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 $\alpha_T \to 0$, we get the consistency result that $\mathbb{P}(E_T^\ell) \to 1$.

(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 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 summarize the main points of the new Section ??.

Informally speaking, both our approach and SiZer (for time series) 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.

There are two versions of SiZer, a global and a row-wise one. We here discuss the global version as it is closer in spirit to our approach. We comment on the row-wise version later on. The test statistic of the global SiZer approach has the form $S_T = \max_{h \in H} S_T(h)$, where $S_T(h) = \max_{u \in I} |s_T(u,h)|$ for each scale h and $s_T(u,h)$ is a test statistic for $H_0(u,h)$. As shown in Chaudhuri & 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 that $H_0(u,h)$ is fulfilled for all

$u \in I$ and $h \in H$. Using this result, one can take the $(1-\alpha)$-quantile of S as a critical value of the SiZer test. (As this quantile can usually not be determined analytically, it has to be approximated, e.g., by the bootstrap procedures of Chaudhuri & Marron (2000) or the extreme value theory based procedure of Hannig & Marron (2006).) Even though related, our methods and theory are very different from those in the SiZer literature:

- Theory for SiZer is derived under the assumption that H is a compact subset of $(0,1)$. As already pointed out in Chaudhuri & Marron (2000), this is a quite severe/substantial 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 & Marron (2000) put it (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 $T^{-\gamma}$. In order to achieve this, we come up with a proof strategy which is very different from that in the SiZer literature: Whereas Chaudhuri and Marron (2000) show that the SiZer statistic S_T weakly converges to a limit process S, our proof technique does not even require our test statistic to have a weak limit and is thus not restricted by the limitations of classic weak convergence theory.

- 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 $S_T = \max_{h \in H} S_T(h)$. As argued in Dümbgen & Spokoiny (2001), this aggregation scheme is not optimal when the set $H = H_T$ contains scales h of many different orders $T^{-\gamma}$ with $\gamma \in [0,c]$ for some $c > 0$. Following their lead, we consider a test statistic of the form $M_T = \max_{h \in H}\{M_T(h) - \lambda(h)\}$ with appropriate additive correction terms $\lambda(h)$. Here, $M_T(h) = \max_{u \in I} |m_T(u,h)|$ and $m_T(u,h)$ is a test statistic for $H_0(u,h)$ which is somewhat different but closely related to $s_T(u,h)$. Deriving distribution theory for the statistic M_T is highly non-trivial. In particular, no theory for nonparametric regression models under general dependence conditions has been available so far. The main technical contribution of our paper is to derive such a theory.

- The main complication in carrying out our multiscale test and SiZer is to determine the critical values, that is, the quantiles of the test statistics under the null. Our theoretical results show that the $(1 - \alpha)$-quantile of the multiscale statistic M_T under the null can be approximated by the $(1-\alpha)$-quantile of a Gaussian version of M_T that can be computed by simulation. It is far from obvious that this approximation is valid. To see this, deep strong approximation theory for dependent data (as derived in Berkes et al. (2014)) is needed. Importantly, our simulation-based procedure is not the same as the bootstrap procedures proposed in Chaudhuri & Marron (1999, 2000) to compute quantiles of the SiZer statistic. Both procedures are of course resampling methods. However, the resampling is done in a quite different way in our case.

- In practice, SiZer is usually implemented in its row-wise rather than its global form, that is, the test is carried out separately for each scale h based on the statistic $S_T(h) = \max_{u \in U} |s_T(u,h)|$. The main reason for applying the row-wise version of SiZer is to gain some power. However, this gain of power comes at a cost: The row-wise version of SiZer ignores the simultaneous test problem across scales h and thus does not allow to make rigorous simultaneous inference across both locations u and scales h.

  The aim of our paper is to provide a multiscale test which allows to make rigorous simultaneous inference across locations u and scales h and, at the same time, has as much power as possible. In the new empirical comparison study of Section ??, we examine the size and power properties of our multiscale test and compare it with SiZer. We in particular show that it has better power properties than a more classical multiscale test which is based on the statistic $M_T = \max_{h \in H} M_T(h)$ without additive correction terms. As explained in more detail in Section ??, this more classical multiscale test can be regarded as a version of global SiZer whose critical values are computed by our simulation-based procedure.

