

## Improved SiZer for Time Series

Cheolwoo Park, Jan Hannig, and Kee-Hoon Kang

*University of Georgia, Colorado State University,  
and Hankuk University of Foreign Studies*

*Abstract:* SiZer (SIgnificant ZERo crossing of the derivatives) is a scale-space visualization tool for statistical inferences. In this paper we improve global inference of SiZer for time series, which was originally proposed by Rondonotti, Marron and Park (2007), in two aspects. First, the estimation of the quantile in a confidence interval is theoretically justified by advanced distribution theory. Second, an improved nonparametric autocovariance function estimator is proposed using a differenced time series. A numerical study is conducted to demonstrate the sample performance of the proposed tool. In addition, asymptotic properties of SiZer for time series are investigated.

*Key words and phrases:* Autocovariance function estimation, Local linear smoothing, Multiple testing adjustment, SiZer, Statistical convergence, Time series.

### 1. Introduction

SiZer (Chaudhuri and Marron, 1999) is a visualization method based on nonparametric curve estimates. SiZer addresses the question of which features observed in a smooth are really present, or represent an important underlying structure, and not simply artifacts of the sampling noise. Thus, SiZer analysis enables statistical inference for the discovery of meaningful structure within a data set, while doing exploratory analysis.

SiZer is based on scale-space ideas from computer vision, see Lindeberg (1994). Scale-space, in our context, is a family of kernel smooths indexed by the scale, which is the smoothing parameter or bandwidth  $h$ . SiZer considers a wide range of bandwidths, which avoids the classical problem of bandwidth selection. The idea is that this approach uses all the information that is available in the data at each given scale. Thus, the target of a SiZer analysis is shifted from finding features in the *true underlying curve* to inferences about the *smoothed*

*version of the underlying curve, i.e. the curve at the given level of resolution.*

Other SiZer tools have been developed and they have proven to be very useful in many applications including Internet traffic data (Park et al., 2005 and 2006), anomaly detection (Park, Marron, and Rondonotti, 2004, and Park et al., 2007a), jump detection (Kim and Marron, 2006), economics data (Chaudhuri and Marron, 1999), outlier identification (Hannig and Lee, 2006), fMRI (functional Magnetic Resonance Imaging) data (Park et al., 2007b), wavelets (Park et al., 2007a), and comparison of regression curves (Park and Kang, 2008). Hannig and Marron (2006) proposed an improved inference version of SiZer to reduce unexpected features in the SiZer map.

Recently, Bayesian versions of SiZer have been proposed as an approach to Bayesian multiscale smoothing. Those include Godtliebsen and Øigård (2005), Erästö and Holmström (2005), and Øigård, Rue and Godtliebsen (2006), etc. They assumed the underlying distribution and prior model for the parameters, and then combine these two to get the posterior distribution. It simplifies the mathematical derivations and also makes it possible fast computation. The inference is based on finite difference quotients or derivatives, which depends on the selected prior model for the underlying curve.

As pointed out by Chaudhuri and Marron (1999), the statistical inference of SiZer makes heavy use of the assumption of independent errors. This assumption is inappropriate in time series contexts. For dependent data, significant features, which are only due to the presence of dependence, appear in the original SiZer. Dependent SiZer, proposed by Park, Marron, and Rondonotti (2004), extends SiZer to time series data. It uses a true autocovariance function of an assumed model and conducts a goodness of fit test. By doing so, one can see how different the behavior of the data is from that of the assumed model. Rondonotti, Marron, and Park (2007) proposed SiZer for time series using an estimated autocovariance function.

The focus of this paper is on SiZer for time series. For SiZer to fulfill its potential to flag significant trends in time series, its underlying confidence intervals must be adjusted to properly account for the correlation structure of the data. This adjustment is not straightforward when the correlation structure is unknown. This is because of the identifiability problem between trend and de-

pendence artifacts. Rondonotti, Marron, and Park (2007) addressed this issue and proposed an approach to this dilemma via a visualization which displays the range of trade-offs.

While the original SiZer for time series is useful, there is still room for improvement. The estimation of the quantile in a confidence interval relies on a heuristic idea rather than on theory, and the estimation of an autocovariance function is not accurate in some situations. Moreover, any theoretical properties of the proposed method are not provided. This paper aims to remedy these problems in a moderately correlated time series. We propose to estimate the quantile by extreme value theory and the autocovariance function based on differenced time series. In addition, weak convergence of the empirical scale-space surface to its theoretical counterpart has been established under appropriate regularity conditions.

This paper is organized as follows. Section 2 reviews the original SiZer for time series proposed by Rondonotti, Marron, and Park (2007). In Section 3, the estimation of the quantile and the autocovariance function is proposed. A simulation study and real example analysis are provided in Section 4. In Section 5, asymptotic properties of SiZer for time series are investigated.

## 2. SiZer for time series

Given the time series data  $\{(i, Y_i), i = 1, \dots, n\}$ , the regression setting is

$$Y_i = f(i) + \epsilon_i, \quad i = 1, \dots, n \quad (2.1)$$

where  $f$  is assumed to be a smooth function and the error is assumed to be a zero mean weakly stationary process, i.e.  $E(\epsilon_i) = 0$ ,  $V(\epsilon_i) = \sigma^2$ , for all  $i = 1, \dots, n$ , and  $Cov(\epsilon_i, \epsilon_j) = \gamma(|i - j|)$  for all  $i, j = 1, \dots, n$ .

In the local linear fit (see Fan and Gijbels, 1996) the function  $f(i)$  is approximated by Taylor expansion of order 1 for  $i$  in the neighborhood of  $i_0$ . The problem to be solved is then

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1(i - i_0))]^2 K_h(i - i_0) \quad (2.2)$$

where  $\boldsymbol{\beta} = (\beta_0 \ \beta_1)^T$ ,  $h$  is the bandwidth controlling the size of the local neighborhood and  $K_h(\cdot) = K(\cdot/h)/h$ , where  $K$  is the Gaussian kernel function. By

Taylor expansion  $\beta_0 = f(i_0)$  and  $\beta_1 = f'(i_0)$ , so the solution to this problem gives estimates of the regression function and its first derivative at  $i_0$ . More specifically,

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

where  $Y = (Y_1, \dots, Y_n)^T$ , the design matrix of the local linear fit at  $i_0$  is

$$X = \begin{pmatrix} 1 & (1 - i_0) \\ 1 & (2 - i_0) \\ \vdots & \vdots \\ 1 & (n - i_0) \end{pmatrix}$$

and  $W = \text{diag}\{K_h(i - i_0)\}$ .

For correlated data, the variance of the local polynomial estimator is given by

$$V(\hat{\beta}|X) = (X^T W X)^{-1} (X^T \Sigma X) (X^T W X)^{-1} \quad (2.3)$$

where, for the assumed correlation structure,  $\Sigma$  is the kernel weighted covariance matrix of the errors where the generic element is given by

$$\sigma_{ij} = \gamma(|i - j|) K_h(i - i_0) K_h(j - i_0). \quad (2.4)$$

Rondonotti, Marron, and Park (2007) proposed an estimate of the variance in (2.3) with estimated  $\gamma$  in (2.4). It is the the sample autocovariance function of the observed residuals from a *pilot bandwidth*,  $h_p$ . A small  $h_p$  assumes independent or weakly correlated errors and a large one corresponds to strongly correlated errors. They consider  $h$  and  $h_p$  separately, which means that in the dependent case, another dimension needs to be added to the SiZer plot. Thus, a series of SiZer plots, indexed by the pilot bandwidth  $h_p$ , represent the different trade-offs available between trend and dependence.

