Nonparametric Estimation of a Log-Variance Function in Scale-Space

Cheolwoo Park

Department of Statistics, University of Georgia, Athens, GA 30602, USA

Jib Huh

Department of Statistics, Duksung Women's University, Seoul 132-714, Republic of Korea

March 18, 2013

Abstract

In a nonparametric regression setting, we consider the kernel estimation of the logarithm of the error variance function, which might be assumed to be homogeneous or heterogeneous. The objective of the present study is to discover important features in the variation of the data at multiple locations and scales based on a nonparametric kernel smoothing technique. Traditional kernel approaches estimate the function by selecting an optimal bandwidth, but it often turns out to be unsatisfying in practice. In this paper, we develop a SiZer (SIgnificant ZERo crossings of derivatives) tool based on a scale-space approach that provides a more flexible way of finding meaningful features in the variation. The proposed approach utilizes local polynomial estimators of a log-variance function using a wide range of bandwidths. We derive the theoretical quantile of confidence intervals in SiZer inference and also study the asymptotic properties of the proposed approach in scale-space. A numerical study via simulated and real examples demonstrates the usefulness of the proposed SiZer tool.

Key words: Asymptotics, Nonparametric regression, Quantile, Scale-Space, SiZer, Variance function

1 Introduction

Suppose we observe a bivariate sample (X_i, Y_i) of (X, Y) where Y_i's are real valued responses associated with covariates X_i's having a density f with support $[0, 1]$, $i = 1 \ldots, n$. Let $m(x) = E(Y|X = x)$ denote the regression mean function. We denote by $v(x)$ the conditional variance function of Y given $X = x$. Then, a nonparametric regression model can be written as

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where, conditional on $X_1, \ldots, X_n$, the ε_i's are independent random variable with mean 0 and variance 1.

The estimation of the variance function $v(x)$ is crucial for statistical inference on the regression model in (1.1) such as construction of confidence and prediction intervals for the mean function. Hence, it is frequently of interest to determine whether $v(x)$ can be assumed to be homogeneous or heterogeneous, and to detect any trends in the variation of the data.

An extensive literature exists for the variance function estimation in nonparametric regression, much of which is based on using squared residuals from a nonparametric fit of the regression mean function. Rice (1984), Gasser et al. (1986), and Hall et al. (1990) estimated the homogeneous error variance. Their estimates are based on the squared differences of the data of various orders. For heteroscedasticity, Müller and Stadtmüller (1987) provided an estimate by using Gasser-Müller type kernel weighted averages of initial local variance estimates, and showed that the estimator is uniformly consistent. Hall and Carroll (1989) studied the influence of the smoothness of the mean function on the convergence rate of the nonparametric variance estimator. Ruppert et al. (1997), Härdle and Tsybakov (1997), and Fan and Yao (1998) considered a local polynomial estimator for the variance function and obtained the asymptotic results for the bias and the variance of the estimator. However, the local polynomial fit having a merit for the rate of convergence cannot be used because its estimate of the variance cannot always be guaranteed to be non-negative. To avoid the restriction of non-negativity and improve the accuracy for the bias term, Yu and Jones (2004) proposed local polynomial estimators of the log-variance function based on a kernel weighted local likelihood. Note that typical nonparametric approaches require a single bandwidth selection process based on asymptotic theory or fine tuning. However, the bandwidth selected by an asymptotic optimal criterion sometimes misses important features with a finite sample size while

a careful trial and error approach might take quite some time even for the skilled data analysts.

In this paper, we also take a nonparametric smoothing approach, but apply the SiZer (SIgnificant ZERo crossings of derivatives) methodology originally developed by Chaudhuri and Marron (1999). SiZer attempts to distinguish the true features from the sampling noise in the data, so that it enables one to find meaningful features during exploratory data analysis. What separates SiZer from other nonparametric smoothing techniques is so called, *scale-space* approach (Lindeberg, 1994), which argues that each level of smoothing is regarded to provide information about the underlying structure at a particular scale. In SiZer, scale-space is a family of kernel smooths indexed by the bandwidth. Therefore, SiZer focuses on smoothed versions of the underlying curve rather than the true curve and utilizes all the information available across a wide range of resolutions. This allows one to avoid selection of an optimal bandwidth and a bias problem that occurs in estimating the true underlying function.

Different versions of SiZer have been developed and have demonstrated their practical usefulness in many applications. The improvement of SiZer inference was done by Hannig and Marron (2006). Hannig and Lee (2006) proposed a robust version of SiZer that examines the median function, and Kim and Marron (2006) developed a tool for detecting jump points in the regression data. Also, SiZer has been extended to survival analysis (Marron and de Uña Álvarez, 2004), generalized linear models (Li and Marron, 2005; Ganguli and Wand, 2007; Park and Huh, 2013), smoothing spline (Marron and Zhang, 2005) and additive models (González-Manteiga et al., 2008). Park and Kang (2008) proposed a SiZer that compares multiple curves based on their differences of smooths. SiZer has been also applied to time series data (Park et al., 2004; Rondonotti et al., 2007; Park et al., 2007, 2009a,b). In addition, various Bayesian multiscale smoothing techniques have been proposed (Erästö and Holmström, 2005; Godtliebsen and Oigard, 2005; Oigard et al., 2006; Erästö and Holmström, 2007; Sørbye et al., 2009). Note that most of these SiZer tools aim to discover the important features in the mean or median structure of the data. As an exception, Park et al. (2010) studied a SiZer which targets the quantile composition of the data.

The main contribution of this article is the proposal of a new SiZer tool that targets the variation of the data. The proposed approach utilizes local polynomial estimators of a log-variance function with a wide range of bandwidths. To determine significance of features SiZer constructs simultaneous confidence intervals for each location and bandwidth and visualizes the testing results

as a color map, called *SiZer map*. In SiZer inference we propose a theoretically justified quantile for the confidence intervals. Therefore, the proposed SiZer serves as a powerful exploratory multiscale analysis tool for the variation of the data, equipped with accurate statistical inference and informative visualization. It can conduct a goodness-of-fit test for the homogeneous assumption or discover significant structures if any at each resolution. Another contribution is to provide the convergence properties of the proposed SiZer tool in scale-space. This theoretical study is different from conventional ones done in nonparametric kernel estimation contexts in the sense that the estimators are considered as a stochastic process with two parameters, location and bandwidth.

The rest of the paper is organized as follows. Section 2 proposes a SiZer tool for estimating a log-variance function based on a local likelihood approach and presents its asymptotic results. Section 3 investigates the performance of the proposed SiZer tool using both simulated and real examples. Then, we discuss some issues in the proposed approach in Section 4. In Section 5, we provide proofs of the asymptotic results introduced in Section 2.3.

2 SiZer for a log-variance function

In Section 2.1, we introduce the local likelihood estimator for a log-variance function, and then propose a SiZer tool based on the estimator in Section 2.2. Section 2.3 presents the convergence properties of the proposed SiZer tool in scale-space.

2.1 Local likelihood estimation of a log-variance function

In the nonparametric regression model given in (1.1), if we assume the error ε_i, $i = 1, \ldots, n$, to be the independent standard normal random variable, then the square of the error $\varepsilon_i^2 = \{Y_i - m(X_i)\}^2/v(X_i)$ follows the χ^2 distribution with one degree of freedom, $\chi^2(1)$. Even when ε_i does not have the exact standard normal distribution, ε_i^2 approximately follows the $\chi^2(1)$ if the distribution of ε_i is symmetric.

Let us define $R = \{Y - m(X)\}^2$. The conditional density of R given $X = x$ is then

$$f_{R|X}(r|x) = \exp\Big\{ -\frac{1}{2}\log 2\pi - \frac{1}{2}\Big(\log r + \log v(x) + \frac{r}{v(x)} \Big) \Big\}, \tag{2.1}$$

which belongs to the one parameter exponential family. Using the local polynomial fit (Fan and Gijbels, 1996) based on the likelihood function in (2.1), Yu and Jones (2004) and Huh (2011)

estimated continuous and discontinuous log-variance function respectively. We also consider the estimator of $v(x)$ by applying the relation $v(x) = \mathrm{e}^{s(x)}$. Similarly, for $k > 0$, the kth derivative of the variance function $v^{(k)}$ can be calculated successively by applying the relation between $s^{(k)}$ and $v^{(k)}$:

$$\begin{aligned} s^{(k)}(x) &= (\log v(x))^{(k)} \\ &= \sum_{j=1}^{k-1} \binom{k-1}{j} \Big(\frac{1}{v(x)}\Big)^{(j)} v^{(k-j)}(x) + \frac{v^{(k)}(x)}{v(x)}. \end{aligned}$$

Write $\ell(z, r)$ for the logarithm of the conditional density in (2.1) with $\log v(x)$ replaced by z. Note that

$$\ell(z, r) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log r - \frac{1}{2}z - \frac{1}{2}r\mathrm{e}^{-z}.$$

Also, define $\widetilde{s}_h^{(k)}(x) = k!\widetilde{\beta}_k$ as the estimator for $s^{(k)}(x)$, where the $(p+1) \times 1$ vector $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \cdots, \widetilde{\beta}_p)^T$ maximizes the following kernel weighted local likelihood function:

$$\sum_{i=1}^{n} \ell\left(\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p, \{Y_i - m(X_i)\}^2\right) K\left(\frac{X_i - x}{h}\right), \tag{2.2}$$

where p is a nonnegative integer satisfying $p \geq k$. Here, K is a kernel function and h is a bandwidth. We use the Gaussian kernel in this paper following the suggestion by the original work in Chaudhuri and Marron (1999). In the case of $\ell(z, r) = (z - r)^2$, $\widetilde{\boldsymbol{\beta}}$ is equivalent to the kernel weighted least squares estimator with $\{Y_i - m(X_i)\}^2$ discussed in Section 3 of Fan and Gijbels (1996).

The estimator $\widetilde{s}_h^{(k)}$ involves the unknown regression mean function m. Hence, we obtain the estimator $\widehat{s}_h^{(k)}$ using the maximizer of (2.2) with $m(X_i)$ being replaced by its estimate $\widehat{m}(X_i)$. It turns out that the estimation of $m(X_i)$ plays an important role in SiZer inference, and we will discuss this issue below with more details.

