Statistical Inference and Visualization in Scale-Space Using Local Likelihood

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

June 25, 2012

Abstract

SiZer (SIgnificant ZERo crossing of the derivatives) is a graphical scale-space visualization tool that allows for exploratory data analysis with statistical inference. Various SiZer tools have been developed in the last decade, but most of them are not appropriate when the response variable takes discrete values. In this paper, we develop a SiZer for finding significant features using a local likelihood approach with local polynomial estimators. This tool improves the existing one (Li and Marron, 2005) by proposing a theoretically justified quantile in a confidence interval using advanced distribution theory. In addition, we investigate the asymptotic properties of the proposed tool. We conduct a numerical study to demonstrate the sample performance of SiZer using Bernoulli and Poisson models using simulated and real examples.

Key words: Generalized linear models, Local likelihood, Local polynomial smoothing, Scale-space, Statistical significance.

1 Introduction

Nelder and Wedderburn (1972) introduced generalized linear models as a means of applying techniques used in ordinary linear regression to more general settings. Green and Silverman (1994) studied an extension of the smoothing spline methodology to generalized linear models. Additionally, Fan et al. (1995) investigated the extension of the nonparametric regression technique of local polynomial fitting with a kernel weight to generalized linear models and quasi-likelihood contexts. In the case of the multiple-covariates, Carroll et al. (1997) and Huh and Park (2002) considered semiparametric and nonparametric versions respectively with kernel regression and a single-index model. Typically these methods produce a nonparametric estimate of the mean function in the form of a smooth curve. Visual inspection of such a nonparametric curve can suggest the existence of trends in the true mean function, but it does not provide statistical inference to determine the statistical significance of such trends.

SiZer (SIgnificant ZERo crossing of the derivatives), originally proposed by Chaudhuri and Marron (1999), is a powerful exploratory data analysis tool equipped with statistical inference based on nonparametric kernel estimates. It provides a new way to look at data in scale-space so that analysts are able to discover any meaningful structure by testing the data against underlying assumptions or potential models while doing exploratory analysis. By doing so, SiZer addresses the critical question in scientific research of which features observed are really there, or represent an important underlying structure.

Several other versions of SiZer have been developed since the seminal work of Chaudhuri and Marron (1999); Hannig and Marron (2006) improved statistical inference of the original SiZer using extreme value theory. Hannig and Lee (2006) proposed a robust version of SiZer which can identify outliers, and Kim and Marron (2006) developed a tool for detecting discontinuities in the data. Park and Kang (2008) proposed a SiZer that compares multiple curves with independent data based on their differences of smooths. Park et al. (2010) studied a SiZer which targets the quantile composition of the data instead of the mean structure. Park et al. (2004) extended the conventional SiZer to dependent errors, which conducts a

goodness-of-fit test by comparing the observed data with an assumed time series model. Rondonotti et al. (2007) developed SiZer for time series that estimates an autocovariance function in order to detect significant features in a time series. Later, Park et al. (2009a) improved this SiZer tool with new quantile and autocovariance function estimator. Park et al. (2009b) introduced a SiZer that puts forth a method for comparing two or more time series. In addition, various Bayesian versions of SiZer have also been proposed as an approach to Baysian multiscale smoothing (Erästö and Holmström, 2005; Godtliebsen and Oigard, 2005; Oigard et al., 2006; Erästö and Holmström, 2007; Sørbye et al., 2009). Note that all of these tools are restricted to data with a continuous response variable, and thus they are not readily applicable to discrete data. In scale-space, Li and Marron (2005) proposed the local likelihood SiZer map that is more efficient in distinguishing features than the original SiZer for discrete data. Ganguli and Wand (2007) considered the problem of determining the significance of features such as peaks or valleys in observed covariate effects under an additive model. They worked with low rank radial spline smoothers to allow for handling of sparse designs and large sample sizes.

In this paper, we develop a SiZer tool using local likelihood, which utilizes a local polynomial estimator with multiple bandwidths to determine the significance of features for discrete data. Because it considers a wide range of bandwidths, it circumvents the classical problem of bandwidth selection, which allows one to do statistical inference and detect all the information that is available at each individual level of resolution. Also, it focuses on smoothed curves depending on bandwidths rather than a true underlying curve because a scale-space approach views that truth exists at each scale. This allows one to avoid a bias problem that occurs in estimating a true underlying function.

Our work is differentiated from that of Li and Marron (2005) in three aspects. First, we improve global inference by proposing a theoretically justified quantile in a confidence interval using advanced distribution theory. This approach was proposed by Hannig and Marron (2006), but we extend their work to discrete data. Second, we provide the asymptotic properties of the proposed SiZer tool, which was not studied in Li and Marron (2005).

Therefore, the proposed SiZer can provide more accurate and informative analysis with statistical inference and visualization for a vast range of statistical problems. Third, we discuss how the proposed SiZer can be extended to multiple-covariate cases.

The rest of the paper is organized as follows. Section 2 reviews a local likelihood approach in generalized linear models and proposes a SiZer tool using local likelihood. Section 3 investigates the performance of the local likelihood SiZer using both simulated and real examples. We study the asymptotic properties of the proposed SiZer in Section 4. We also briefly discuss how a local likelihood SiZer for multiple covariates can be constructed in Section 5. Finally, details on the new quantile estimator in the SiZer based on advanced distribution theory are provided in Section 6.

2 Local Likelihood SiZer

SiZer is based on scale-space ideas from computer vision, see Lindeberg (1994), where it refers to a family of smooths of a digital image. A scale-space approach regards no particular level of smoothing as correct and considers that each smooth provides information about the underlying image structure at a particular scale. In SiZer, scale-space is a family of kernel smooths indexed by the bandwidth. The idea is that this approach uses all the information that is available in the data at each given bandwidth. SiZer extends the usefulness of a family of smooths plot by adding a SiZer map, which displays results of statistical inference. A SiZer map visually displays the significance of features over both location and scale (i.e., bandwidth). Multiple comparison tests based on confidence intervals for the derivatives of the underlying curve are involved in flagging significant features. Therefore, SiZer is a more advanced version of a basic statistical graphic, such as a plot or chart, that simultaneously looks at data with different scopes with statistical inference.