The SiZer inference is based on confidence intervals for the derivative of the smoothed underlying function. These are of the form

$$\hat{f}'_h(i) \pm q(h) \times \hat{sd}(\hat{f}'_h(i)) \quad (2.5)$$

where  $q(h)$  is an appropriate quantile depending on  $h$ . Rondonotti, Marron, and Park (2007) suggested to use the quantile

$$q(h) = \Phi^{-1}\left(\frac{1 + (1 - \alpha)^{1/l(h)}}{2}\right),$$

where  $\Phi$  is the cumulative distribution function of the standard normal,  $\alpha$  is a significance level, and  $l(h)$  reflects the number of independent blocks at the scale  $h$ .

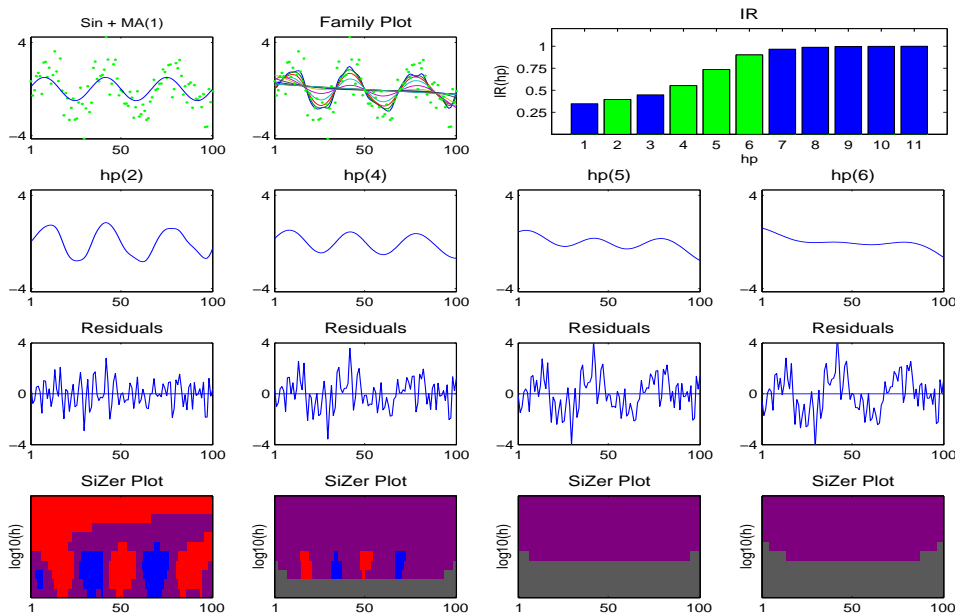


Figure 1: Original SiZer for time series: Sine plus MA(1)

In order to motivate our work, we consider an example of SiZer plots, which is shown in Figure 1. We generate MA(1) time series with a signal  $f(i) = \sin(6\pi i/n)$  where  $n = 100$ . The generated data are shown in the first plot above on the left (the continuous line shows  $f(i)$ , the deterministic part of the simulated time series), while the next graphic on the right is the family plot. The family of smooths is constructed by considering a very wide range of bandwidths in the log scale and, in particular, the number of curves is here taken to be 11. We use this 11 bandwidths as  $h_p$  for estimating  $\gamma$  in (2.4). But showing the complete SiZer maps for these 11 bandwidths is too long and inefficient. Thus, only 4 bandwidths are chosen by a simple measure of *Indicator of the Residual component (IR)* defined as

$$IR(h_p) = \frac{\sum_{i=1}^n \hat{\epsilon}_{h_p,i}^2}{\max_{h_p} \sum_{i=1}^n \hat{\epsilon}_{h_p,i}^2},$$

where  $\hat{\epsilon}_{h_p, i}$ 's are the residuals obtained from the pilot bandwidth  $h_p$ . Further right of the top in Figure 1 is the bar diagram using this information and in this case the second, fourth, fifth, and sixth bandwidths are selected. For more details on this choice, see Rondonotti, Marron, and Park (2007). The series of plots in the second and third rows represent, respectively, the local linear fits and the residuals corresponding with the selected bandwidths.

SiZer extends the usefulness of the family plot by visually displaying the statistical significance of features over both location  $x$  and scale  $h$ . Inference is based on confidence intervals in (2.5) for the derivative of the underlying function. The graphical device is a color map, reflecting statistical significance of the slope at  $(x, h)$  locations in scale-space. At each  $(x, h)$  location, the curve is significantly increasing (decreasing) if the confidence interval is above (below) 0, so that map location is colored blue (red). If the confidence interval contains 0, the curve at the level of resolution  $h$  and at the point  $x$  does not have a statistically significant slope, so purple is used. Finally, if there is not enough information in the data set, at this scale space  $(x, h)$  location, then no conclusion can be drawn, so gray is used to indicate that the data are too sparse.

The four plots in the bottom of Figure 1 are the SiZer maps using each  $\gamma$  estimated from the selected bandwidths. The first SiZer map (corresponding to  $h_p(2)$ ) shows significant features along the sine curve. Note that as we move to the other SiZer maps (i.e.  $h_p(4)$ ,  $h_p(5)$ , and  $h_p(6)$ ), an increasing amount of correlation appears in the error component, so that less features are significant at every level of resolution. Also, at the fine levels of resolution of the third and fourth maps there is less perceived useful information in the data, which means more data sparsity, thus more bottom lines of the SiZer plots are shaded gray. Since MA(1) is weakly correlated, it is reasonable to interpret the first or second SiZer map.

However, a deeper look creates some concerns. While SiZer maps flag the sine trend reasonably well, some spurious features are flagged significant in the first SiZer map. For example, the global downward trend is flagged as significant since the red color appears at large resolutions. Since this is not a deterministic trend but instead created by MA(1), it should be colored as purple. The SiZer map for  $h_p(4)$  does not have these spurious features but shows less features than

expected. Furthermore, some areas are colored gray and no decision is made. The map for  $h_p(3)$  would be appropriate in this example, but it is not selected by the  $IR$  statistic.

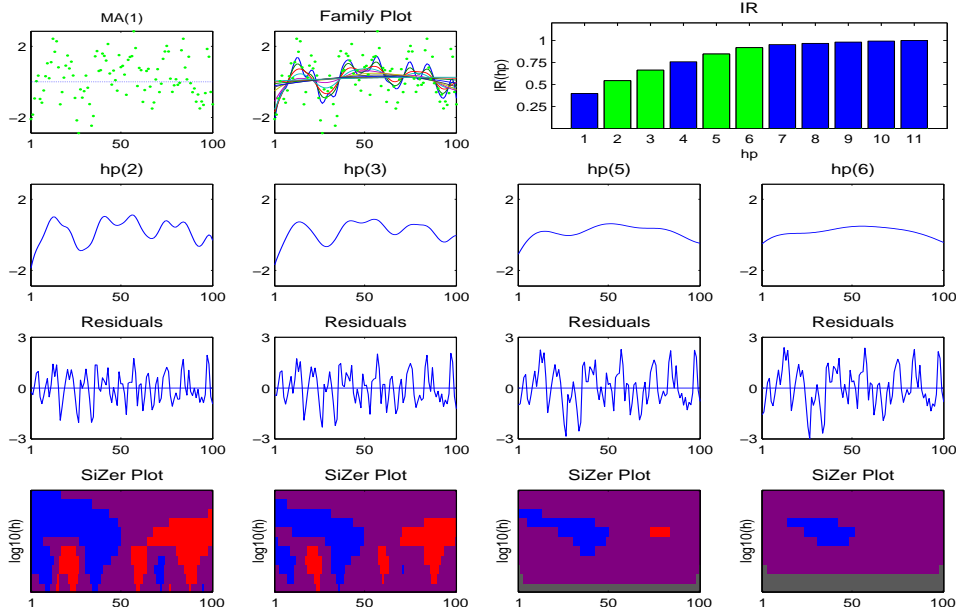


Figure 2: Original SiZer for time series: MA(1)

The problem becomes clearer when we remove the sine curve from the time series. Figure 2 shows SiZer plots for MA(1) only. Since no trend is added to MA(1), true SiZer maps would show only purple at all pilot bandwidths. However, the four SiZer maps show some serious significant features. The first two SiZer maps, which correspond to weakly correlated errors, flag many features as significant. This motivates us to improve SiZer inference, which is detailed in the following section.