Define $\ell_i(z, r) = \partial^i \ell(z, r)/\partial z^i$, $i = 1, 2$. Note that ℓ_i is linear in r for fixed z and that

$$\ell_1(s(x), v(x)) = 0, \quad \ell_2(s(x), v(x)) = -\frac{1}{2}.$$

Since $\ell_2(z, r) = -\frac{1}{2}r\mathrm{e}^{-z} < 0$ for all real z and positive r, the kernel weighted log likelihood function in (2.2) is concave in $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$, which ensures the uniqueness of the maximizer. Note that we do not have explicit solutions to the maximization of (2.2) except for the case $p = 0$ and $k = 0$ where the estimator $v(x)$ is the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964):

$$\widehat{v}_h(x) = \mathrm{e}^{\widehat{s}_h(x)} = \frac{\sum_{i=1}^n K((X_i - x)/h)\{Y_i - \widehat{m}(X_i)\}^2}{\sum_{i=1}^n K((X_i - x)/h)}.$$

From here on, we consider a local linear estimator when $p = 1$ in this paper.

2.2 Statistical inference in scale-space

The objective of the study is to determine significance of trends in the function s at a particular location x by examining the first derivative of s. The null hypothesis of interest is then $H_0 : s'(x) = 0$. However, because the proposed method attempts to find significant features in the noisy data at a given level of resolution, the null hypothesis of SiZer inference is given as $H_0 : s'_h(x) = 0$ where $s'_h(x) = E[\hat{s}'_h(x)]$. This is motivated by the scale-space idea that the true exists at each scale (Lindeberg, 1994).

A SiZer map visually displays the inference results from the simultaneous $100(1-\alpha)\%$ confidence intervals for $s'_h(x)$. In a map, the horizontal (row) and vertical (column) axes represent the location x and the bandwidth h, respectively. Proposition 1 illustrates how to address the multiple comparisons issue in SiZer inference.

Proposition 1 *Let g be the number of pixels in each row, and $Q(x_1), \ldots, Q(x_g)$ denote the test statistics of a row in the SiZer map. Suppose that the assumptions (A.1)–(A.4) in Section 5 are satisfied. Define the cluster index*

$$\theta = 2\Phi\left(\sqrt{3\log g}\,\frac{\tilde{\Delta}}{2h}\right) - 1$$

where Φ is the cumulative distribution function of the standard normal and $\tilde{\Delta}$ denotes the distance between the pixels of the SiZer map. Then,

$$P\left[\max_{i=1,\ldots,g} Q(x_i) \le x\right] \approx \Phi(x)^{\theta g}.$$

The proof of Proposition 1 is given in Section 5.2. Therefore, the simultaneous $100(1-\alpha)\%$ confidence intervals for $s'_h(x)$ for a given h are expressed as

$$\hat{s}'_h(x) \pm q(h)\widehat{SD}(\hat{s}'_h(x)) \tag{2.3}$$

where $q(h)$ is the quantile at a nominal level α given as

$$q(h) = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{1/(\theta g)}\right). \tag{2.4}$$

The nominal level of $\alpha = 0.05$ used for all of the numerical examples in this paper. It can be seen that $0 \leq \theta \leq 1$ and θg can be interpreted as the equivalent number of independent observations. Therefore, for a small (large) scale, the cluster index θ becomes large (small) when g and $\tilde{\Delta}$ are fixed, which yields a large (small) quantile since there are a large (small, respectively) number of independent comparisons. Note that the quantile $q(h)$ defined in (2.4) takes multiple comparisons adjustment into account across different locations for a given h. One can derive the quantile that accounts for both location and scale by following the global inference steps in Hannig and Marron (2006).

The estimate of the standard deviation (SD) is also given in Section 5.2. For a local linear estimator with the Gaussian kernel function, the standard deviation is given as

$$SD(\hat{s}'_h(x)) = \sqrt{\frac{1}{4\sqrt{\pi}nh^3 f(x)}\Big(\frac{\kappa(x)}{v^2(x)} - 1\Big)}, \tag{2.5}$$

where $\kappa(x) = E[\{Y - m(X)\}^4 | X = x]$ is the central fourth moment of Y given $X = x$.

In a SiZer map, each pixel shows a color that gives the result of the hypothesis test $H_0 : s'_h(x) = 0$ at x and h based on (2.3). For a given h, if the curve is increasing at x, i.e., the confidence interval for $s'_h(x)$ is above 0, then that particular map location is colored black. Similarly, if the curve is decreasing at x, i.e., the confidence interval for $s'_h(x)$ is below 0, then that particular map location is colored white. On the other hand, if the curve does not have a statistically significant slope, i.e., the confidence interval includes 0, then that map location is flagged intermediate gray. Finally, if the sample size around x is not sufficiently large, then no test is conducted and the pixel is colored dark gray. In order to determine the dark gray areas, we use the estimated effective sample size (ESS) suggested by Chaudhuri and Marron (1999); for each (x, h)

$$\mathrm{ESS}(x, h) = \frac{\sum_{i=1}^{n} K_h(X_i - x)}{K_h(0)},$$

where $K_h(\cdot) = K(\cdot/h)/h$. If $\mathrm{ESS}(x, h) < 5$, then the corresponding pixel is colored dark gray.

To construct a confidence interval in (2.3), the functions m, v, κ and f in (2.5) should be estimated. Because the proposed approach is scale-dependent, these estimates are also dependent on the bandwidth. For v and f, we use the same bandwidth h in a SiZer map for estimating a log-variance function, in other words, $\hat{v}_h(x) = \exp(\hat{s}_h(x))$, and

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x).$$

For $\kappa(x)$, one could estimate it separately by applying a kernel technique to $(Y_i - \hat{m}(X_i))^4$, or approximate it by setting $\kappa(x)/v^2(x) - 1 = 2$ under the assumption that ϵ_i^2 follows $\chi^2(1)$. We use the latter approach, $\hat{\kappa}_h(x) = 3\hat{v}_h^2(x)$ in Section 3 so that one only needs to estimate $f(x)$ for the standard deviation.

For the mean function m, we apply a local linear estimator, but do not use the same bandwidth h that is used for estimating the function v (or s). If one uses the same bandwidth for both m and v, a trend in m could be incorrectly carried over to the estimation of v for large bandwidths and vice versa for small bandwidths. Hence, we use a pilot bandwidth h_p for estimating the mean function m, that is different from the bandwidth h used for constructing a SiZer map. In light of the scale-space idea, we also consider a wide range of pilot bandwidths h_p in this paper. A large h_p assumes little structure and a small one corresponds to complicated structure in the mean function. This means the addition of another dimension to the SiZer plot, and thus we draw a series of SiZer plots indexed by the pilot bandwidth h_p. In our numerical analysis, we try various h_ps, but choose three sets of pilot bandwidths because the complete series of SiZer plots would be unnecessarily long. The simultaneous view of all these SiZer plots is hard to comprehend and the information contained in several such plots is often redundant. In the implementation of a SiZer plot, we calculate the range of the covariates in the given data, and divide it into 11 equally spaced values on a logarithmic scale. Among the 11 pilot bandwidths, we choose the second, fifth, and eighth smallest bandwidths ($h_p(2), h_p(5)$ and $h_p(8)$) that reflect the effects of small, medium and large bandwidths. We do not recommend to choose the smallest pilot bandwidth $h_p(1)$ because it sometimes produces spurious features due to oversmoothing.

2.3 Asymptotic results

In this section, we derive asymptotic properties of the proposed SiZer in scale-space while fixing (x, h) at a certain location x and a certain resolution level h. This theoretical study is different from conventional ones done in nonparametric kernel estimation contexts in the sense that the estimators are considered as a stochastic process with two parameters, location and bandwidth. In a similar fashion to Chaudhuri and Marron (2000) whose work is done on the conventional SiZer, we study the statistical convergence of the difference between the kth derivative of the log-variance function $\widehat{s}_h^{(k)}(x)$ and its theoretical scale-space surfaces $E\{\widehat{s}_h^{(k)}(x)\}$ for $k = 0, 1, \ldots$.

Theorem 1 provides the weak convergence of the empirical scale-space to their theoretical scale-space surfaces. Let I and H be compact subintervals of $[0,1]$ and $(0,\infty)$, respectively.

Theorem 1 *Suppose that the assumptions (A.1)–(A.7) in Section 5 are satisfied. Then as* $n \longrightarrow \infty$*, the two parameter stochastic process*

$$U_n(h,x) = n^{1/2}\left[\widehat{s}_h^{(k)}(x) - E\{\widehat{s}_h^{(k)}(x)\}\right]$$

with $(h,x) \in H \times I$ *converges weakly to a Gaussian process on* $H \times I$ *with zero mean and covariance function* $cov(h_1,x_1,h_2,x_2)$.

Theorem 2 investigates the difference between the empirical and the theoretical scale-space surfaces under the supremum norm and their uniform convergence.

Theorem 2 *Suppose that the assumptions (A.1)–(A.8) in Section 5 are satisfied. Then as* $n \longrightarrow \infty$*,*

$$\sup_{x\in I, h\in H} n^{1/2}\left|\widehat{s}_h^{(k)}(x) - E\{\widehat{s}_h^{(k)}(x)\}\right|$$

converges weakly to a random variable that has the same distribution as that of $\sup_{x\in I,h\in H}|Z(h,x)|$. *Here* $Z(h,x)$ *with* $(h,x) \in H \times I$ *is a Gaussian process with zero mean and covariance function* $cov(h_1,x_1,h_2,x_2)$ *as defined in Theorem 1 so that*

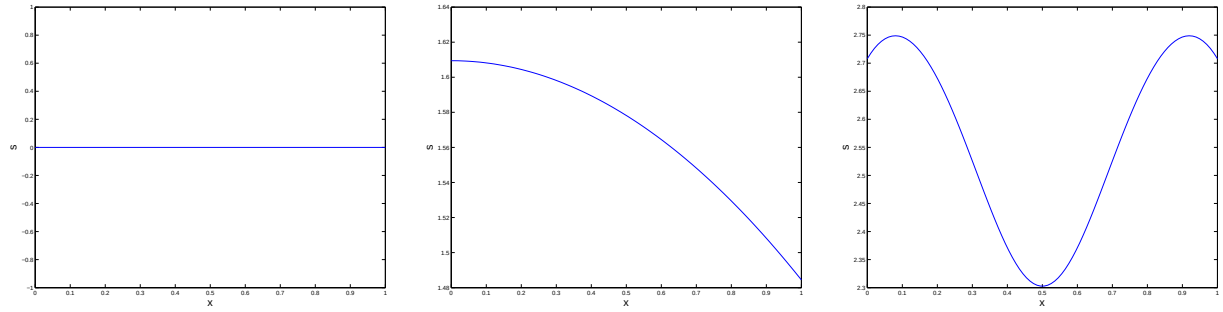
$$Pr\{Z(h,x) \text{ is continuous for all } (h,x) \in H \times I\} = 1,$$

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

Remark Although we consider a wider range for h_p in our numerical analysis, we assume that $h_p \to 0, \quad nh_p \to \infty$ when $n \to \infty$ in the proof of Lemma 1 (see (5.2)). A further discussion on the pilot bandwidth follows in Section 4.