In what follows we propose the local likelihood SiZer for one covariate cases. We illustrate how to extend the proposed tool to the multiple-covariate cases in Section 5.

Suppose we observe a random sample (X_i, Y_i) of (X, Y) where Y_i's are real valued re-

sponses associated with covariates X_i's having density f with support $[0,1]$ without loss of generality for $i = 1, 2, \ldots, n$. Assume that the conditional distribution of $Y|X = x$ belongs to the following one-parameter exponential family:

$$f_{Y|X}(y|x) = \exp\{y\tau(x) - b(\tau(x)) + c(y)\} \tag{1}$$

where b and c are some known functions. Of interest is to estimate the regression function $m(x) \equiv E(Y|X = x) = b'(\tau(x))$. In parametric generalized linear models, the function $m(x)$ is modeled linearly via a link function g by

$$\eta(x) \equiv g(m(x)) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p,$$

where p is the degree of a polynomial function. If $g = (b')^{-1}$, then g is called the canonical link (McCullagh and Nelder, 1989). Then, the conditional density $f_{Y|X}(y|x)$ in (1) can be written in terms of $\eta(x)$ as

$$f_{Y|X}(y|x) = \exp\{y(g \circ b')^{-1}(\eta(x)) - b((g \circ b')^{-1}(\eta(x))) + c(y)\} \tag{2}$$

where $\circ$ denotes the composition of functions.

Let us write $\ell(z, y)$ for the logarithm of the conditional density in (2) with $\eta(x)$ replaced by z. Also, define $\widehat{\eta}_h(x) = \widehat{\beta}_0$ and $\widehat{\eta}'_h(x) = \widehat{\beta}_1$ as the estimators for $\eta(x)$ and $\eta'(x)$, where $(\widehat{\beta}_0, \widehat{\beta}_1, \cdots \widehat{\beta}_p)$ maximizes the following kernel weighted local-likelihood function (Fan and Gijbels, 1996):

$$\sum_{i=1}^{n} \ell\left(\beta_0 + \beta_1(X_i - x) + \cdots + \beta_p(X_i - x)^p, Y_i\right) K\left(\frac{X_i - x}{h}\right). \tag{3}$$

Here, K is a kernel function and h is a bandwidth. We use the Gaussian kernel in this paper by the suggestion of Chaudhuri and Marron (1999). Also, p=1 is used for our numerical study. Note that we do not have explicit solutions to the maximization (3). This estimation also applies to quasi-likelihood models where only the relationship between the mean and the variance is specified. In this circumstance our estimators can be achieved by replacing the log-likelihood by a quasi-likelihood (Li and Marron, 2005).

The objective of the proposed method is to determine the significance of trends in the function η at a particular location x and scale h. Since we change our emphasis from finding significant features in noisy data of the true underlying curve to finding them in the curve at a given level of resolution, the null hypothesis of SiZer inference in scale-space is given as $H_0 : \eta'_h(x) = 0$ where $\eta'_h(x) \equiv E[\widehat{\eta}'_h(x)]$ rather than $H_0 : \eta'(x) = 0$. The corresponding confidence interval for $\eta'_h(x)$ given as

$$\widehat{\eta}'_h(x) \pm q(h)\widehat{SD}(\widehat{\eta}'_h(x)) \tag{4}$$

where $q(h)$ is an appropriate quantile depending on h given as

$$q(h) = \Phi^{-1}\left(\left(1 - \frac{\alpha}{2}\right)^{1/(\theta r)}\right). \tag{5}$$

Here, Φ is the cumulative distribution function of the standard normal, r is the number of pixels in each row of a SiZer map, and the cluster index θ is given as

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

which measures the equivalent number of independent observations. Here, $\tilde{\Delta}$ denotes the distance between the pixels of the SiZer map. The nominal level used in all of the numerical examples in this paper is α=0.05. The detail of (5) using the canonical link is given in the Appendix. For the estimate of the standard deviation (SD), it is given in Fan et al. (1995). In the case of the local linear estimator and the canonical link,

$$\widehat{SD}(\widehat{\eta}'_h(x)) = \sqrt{\frac{\int z^2 K^2(z)dz}{nh^3 v(x) f(x)}}$$

where $v(x) \equiv Var(Y|X = x)$ is the conditional variance function at x. Note that $v(x) = (g^{-1})'(\eta(x))$ if the canonical link is chosen. To construct a confidence interval in (4), the functions v and f should be estimated. Because the proposed approach is scale-dependent, these estimates are also dependent on the scale. Hence, we use the same bandwidth h and the canonical link in a SiZer map, i.e.,

$$\hat{v}_h(x) = (g^{-1})'(\hat{\eta}_h(x)), \quad \hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x),$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$.

SiZer visually displays the significance of trends in a family of smooths $\widehat{\eta}_h(x)$ over both location x and scale h, using a color map. It is based on confidence intervals for $\widehat{\eta}'_h(x)$ defined in (4), and uses multiple comparison level adjustment. Each pixel shows a color that gives the result of the hypothesis testing of $H_0 : \eta'_h(x) = 0$ at the point indexed by the horizontal location x, and by the bandwidth corresponding to the row h. At each (x, h), if the confidence interval is above (below) 0, meaning that the curve is increasing (decreasing) at x, i.e., $\eta'_h(x) > 0$ ($\eta'_h(x) < 0$), then that particular map location is colored black (white, respectively). On the other hand, if the confidence interval contains 0, meaning that the curve does not have a statistically significant slope, then that map location is given intermediate gray. Finally, if there are not enough data points to carry out the test, then no decision can be made and the location is colored dark gray. To determine the dark gray areas, as in Chaudhuri and Marron (1999), we define the estimated effective sample size (ESS), for each (x, h) as

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

If $\mathrm{ESS}(x, h) < 5$, then the corresponding pixel is colored dark gray.

Remark. Li and Marron (2005) adopted an approach for estimating the quantile proposed by Chaudhuri and Marron (1999). They chose the number of independent blocks of average size available from the data, $s(h)$:

$$s(h) = \frac{n}{avg_x \mathrm{ESS}(x, h)},$$

and the simultaneous quantile is given as

$$q(h) = \Phi^{-1}\left(\left(1 + (1 - \alpha)^{1/s(h)}\right)/2\right).$$

Hannig and Marron (2006), however, showed that this quantile tends to produce spurious pixels in a SiZer map and fails to control Type I errors. Hence, they improved the quantile using advanced distribution theory. In the Appendix, we illustrate how to obtain the quantile $q(h)$ in (5) by extending the result of Hannig and Marron (2006) to discrete data using the local linear estimator with $p = 1$ in (3).