### 3. Improved inference for time series data

#### 3.1 Quantile estimation

In this section we extend the result of Hannig and Marron (2006) to the time series context. We omit most of the technical details as the derivations are

similar.

SiZer uses the local linear smoother defined by (2.2). To color the pixels SiZer checks whether the estimate of the first derivative

$$\begin{aligned} \hat{\beta}_1 &= -c^{-1} \left[ \sum_{i=1}^n K_h(x - X_i) \right] \left[ \sum_{i=1}^n (x - X_i) K_h(x - X_i) Y_i \right] \\ &\quad + c^{-1} \left[ \sum_{i=1}^n (x - X_i) K_h(x - X_i) \right] \left[ \sum_{i=1}^n K_h(x - X_i) Y_i \right], \\ c &= \left[ \sum_{i=1}^n K_h(x - X_i) \right] \left[ \sum_{i=1}^n (x - X_i)^2 K_h(x - X_i) \right] - \left[ \sum_{i=1}^n (x - X_i) K_h(x - X_i) \right]^2. \end{aligned} \quad (3.1)$$

is significantly different from 0. In the particular case of fixed design regression the design points  $X_i$  satisfy  $X_i = i\Delta$ , where  $\Delta > 0$  is the distance between design points. If  $x$  is away from the boundary, it follows from symmetry of the kernel that

$$\sum_{i=1}^n (x - X_i) K_h(x - X_i) \approx 0.$$

This means that the second term in (3.1) disappears.

Let  $\tilde{\Delta}$  denote the distance between the pixels of the SiZer map and  $p = \tilde{\Delta}/\Delta$  denote the number of data points per SiZer column. For simplicity of notation we can assume that  $p$  is a positive integer. Let  $g$  be the number of pixels in each row, and  $T_1, \dots, T_g$  denote the test statistics of a row in the SiZer map. Then  $T_j$  is proportional to the estimate of the first derivative  $\hat{\beta}_1$  calculated for  $x = j\tilde{\Delta} = jp\Delta$ . In particular,  $T_j \approx \sum_{q=1}^n W_{jp-q}^h Y_q$ . The exact form of the  $W_{jp-q}^h$  is given in the first term of (3.1). For our purpose it suffices to realize that  $W_{jp-q}^h$  is proportional to  $-(jp - q) K_{h/\Delta}(jp - q)$ . Thus the weights  $W_q^h$  are proportional to the derivative of the Gaussian kernel with standard deviation  $h/\Delta$ .

If the null hypothesis of no signal is true, then the  $Y_i$ 's are identically distributed Gaussian random variables with mean zero and covariance  $E(Y_i Y_j) = \gamma(i - j)$ . We assume that  $\gamma$  is an even function.

If the  $Y_i$ 's are not Gaussian but have two finite moments and the covariance  $\gamma$  decays fast enough, the linear approximation of  $T_j$  greatly simplifies the distribution theory, because for  $h/\Delta$  large enough the Cramèr-Wold device and Lindeberg-Feller Central Limit Theorem (see for example Durrett, 2005) give an

approximate Gaussian distribution, with mean 0 (under the SiZer null hypothesis) and variance 1, by appropriate scaling.

The full joint distribution of  $T_1, \dots, T_g$  also depends on the correlation between them. This correlation is approximated by

$$\begin{aligned} \rho_{j-i} = \text{corr}(T_i, T_j) &= \frac{\sum_q \sum_r W_{ip-q}^h W_{jp-r}^h \gamma(q-r)}{\sum_q \sum_r W_q^h W_r^h \gamma(q-r)} \\ &\approx \frac{\iint (ip-x)K_{h/\Delta}(ip-x)(jp-y)K_{h/\Delta}(ip-y)\gamma(x-y) dx dy}{\iint xK_{h/\Delta}(x)yK_{h/\Delta}(y)\gamma(x-y) dx dy} \\ &= \frac{\int \gamma(r) \int (ip-r-y)K_{h/\Delta}(ip-r-y)(jp-y)K_{h/\Delta}(ip-y) dy dr}{\int \gamma(r) \int (r+y)K_{h/\Delta}(r+y)(y)K_{h/\Delta}(y) dy dr} \\ &= \frac{\int \gamma(r) e^{-[(i-j)\tilde{\Delta}-r\Delta]^2/(4h^2)} \left\{ 1 - \frac{[(i-j)\tilde{\Delta}-r\Delta]^2}{2h^2} \right\} dr}{\int \gamma(r) e^{-r^2\Delta^2/(4h^2)} \left[ 1 - \frac{r^2\Delta^2}{2h^2} \right] dr}, \end{aligned}$$

where the second line follows by replacing the sums by integral approximations and the last step follows by observing that  $p\Delta = \tilde{\Delta}$ .

To find an asymptotic distribution of the maximum we use the method of Hsing, Husler, and Riess (1996), that is based on the observation that for dependent stationary, mean zero, variance one Gaussian process it is often numerically better to approximate  $P[\max(T_1, \dots, T_g) \leq x]$  by  $\Phi(x)^{\theta g}$  where  $\theta < 1$ , than by quantities based on the limiting Gumbel distribution. This is due to the extremely slow rate of convergence of the maximum to the limiting Gumbel distribution.

Since we are dealing with stationary Gaussian sequence, direct computation of the limit as  $g \rightarrow \infty$  would lead to  $\theta = 1$  (Berman, 1964). In order to get  $\theta < 1$ , they need the correlation  $\rho_j$  to increase to 1 with  $g$  for each fixed  $j$ . To achieve this Hsing, Husler, and Riess (1996) embed the series in a triangular array  $\hat{T}_{j,g}$ , where rows are indexed by  $g$ . For each fixed  $g$ , the random variables  $\hat{T}_{j,g}, j = 1, 2, \dots$  comprise a mean zero, variance one, stationary Gaussian series with the  $j$  step correlations  $\rho_{j,g}$  satisfying

$$\log(g)(1 - \rho_{j,g}) \rightarrow \delta_j \text{ as } g \rightarrow \infty, \text{ for all } j,$$

where  $\delta_j \in (0, \infty]$ . They define

$$\vartheta = P \left[ V/2 + \sqrt{\delta_k} H_k \leq \delta_k \text{ for all } k \geq 1 \right],$$

where  $V$  is a standard exponential random variable and  $H_k$  is a mean zero Gaussian process independent of  $V$  that satisfies  $E(H_i H_j) = (\delta_i + \delta_j - \delta_{|i-j|}) / (2\sqrt{\delta_i \delta_j})$ . The authors then claim that under certain technical conditions on  $\rho_{j,g}$  the distribution function  $P[\max(\hat{T}_{1,g}, \dots, \hat{T}_{g,g}) \leq x]$  could be approximated by  $\Phi(x)^{\theta g}$ . The parameter  $\theta$  has been called the ‘‘cluster index’’.