3 A numerical study

In this section, we demonstrate the practical aspects of the proposed SiZer using both simulated and real examples.



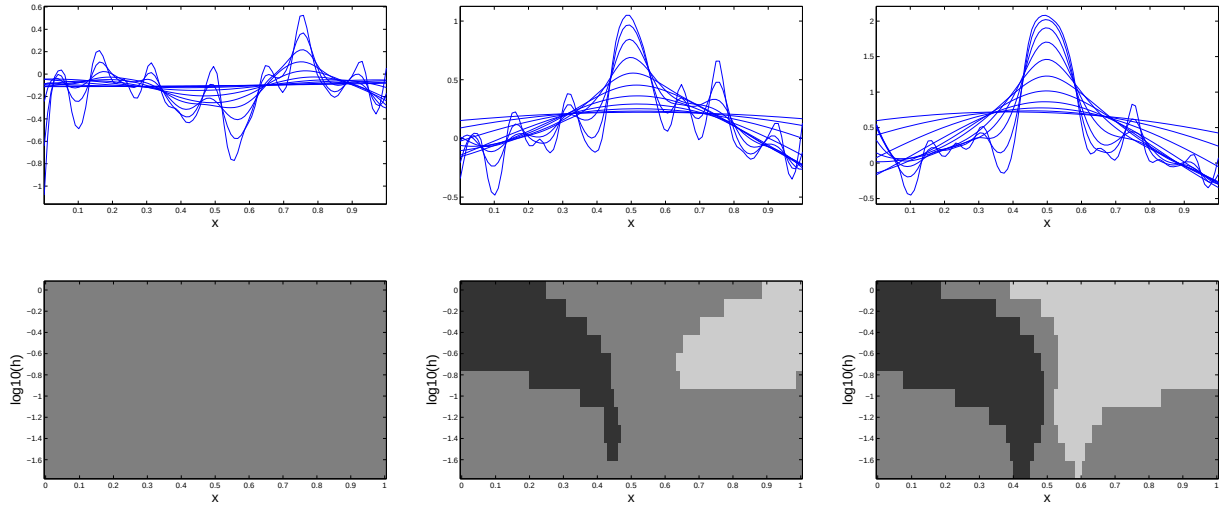
(a) Constant (b) Decreasing (c) Sinusoidal

Figure 1: Three log-variance functions $s(x) = \log v(x)$ are drawn: (a) $v(x) = 1$, (b) $v(x) = 5\exp(-x^2/8)$ and (c) $v(x) = 5(2 + \sin(\pi x) + \cos(2\pi x))$.

3.1 Simulated examples

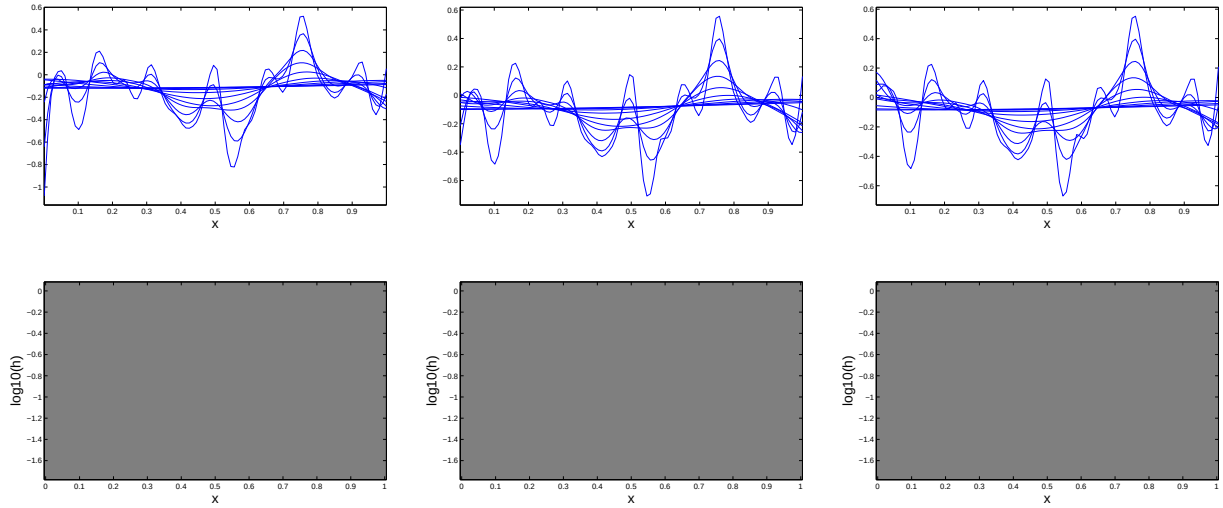
We test three different variance functions, $v(x) = 1$ (homogeneous), $v(x) = 5\exp(-x^2/8)$ (heteroscedastic with a decreasing function), and $v(x) = 5(2 + \sin(\pi x) + \cos(2\pi x))$ (heteroscedastic with a sinusoidal function), to see if SiZer analysis correctly detect their trends. For the constant variance case, one can conduct a goodness-of-fit test for the homogeneous assumption. Yu and Jones (2004) used similar examples for the heteroscedastic setups where smoothing cannot easily distinguish between structure in the mean and that in the variance. The three log-variance functions $s(x) = \log v(x)$ are graphed in Figure 1. We choose two mean functions, $m(x) = 0$ (no signal) and $m(x) = 4x + 4\exp(-100(x - 0.5)^2)$ (complex signal). It would be more challenging for SiZer to extract the variance component when a signal is present. Each example has the sample size $n = 500$, X's are generated from $U(0, 1)$ and ϵ's are generated from $N(0, 1)$.

Figure 2 shows SiZer plots for $m(x) = 4x + 4\exp(-100(x - 0.5)^2)$ and $v(x) = 1$ with three pilot bandwidths $h_p(2), h_p(5)$ and $h_p(8)$. In the top panels, the thin curves display the family of smooths, which are local linear smooths $\hat{s}_h(x)$ with different h by maximizing (2.2). In Figure 2(a), for $h_p(2)$, these curves are located around 0 as expected. By contrasts the families of smooths plots for $h_p(5)$ and $h_p(8)$ show the global increasing and decreasing trends at most of the resolutions. The bottom panels show SiZer maps that are arrays of colored pixels. The horizontal locations correspond to the horizontal locations in the family of smooths in the top panel. The vertical locations correspond to the level of smoothing on a log scale, and each row of the SiZer map provides statistical inference



(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 2: $m(x) = 4x + 4\exp(-100(x-0.5)^2)$ and $v(x) = 1$. Three selected SiZer plots with the pilot bandwidths $h_p(2), h_p(5), h_p(8)$ are displayed.

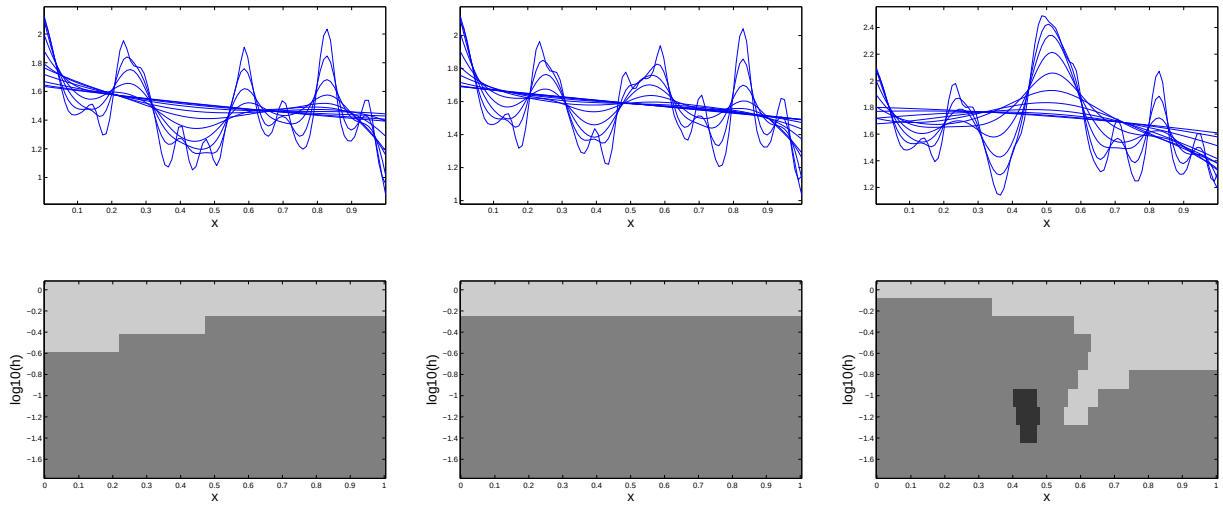


(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 3: $m(x) = 0$ and $v(x) = 1$.

for one of the thin curves in the top panel. The inference focuses on the slope of the curve, and each pixel shows a color that gives the result of a hypothesis test for the sign of the first derivative of the thin curve at each (x, h) as explained in Section 2.2. The result shows mostly intermediate gray, and thus it correctly indicates no significant trends in the log-variance function. However, for $h_p(5)$ and $h_p(8)$ in Figures 2(b) and (c) respectively, the SiZer maps find the global trends in the family of smooth plot significant. These features are artificially created by a rough estimate of the mean function and some remaining structures from the mean estimation are consequently carried over to the estimation of the log-variance function.