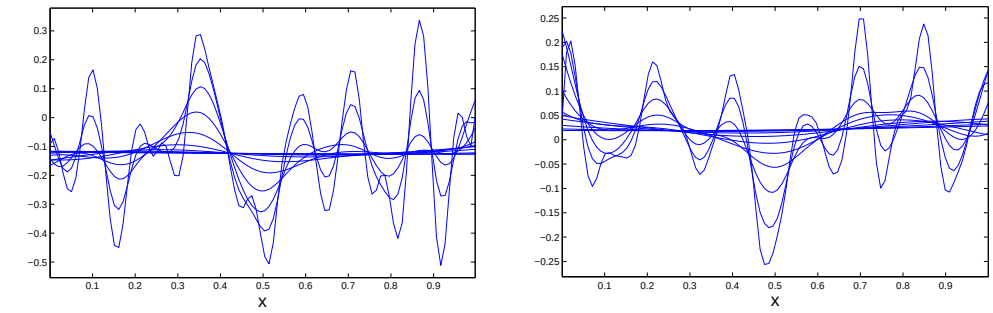
3 A numerical study

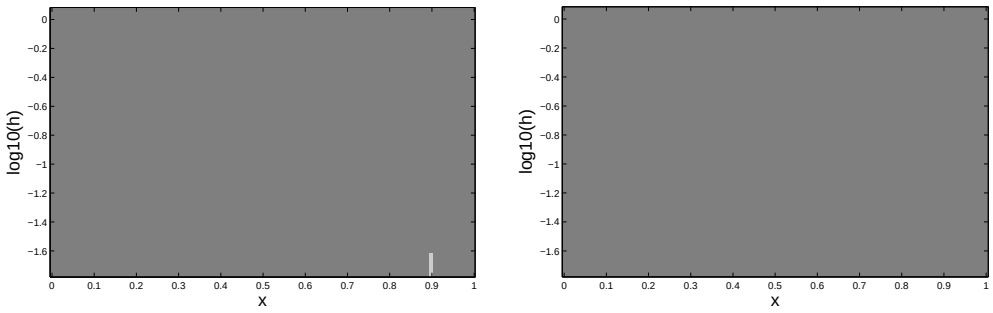
In this section, we demonstrate the practical aspects of the proposed SiZer using both simulated and real examples for one covariate cases.

3.1 Simulated data

This simulation study considers two likelihoods, Bernoulli and Poisson. For the Bernoulli likelihood the logit link function $\eta(x) = \text{logit}\{P(Y = 1|X = x)\} = \ln\{m(x)/(1 - m(x))\}$ is used, and for the Poisson case the log link function $\eta(x) = \ln m(x)$ is used. Using these two likelihoods, we test three different functions, $\eta(x) = 0$, $\eta(x) = x - 0.5$, and $\eta(x) = \cos(6\pi x)$. Each example has the sample size $n = 1000$ and X's are generated from either $U(0, 1)$ or the truncated normal between 0 and 1 with the location parameter 0.5 and the scale parameter 0.6.

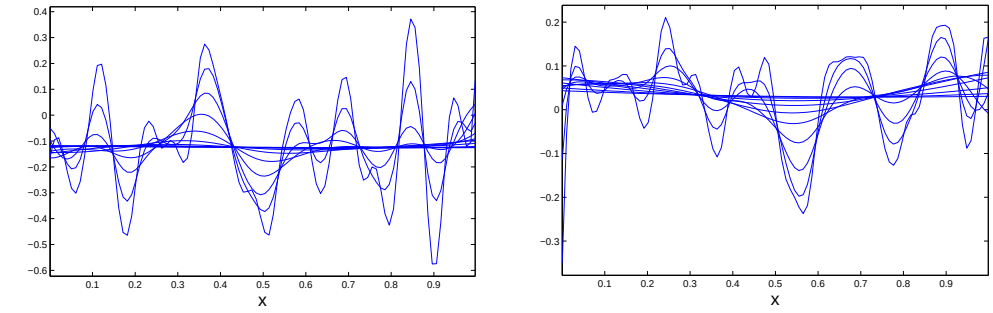
Figure 1 shows SiZer plots for $\eta(x) = 0$. In Figures 1 (a) and (b), X's are generated from $U(0, 1)$, and in (c) and (d), X's are generated from truncated normal. In addition, the Bernoulli likelihood is used for (a) and (c) and the Poisson likelihood for (b) and (d). In the top panels, the thin curves display the family of smooths, which are local linear smooths $\widehat{\eta}_h(x)$ with different h. These curves are located around 0 for both densities and both likelihoods. 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. In particular, each row of the SiZer map does statistical inference for one of the thin curves in the top panel. The inference focuses on the slope of the curve, i.e. the test of $\eta'_h(x) = 0$ by investigating the confidence intervals in (4) at each (x, h). Each pixel shows a color that gives the result of a hypothesis test for the sign of the thin curve, at the point indexed by the horizontal location, and at the bandwidth corresponding to that row as explained in Section 2. The result shows mostly intermediate gray except a few spurious pixels for all cases, meaning no significant trends for both densities and both likelihoods, as expected.