We hope this summary shows that the methodological and theoretical contributions of our paper are quite substantial relative to the SiZer methodology. As already mentioned above, more details are provided in the new Section ??.

Reply to Referee 1

- A consistency result of Proposition 3.3.

  I believe that the following type of result can be obtained: $\Pr(E_T^l) \to 1$. 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 $\alpha \to 0$.

- 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 $\widehat{a}_1, \ldots, \widehat{a}_p$ can be obtained from the results in the following papers

  - W. B. Wu and Mohsen Pourahmadi (2009): Banding Sample Covariance Matrices of Stationary Processes, Statistica Sinica 19 1755-1768.
  - H. Xiao and W. B. Wu (2012). Covariance Matrix Estimation for Stationary Time Series. Annals of Statistics, Volume 40, Number 1 (2012), 466-493.

  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 $c < 1/2$.

- 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 "Isotonic regression: another look at the change point problem. Biometrika, 88, 793-804, 2001", an increasing trend function is fitted. It will be important to know which period the sequence in increasing/decreasing.

- 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_Ts 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 **?**. 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.

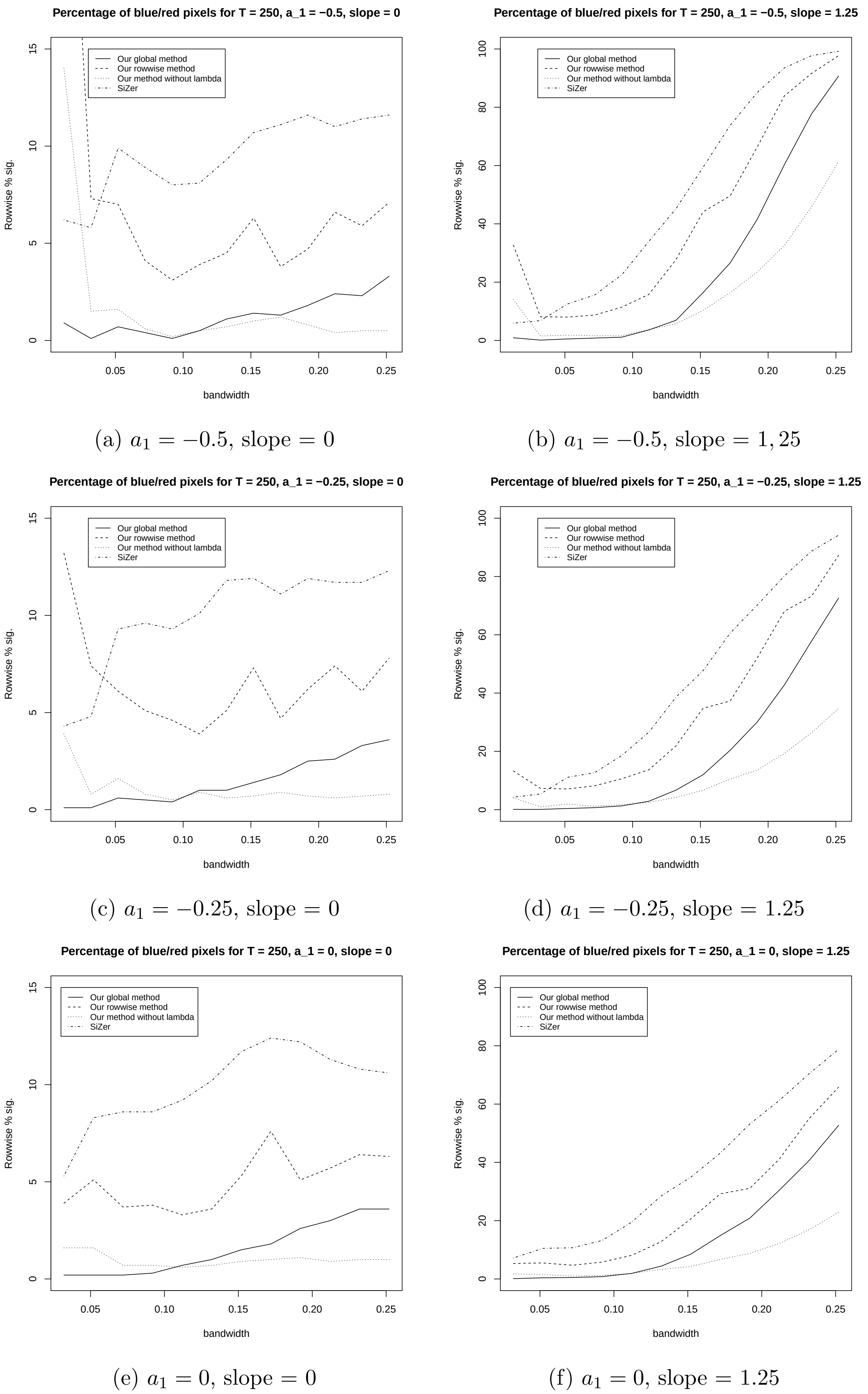
6. Page 25, line 1-26: I do not quite understand this figure. Would it be possible to rather reproduce the colorful SiZer figures that show the results of the test at various scales and locations? Also you should use several different signals. I believe that a single relatively large bump is not sufficient 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 $\sigma^2_\eta = 1$.
   - 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 **?**) for sample size $T = 250$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma^2_\eta = 1$.
     (b) Plots comparing SiZer and global method and rowwise methods respectively for blocks plus AR(1) noise (signal from **?**, recentered and normalized as in **?**) for sample size $T = 250$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma^2_\eta = 0.01$.
     (c) Plots comparing SiZer and global method and rowwise methods respectively for blocks plus AR(1) noise (signal from **?**, recentered and normalized as in **?**) for sample size $T = 1000$. AR(1) parameters are $a_1 = -0.25, 0, 0.25$ and $\sigma^2_\eta = 0.01$.