Figure 3 displays SiZer plots for $m(x) = 0$ and $v(x) = 1$ with three pilot bandwidths $h_p(2), h_p(5)$ and $h_p(8)$. In this case, since there is no trend in the mean function, all three SiZer plots correctly do not show any significant trend in the SiZer plots for the log-variance function.

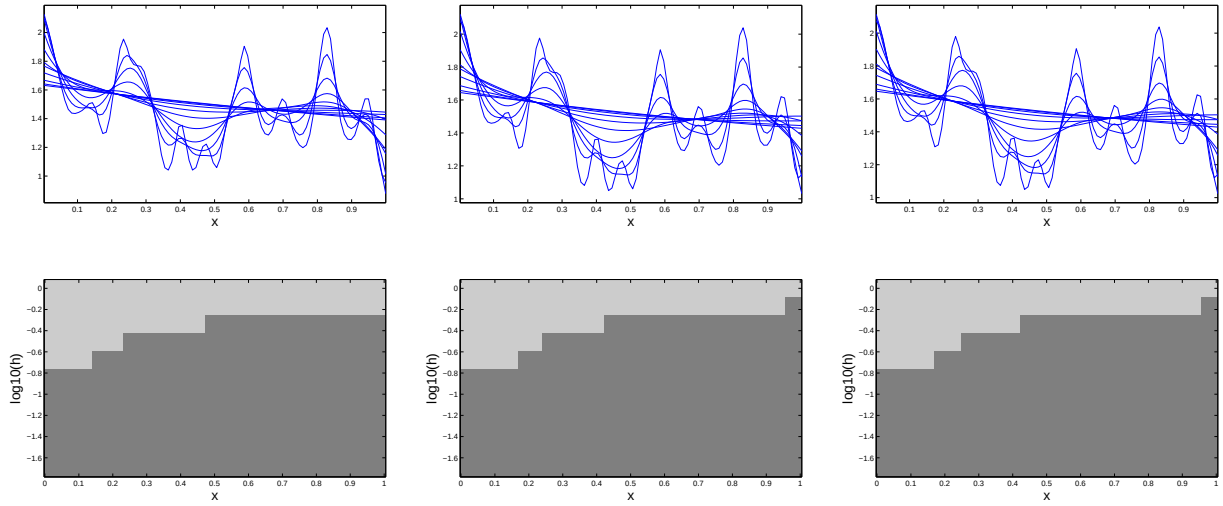


(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 4: $m(x) = 4x + 4\exp(-100(x - 0.5)^2)$ and $v(x) = 5\exp(-x^2/8)$.

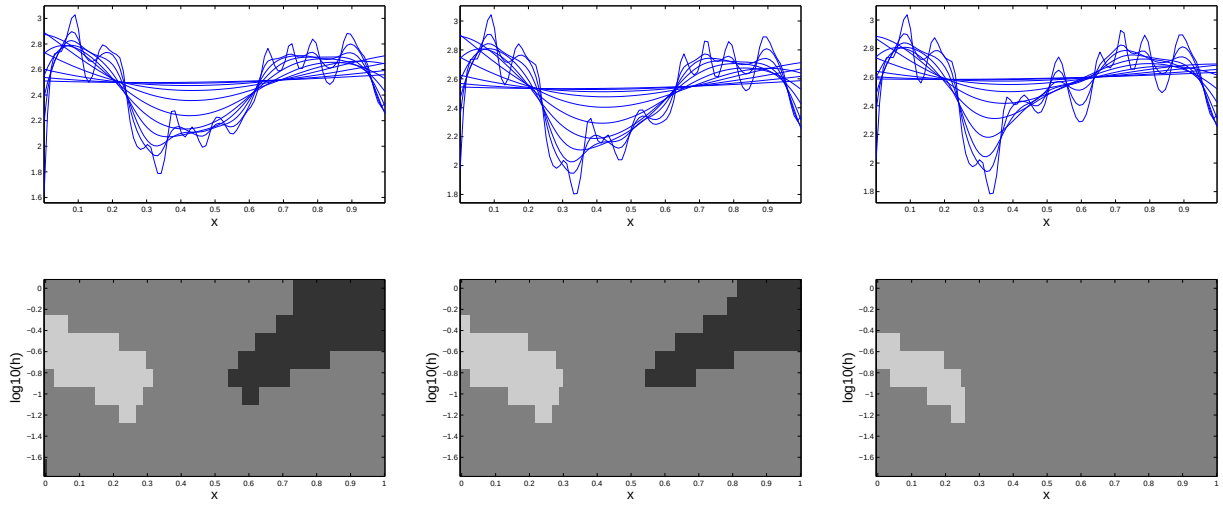
Looking at both Figures 4 and 5 when $v(x) = 5(2 + \sin(\pi x) + \cos(2\pi x))$, one can see that all the SiZer plots except for Figure 4(c) detect a global decreasing trend in their SiZer maps for both $m(x) = 4x + 4\exp(-100(x - 0.5)^2)$ and $m(x) = 0$. In Figure 4(c) with the combination of $m(x) = 4x+4\exp(-100(x-0.5)^2)$ and $h_p(8)$, a spurious increasing trend created by undersmoothing is observed in the SiZer plot.

In Figures 6 and 7, SiZer plots with $v(x) = 5(2+\sin(\pi x)+\cos(2\pi x))$ are graphed with the two



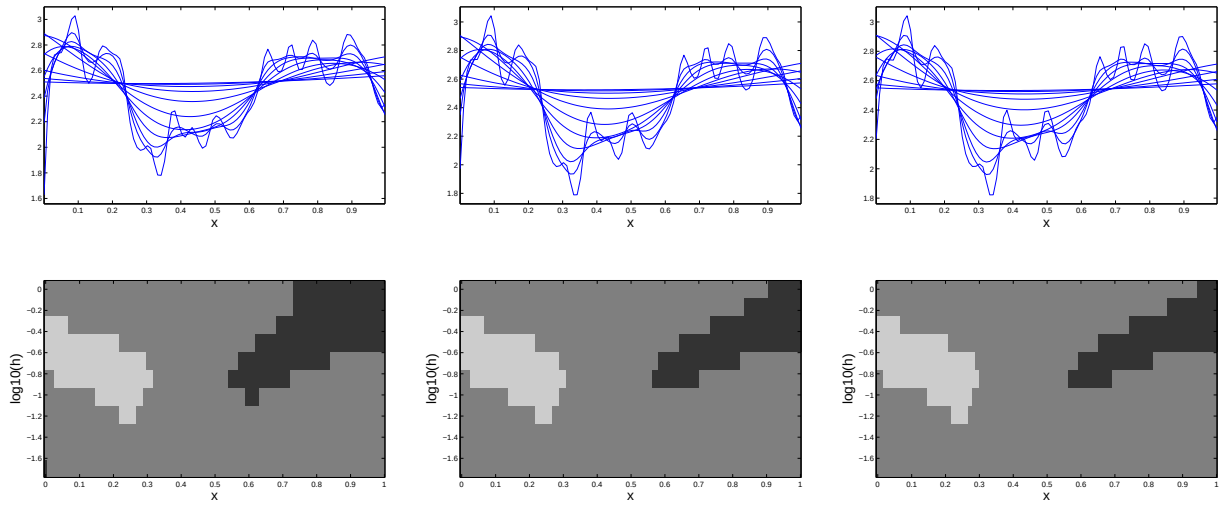
(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 5: $m(x) = 0$ and $v(x) = 5\exp(-x^2/8)$.



(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 6: $m(x) = 4x + 4\exp(-100(x - 0.5)^2)$ and $v(x) = 5(2 + \sin(\pi x) + \cos(2\pi x))$.



(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

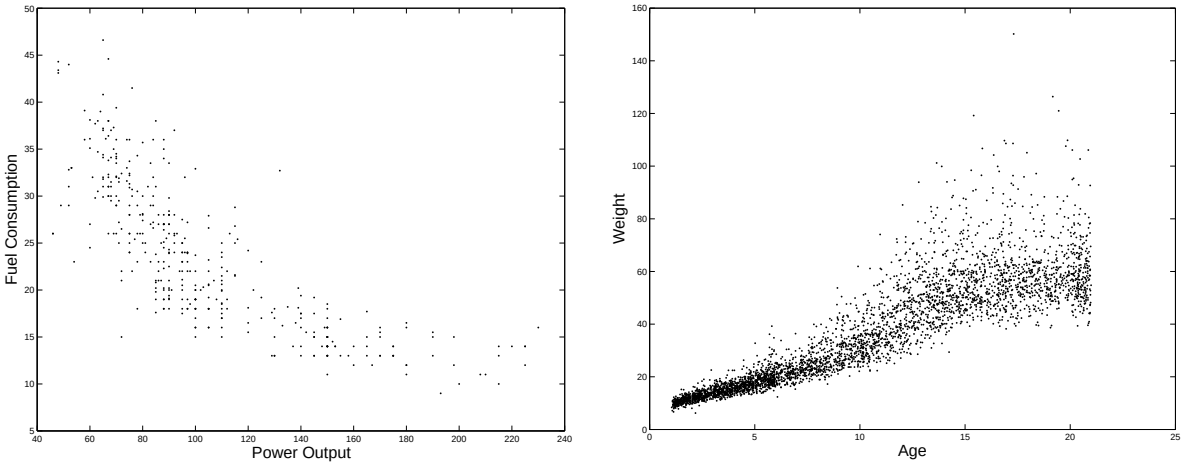
Figure 7: $m(x) = 0$ and $v(x) = 5(2 + \sin(\pi x) + \cos(2\pi x))$.

different mean functions. All the SiZer plots correctly flag the decreasing and increasing trends in the log-variance function except for Figure 6(c) with the combination of $m(x) = 4x+4\exp(-100(x-0.5)^2)$ and $h_p(8)$. In that case, the trends in the mean and log-variance functions are compounded, and thus the increasing trend in the log-variance function is removed when $(Y_i - \hat{m}(X_i))^2$ is evaluated in the equation (2.2).