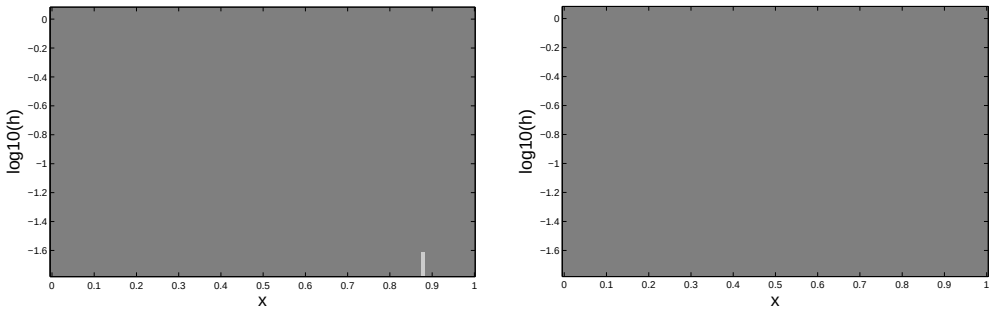




(a) Uniform, Bernoulli

(b) Uniform, Poisson

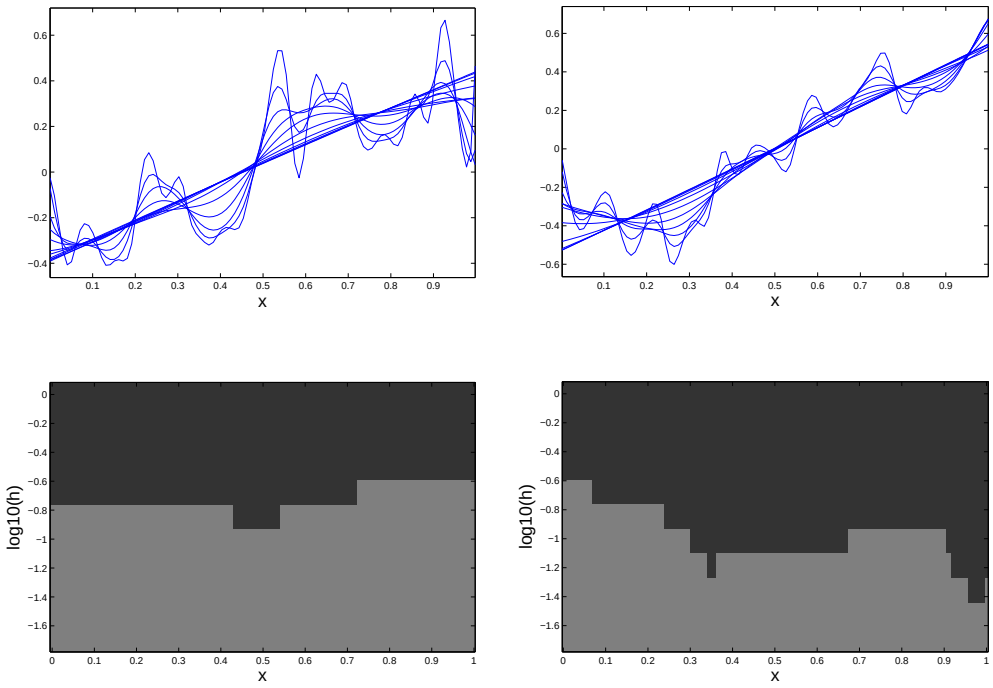




(c) T-Normal, Bernoulli

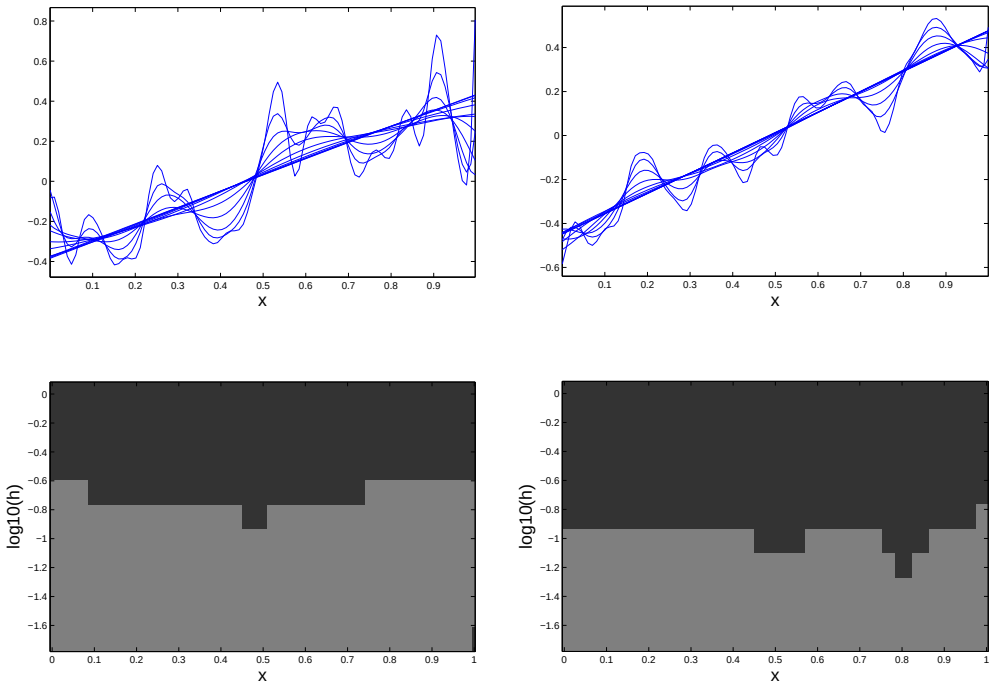
(d) T-Normal, Poisson

Figure 1: Example 1: $\eta(x) = 0$. The X's are generated from $U(0,1)$ for the first row ((a) and (b)), and truncated normal for the second row ((c) and (d)). The Bernoulli likelihood is used in the first column ((a) and (c)) and the Poisson likelihood is used in the second column ((b) and (d)).



(a) Uniform, Bernoulli

(b) Uniform, Poisson



(c) T-Normal, Bernoulli

(d) T-Normal, Poisson

Figure 2: Example 2: $\eta(x) = x - 0.5$. Caption is the same as Figure 1.