In the particular case of SiZer, it is reasonable to assume that under the null hypothesis  $T_1, \dots, T_g$  are Gaussian, with mean 0 and variance 1 and  $j$  step correlation  $\rho_j$ . A natural way to embed our SiZer row into a triangular array compatible with Hsing, Husler, and Riess (1996) is to assume that  $\tilde{\Delta}/h = C/\sqrt{\log g}$ . In order to keep the presence of the correlation between observations we assume that  $\gamma_g(i) = r(i\Delta/h)$ , where  $r$  is a suitable function. Then, we calculate

$$\rho_{k,g} = \frac{\int r(s) e^{-(Ck/\sqrt{\log g} - s)^2/4} \left\{ 1 - \frac{(Ck/\sqrt{\log g} - s)^2}{2} \right\} ds}{\int r(s) e^{-s^2/4} \left\{ 1 - \frac{s^2}{2} \right\} ds}.$$

Since  $r(\cdot)$  is an even function we get by dominated convergence theorem

$$\lim_{g \rightarrow \infty} \log(g)(1 - \rho_{k,g}) = k^2 \frac{C^2 \int r(s) e^{-s^2/4} \frac{12 - 12s^2 + s^4}{16} ds}{\int r(s) e^{-s^2/4} \left\{ 1 - \frac{s^2}{2} \right\} ds}.$$

Therefore just as in Hannig and Marron (2006) we conclude that in the case of SiZer

$$P \left[ \max_{i=1, \dots, g} T_i \leq x \right] \approx \Phi(x)^{\theta g},$$

where the cluster index

$$\theta = 2\Phi \left( \sqrt{I_\gamma \log g} \frac{\tilde{\Delta}}{h} \right) - 1$$

and

$$I_\gamma = \frac{\int \gamma(sh/\Delta) e^{-s^2/4} \frac{12 - 12s^2 + s^4}{16} ds}{\int \gamma(sh/\Delta) e^{-s^2/4} \left\{ 1 - \frac{s^2}{2} \right\} ds}.$$

Finally,

$$q(h) = \Phi^{-1} \left( \left( 1 - \frac{\alpha}{2} \right)^{1/(\theta g)} \right).$$

Figures 3 (a) and (b) show SiZer maps using the new quantile proposed above. Figure 3 (a) corresponds to Figure 1 and shows only SiZer maps to save

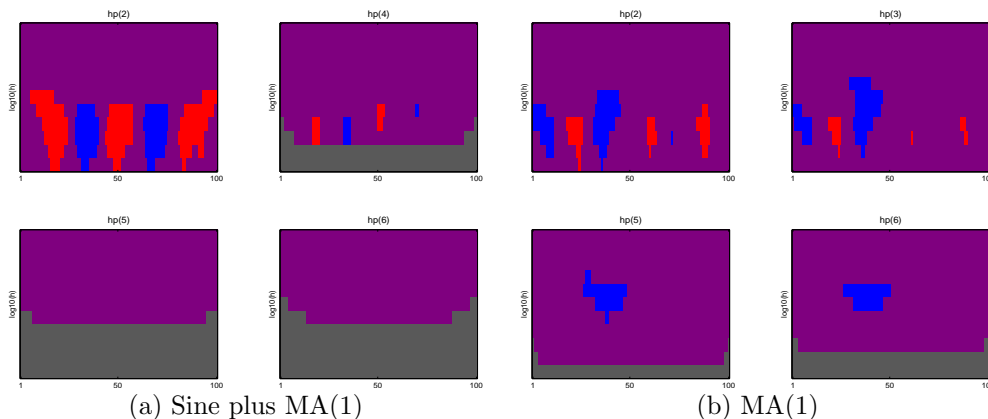


Figure 3: SiZer plots using the proposed quantile.

space. The first SiZer map shows the sine curve trend and less spurious features appear throughout all the SiZer maps compared to the ones in Figure 1. Figure 3 (b) corresponds to Figure 2. Again many spurious features disappear but there still remain some significant features which are not supposed to show up in the map.

This simulation confirms that the proposed quantile works better, but there is still room for improvement.

### 3.2. Autocovariance function estimation

This section explains why there still remain unexpected features in SiZer maps in Section 3.1 and proposes a new autocovariance estimator to fix this problem when a time series has a moderate correlation. Since the proposed estimator does not require a pilot bandwidth, there is no need to select bandwidths to display.

The original SiZer for time series uses residuals to estimate an autocovariance function. The residuals can be written as

$$\hat{\epsilon}_i = Y_i - \hat{f}_{h_p}(i) = Y_i - \frac{1}{n} \sum_{k=1}^n w_n(h_p, i, k) Y_k = \sum_{k=1}^n b_{ik} Y_k$$

where  $w_n(h_p, i, k)/n$  is the weight in a local linear estimate of  $f$  with the pilot

bandwidth  $h_p$ ,  $b_{ii} = 1 - w_n(h_p, i, i)/n$  and  $b_{ik} = -w_n(h_p, i, k)$  for  $i \neq k$ . Then,

$$\gamma^*(|i - j|) = \text{Cov}(\hat{\epsilon}_i, \hat{\epsilon}_j) = \text{Cov}\left(\sum_{k=1}^n b_{ik} Y_k, \sum_{l=1}^n b_{jl} Y_l\right) = \sum_{k=1}^n \sum_{l=1}^n b_{ik} b_{jl} \gamma(|k - l|).$$

Thus, the autocovariance estimate  $\gamma^*$  from the residuals is not equal to the original  $\gamma$ , which is responsible for the spurious features in Figure 3. Therefore, we need to either do a proper adjustment for  $\gamma^*$  or find a reliable estimate of the covariance function of the residuals.

Recall that in our model (2.1)  $\epsilon_i$  is a mean zero stationary process with a autocovariance function  $\gamma(|i - j|) = E(\epsilon_i \epsilon_j)$ . Since we do not observe the  $\epsilon_i$ , we need to estimate the autocovariance function from  $Y_i$ . A particular care has to be taken to remove the effects of the smooth mean on the estimation as much as possible. This is because smooth biases could introduce a spurious long range dependence in the estimated covariance function.

To address this issue we do not estimate the covariance from the residuals  $\hat{\epsilon}_i$ . Instead we estimate the covariance structure directly from a (possibly several times) differenced time series. One of the advantages of this approach is that the estimator of the covariance no longer depends on the bandwidth.

Let  $e_i$  be the differenced time series, i.e.,  $\mathbf{e} = \mathbf{A}\mathbf{y}$  where  $A = (a_{i,k})$  is the difference matrix, e.g.,

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

if the first difference is used. A simple calculation shows for all  $i, j$

$$\begin{aligned} \text{Cov}(e_i, e_j) &= \sum_{k=1}^n a_{i,k} a_{j,k} \gamma(0) + \sum_{k=1}^{n-1} (a_{i,k} a_{j,k+1} + a_{i,k+1} a_{j,k}) \gamma(1) \\ &\quad + \cdots + (a_{i,1} a_{j,n} + a_{i,n} a_{j,1}) \gamma(n-1). \end{aligned}$$

From this we can set

$$\begin{aligned} e_i e_j &= \sum_{k=1}^n a_{i,k} a_{j,k} \gamma(0) + \sum_{k=1}^{n-1} (a_{i,k} a_{j,k+1} + a_{i,k+1} a_{j,k}) \gamma(1) \\ &\quad + \cdots + (a_{i,1} a_{j,n} + a_{i,n} a_{j,1}) \gamma(n-1) + \delta_{ij}. \end{aligned}$$

We assume that the regression function was smooth enough so that  $E(\delta_{ij}) \approx 0$ . Thus we have  $n^2$  equations and  $n$  variables. Estimating  $\gamma$  by the least squares method, i.e. by minimizing

$$\sum_{i,j} \left( e_i e_j - \sum_{k=1}^n a_{i,k} a_{j,k} \gamma(0) - \sum_{k=1}^{n-1} (a_{i,k} a_{j,k+1} + a_{i,k+1} a_{j,k}) \gamma(1) \right. \\ \left. - \cdots - (a_{i,1} a_{j,n} + a_{i,n} a_{j,1}) \gamma(n-1) \right)^2 \quad (3.2)$$

fails because the least square problem in (3.2) does not lead to a unique solution.

We therefore need to regularize the problem (3.2). First, since  $\gamma(0) \geq |\gamma(i)|$  for each  $i$ , we consider only such solutions. Additionally we regularize the least square problem by introducing the penalty  $\lambda \sum_{i=1}^{n-1} i \gamma(i)^2$ . The weight  $i$  is motivated by the belief that the covariance  $\gamma(i)$  should be decaying as  $i$  increases. It has a similar effect as using the  $n$  denominator instead of  $n-j$  in the estimator  $n^{-1} \sum_{i=1}^{n-j} \hat{\epsilon}_i \hat{\epsilon}_{i+j}$  of  $\gamma(j)$ .