   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?



(a) $a_1 = -0.5$, slope $= 0$

(b) $a_1 = -0.5$, slope $= 1, 25$

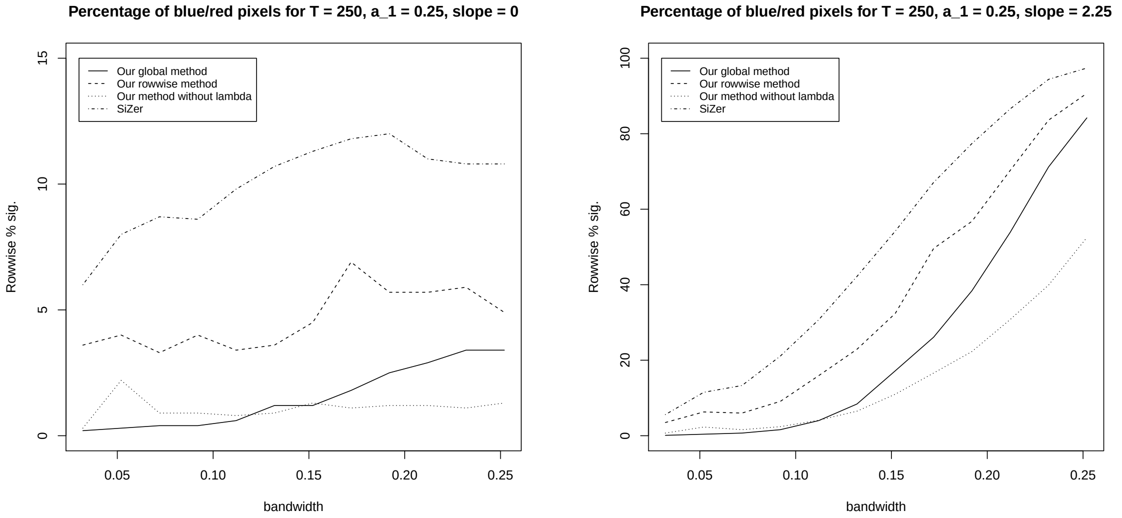
(c) $a_1 = -0.25$, slope $= 0$

(d) $a_1 = -0.25$, slope $= 1.25$

(e) $a_1 = 0$, slope $= 0$