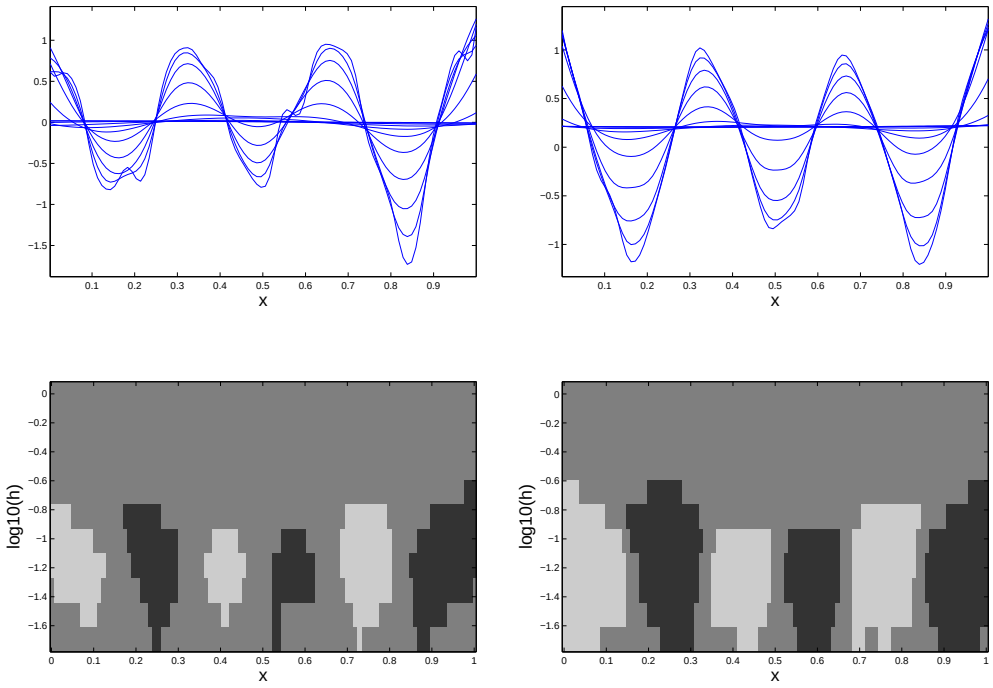
For the second example, $\eta(x) = x - 0.5$ and Figure 2 displays the corresponding SiZer plots. The upper panel shows a clear increasing linear trend for all cases. The SiZer map shows exclusively black for all of the larger windows (coarser scales), i.e. the upper rows in the SiZer map, which correctly identifies the global increasing trend.

The result of the last example with $\eta(x) = \cos(6\pi x)$ is displayed in Figure 3. The family of smooths shows a cosine trend in the top panels, and this trend is significantly flagged (black and white alternate each other) in the SiZer maps in the bottom panels. The three examples demonstrate that the proposed SiZer correctly identifies significant features with both densities and both likelihoods.

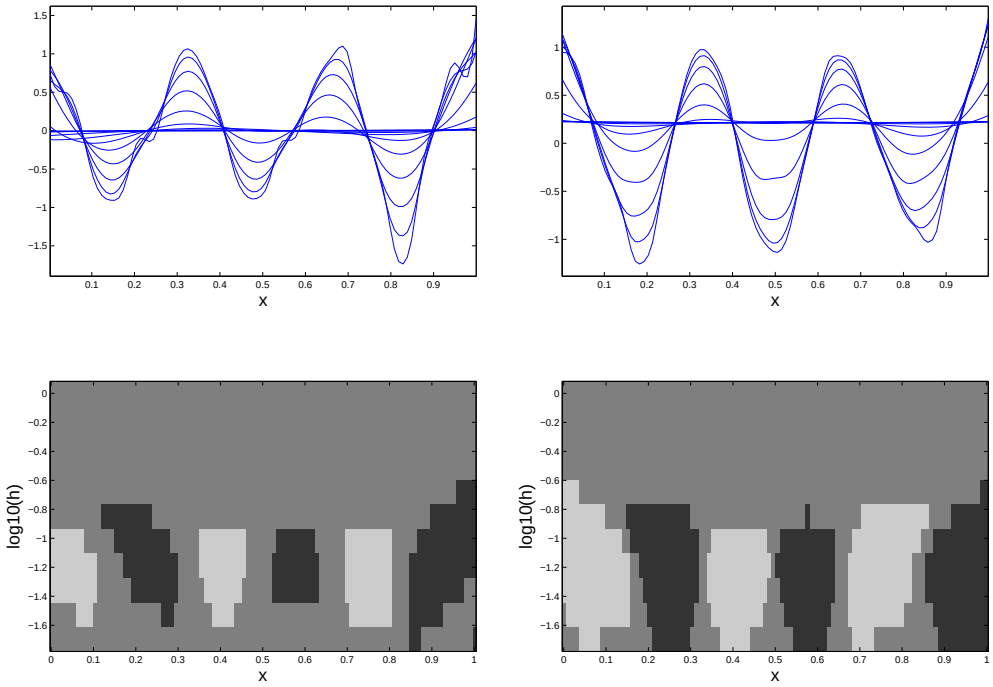
3.2 Real examples

In this subsection we analyze two real examples. The first data considers the relationship between wages (continuous) and union membership (binary) in 1985 (Berndt, 1991). Since the union membership is binary, the Bernoulli likelihood is used. Figure 4(a) displays its SiZer plot. The family of smooths shows an increasing trend for large bandwidths, but a decreasing trend after the location \$15/hour for medium bandwidths. The corresponding SiZer map in the bottom panel of Figure 4(a) shows exclusively black for all of the larger windows. This is evidence of a strong upwards trend. Substantial white regions appear after the location \$15/hour in the middle part of the map, i.e. the finer levels of resolution or smaller window widths. These correspond to short term decreases in the more wiggly thin curves. At very small scales black and white colors alternate, which suggests that the wiggles in the top panel are statistically significant for very small bandwidths. Also, substantial dark gray regions on the right side of the map indicate that insufficient amount of data is available for statistical inference.