3.2 Real examples

In this subsection, we analyze two real data sets with the proposed SiZer tool for a log-variance function.

The first one is the Cars data set analyzed for example by Hawkins (1994); Ng (1996); Park et al. (2010). This data set has 392 observations on the response, fuel consumption (measured in miles/gallon), and the covariate, power output (measured in HP). This data set is plotted in Figure 8(a) and it is known that the mean function should be monotonically decreasing (Hawkins, 1994; Ng, 1996). Note that the variation in the data seems to be decreasing as the covariate grows. In the top panels of Figure 9, the fitted log-variance curves with different bandwidths seem to decrease towards the right end of the data, and this trend is statistically significant according to the corresponding SiZer maps in the bottom panels for all three pilot bandwidths. This is a

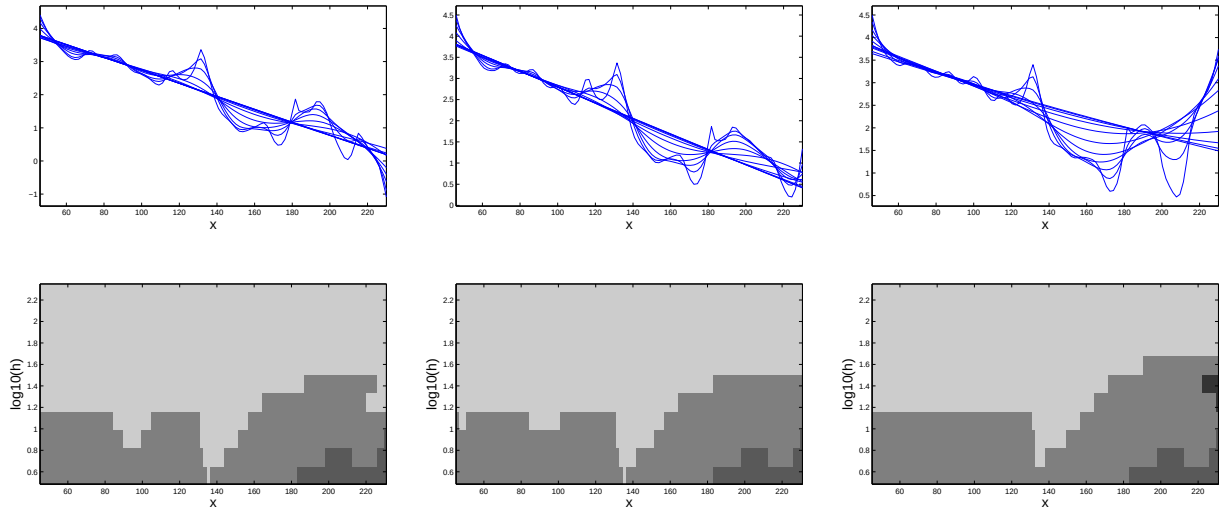


(a) Cars (b) US Girls

Figure 8: Scatter plots of two real data sets.

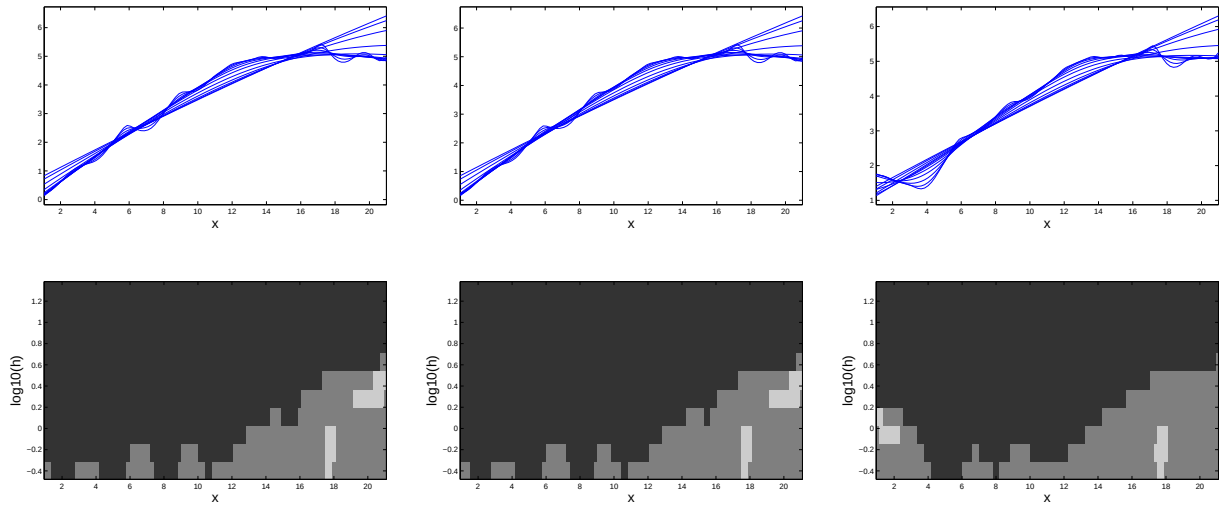
consistent observation in Ng (1996) where they pointed out that dispersion in fuel efficiency is greater for small engine cars than those of larger engines. Note that there is a small increasing trend at the end only for $h_p(8)$. If one believes that the mean function should be monotonic, the increasing trend at the end in the variance might be an important feature which is not captured with the other two pilot bandwidths. A further investigation would be necessary to resolve this issue.

The second data set is the US Girls data set analyzed in Yu et al. (2003) and Park et al. (2010). This data set contains the weights (kg) and ages for a sample of 4011 US girls. The data are plotted in Figure 8(b), and it indicates that both mean and variance functions increase with ages. The three SiZer plots in Figure 10 show that this global trend in the log-variance function is strong and statistically significant. Note that the fitted log-variance curves at small scales in Figure 8(b) exhibit a decreasing trend around $x = 17$. In Yu et al. (2003), some running regression quantiles exhibited a decreasing trend for age greater than 17, but Park et al. (2010) argued that the decreasing trend seemed to be a spurious feature as their quantile SiZer inference did not support the existence. Combining their conclusion with our finding with the proposed SiZer for variance, we conclude that the girl's weights do not start decreasing around age 17, but in fact their variation does.



(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 9: Three selected SiZer plots with the pilot bandwidths $h_p(2), h_p(5), h_p(8)$ are displayed for Cars data.



(a) $h_p(2)$ (b) $h_p(5)$ (c) $h_p(8)$

Figure 10: Three selected SiZer plots with the pilot bandwidths $h_p(2), h_p(5), h_p(8)$ are displayed for US Girls data.

4 Discussion

In the proposed SiZer, there are two bandwidths; mean function estimation is characterized by h_p, and variance estimation by h. As mentioned in Section 2.2, a large h_p assumes little structure and a small one corresponds to complicated structure in the mean function in the viewpoint of a scale-space approach. Therefore, we present our SiZer plots for several pilot bandwidths to summarize all the information available at each scale for both mean and variance functions. From the simulation results in Section 3.1, it can be seen that $h_p(8)$ is a proper choice for $m = 0$, but introduces spurious features when m is nonlinear. On the contrary, $h_p(2)$ provides more accurate SiZer inference for the nonlinear case. In general, $h_p(1)$ is not recommend due to oversmoothing. Therefore, if prior knowledge about the trend of the mean function is available, it can be used for the selection of the pilot bandwidth. In the simulation, $h_p(2)$ provides accurate SiZer inference for both cases when $m = 0$ and m is nonlinear, which might indicate that $h_p(2)$ could be appropriate for many examples in practice. However, as seen in Section 3.2, a large pilot bandwidth can find an interesting feature that might not be detected at small and/or medium bandwidths. This suggests that a full scale-space approach would provide richer information about features in the data.

As mentioned in the Remark in Section 2.3, however, we assume that $h_p \to 0$ and $nh_p \to \infty$ when $n \to \infty$ for our theoretical study due to difficulty of handling the two bandwidths. It is a challenging task to show asymptotic properties of the proposed SiZer for a variance function without assuming $\hat{m}_{h_p}$ being close to m. This is a limitation of the proposed SiZer and we suggest a theoretical investigation using a full scale-space approach for both mean and variance functions as our future work.

5 Appendix

In Section 5.1, we introduce Lemma 1 which is a key for the proofs of the asymptotic properties. In Section 5.2, we provide the proofs of Proposition 1 and details on the the standard deviation of $\widehat{s}'_h$ given in Section 2 using the local linear estimator in (2.2) with $p = 1$. In Section 5.3, we prove Theorems 1 and 2 presented in Section 2.3.

5.1 Preparatory result

Let $S_{n,x,h}$ be the $(p+1)\times(p+1)$ matrix having their (i,j)th entry equal to

$$\frac{\mathrm{e}^{-s(x)}}{(i-1)!(j-1)!}\int K(u)u^{i+j-2}vf(x+hu)du \tag{5.1}$$

where p is a degree of a local polynomial estimator and $vf(x)=v(x)f(x)$. And let $\boldsymbol{e}_j$ be the $(p+1)\times 1$ vector with 1 appearing at the jth position and 0 otherwise. Define

$$W_{n,k}(h,x,u)=\frac{2}{h^{k+1}}\boldsymbol{e}_{k+1}^T K\Big(\frac{u-x}{h}\Big)S_{n,x,h}^{-1}\begin{pmatrix}1\\(u-x)/h\\\vdots\\(u-x)^p/(h^p p!)\end{pmatrix}.$$

Recall that $\kappa(x)=E[\{Y-m(X)\}^4|X=x]$ is the central fourth moment of Y given $X=x$. In order to obtain the approximation form of $\widehat{s}_h^{(k)}(x)$ uniformly in $x\in I$ and $h\in H$ in Lemma 1, the following conditions are needed.

(A.1) The function $s^{(p+1)}$ is uniformly continuous over I, which is the support of X.

(A.2) The function κ is uniformly continuous over I.

(A.3) $\inf_{x\in I} v(x)>0$.

(A.4) $\sup_{x\in I} E(|Y-m(X)|^{4+\rho}|X=x)<\infty$ for some positive ρ.

