and can be summarized as follows: For the negative parameter $a = -0.9$, the size of the multiscale test is close to the nominal target α for sample sizes $T \geq 1000$. However, for the smaller sample sizes $T = 250$ and $T = 500$ there are substantial size distortions. In the case of $a = 0.9$, a similar picture arises, however, with somewhat smaller size distortions. Hence, when the error terms are strongly (positively or negatively) autocorrelated, our test method has good size properties only for sufficiently large sample sizes. In our opinion, this is not so surprising: Statistical inference in the presence of strongly autocorrelated data is a very difficult problem in general and satisfying results can only be expected for sufficiently large sample sizes.

Reply to Referee 2

Thank you very much for the constructive and useful suggestions. In our revision, we have addressed all of them. In particular, we have thoroughly revised the simulation study which compares our multiscale test with SiZer according to your suggestions. Here are our point-by-point responses to your comments.

(1) *Section 3.2: The authors recommend computing the quantiles for the independent Gaussian case by simulation. This suggestion is already in the original SiZer paper (Chaudhuri and Marron, 1999). However in the late 1990s computing power was not sufficient to make this suggestion feasible. This led to the use of approximation such as in Hannig and Marron (2006). I would like to ask how does the simulation based quantile compare to the approximation in Hannig and Marron (2006).*

In the revised simulation study of Section ??, we consider a row-wise version $\mathcal{T}_{\text{RW}}$ of our multiscale test, which is the direct counterpart to the row-wise dependent SiZer method $\mathcal{T}_{\text{SiZer}}$ developed in Park et al. (2004), Rondonotti et al. (2007) and Park et al. (2009). The quantiles of $\mathcal{T}_{\text{RW}}$ under the null are approximated by our simulation-based procedure, whereas the quantiles of $\mathcal{T}_{\text{SiZer}}$ are approximated by the method of Hannig and Marron (2006) and Park et al. (2009) which is based on extreme value theory. In our simulation study, we compare the size and power properties of $\mathcal{T}_{\text{RW}}$ and $\mathcal{T}_{\text{SiZer}}$. This shows how our simulation-based procedure compares to the approximation in Hannig and Marron (2006) and Park et al. (2009).

At this point, we would like to add two remarks on our simulation-based procedure to compute the quantiles of the multiscale test under the null:

- As already mentioned when summarizing the main differences between our multiscale approach and SiZer at the beginning of this letter, the Gaussian approximation procedure that we use for simulating the quantiles of the multiscale statistic under the null is not the same as the bootstrap procedures proposed in Chaudhuri and Marron (1999, 2000). Both procedures are of course resampling methods. However,the resampling is done in a quite different way.
- It far from clear that our simulation-based procedure is theoretically valid and provides an adequate critical value such that our test has asymptotically the correct size under the null. One of the main theoretical contributions of the paper is that this is indeed the case.

(2) *Page 12, line 15: What is random here? After a spending some time I believe that it is the* Π_T *but on first reading I thought* E_T*s are non-random. Please explain these various objects better.*

We have rewritten the text concerning the objects Π_T^ℓ and E_T^ℓ for $\ell \in \{\pm, +, -\}$ as suggested. We have in particular attempted to explain these objects in a clearer way and to clarify the question what is random here. (The collection of intervals Π_T^ℓ is indeed random.) Please see p.?? for the details. We hope the new exposition of the material makes things clearer.

(3) *Page 18, line 52: Please remove the speculative statements about what can be shown unless you actually show it in this paper.*

We have removed the speculative statements following Proposition 4.1.

(4) *Section 4.1: This section does not contain any truly new material and should be removed.*

We have removed Section 4.1 from the paper. Moreover, we have thoroughly revised Section 4 on the estimation of the long-run error variance according to the suggestions of Referee 1. Please see our reply to comment (2) of Referee 1 for the details.

(5) *Section 5.2: I understand that you are doing comparisons to SiZer out of the box. However, some of the comparison might not be quite fair. SiZer is adjusting multiplicity row-wise while the proposed method is attempting a global multiple control. What would happen if your G_T only focused on one scale?*

We have thoroughly revised the comparison study with SiZer, attempting to give a better and fairer picture of the methods. The revised study compares the following versions of our multiscale test and SiZer:

- our multiscale test $\mathcal{T}_{\text{MS}}$ as defined in Section ?? of the paper
- an uncorrected version $\mathcal{T}_{\text{UC}}$ of our multiscale test without the additive correction terms $\lambda(h)$
- a row-wise version of $\mathcal{T}_{\text{RW}}$ of our multiscale test
- the row-wise dependent SiZer $\mathcal{T}_{\text{SiZer}}$ of Park et al. (2004), Rondonotti et al. (2007) and Park et al. (2009).

We have analyzed the size and power properties of the four methods (not only the global but also the row-wise size and power properties to be fairer with regard to SiZer). As already mentioned in our answer to your comment (1), the row-wise version $\mathcal{T}_{\text{RW}}$ of our multiscale test focuses on one scale at a time and is thus the direct counterpart to $\mathcal{T}_{\text{SiZer}}$. Please see Section ?? for the full simulation study, which in particular shows how $\mathcal{T}_{\text{RW}}$ and $\mathcal{T}_{\text{SiZer}}$ compare to each other. We hope you find the revised study more accurate.

(6) *Page 25, line 1-26: I do not quite understand this figure. Would it be possible to rather reproduce the colorful SiZer figures that show the results of the test at various scales and locations? Also you should use several different signals. I believe that a single relatively large bump is not sufficient test bed. A good collection of signals can be found in Donoho and Johnstone (1995). Also, would Hannig et al. (2013) be helpful in comparing the results?*

(7) *Page 31, line 1-39: Can you plot the SiZer results on this data?*

References

Berkes, I., Liu, W. and Wu, W. B. (2014). Komlós-Major-Tusnády approximation under dependence. *Annals of Probability*, **42** 794–817.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for the exploration of structures in curves. *Journal of the American Statistical Association*, **94** 807–823.

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Annals of Statistics*, **28** 408–428.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via Wavelet shrinkage. *Journal of the American Statistical Association*, **90** 1200–1224.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29** 124–152.

Hannig, J., Lee, T. C. and Park, C. (2013). Metrics for SiZer map comparison. *Stat*, **2** 49–60.

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, **101** 484–499.

Park, C., Hannig, J. and Kang, K.-H. (2009). Improved SiZer for time series. *Statistica Sinica*, **19** 1511–1530.

Park, C., Marron, J. S. and Rondonotti, V. (2004). Dependent SiZer: goodness-of-fit tests for time series models. *Journal of Applied Statistics*, **31** 999–1017.

Rondonotti, V., Marron, J. S. and Park, C. (2007). SiZer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, **1** 268–289.

Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, **19** 1755–1768.

Wu, W. B., Woodroofe, M. and Mentz, G. (2001). Isotonic regression: another look at the changepoint problem. *Biometrika*, **88** 793–804.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *Annals of Statistics*, **40** 466–493.