In the second example, daily data over 10 years are available on mortality, air pollution, and several meteorological variables for the city of Milan, Italy (Ruppert et al., 2003). In our analysis relative humidity is the explanatory variable and total mortality is the response. We use the Poisson likelihood for SiZer analysis. The family of smooths suggests a decreasing



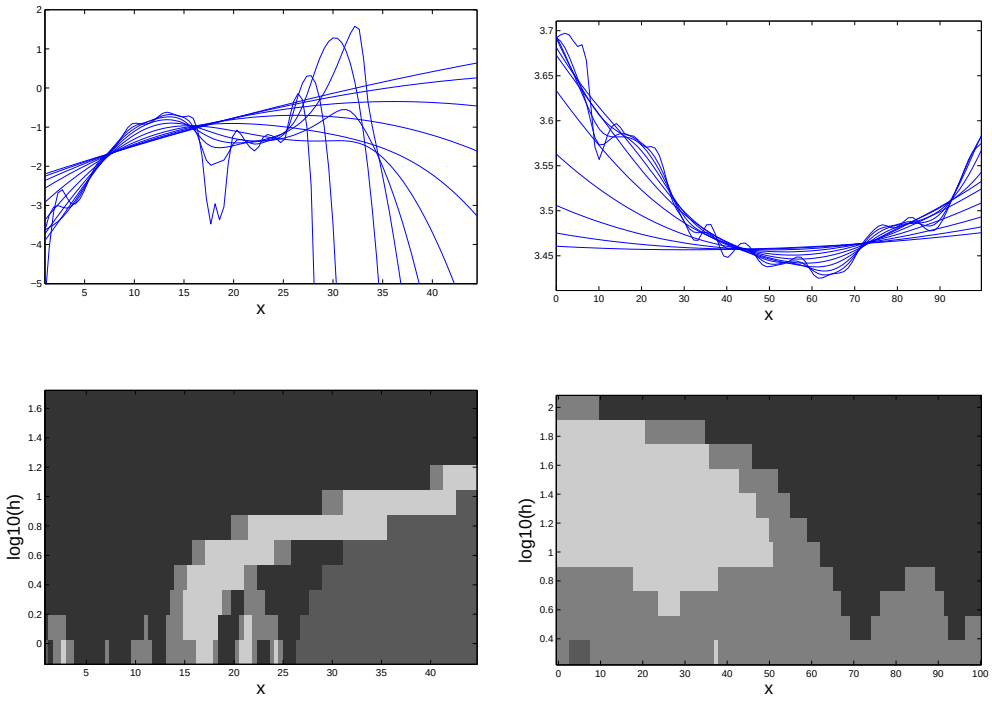
(a) Uniform, Bernoulli (b) Uniform, Poisson



(c) T-Normal, Bernoulli (d) T-Normal, Poisson

Figure 3: Example 3: $\eta(x) = \cos(6\pi x)$. Caption is the same as Figure 1.

trend in the first half and then an increasing trend in the second. These trends are significantly flagged in the corresponding SiZer map as it shows black in the first half and white in the second at large and medium resolution levels.

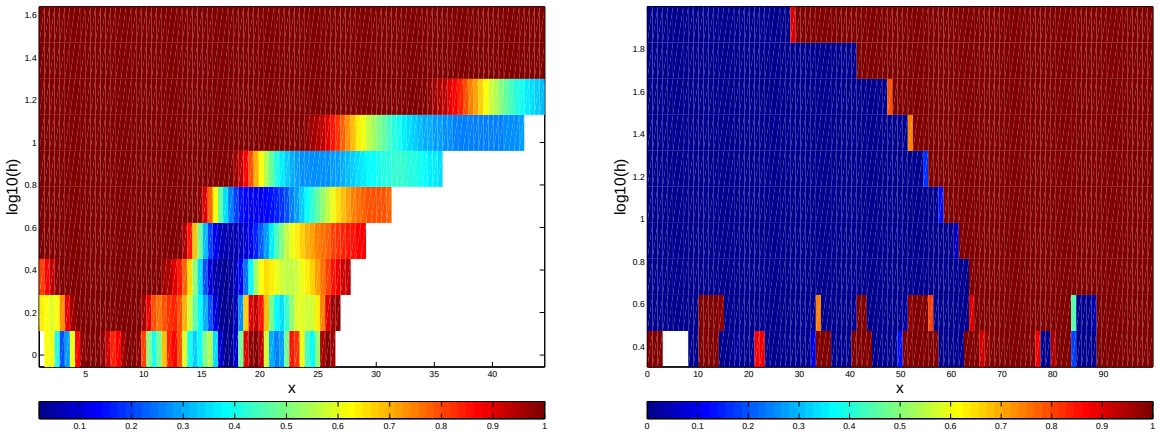


(a) Trade union data (b) Mortality data

Figure 4: SiZer plots for (a) trade union data with the Bernoulli likelihood and (b) mortality data from Milan with the Poisson likelihood

The current SiZer maps report the testing results using only four categories: increasing trend, decreasing trend, insignificant trend, and insufficient data. Following the referee's suggestion, we create a different type of SiZer maps using the one-sided p-values of the test statistics in Figure 5. Dark red colors (p-values close to 1) suggest that the trend is significantly increasing, dark blue (p-values close to 0) significantly decreasing, and the middle colors (medium p-values) insignificant. Also, white colors indicate that there are insufficient data points for the statistical tests. Although the main findings from Figure 4 remain the same, the new maps in Figure 5 provide richer information about the degree of

significance of the local features since the significant levels are drawn on a continuous scale.



(a) Trade union data (b) Mortality data

Figure 5: SiZer maps based on the one-sided p-values of the test statistics for the examples in Figure 4.

4 Asymptotic results

In this section, we describe statistical convergence of the difference between the empirical and the theoretical scale space surfaces $(\widehat{\eta}_h^{(m)}(x), E\{\widehat{\eta}_h^{(m)}(x)\})$ for any fixed non-negative integer m, which provides theoretical justification of the proposed SiZer tool in scale-space. Chaudhuri and Marron (2000) addressed these issues for the nonparametric density and regression models. Note that the canonical link is used in this section.

We use an approximation form of $\widehat{\eta}_h^{(m)}(x)$ described in Lemma 1 to show the statistical convergence. To derive Lemma 1, let $S_{n,x}$ be the $(p+1)\times(p+1)$ matrix having their (j,k)th entry equal to

$$\frac{1}{(j-1)!(k-1)!}\int K(u)u^{j+k-2}f(x+hu)du.$$

And let $\boldsymbol{e}_k$ be the $(p+1)\times 1$ vector with 1 appearing at the kth position and 0 otherwise.

Define

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

In usual kernel type estimation when $h \to 0$, the (j,k)th entry of $S_{n,x}$ can be expressed approximately as follows: $f(x)\int K(u)u^{j+k-2}du/((j-1)!(k-1)!)$ when f has a regularity condition.

In order to obtain the approximation form of $\widehat{\eta}_h^{(m)}(x)$ uniformly in x in Lemma 1, the following conditions are needed.