(f) $a_1 = 0$, slope $= 1.25$

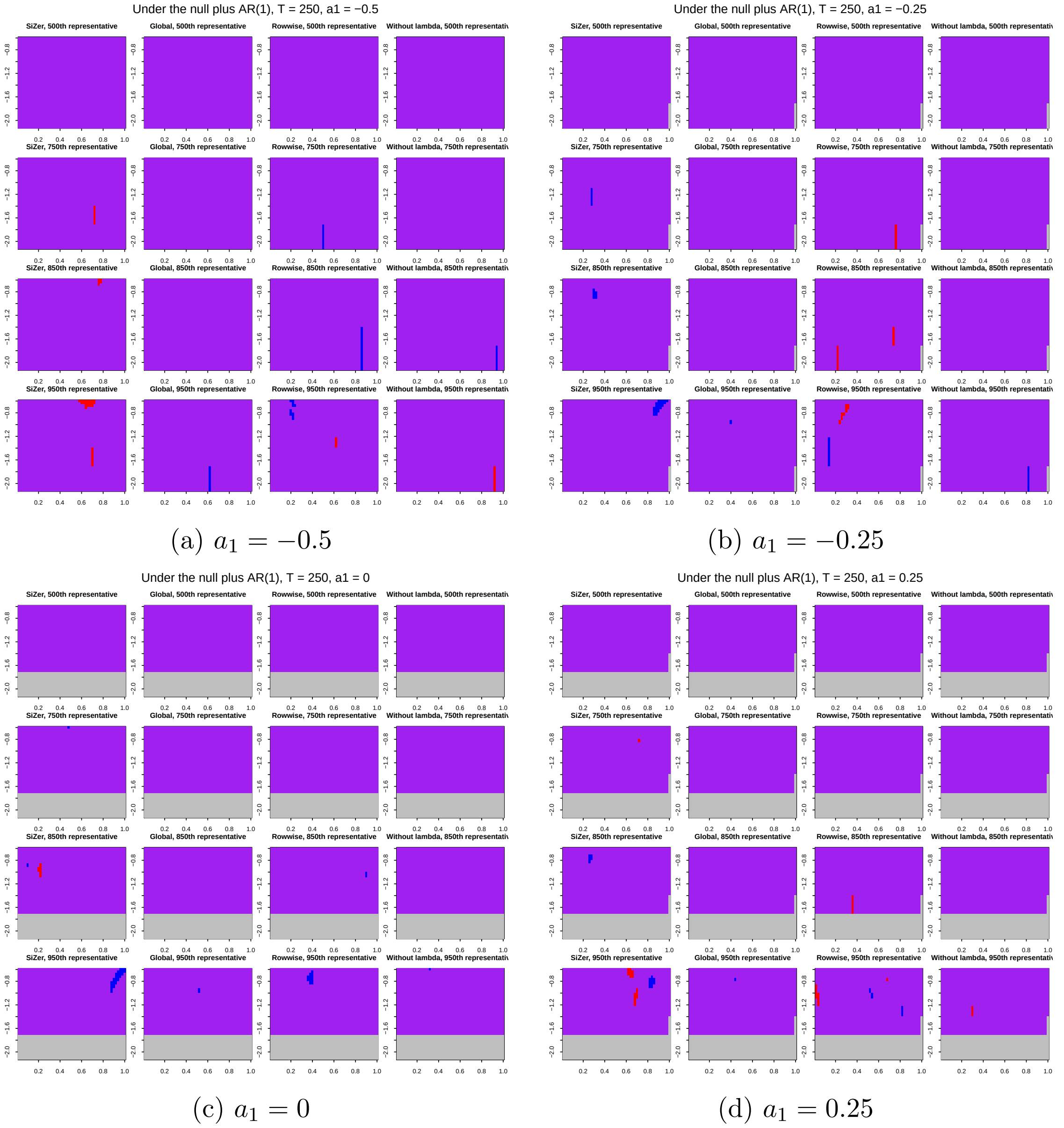
Figure 1: Parallel coordinate plots for $T = 250$ and different AR(1) parameters $a_1 = -0.5, -0.25, 0, 0.25$ in the simulation scenarios with a pronounced trend and without the trend respectively.