This leads to the following constrained ridge regression

$$\arg \min_{\gamma \in R} \left\{ \sum_{i,j} \left( e_i e_j - \sum_{k=1}^n a_{i,k} a_{j,k} \gamma(0) - \sum_{k=1}^{n-1} (a_{i,k} a_{j,k+1} + a_{i,k+1} a_{j,k}) \gamma(1) \right. \right. \\ \left. \left. - \cdots - (a_{i,1} a_{j,n} + a_{i,n} a_{j,1}) \gamma(n-1) \right)^2 + \lambda \sum_{i=1}^{n-1} i \gamma(i)^2 \right\},$$

where  $R = \{\gamma : \gamma(0) \geq |\gamma(i)|, i = 1, \dots, n-1\}$ . We implemented this minimization using the MATLAB function `lsqlin`. This function uses methods of quadratic programming to find the minimum.

We have investigated several choices of  $\lambda$  and found that  $\lambda = 1$  works well as long as the time series is weakly to moderately dependent. An extensive study of the statistical properties of the proposed estimator and its possible modifications goes beyond the scope of this paper and we suggest this as future work. In particular it would be interesting to allow for either LASSO type  $L_1$  penalty or different weights, e.g.,  $i^\alpha$  to better match the decay of the covariance function.

Figure 4 compares the estimates of autocorrelation functions using the original and the proposed estimators for  $N(0, 1)$ , MA(1), and MA(5). For the original method 11 different pilot bandwidths are used. For  $N(0, 1)$ , the original estimate has a small deviation from the 95% confidence interval of no correlation. On the

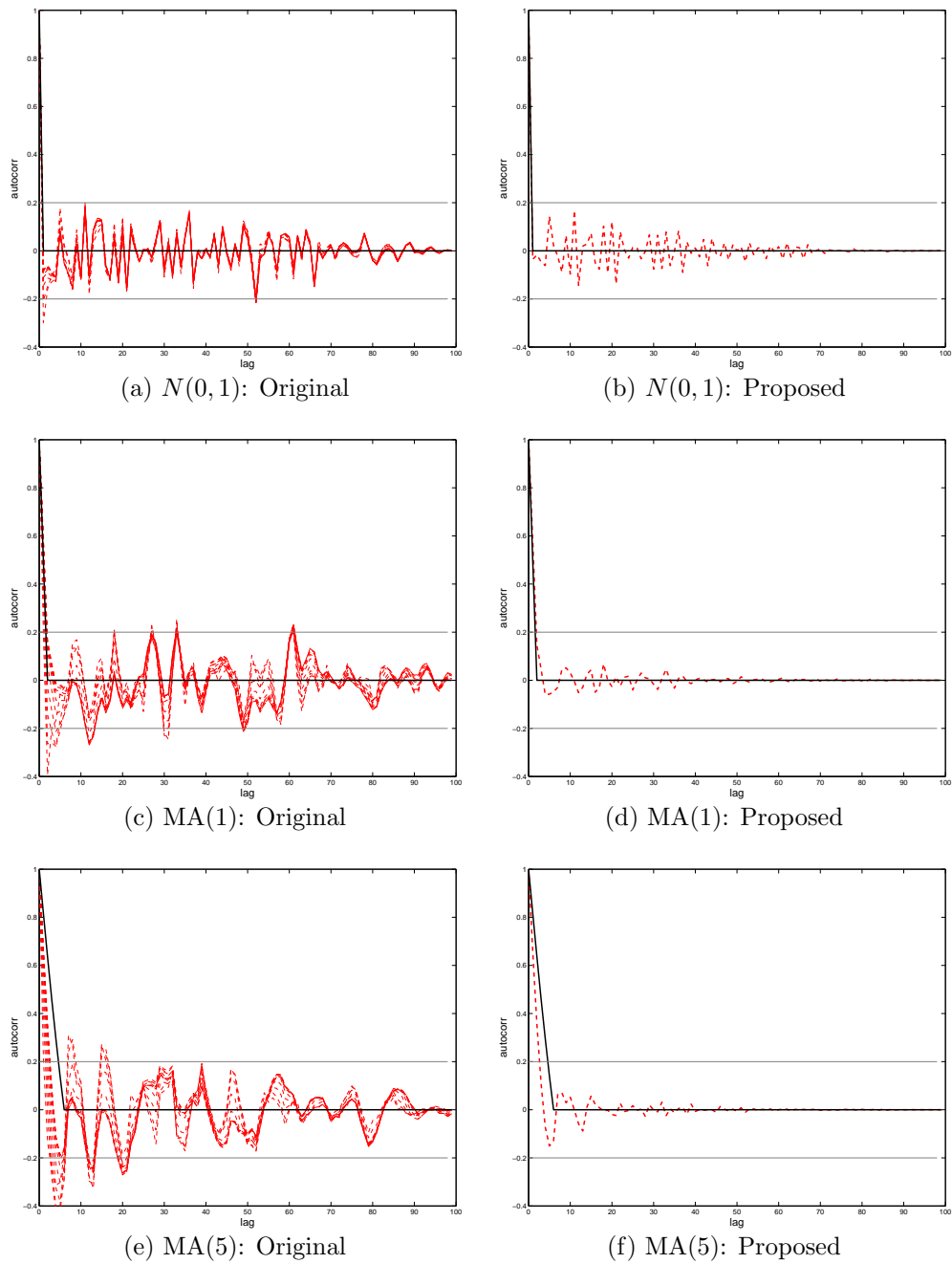


Figure 4: Comparison of estimated autocorrelation functions for  $N(0,1)$ ,  $MA(1)$ , and  $MA(5)$ : original (with 11 pilot bandwidths) versus proposed.

contrary, the proposed estimate stays within the confidence interval and looks more stable. For MA(1) and MA(5), the original estimate has a deeper deviation as the degree of dependency increases but the proposed once again stays within the interval and looks very stable in both examples.

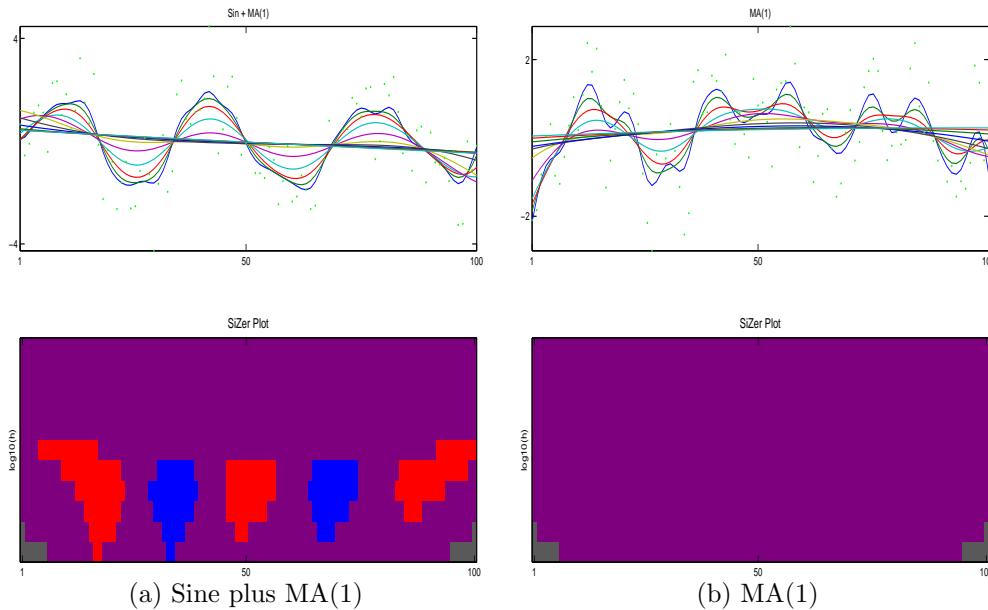


Figure 5: SiZer plots using the proposed quantile and autocovariance function estimate.