(A.1) The function $\eta^{(p+1)}$ is uniformly continuous over $[0,1]$, which is the support of X.

(A.2) The function $(g^{-1})'$ is uniformly continuous over $(-\infty,\infty)$.

(A.3) $\inf_{x\in[0,1]} v(x) > 0$.

(A.4) $\sup_{x\in[0,1]} E(|Y - E(Y|X=x)|^{2+\rho}|X=x) < \infty$ for some positive ρ.

Lemma 1 *Suppose that the assumptions (A.1)–(A.4) are satisfied.*
(i) *If $m=0$, then*

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

uniformly in x as $n\to\infty$. Here, $Y_{i,x}^* = Y_i - g^{-1}(\eta(x))$ for $i=1,\cdots,n$.
(ii) *If $m>0$, then*

$$\widehat{\eta}_h^{(m)}(x) = \frac{1}{n}\sum_{i=1}^n W_{n,m}(h,x,X_i)Y_{i,x}^*(1+o_P(1))$$

uniformly in x as $n\to\infty$.

Proof. The proof follows the similar line to that of Lemma 1 in Huh (2010). Let $\bar{\boldsymbol{\beta}}$ and

$\boldsymbol{Z}_i$ be the $(p+1)\times 1$ vectors as follows:

$$\bar{\boldsymbol{\beta}} = \sqrt{nh}\begin{pmatrix} \widehat{\beta}_0 - \eta(x) \\ h\widehat{\beta}_1 \\ \vdots \\ h^p p!\widehat{\beta}_p \end{pmatrix}, \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 \\ (X_i - x)/h \\ \vdots \\ (X_i - x)^p/(h^p p!) \end{pmatrix},$$

for $i = 1, \cdots, n$. Note that $\sum_{j=0}^{p} \widehat{\beta}_j (X_i - x)^j = \eta(x) + \bar{\boldsymbol{\beta}}^T \boldsymbol{Z}_i/\sqrt{nh}$. Thus, $\bar{\boldsymbol{\beta}}$ maximizes $\sum_{i=1}^{n} \ell\big(\eta(x) + \boldsymbol{\beta}^{*T}\boldsymbol{Z}_i/\sqrt{nh}, Y_i\big)K((X_i - x)/h)$ as a function of $\boldsymbol{\beta}^*$. Consider the normalized function

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

which is also maximized by $\boldsymbol{\beta}^*$. Note that L_n is concave in $\boldsymbol{\beta}^*$ since the canonical function is used. Define

$$A_n = \frac{1}{nh}(g^{-1})'(\eta(x)) \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)\boldsymbol{Z}_i\boldsymbol{Z}_i^T$$

and $\ell_i = \partial^i \ell(z, y)/\partial^i z$, $i = 1, 2$. Then, we obtain

$$\ell_1(z, y) = y - g^{-1}(z) \quad \text{and} \quad \ell_2(z, y) = -(g^{-1})'(z). \tag{6}$$

Using a Taylor series expansion of the function $\ell(\cdot, Y_i)$ and (6), we obtain $L_n(\boldsymbol{\beta}^*) = \boldsymbol{W}_n^T\boldsymbol{\beta}^* - \frac{1}{2}\boldsymbol{\beta}^{*T}A_n\boldsymbol{\beta}^* + o_P(1/(nh))$ where $\boldsymbol{W}_n = (1/\sqrt{nh})\sum_{i=1}^{n} K((X_i - x)/h)\boldsymbol{Z}_i Y_{i,x}^*$. Let

$$\xi(x) = \frac{1}{(j-1)!(k-1)!}\frac{1}{nh}(g^{-1})'(\eta(x)) \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)\Big(\frac{X_i - x}{h}\Big)^{j+k-2}$$

be the (j, k)th element of A_n where $1 \le j, k \le (p+1)$. By similarly following the proof of Lemma 1 in Huh (2010), we have $\sup_{x\in[0,1]} |\xi(x) - E(\xi(x))| = O_P(\sqrt{\log n/(nh)})$ and

$$L_n(\boldsymbol{\beta}^*) = \boldsymbol{W}_n^T\boldsymbol{\beta}^* - \frac{1}{2}\boldsymbol{\beta}^{*T}E(A_n)\boldsymbol{\beta}^* + O_P\Big(\sqrt{\frac{\log n}{nh}}\Big) + o_P\Big(\frac{1}{nh}\Big).$$

Since the canonical link is used, we have the following relation $v(x) = (g^{-1})'(\eta(x))$. The expected value of the (j, k)th element of A_n is then

$$E(\xi(x)) = -\frac{v(x)}{(j-1)!(k-1)!}\int K(u)u^{j+k-2}f(x + hu)du.$$

Define $L(\boldsymbol{\beta}^*) = \boldsymbol{W}_n^T\boldsymbol{\beta}^* - \frac{1}{2}v(x)\boldsymbol{\beta}^{*T}S_{n,x}\boldsymbol{\beta}^*$. By the Convexity Lemma in Pollard (1991), $\sup_{\boldsymbol{\beta}^*\in\mathcal{C}}|L_n(\boldsymbol{\beta}^*) - L(\boldsymbol{\beta}^*)| = o_P(1)$ for any compact set $\mathcal{C}$, and the maximizer of $L(\boldsymbol{\beta}^*)$ is given as

$$\widehat{\boldsymbol{\beta}}^* = \frac{1}{v(x)}S_{n,x}^{-1}\boldsymbol{W}_n.$$

Then, $\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^* = o_P(1)$ by the Quadratic Approximation Lemma in Carroll et al. (1997), which completes the proof. ■

For simplicity, define the approximation form of $\widehat{\eta}_h(x) - \eta(x)$ or $\widehat{\eta}_h^{(m)}(x)$ as

$$\widetilde{\eta}_h^{(m)}(x) = n^{-1}\sum_{i=1}^{n} W_{n,m}(h, x, X_i)Y_{i,x}^*. \tag{7}$$

We need the following additional set of assumptions for the weight $W_{n,m}(h, x, u)$ in (7) to describe Theorems 1 and 2. Let I and H be compact subintervals of $[0, 1]$ and $(0, \infty)$, respectively.