Lemma 1 *Suppose that the assumptions (A.1)–(A.4) are satisfied and $h_p\to 0$, $nh_p\to\infty$ as $n\to\infty$.*

(i) *If $k=0$, then*

$$\widehat{s}_h(x)-s(x)=\frac{1}{n}\sum_{i=1}^n W_{n,0}(h,x,X_i)\ell_1(s(x),R_i)(1+o_P(1))$$

uniformly in $x\in I$ and $h\in H$ as $n\to\infty$. Here, $R_i=\{Y_i-m(X_i)\}^2$ for $i=1,\ldots,n$.

(ii) *If $k>0$, then*

$$\widehat{s}_h^{(k)}(x)=\frac{1}{n}\sum_{i=1}^n W_{n,k}(h,x,X_i)\ell_1(s(x),R_i)(1+o_P(1))$$

uniformly in $x\in I$ and $h\in H$ as $n\to\infty$.

Proof. Recall that $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ maximizes

$$\sum_{i=1}^{n} \ell\Big(\sum_{j=0}^{p} \beta_j (X_i - x)^j, \widehat{R}_i\Big) K\Big(\frac{X_i - x}{h}\Big)$$

with respect to β_j's where $\widehat{R}_i = \{Y_i - \widehat{m}(X_i)\}^2$, $i = 1, \ldots, n$. Here, $\widehat{m}(\cdot)$ is the local linear estimator depending on the pilot bandwidth h_p, which is different from the given h. Note that if $nh_p \to \infty$ and $h_p \to 0$, then

$$\sup_{x \in I} |\widehat{m}(x) - m(x)| = o_P(1) \tag{5.2}$$

by Mack and Silverman (1982). The $(p+1) \times 1$ vector $\widehat{\boldsymbol{\beta}}$ maximizes (2.2) with R_i being replaced by $\widehat{R}_i$. Let $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{Z}_i$ be the $(p+1) \times 1$ vectors as follows:

$$\bar{\boldsymbol{\beta}} = \sqrt{nh} \begin{pmatrix} \widehat{\beta}_0 - s(x) \\ h\widehat{\beta}_1 \\ \vdots \\ h^p p! \widehat{\beta}_p \end{pmatrix} \quad \text{and} \quad Z_i = \begin{pmatrix} 1 \\ (X_i - x)/h \\ \vdots \\ (X_i - x)^p / (h^p p!) \end{pmatrix}$$

for $i = 1, \ldots, n$. Note that $\sum_{j=0}^{p} \widehat{\beta}_j (X_i - x)^j = s(x) + \bar{\boldsymbol{\beta}}^T \boldsymbol{Z}_i / \sqrt{nh}$. Thus, $\bar{\boldsymbol{\beta}}$ maximizes $\tilde{L}_n(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \ell(s(x) + \boldsymbol{\beta}^{*T} \boldsymbol{Z}_i / \sqrt{nh}, \widehat{R}_i) K((X_i - x)/h)$ as a function of $\boldsymbol{\beta}^*$. Let $L_n(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \ell(s(x) + \boldsymbol{\beta}^{*T} \boldsymbol{Z}_i / \sqrt{nh}, R_i) K((X_i - x)/h)$. Since $\ell(z, \cdot)$ is continuous,

$$\sup_{x \in I} |\tilde{L}_n(\boldsymbol{\beta}^*) - L_n(\boldsymbol{\beta}^*)| = o_P(1) \tag{5.3}$$

by (5.2). Consider the normalized function

$$L_{n,c}(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \big\{\ell\big(s(x) + \boldsymbol{\beta}^{*T} \boldsymbol{Z}_i / \sqrt{nh}, R_i\big) - \ell\big(s(x), R_i\big)\big\} K\Big(\frac{X_i - x}{h}\Big)$$

which is maximized by $\bar{\boldsymbol{\beta}}$ as well. Note that $L_{n,c}$ is concave in $\boldsymbol{\beta}^*$. Define

$$\begin{aligned} \boldsymbol{V}_n &= \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \ell_1(s(x), R_i) K\Big(\frac{X_i - x}{h}\Big) \boldsymbol{Z}_i, \\ A_n &= \frac{1}{nh} \sum_{i=1}^{n} \ell_2(s(x), R_i) K\Big(\frac{X_i - x}{h}\Big) \boldsymbol{Z}_i \boldsymbol{Z}_i^T. \end{aligned}$$

Using a Taylor series expansion of the function $\ell(\cdot, R_i)$, we obtain

$$L_{n,c}(\boldsymbol{\beta}^*) = \boldsymbol{V}_n^T \boldsymbol{\beta}^* + \frac{1}{2} \boldsymbol{\beta}^{*T} A_n \boldsymbol{\beta}^* + o_P\Big(\frac{1}{n}\Big). \tag{5.4}$$

Define $H = [\underline{h}, \overline{h}]$ for $0 < \underline{h} \leq \overline{h} < \infty$. Let $\xi(x,h)$ be the (i,j)th element of A_n where $1 \leq i,j \leq (p+1)$. And let D_{1n} and D_{2n} be discretized grids of I and H respectively, which are given by $D_{1n} = \{j\delta_{1n} : j = 0,\ldots,[1/\delta_{1n}]\}$ and $D_{2n} = \{\underline{h} + j\delta_{2n} : j = 0,\ldots,[(\overline{h}-\underline{h})/\delta_{2n}]\}$ where $\delta_{in} = O(n^{-\gamma}/2)$, $i = 1,2$, for some positive γ and $[x]$ is the largest integer not exceeding x. Then, we obtain

$$\begin{aligned}\sup_{x\in I}\sup_{h\in H}|\xi(x,h) - E(\xi(x,h))| \leq & \sup_{x'\in D_{1n}}\sup_{h'\in D_{2n}}|\xi(x',h') - E(\xi(x',h'))| \\ & + \sup_{x,x',h,h'}|\{\xi(x,h) - E(\xi(x,h))\} - \{\xi(x',h') - E(\xi(x',h'))\}| \qquad (5.5)\end{aligned}$$

where $\sup_{x,x',h,h'}$ denotes supremum over $x,x' \in I$ and $h,h' \in H$ satisfying $\sqrt{(x-x')^2 + (h-h')^2} \leq \sqrt{\delta_{1n}^2 + \delta_{2n}^2}$. We take γ to be large enough to ensure that the second term of righthand side in (5.5) is negligible compared to the first one. If we define $\bar{R}_i = R_i 1_{[R_i \leq \sqrt{n^{1-\zeta}}]}$, $i = 1,\ldots,n$, for an arbitrary small positive ζ, it follows that

$$\begin{aligned}&\sup_{x'\in D_{1n}}\sup_{h'\in D_{2n}}|\xi(x',h') - E(\xi(x',h'))| \\ &\quad \leq \sup_{x'\in D_{1n}}\sup_{h'\in D_{2n}}|\xi(x',h') - \bar{\xi}(x',h') - \{E(\xi(x',h')) - E(\bar{\xi}(x',h'))\}| \\ &\quad\quad + \sup_{x'\in D_{1n}}\sup_{h'\in D_{2n}}|\bar{\xi}(x',h') - E(\bar{\xi}(x',h'))| \qquad (5.6)\end{aligned}$$

where $\bar{\xi}$ is ξ with R_i being replaced by $\bar{R}_i$. By Proposition 1 in Mack and Silverman (1982) with the assumptions in Lemma 1

$$\sup_{x'\in D_{1n}}|\xi(x',h') - \bar{\xi}(x',h') - \{E(\xi(x',h')) - E(\bar{\xi}(x',h'))\}| = O_p((n^{1-\zeta})^{-(1+\rho)/2}) \qquad (5.7)$$

where ρ is some positive constant in (A.4). According to the proof of Proposition 1 in Mack and Silverman (1982), we can easily show that the supremum (5.7) extends over the discretized set D_{2n}. Then, the first term of righthand side in (5.6) is $O_p((n^{1-\zeta})^{-(1+\rho)/2})$.

Let us consider the second term of righthand side in (5.6). Define $\bar{\xi}(x',h') - E(\bar{\xi}(x',h')) = \sum_{l=1}^n \{U_{l,n}(x',h') - E(U_{l,n}(x',h'))\}$ where

$$U_{l,n}(x',h') = \frac{1}{nh}\ell_2(s(x'),\bar{R}_l)K\Big(\frac{X_l - x'}{h'}\Big)\Big(\frac{X_l - x'}{h'}\Big)^{i+j-2},$$

$l = 1\ldots,n$. Note that, for some positive constants c_0 and c_1,

$$P\Big(\sup_{x'\in D_{1n}}\sup_{h'\in D_{2n}}|\bar{\xi}(x',h') - E(\bar{\xi}(x',h'))| > c_0\left(n^{1-\zeta}\right)^{-1/2}\Big)$$

$$\leq c_1 n^{\gamma} \max_{x' \in D_{1n}} \max_{h' \in D_{2n}} P\Big(|\bar{\xi}(x', h') - E(\bar{\xi}(x', h'))| > c_0 \Big(n^{1-\zeta} \Big)^{-1/2} \Big) \tag{5.8}$$

and that

$$E(U_{l,n}(x', h') - E(U_{l,n}(x', h'))) = 0 \quad \text{and} \quad Var(U_{l,n}(x', h')) = O(n^{-2})$$

for all $x' \in D_{1n}$, $h' \in D_{2n}$ and $l = 1, \ldots, n$. Since $|U_{l,n}(x', h')| \leq c_2 \sqrt{n^{-1-\zeta}}$, the probability in (5.8) is then bounded by Bernstein's inequality

$$\begin{aligned} & P\Big(|\bar{\xi}(x', h') - E(\bar{\xi}(x', h'))| > c_0 \Big(n^{1-\zeta} \Big)^{-1/2} \Big) \\ & \leq 2 \exp\Big(- \frac{c_0^2 n^{-(1-\zeta)}}{2nVar(U_{1,n}(x', h')) + \frac{2}{3} c_3 \sqrt{n^{-1-\zeta}} \sqrt{n^{1-\zeta}}} \Big) \end{aligned} \tag{5.9}$$

with suitable constants c_2 and c_3. By (5.9) with sufficiently large c_0, the term (5.8), $2c_1 n^{\gamma} e^{-cn^{\zeta}}$ for some c, goes to 0. Therefore, we immediately obtain

$$\sup_{x' \in D_{1n}} \sup_{h' \in D_{2n}} |\bar{\xi}(x', h') - E(\bar{\xi}(x', h'))| = o_P\Big(\frac{1}{\sqrt{n^{1-\zeta}}} \Big). \tag{5.10}$$

Thus, $\sup_{x \in I} \sup_{h \in H} |\xi(x, h) - E(\xi(x, h))| = o_P(1/\sqrt{n^{1-\zeta}})$ and

$$L_{n,c}(\boldsymbol{\beta}^*) = \boldsymbol{V}_n^T \boldsymbol{\beta}^* + \frac{1}{2} \boldsymbol{\beta}^{*T} E(A_n) \boldsymbol{\beta}^* + o_P\Big(\frac{1}{\sqrt{n^{1-\zeta}}} \Big).$$

uniformly in $x \in I$ and $h \in H$.