(g) $a_1 = 0.25$, slope $= 0$

(h) $a_1 = 0.25$, slope $= 2.25$

Figure 1: Parallel coordinate plots for $T = 250$ and different AR(1) parameters $a_1 = -0.5, -0.25, 0, 0.25$ in the simulation scenarios with a pronounced trend and without the trend respectively.



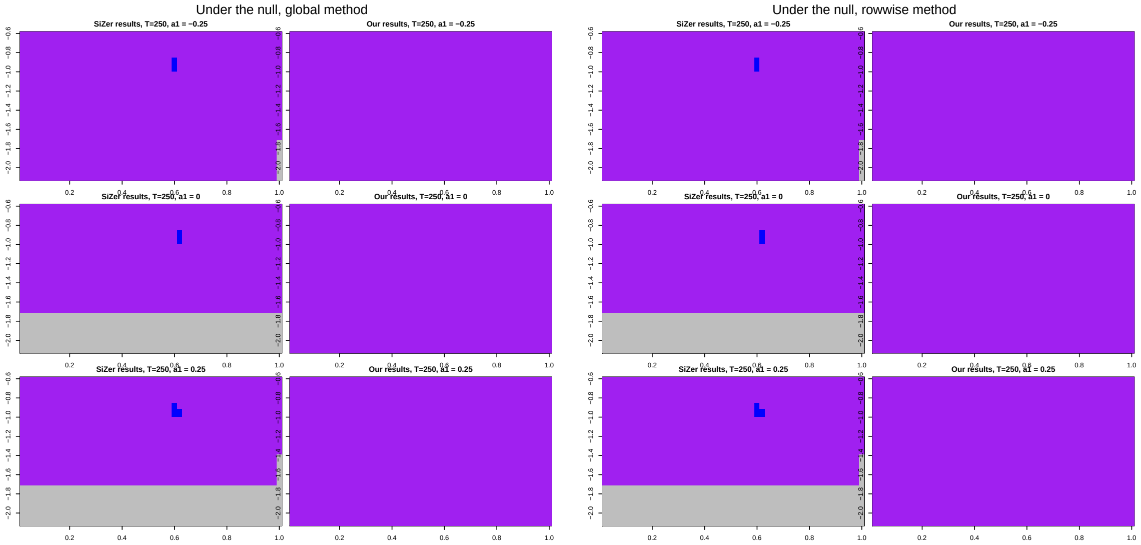
(a) $a_1 = -0.5$

(b) $a_1 = -0.25$

(c) $a_1 = 0$

(d) $a_1 = 0.25$

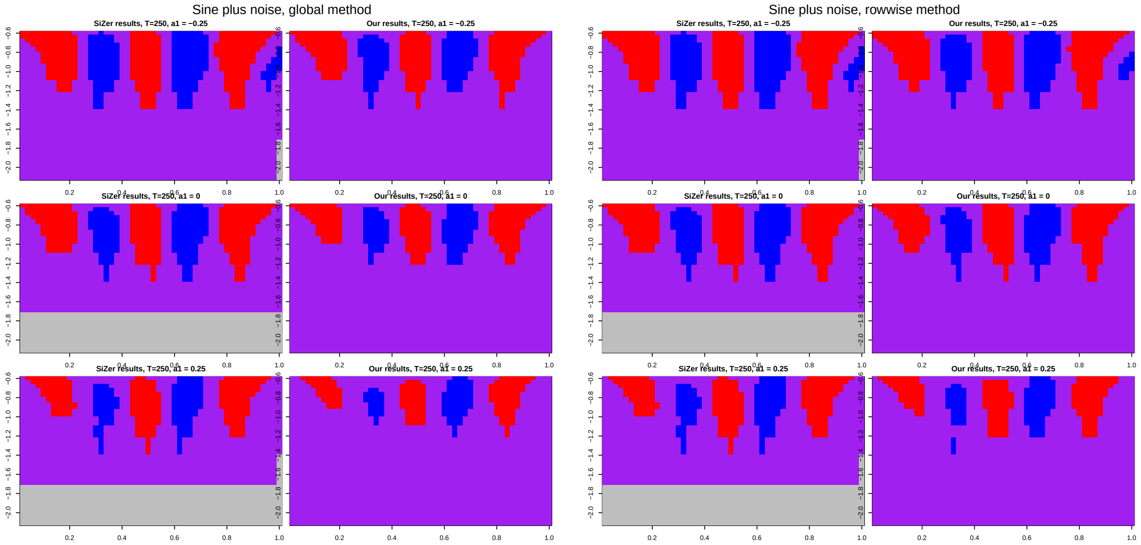
Figure 2: Representative SiZer maps $T = 250$ and different AR(1) parameters $a_1 = -0.5, -0.25, 0, 0.25$ in the simulation scenario under the null.



(a) Global method

(b) Rowwise method