Figure 5 shows SiZer maps using the proposed autocovariance function estimate and the quantile introduced in Section 3.1. Figure 5 (a) is the SiZer plot corresponding to Figure 1. Note that there is only one SiZer plot since the proposed method does not rely on pilot bandwidths. This is a big advantage because we neither need to interpret several SiZer maps at the same time nor to select some bandwidths we should look at. The SiZer map clearly shows the sine curve trend and no spurious features appear compared to the ones in Figure 1. Figure 5 (b) corresponds to Figure 2. At first glance the data seem to have a nonlinear trend according to the family plot, but it was created by MA(1) dependence structure. The proposed SiZer is able to recognize it and shows only purple, which an ideal SiZer map would do. Needless to say spurious features disappear compared to the ones in Figure 2.

This simulation demonstrates that the proposed estimation of the autocovariance function removes unexpected features for MA(1).

## 4. Numerical study

### 4.1 Simulated data

In this section, we extend our simulation in Section 3 to other error structures. To save space we exclude the family plots for the original method.

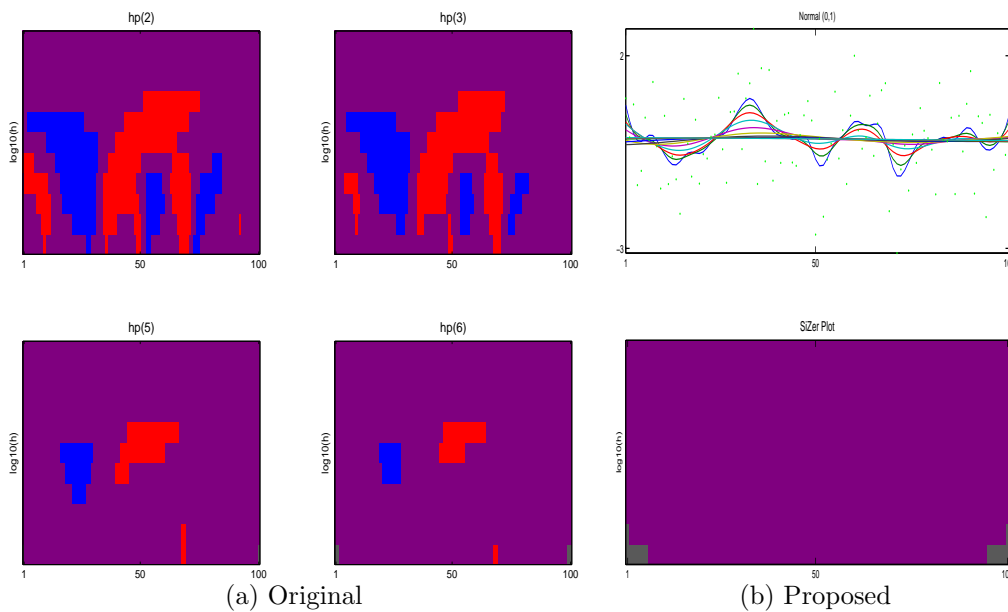
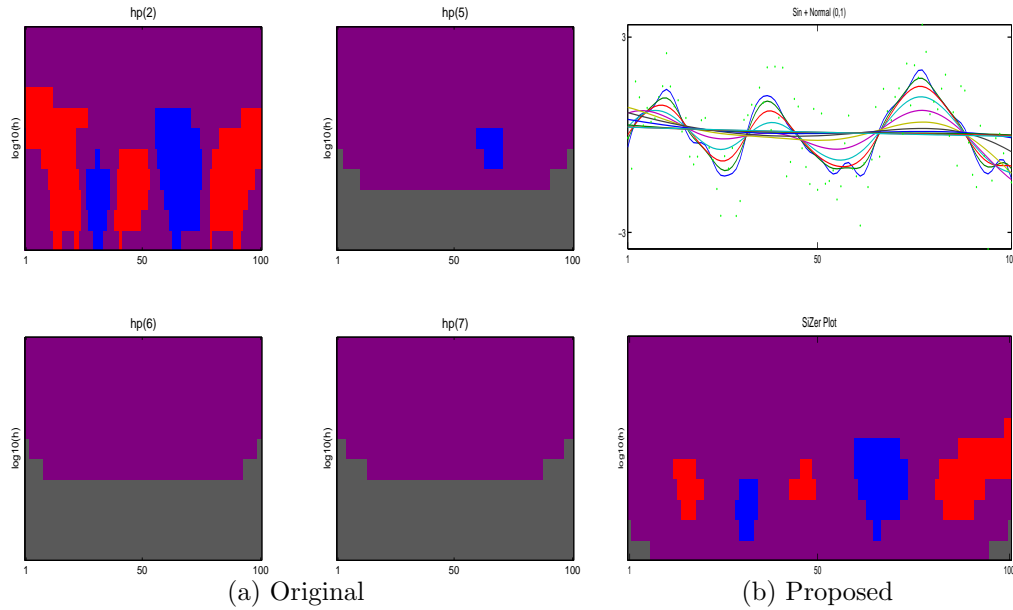


Figure 6: SiZer plots for time series:  $N(0, 1)$ .

We generate  $n = 100$  observations with zero mean function ( $f = 0$ ) and  $N(0, 1)$  error. Figure 6 draws the corresponding SiZer maps. An ideal SiZer map would show no significant features since there is no trend in this time series. Similar to MA(1) the original SiZer maps in (a) show many false significant features. On the contrary, the proposed method in (b) using both the new quantile and the autocovariance function estimate removes spurious features and shows no significant ones.

Figure 7 compares SiZer maps for  $N(0, 1)$  with the sine curve added. The

Figure 7: SiZER for time series: Sine plus  $N(0, 1)$ .

original SiZER in (a) captures the sine curve well by displaying red (decreasing) and blue (increasing) in the first SiZER map. Since the data are generated from  $N(0, 1)$ , the first SiZER map would be informative in this case. However, it would be hard to choose the right pilot bandwidth for real examples since their covariance structures are unknown in advance. The proposed method offers only one SiZER map in Figure 7 (b) and it also catches the sine curve well.

Although we assume a time series to be weakly correlated, the improved SiZER works reasonably well for some strongly correlated time series such as AR(1) with the coefficient 0.9 (the result is not reported to save space). However, it does not work well for fractional Gaussian noise with large Hurst parameters. We propose a thorough study of the proposed autocovariance estimator as future work.

## 4.2 Real examples

The real data sets shown here are the Deaths data set and the Chocolate data set, that were analyzed in Rondonotti, Marron, and Park (2007). The Deaths data set contains the monthly number of accidental deaths in US from 1973 to

1978 (in thousands) and the Chocolate data set contains the monthly production of chocolate in Australia from July of 1957 to October of 1990 (in kilotonnes). Both data sets come with the software companion to Brockwell and Davis (1996). Figure 8 shows the proposed SiZer plots for time series for these data sets. The dots in Figure 8 show number of accidental deaths and the Chocolate production after deseasonalising and linearly detrending the original time series.

