

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 provide a generalization of Proposition 3.3 which shows that $\mathbb{P}(E_T^\ell) = (1 - \alpha_T) + o(1)$ for any sequence $\{\alpha_T\}$ of significance levels $\alpha_T \in (0, 1)$. In particular, for $\alpha_T \rightarrow 0$, we get the consistency result that $\mathbb{P}(E_T^\ell) \rightarrow 1$.
- (ii) We have generalized our estimator of the long-run error variance. The estimation procedure is shown to be valid not only for $\text{AR}(p)$ processes of known order p but for any stationary error process $\{\varepsilon_t\}$ with an $\text{AR}(\infty)$ 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 aggregates the individual statistics $s_T(u, h)$ into the overall statistic $S_T = \max_{h \in H} S_T(h)$, where $S_T(h) = \max_{u \in U} |s_T(u, h)|$. 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 $\hat{\Psi}_T$ 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 $\widehat{\Psi}_{T, \text{uncorrected}} = \max_{h \in H} \widehat{\Psi}_T(h) = \max_{(u, h) \in U \times H} |\widehat{\psi}_T(u, h)/\widehat{\sigma}|$. 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 $\widehat{\Psi}_T = \max_{(u, h) \in U \times H} \{|\widehat{\psi}_T(u, h)/\widehat{\sigma}| - \lambda(h)\}$ 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.

Reply to Referee 1

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

I believe that the following type of result can be obtained: $\mathbb{P}(E_T^\ell) \rightarrow 1$. 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 $\alpha \rightarrow 0$.

- (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 $\hat{a}_1, \dots, \hat{a}_p$ 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 $c < 1/2$.

- (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 is 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?

Reply to Referee 2

- (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).*
 – Compare multiscale test with and without additive correction terms $\lambda(h)$ by “parallel coordinate plots” as in Figure 5 of Hannig & Marron (2006).
- (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.*
- (3) *Page 18, line 52: Please remove the speculative statements about what can be shown unless you actually show it in this paper.*
- (4) *Section 4.1: This section does not contain any truly new material and should be removed.*
- (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?*

What’s done (here α is always taken to be 0.05):

- For each setting, each sample size, and each bandwidth row we report the percentage of realizations of the data in which there were some red or blue pixels in that row. The reports are given in “parallel coordinate plots” as in Figure 5 of Hannig and Marron (2006). Each plot has the results of SiZer, the results of our global method and the results of our rowwise method respectively. The settings are as follows:
 - (a) Under the null for $a_1 = -0.25, 0, 0.25$ and sample size $T = 250$.
 - (b) Under the alternative for $a_1 = -0.25, 0$ and sample size $T = 250$. The trend function is linear increasing on the whole interval with the slope $\beta = 1.25$.
 - (c) Under the alternative for $a_1 = 0.25$ and sample size $T = 250$. The trend function is linear increasing on the whole interval with the slope $\beta = 2.25$.

- Representative SiZer maps (for SiZer, our global method and our rowwise method) for simulated data from the null distribution for $a_1 = -0.25, 0, 0.25$. Sample size is $T = 250$. 1,000 such SiZer maps were simulated for each setting and the population of them was ordered in terms of number of pixels that flag significant structure by being red or blue. For each setting and for each method we show 500th of these maps (essentially the median of the population), the 750th (the third quartile), the 850th and the 950th member of the ordered population.

(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?*

What's done (here α is always taken to be 0.05):

- Plots comparing SiZer maps under the null:
 - (a) Plots comparing SiZer and global method and rowwise methods respectively for sample size is $T = 250$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma_\eta^2 = 1$.
- Plots comparing SiZer maps under the alternative:
 - (a) Plots comparing SiZer and global method and rowwise methods respectively for sine curve plus AR(1) noise (as in Park, Hannig and Kang (2009)) for sample size $T = 250$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma_\eta^2 = 1$.
 - (b) Plots comparing SiZer and global method and rowwise methods respectively for blocks plus AR(1) noise (signal from Donoho, Johnstone, Kerkycharian and Picard (1995), recentered and normalized as in Hannig and Marron (2006)) for sample size $T = 250$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma_\eta^2 = 0.01$.
 - (c) Plots comparing SiZer and global method and rowwise methods respectively for blocks plus AR(1) noise (signal from Donoho, Johnstone, Kerkycharian and Picard (1995), recentered and normalized as in Hannig and Marron (2006)) for sample size $T = 1000$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma_\eta^2 = 0.01$.

Different signals:

- Power results could be reported in terms of ability to find the jump points in the blocks example.

(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.
- DONOHOO, D., JOHNSTONE, I., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B*, **57** 301–369.
- DONOHOO, 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.
- 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.