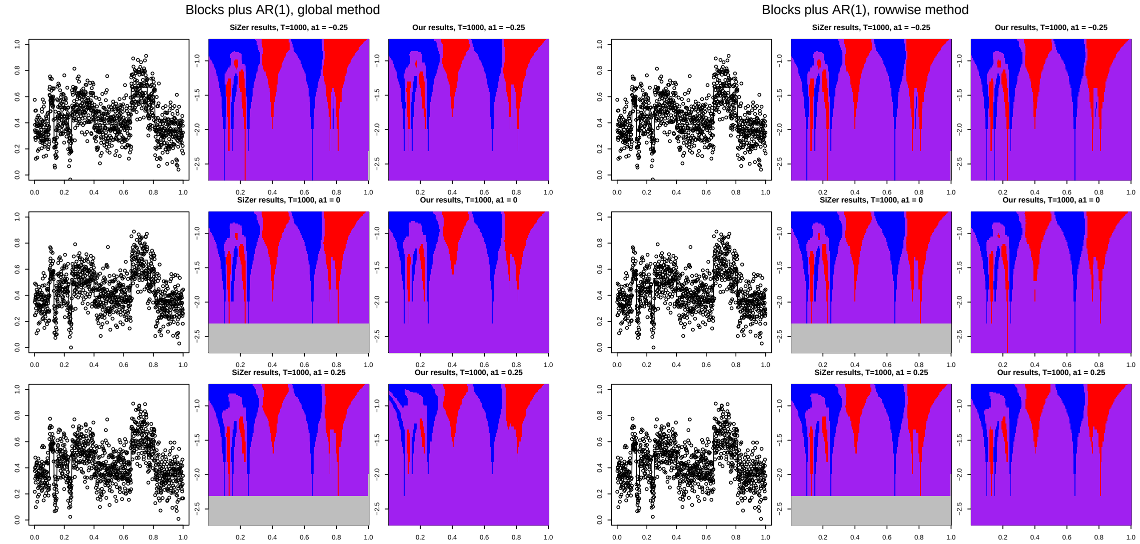
Figure 3: Comparative SiZer maps for $T = 250$ and different AR(1) parameters $a_1 = -0.25, 0, 0.25$ in the simulation scenario under the null



(a) Global method

(b) Rowwise method

Figure 4: Comparative SiZer maps for $T = 250$ and different AR(1) parameters $a_1 = -0.25, 0, 0.25$ in the simulation scenario with sine curve as the trend function



(a) Global method

(b) Rowwise method

Figure 5: Comparative SiZer maps for $T = 1000$ and different AR(1) parameters $a_1 = -0.25, 0, 0.25$ in the simulation scenario with blocks as the trends function