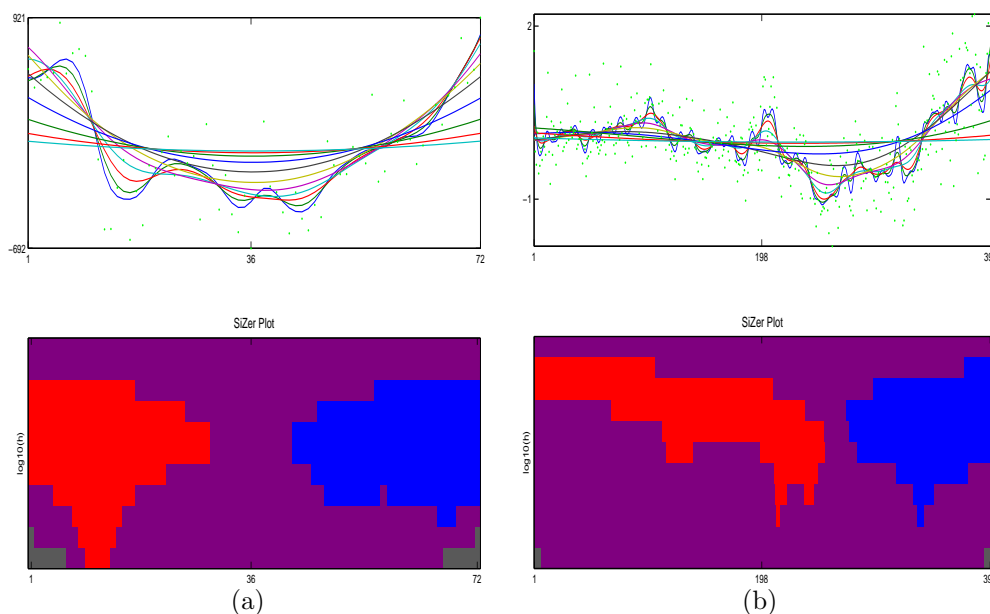


Figure 8: SiZer for time series for (a) the Deaths data set (b) the Chocolate data set.

For the Deaths data (Figure 8 (a)), the only feature that appears to be significant at most of the levels of resolution, is the valley around the third year of observation. This is a similar result to that of Rondonotti, Marron, and Park (2007) but we can see the same improvement observed in Section 4.1; some spurious features disappear. Also, the gray area has been reduced. Rondonotti, Marron, and Park (2007) found a significant increase near  $i = 20$  for smaller bandwidths in the first SiZer map, but it is not flagged as significant in Figure 8 (a).

The SiZer plots for the Chocolate data are depicted in Figure 8 (b). The significant feature is the major minimum around  $i = 250$  (which corresponds

to the year 1978), which matches the strongest feature of the SiZer plots in Rondonotti, Marron, and Park (2007). They also concluded that many peaks and valleys were significant for the smallest values of the bandwidth, but they are not flagged as significant in Figure 8 (b).

## 5. Asymptotic results

In this section, we study statistical convergence of the difference between the empirical and the theoretical scale space surfaces ( $\hat{f}_h(x)$  and  $E\hat{f}_h(x)$ ), which provides theoretical justification of SiZer for time series in scale space. Chaudhuri and Marron (2000) addressed this issue based on independent observations and we extend it to correlated data. The first theorem provides the weak convergence of the empirical scale space surfaces and their derivatives to their theoretical counterpart. The second theorem states the behavior of the difference between the empirical and the theoretical scale space surfaces under the supremum norm and the uniform convergence of the empirical version to the theoretical one.

Let  $I$  and  $H$  be compact subintervals of  $[0, \infty)$  and  $(0, \infty)$ , respectively and let  $\hat{f}_h(x) = \sum_{i=1}^n Y_i w_n(h, x, i)/n$ . We need the following set of assumptions.

**(A.1)** The errors  $(\epsilon_1, \epsilon_2, \dots)$  in (2.1) are stationary,  $\phi$ -mixing with the mixing function  $\phi(i)$  satisfying  $\sum_{i=1}^{\infty} \phi(i)^{1/2} < \infty$ . (See for example Doukhan (1994) for definition of  $\phi$  mixing.)

**(A.2)** The errors have a bounded moment  $E\{|\epsilon_i|^{2+\rho}\} < \infty$  for some  $\rho > 0$ .

**(A.3)** For integer  $m \geq 0$ , as  $n \rightarrow \infty$

$$n^{-1} \left[ \sum_{i=1}^n \sum_{j=1}^n \gamma(|j-i|) \frac{\partial^m w_n(h_1, x_1, i)}{\partial x_1^m} \frac{\partial^m w_n(h_2, x_2, j)}{\partial x_2^m} \right]$$

converges to a covariance function  $cov(h_1, x_1, h_2, x_2)$  for all  $(h_1, x_1)$  and  $(h_2, x_2) \in H \times I$ .

**(A.4)**  $n^{-(1+\rho/2)} \left\{ \max_{1 \leq i \leq n} \left| \frac{\partial^m w_n(h, x, i)}{\partial x^m} \right|^\rho \right\} \sum_{i=1}^n \left\{ \frac{\partial^m w_n(h, x, i)}{\partial x^m} \right\}^2 \rightarrow 0$  for all  $(h, x) \in H \times I$ .

**(A.5)**  $\frac{\partial^{m+2} w_n(h, x, i)}{\partial h \partial x^{m+1}} \frac{\partial^{m+2} w_n(h, x, j)}{\partial h \partial x^{m+1}}$  will be uniformly dominated by a positive finite number  $M$ .

(A.6)

$$\frac{\partial^{m+1}w_n(h, x, i)}{\partial x^{m+1}} \frac{\partial^{m+1}w_n(h, x, j)}{\partial x^{m+1}}, \frac{\partial^{m+1}w_n(h, x, i)}{\partial h \partial x^m} \frac{\partial^{m+1}w_n(h, x, j)}{\partial h \partial x^m}$$

and

$$\frac{\partial^{m+1}w_n(h, x, i)}{\partial x^{m+1}} \frac{\partial^{m+1}w_n(h, x, j)}{\partial h \partial x^m}$$

will be uniformly dominated by a positive finite number  $M^*$ .

**Theorem 1** *Suppose that assumptions (A.1)-(A.5) are satisfied. Define*

$$U_n(h, x) = n^{1/2} \left[ \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right], \quad (h, x) \in H \times I.$$

As  $n \rightarrow \infty$ ,  $U_n(h, x)$  converges to Gaussian process on  $H \times I$  with zero mean and covariance function  $\text{cov}(h_1, x_1, h_2, x_2)$ .

**Proof.** It is enough to show that all the finite dimensional distribution of the process converges weakly to the normal distribution and the process satisfies a tightness condition.

Fix  $(h_1, x_1), (h_2, x_2), \dots, (h_k, x_k) \in H \times I$  and  $(t_1, \dots, t_k) \in (-\infty, \infty)$ . Define

$$\begin{aligned} Z_n &= n^{1/2} \sum_{i=1}^k t_i \left[ \frac{\partial^m \hat{f}_{h_i}(x_i)}{\partial x_i^m} - \frac{\partial^m E\{\hat{f}_{h_i}(x_i)\}}{\partial x_i^m} \right] \\ &= n^{-1/2} \sum_{j=1}^n \epsilon_j \sum_{i=1}^k t_i \frac{\partial^m w_n(h_i, x_i, j)}{\partial x_i^m}. \end{aligned}$$

From here  $E(Z_n) = 0$  and

$$\begin{aligned} \text{Var}(Z_n) &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^k t_i t_j \left[ \sum_{l=1}^n \sum_{p=1}^n \gamma(|p-l|) \frac{\partial^m w_n(h_i, x_i, l)}{\partial x_i^m} \frac{\partial^m w_n(h_j, x_j, p)}{\partial x_j^m} \right] \\ &\longrightarrow \sum_{i=1}^k \sum_{j=1}^k t_i t_j \text{cov}(h_i, x_i, h_j, x_j) \end{aligned} \quad (5.1)$$

as  $n \rightarrow \infty$  by assumption (A.3).

Assumptions (A.2) and (A.4) imply that Lyapunov's and hence Lindeberg's condition holds for the terms in  $Z_n$ . This and assumption (A.1) verify the conditions of the main theorem in Utev (1990) allowing us to conclude that  $Z_n$  converges in distribution to a normal random variable with variance given by (5.1).