The expected value of the (i, j)th element of A_n is then

$$\begin{aligned} E(\xi(x, h)) &= \frac{1}{(i-1)!(j-1)!} \frac{1}{h} \int_0^1 K\Big(\frac{u - x}{h} \Big) \Big(\frac{u - x}{h} \Big)^{i+j-2} \Big(- \frac{1}{2} \Big) v(u) \mathrm{e}^{-s(x)} f(u) du \\ &= -\frac{1}{2} \frac{\mathrm{e}^{-s(x)}}{(i-1)!(j-1)!} \int K(u) u^{i+j-2} v f(x + hu) du. \end{aligned}$$

Define $L_c(\boldsymbol{\beta}^*) = \boldsymbol{V}_n^T \boldsymbol{\beta}^* - \frac{1}{4} \boldsymbol{\beta}^{*T} S_{n,x,h} \boldsymbol{\beta}^*$. By Convexity Lemma in Pollard (1991),

$$\sup_{\boldsymbol{\beta}^* \in \mathcal{C}} |L_{n,c}(\boldsymbol{\beta}^*) - L_c(\boldsymbol{\beta}^*)| = o_P(1) \tag{5.11}$$

for any compact set $\mathcal{C}$, and the maximizer of $L_c(\boldsymbol{\beta}^*)$ is given as $\widehat{\boldsymbol{\beta}}^* = 2 S_{n,x,h}^{-1} \boldsymbol{V}_n$. Then, $\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^* = o_P(1)$ uniformly in x and h by Quadratic Approximation Lemma in Carroll et al. (1997) with (5.3) and (5.11). Therefore, we obtain the result. ■

5.2 Derivation of simultaneous confidence intervals in SiZer inference

5.2.1 Proof of Proposition 1

In order to show (2.4), we consider the particular case of fixed design regression and the design points X_i's satisfy $X_i = i\Delta$, where $\Delta > 0$ is the distance between design points. In the proof of Lemma 1 in Section 5.1, if we take $L_c(\boldsymbol{\beta}^*) = \boldsymbol{V}_n^T\boldsymbol{\beta}^* + \frac{1}{2}\boldsymbol{\beta}^{*T}A_n\boldsymbol{\beta}^*$, the maximizer of $L_c(\boldsymbol{\beta}^*)$ is then $\widehat{\boldsymbol{\beta}}^* = -A_n^{-1}\boldsymbol{V}_n$. Using the integral approximation with $X_i = i\Delta$, the second element of $\widehat{\boldsymbol{\beta}}^*$ is approximated as follows:

$$\widehat{\beta}_1^* \approx \frac{1}{\sqrt{nh}}\Big(\frac{2\mathrm{e}^{s(x)}\mu_{0,v}^*(x)}{\mu_{2,v}^*(x)\mu_{0,v}^*(x) - (\mu_{1,v}^*(x))^2}\Big)\sum_{i=1}^{n}\Big(\frac{X_i - x}{h}\Big)K\Big(\frac{X_i - x}{h}\Big)\ell_1(s(x), R_i)$$

where $\mu_{i,v}^*(x) = \int u^i K(u)v(x+hu)du$ for $i = 0, 1, 2$. If x is away from the boundary, it follows from symmetry of the kernel that

$$\sum_{i=1}^{n}\Big(\frac{X_i - x}{h}\Big)K\Big(\frac{X_i - x}{h}\Big) \approx 0.$$

Therefore,

$$\widehat{s}_h'(x) \approx \frac{1}{nh^2}\Big(\frac{\mu_{0,v}^*(x)}{\mu_{2,v}^*(x)\mu_{0,v}^*(x) - (\mu_{1,v}^*(x))^2}\Big)\sum_{i=1}^{n}\Big(\frac{X_i - x}{h}\Big)K\Big(\frac{X_i - x}{h}\Big)R_i. \tag{5.12}$$

Let $t = \tilde{\Delta}/\Delta$ denote the number of data points per SiZer column. For simplicity of notation, we can assume that t is a positive integer. Let g be the number of pixels in each row, and $Q_1, \cdots, Q_g$ denote the test statistics of a row in the SiZer map. Then Q_j is proportional to the estimate of s' calculated for $x = j\tilde{\Delta} = jt\Delta$. In particular,

$$Q_j \approx \sum_{l=1}^{n} W_{jt-l}^h R_l. \tag{5.13}$$

The exact form of the W_{jt-l}^h is the weight in the sum of the right-hand side in (5.12). Note that W_{jt-l}^h is proportional to $-(jt - l)K_{h/\Delta}(jt - l)$. Thus, the weights W_t^h are proportional to the derivative of the Gaussian kernel with standard deviation h/Δ.

If there is no signal, then the R_i are independent and identically distributed random variables. If in addition the R's have two finite moments, then for sufficiently large h/Δ, the Cramér-Wold device and Lindeberg-Feller central limit theorem provide an approximate Gaussian distribution, with mean 0 and variance 1 after appropriate scaling. Following Appendix in Hannig and Marron

(2006) with $t\Delta = \tilde{\Delta}$, we have

$$Corr(Q_i, Q_{i+j}) \approx \exp\Big(-\Big(\frac{j\tilde{\Delta}}{2h}\Big)^2\Big)\Big\{1 - \frac{1}{2}\Big(\frac{j\tilde{\Delta}}{h}\Big)^2\Big\}.$$

Then, applying Theorem 1 of Hannig and Marron (2006), one can get j step correlation $\rho_{j,g}$ for each g

$$\rho_{j,g} = e^{-j^2C^2/(4\log g)}\left[1 - \frac{j^2C^2}{2\log g}\right],$$

by setting $\tilde{\Delta}/h = C/\sqrt{\log g}$, and obtain

$$\lim_{g\to\infty} \log(g)(1 - \rho_{j,g}) = \frac{3j^2C^2}{4}.$$

Using the similar arguments in Hannig and Marron (2006), one can show that

$$P\left[\max_{i=1,\ldots,g} Q(x_i) \le x\right] \approx \Phi(x)^{\theta g},$$

where

$$\theta = 2\Phi\left(\sqrt{3\log g}\,\frac{\tilde{\Delta}}{2h}\right) - 1. \quad \blacksquare$$

5.2.2 Derivation of the standard deviation

In the case of local linear estimators with $p = 1$, the (i, j)th entry of 2×2 matrix $S_{n,x}$ in (5.1) can be expressed approximately as follows: $f(x)\mu_{i+j-2,K}$, $i, j = 1, 2$ when the additional assumption that f is uniformly continuous in $x \in I$, a compact subinterval of $[0, 1]$, is satisfied. By symmetry of the kernel function,

$$S_{n,x,h} = f(x) \cdot \mathrm{diag}(\mu_{0,K}, \mu_{2,K}) \tag{5.14}$$

due to $\mu_{1,K} = 0$. By Lemma 1 and by (5.14), we have the approximation form of $\widehat{s}'_h(x)$ as follows:

$$\widehat{s}'_h(x) \approx \frac{1}{nh^2}\frac{2}{f(x)\mu_{2,K}}\sum_{i=1}^{n}\Big(\frac{X_i - x}{h}\Big)K\Big(\frac{X_i - x}{h}\Big)\ell_1(s(x), R_i). \tag{5.15}$$

We note that (5.15) is the leading term in stochastic expansion of $\widehat{s}'_h(x)$ when $h \to 0$ and $nh \to \infty$ as $n \to \infty$. We choose to use this approximation because it is rather complicated to derive the exact standard deviation when h is fixed, and the current form in (5.15) is shown to practically work well in our numerical study. According to the null hypothesis, the expected value of $\widehat{s}'_h(x)$ is 0. Note that

$$\sup_{u\in[x-h,x+h]} |s(u) - s(x)| \le (\text{constant})h^2 \tag{5.16}$$

for all x. By (5.16), we have

$$E(\ell_1^2(s(x), R_1)|X_1 = u) \approx E(\ell_1^2(s(X_1), R_1)|X_1 = u) \approx \frac{1}{4}\Big(\frac{\kappa(x)}{v^2(x)} - 1\Big)$$

for all $u \in [x-h, x+h]$. Therefore, it can be shown that the variance of $\widehat{s}_h'(x)$ is approximated by

$$\frac{1}{nh^3}\frac{1}{f(x)\mu_{2,K}}\Big(\frac{\kappa(x)}{v^2(x)} - 1\Big)\int u^2K^2(u)du.$$

Note that $\int u^2K^2(u)du = (4\sqrt{\pi})^{-1}$ and $\mu_{2,K} = 1$ for the Gaussian kernel. Then, the standard deviation in (2.5) is obtained.

5.3 Proof of Theorems 1 and 2

We further introduce the set of assumptions for the weight $W_{n,k}(h, x, u)$ in (5.17) to describe Theorems 1 and 2. For simplicity, we define the approximation form of $\widehat{s}_h(x) - s(x)$ or $\widehat{s}^{(k)}$ as

$$\check{s}_h^{(k)}(x) = \frac{1}{n}\sum_{i=1}^{n} W_{n,k}(h, x, X_i)\ell_1(s(x), R_i). \tag{5.17}$$

(A.5) For integer $k \geq 0$, as $n \longrightarrow \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{4}\{\kappa(X_i) - v^2(X_i)\}\tilde{W}_{n,k}(h_1, x_1, X_i)\tilde{W}_{n,k}(h_2, x_2, X_i)$$

converges in probability to a covariance function $cov(h_1, x_1, h_2, x_2)$ for all (h_1, x_1) and $(h_2, x_2) \in H \times I$, where $\tilde{W}_{n,k}(h, x, X) = \mathrm{e}^{-s(x)}W_{n,k}(h, x, X)$.

(A.6) As $n \longrightarrow \infty$,

$$n^{-(1+\rho/2)}\Big\{\max_{1\leq i\leq n}|W_{n,k}(h, x, X_i)|^\rho\Big\}\sum_{i=1}^{n}\{W_{n,k}(h, x, X_i)\}^2$$

converges in probability to zero for all $(h, x) \in H \times I$.

(A.7) As h varies in H and x varies in I, $\{\partial^2(W_{n,k}(h, x, X_i))/(\partial h\partial x)\}^2$ are uniformly dominated by a positive function $M(X_i)$ such that $E\{M(X_i)\} < \infty$.

(A.8) As h varies in H and x varies in I, $\{\partial(W_{n,k}(h, x, X_i))/\partial x\}^2$ and $\{\partial(W_{n,k}(h, x, X_i))/\partial h\}^2$ are uniformly dominated by a positive function $M^*(X_i)$ such that $E\{M^*(X_i)\} < \infty$.

Note that many standard kernels including the Gaussian kernel satisfy the assumptions (A.5)–(A.8) for the weight function of the local polynomial fit (Chaudhuri and Marron, 2000).

Proof of Theorem 1. In what follows we show that all the finite dimensional distribution of the process converges weakly to the normal distribution and the process satisfies a tightness condition as similarly done in Chaudhuri and Marron (2000).

Fix $(h_1, x_1), (h_2, x_2), \cdots, (h_l, x_l) \in H \times I$ and $t_1, t_2, \ldots, t_l \in (-\infty, \infty)$. Define

$$Z_n = n^{1/2} \sum_{i=1}^{l} t_i \Big[\widehat{s}_{h_i}^{(k)}(x_i) - E\{\widehat{s}_{h_i}^{(k)}(x_i)\} \Big] \text{ and } \tilde{Z}_n = n^{1/2} \sum_{i=1}^{l} t_i \Big[\check{s}_{h_i}^{(k)}(x_i) - E\{\check{s}_{h_i}^{(k)}(x_i)\} \Big].$$

One can easily show that $Z_n = \tilde{Z}_n(1 + o_P(1))$ for any fixed $x_i \in I$ since we have the approximation form of $\widehat{s}_h^{(k)}(x)$ uniformly in $x \in I$ and $h \in H$ by Lemma 1. The conditional mean and variance of $\tilde{Z}_n$ are zero and

$$\frac{1}{n} \sum_{i=1}^{l} \sum_{j=1}^{l} t_i t_j \sum_{a=1}^{n} \frac{1}{4} (\kappa(X_a) - v^2(X_a)) \tilde{W}_{n,k}(h_i, x_i, X_a) \tilde{W}_{n,k}(h_j, x_j, X_a),$$

which converges in probability to $\sum_{i=1}^{l} \sum_{j=1}^{l} t_i t_j cov(h_i, x_i, h_j, x_j)$. Also, the assumption (A.4) and (A.6) imply that Lindeberg's condition holds for $\tilde{Z}_n$ and consequently its limiting distribution must be normal. Define

$$\tilde{U}_n(h_i, x_i) = n^{1/2} \Big[\check{s}_{h_i}^{(k)}(x_i) - E\{\check{s}_{h_i}^{(k)}(x_i)\} \Big]$$

for $i = 1, \ldots, l$. Using Cramer-Wold device, one can show that the joint limiting distribution of $\tilde{U}_n(h_i, x_i)$ for $1 \le i \le l$ converges to multivariate normal with zero mean and $cov(h_i, x_i, h_j, x_j)$ as the limiting covariance matrix for $1 \le i, j \le l$.

In order to show tightness, we fix $h_1 < h_2$ in H and $x_1 < x_2$ in I. Then, we obtain

$$\begin{aligned}
& E\big\{\tilde{U}_n(h_2, x_2) - \tilde{U}_n(h_2, x_1) - \tilde{U}_n(h_1, x_2) + \tilde{U}_n(h_1, x_1)\big\}^2 \\
&= \frac{1}{n} E\Big[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{4} (\kappa(X_i) - v^2(X_i)) \big\{ \tilde{W}_{n,k}(h_2, x_2, X_i) - \tilde{W}_{n,k}(h_2, x_1, X_i) \\
&\qquad\qquad - \tilde{W}_{n,k}(h_1, x_2, X_i) + \tilde{W}_{n,k}(h_1, x_1, X_i) \big\}^2 \Big] \\
&\le C_2 (h_2 - h_1)^2 (x_2 - x_1)^2 E\Big\{ \frac{1}{n} \sum_{i=1}^{n} M(X_i) \Big\} \qquad (5.18) \\
&\le C_3 (h_2 - h_1)^2 (x_2 - x_1)^2
\end{aligned}$$

for some constants C_2 and $C_3 > 0$. The assumptions (A.1), (A.2), and (A.3) imply that κ and v are bounded on the compact interval I. Then, by (A.7), we have the inequality (5.18). It now follows by Bickel and Wichura (1971) that the sequence of process $\tilde{U}_n(h, x)$ on $H \times I$ will have the tightness property. The difference between $\hat{s}_{h_i}^{(k)}(x_i)$ and $\check{s}_{h_i}^{(k)}(x_i)$, for $1 \leq i \leq l$, is stochastically negligible by Lemma 1. It implies that the joint limiting distribution of $U_n(h_i, x_i)$, for $1 \leq i \leq l$, converges to the same multivariate normal distribution of $\tilde{U}_n(h_i, x_i)$. Consequently the theorem follows. ■

Proof of Theorem 2. By (A.8) with some appropriate choice of C_4, one obtains that

$$E\{\tilde{U}_n(h_2, x_2) - \tilde{U}_n(h_1, x_1)\}^2 \leq C_4\{(h_2 - h_1)^2 + (x_2 - x_1)^2\}$$

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

Acknowledgement

The authors are grateful to the reviewers and the associate editor for many helpful comments. The authors are also grateful to Dr. Keming Yu for sending us the real data set used in Section 3.2. The research of the corresponding author was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0002748).

References

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics*, 42:1656–1670.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for 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. *The Annals of Statistics*, 28:408–428.

Erästö, P. and Holmström, L. (2005). Bayesian multiscale smoothing for making inferences about features in scatter plots. *Journal of Computational and Graphical Statistics*, 14:569–589.

Erästö, P. and Holmström, L. (2007). Bayesian analysis of features in a scatter plot with dependent observations and errors in predictors. *Journal of Statistical Computation and Simulation*, 77:421–434.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 17:179–192.

Ganguli, B. and Wand, M. P. (2007). Feature significance in generalized additive models. *Statistics and Computing*, 17:179–192.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73:625–634.

Godtliebsen, F. and Oigard, T. A. (2005). A visual display device for significant features in complicated signals. *Computational Statistics and Data Analysis*, 48:317–343.

González-Manteiga, W., Martínez-Miranda, M., and Raya-Miranda, R. (2008). SiZer map for inference with additive models. *Statistics and Computing*, 18:297–312.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society Series B*, 51:3–14.

Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77:521–528.

Hannig, J. and Lee, T. (2006). Robust SiZer for exploration of regression structures and outlier detection. *Journal of Computational and Graphical Statistics*, 15:101–117.

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

Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81:223–242.

Hawkins, D. M. (1994). Fitting Monotonic Polynomials to Data. *Computational Statistics*, 9:233–247.

Huh, J. (2011). Likelihood based estimation of the discontinuous log-variance function. Unpublished manuscript.

Kim, C. S. and Marron, J. S. (2006). Sizer for jump detection. *Journal of Nonparametric Statistics*, 18:13–20.

Li, R. and Marron, J. S. (2005). Local likelihood SiZer map. *Sankhya*, 67:476–498.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer, Boston.

Mack, Y. and Silverman, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 61:405–415.

Marron, J. and de Uña Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. *Journal of Statistical Planning and Inference*, 121:149–161.

Marron, J. and Zhang, J. (2005). SiZer for smoothing splines. *Computational Statistics*, 20:481–502.

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15:610–625.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Ng, P. T. (1996). An algorithm for quantile smoothing splines. *Computational Statistics and Data Analysis*, 22:99–118.

Oigard, T. A., Rue, H., and Godtliebsen, F. (2006). Bayesian multiscale analysis for time series data. *Computational Statistics and Data Analysis*, 51:1719–1730.

Park, C., Godtliebsen, F., Taqqu, M., Stoev, S., and Marron, J. S. (2007). Visualization and inference based on wavelet coefficients, SiZer and SiNos. *Computational Statistics and Data Analysis*, 51:5994–6012.

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

Park, C. and Huh, J. (2013). Statistical inference and visualization in scale-space using local likelihood. *Computational Statistics and Data Analysis*, 57:336–348.

Park, C. and Kang, K. (2008). SiZer analysis for the comparison of regression curves. *Computational Statistics and Data Analysis*, 52:3954–3970.

Park, C., Lee, T., and Hannig, J. (2010). Multiscale exploratory analysis of regression quantiles using quantile SiZer. *Journal of Computational and Graphical Statistics*, 19:497–513.

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.

Park, C., Vaughan, A., Hannig, J., and Kang, K. (2009b). Sizer analysis for the comparison of time series. *Journal of Statistical Planning and Inference*, 139:3974–3988.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12:1215–1230.

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.

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technomtrics*, 39:262–273.

Sørbye, S., Hindberg, K., Olsen, L., and Rue, H. (2009). Bayesian multiscale feature detection of log-spectral densities. *Computational Statistics and Data Analysis*, 53:3746–3754.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyã Series A*, 26:359–372.

Yu, K. and Jones, M. C. (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99:139–144.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society Series D*, 52:331–350.