By Cramèr-Wold device, the limiting distribution of  $U_n(h_i, x_i)$  ( $i = 1, \dots, k$ ) is the multivariate normal distribution with zero mean and  $\text{cov}(h_i, x_i, h_j, x_j)$  as the  $(i, j)$ th entry of the limiting variance-covariance matrix.

We now proceed to the issue of tightness. Fix  $h_1 < h_2$  in  $H$  and  $x_1 < x_2$  in  $I$ . Then, by Bickel and Wichura (1971) the second moment of increment of  $U_n$  is defined by

$$\begin{aligned} E_{g_n} \{U_n(h_2, x_2) - U_n(h_2, x_1) - U_n(h_1, x_2) + U_n(h_1, x_1)\}^2 \\ = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \gamma(|j-i|) D_i D_j, \end{aligned} \quad (5.2)$$

where

$$D_i = \frac{\partial^m w_n(h_2, x_2, i)}{\partial x_2^m} - \frac{\partial^m w_n(h_2, x_1, i)}{\partial x_1^m} - \frac{\partial^m w_n(h_1, x_2, i)}{\partial x_2^m} + \frac{\partial^m w_n(h_1, x_1, i)}{\partial x_1^m}.$$

Then, by the assumption (A.5) equation (5.2) is bounded by

$$C_1(x_2 - x_1)^2 (h_2 - h_1)^2 \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \gamma(|i-j|),$$

which is again bounded by  $C_2(x_2 - x_1)^2 (h_2 - h_1)^2$ , since conditions (A.1) and (A.2) imply that  $\sup_n n^{-1} \sum_{i=1}^n \sum_{j=1}^n \gamma(|i-j|) < \infty$ , c.f., Doukhan (1994), page 45. Then the tightness property of the sequence of processes

$$n^{1/2} \left[ \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right]$$

on  $H \times I$  is implied by the Theorem 3 in Bickel and Wichura (1971). Together with the finite dimensional convergence property, this implies that the theorem holds.

**Theorem 2** *Suppose that assumptions (A.1)-(A.6) are satisfied. As  $n \rightarrow \infty$ ,*

$$\sup_{x \in I, h \in H} n^{1/2} \left| \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right|$$

*converges weakly to a random variable that has the same distribution as that of  $\sup_{x \in I, h \in H} |Z(h, x)|$ , where  $Z(h, x)$  is a Gaussian process with zero mean and covariance function  $\text{cov}(h_1, x_1, h_2, x_2)$  so that*

$$P\{Z(h, x) \text{ is continuous for all } (h, x) \in H \times I\} = 1,$$

and consequently  $P \left\{ \sup_{x \in I, h \in H} |Z(h, x)| < \infty \right\} = 1$ .

**Proof.** Let us denote  $D_i^*$  by

$$D_i^* = \frac{\partial^m w_n(h_2, x_2, i)}{\partial x_2^m} - \frac{\partial^m w_n(h_1, x_1, i)}{\partial x_1^m}.$$

Then just as in Chaudhuri and Marron (2000),

$$\begin{aligned} E\{U_n(h_2, x_2) - U_n(h_1, x_1)\}^2 &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \gamma(|j-i|) D_i^* D_j^* \\ &\leq C_3 \{(h_2 - h_1)^2 + (x_2 - x_1)^2\}. \end{aligned}$$

Define the pseudo metric  $d$  by  $d\{(h_2, x_2), (h_1, x_1)\} = [E\{Z(h_2, x_2) - Z(h_1, x_1)\}^2]^{1/2}$ . Then, the rest of the proof can be done by the same way in Chaudhuri and Marron (2000).

### Acknowledgment

We would like to thank Taewook Lee for his helpful comments. The first author was supported by National Security Agency Grant No. H982300810056. The second author was supported in part by the National Science Foundation under Grants No. 0504737 and 0707037. The third author was supported by Korea Research Foundation Grant funded by Korea Government (MOEHRD) (KRF-2007-013-C00013).

### References

- Berman, S. M. (1964). Limit theorems for the maximum term in stationary sequences. *Ann. Math. Statist.* **35**, 502–516.
- Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42**, 1656–1670.
- Brockwell, P. J. and Davis, R. A. (1996). *Introduction to time series and forecasting*. Springer–Verlag, New York.

- Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.* **94**, 807–823.
- Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.* **28**, 408–428.
- Doukhan, P. (1994). *Mixing: Properties and Examples*. Lecture Notes in Statistics **85**. Springer, New York.
- Durrett, R. (2005). *Probability: Theory and Examples (3rd ed.)*. Belmont, CA: Duxbury Press.
- Erästö, P. and Holmström, L. (2005). Bayesian Multiscale Smoothing for Making Inferences about Features in Scatterplots. *J. Comput. Graph. Statist.* **14**, 569–589.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.
- Godtliebsen, F. and Øigård, T. A. (2005). A visual display device for significant features in complicated signals. *Comput. Statist. Data Anal.* **48**, 317–343.
- Hannig, J. and Lee, T. C. M. (2006). Robust SiZer for exploration of regression structures and outlier detection. *J. Comput. Graph. Statist.* **15**, 101–117.
- Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *J. Amer. Statist. Assoc.* **101**, 484–499.
- Hsing, T., Husler, J., and Riess, R. D. (1996). The Extremes of a Triangular Array of Normal Random Variables. *Ann. Appl. Probab.* **6**, 671–686.
- Kim, C. S. and Marron, J. S. (2006). SiZer for jump detection. *J. Nonpara. Statist.* **18**, 13–20.
- Lindeberg, T. (1994). *Scale Space Theory in Computer Vision*. Kluwer, Boston.
- Øigård, T. A., Rue, H. and Godtliebsen, F. (2006). Bayesian multiscale analysis for time series data. *Comput. Statist. Data Anal.* **51**, 1719–1730.

- Park, C., Hernández-Campos, F., Marron, J. S., and Smith, F. D. (2005). Long-range dependence in a changing Internet traffic mix. *Computer Networks* **48**, 401–422.
- Park, C., Hernández Campos, F., Le, L., Marron, J. S., Park, J., Pipiras, V., Smith F. D., Smith, R. L., Trovero, M., and Zhu, Z. (2006). Long-range dependence analysis of Internet traffic. *Under revision, Technometrics*. Web-available at <http://www.stat.uga.edu/~cpark/papers/LRDWebPage5.pdf>
- Park, C., Godtlielsen, F., Taqqu, M., Stoev, S., and Marron, J. S. (2007a). Visualization and inference based on wavelet coefficients, SiZer and SiNos. *Comput. Statist. Data Anal.* **51**, 5994–6012.
- Park, C., Lazar, N. A., Ahn, J., and Sornborger, A. (2007b). Do Different Parts of the Brain Have the Same Dependence Structure? A Multiscale Analysis of the Temporal and Spatial Characteristics of Resting fMRI Data. *Submitted to Ann. Appl. Statist.*
- Park, C. and Kang, K. -H. (2008). SiZer Analysis for the Comparison of Regression Curves. *Comput. Statist. Data Anal.* **52**, 3954–3970.
- Park, C., Marron, J. S. and Rondonotti, V. (2004). Dependent SiZer: goodness of fit tests for time series models. *J. Appl. Statist.* **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 J. Statist.* **1**, 268–289.
- Utev, S. A. (1990), The central limit theorem for  $\varphi$ -mixing arrays of random variables, *Theory Probab. Appl.*, **35**, 110–117.

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: [cpark@stat.uga.edu](mailto:cpark@stat.uga.edu)

Department of Statistics, Colorado State University, Fort Collins, CO 80526, U.S.A.

E-mail: [jan.hannig@colostate.edu](mailto:jan.hannig@colostate.edu)

Department of Statistics, Hankuk University of Foreign Studies, Yongin 449–791, Korea.

E-mail: [khkang@hufs.ac.kr](mailto:khkang@hufs.ac.kr), corresponding author