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

$$n^{-1}\sum_{i=1}^{n} v(X_i)W_{n,m}(h_1, x_1, X_i)W_{n,m}(h_2, x_2, X_i)$$

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

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

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

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, $v(X_i)\{\partial^2 W_{n,m}(h, x, X_i)/(\partial h\partial x)\}^2$ 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, $v(X_i)\{\partial W_{n,m}(h, x, X_i)/\partial x\}^2$ and $v(X_i)\{\partial W_{n,m}(h, x, X_i)/\partial h\}^2$ are uniformly dominated by a positive function $M^*(X_i)$ such that $E\{M^*(X_i)\} < \infty$.

As mentioned in Chaudhuri and Marron (2000), the assumptions (A.5)–(A.8) for the weight function of the local polynomial fit are satisfied for many standard kernels including the Gaussian kernel. The assumption (A.3) for the denominator in the approximation form is needed to assure the assumptions (A.5)–(A.8).

Theorem 1 provides the weak convergence of the empirical scale-space surfaces to their theoretical counterpart.

Theorem 1 *Suppose that the assumptions (A.1)–(A.8) are satisfied. Then as* $n \longrightarrow \infty$, *the two-parameter stochastic process*

$$U_n(h,x) = n^{1/2}\Big[\widehat{\eta}_h^{(m)}(x) - E\{\widehat{\eta}_h^{(m)}(x)\}\Big]$$

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

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

For the weak convergence to the normal distribution, let us fix $(h_1,x_1),(h_2,x_2),\cdots,(h_k,x_k) \in H \times I$ and $t_1,t_2,\cdots,t_k \in (-\infty,\infty)$. Define

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

Note that $Z_n = \tilde{Z}_n(1+o_P(1))$ for any fixed $x_i \in I$ since we have the approximation form of $\widehat{\eta}_h^{(m)}(x)$ uniformly in $x \in I$ by Lemma 1. The conditional mean and variance of $\tilde{Z}_n$ are 0 and

$$n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{k} t_i t_j \sum_{l=1}^{n} v(X_l) W_{n,m}(h_i,x_i,X_l) W_{n,m}(h_j,x_j,X_l)$$

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

$$\tilde{U}_n(h,x) = n^{1/2}\Big[\widetilde{\eta}_h^{(m)}(x) - E\{\widetilde{\eta}_h^{(m)}(x)\}\Big].$$

Then, we have $U_n(h,x) = \tilde{U}_n(h,x)(1+o_P(1))$ uniformly in $x \in I$ and $h \in H$. Finally, it follows using the Cramer-Wold device that as $n \to \infty$, the joint limiting distribution of $\tilde{U}_n(h_i,x_i)$ for $1 \le i \le k$ is multivariate normal with zero mean and $cov(h_i,x_i,h_j,x_j)$ as the (i,j)th entry of the limiting covariance matrix for $1 \le i,j \le k$.

For tightness, fix $h_1 < h_2$ in H and $x_1 < x_2$ in I. Then,

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

for some constants C_2 and $C_3 > 0$. 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, and consequently the theorem follows. ■

Theorem 2 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.

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

$$\sup_{x\in I, h\in H} n^{1/2}\Big|\widehat{\eta}_h^{(m)}(x) - E\{\widehat{\eta}_h^{(m)}(x)\}\Big|$$

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

Proof. 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 way in Chaudhuri and Marron (2000). ■

5 Discussion

While the proposed tool in Section 2 is useful for detecting statistically meaningful features when there is one covariate, it has limitations in applying to multiple-covariate cases. In this section, we briefly discuss how to extend the local likelihood SiZer tool to the case of multiple covariates in a straightforward way using the local linear estimator.

Define $\boldsymbol{X} = (X_1, \cdots, X_d)^T$ as the vector of covariates. Suppose that we observe a random sample $(\boldsymbol{X}_i^T, Y_i)$, $i = 1, \cdots, n$, of $(\boldsymbol{X}^T, Y)$ where the Y_i's are real valued responses associated with d-dimensional covariates $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{id})^T$'s. Let $\eta(\boldsymbol{x})$ be the d-variate regression function transformed by a link function g where $\boldsymbol{x} = (x_1, \cdots, x_d)^T$ is a point in $[0, 1]^d$.

Let K be a d-variate kernel function and $B = \text{diag}(h_1, h_2, \cdots, h_d)$ be a $d \times d$ diagonal bandwidth matrix. We assume that the bandwidth matrix B is diagonal for simplicity, but non-diagonal elements could be nonzero if one considers the relationships between covariates. The d-variate extension of the local linear regression is then based on the following kernel weighted local-likelihood function:

$$\sum_{i=1}^{n} \ell(\boldsymbol{\beta}^T(\boldsymbol{X}_i - \boldsymbol{x}), Y_i) K(B^{-1}(\boldsymbol{X}_i - \boldsymbol{x})) \tag{8}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_d)^T$ and $\ell(\eta(\boldsymbol{x}), y)$ is the logarithm of the conditional density in (2) with $\eta(x)$ replaced by $\eta(\boldsymbol{x})$. Define

$$\widehat{\eta}(\boldsymbol{x}) = \widehat{\beta}_0, \quad \frac{\partial}{\partial x_i}\widehat{\eta}(\boldsymbol{x}) = \widehat{\beta}_i, \quad i = 1, \cdots, d,$$

as the estimators for $\eta(\boldsymbol{x})$ and $\frac{\partial}{\partial x_i}\eta(\boldsymbol{x})$ where $(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_d)$ maximizes (8). We suggest the actual implementation of the multiple-covariate SiZer tool as our future work.

Acknowledgement

The authors are grateful to the reviewers and the associate editor for many helpful comments. The research of the second 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 (2011-0010856).

6 Appendix

In this section we provide the details on the quantile given in (5) using the local linear estimator in (3) with $p = 1$. In the particular case of fixed design regression, 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 4, if we take

$$L(\boldsymbol{\beta}^*) = \boldsymbol{W}_n^T \boldsymbol{\beta}^* - \frac{1}{2}\boldsymbol{\beta}^{*T} A_n \boldsymbol{\beta}^*,$$

the maximizer of $L(\boldsymbol{\beta}^*)$ is then

$$\widehat{\boldsymbol{\beta}}^* = A_n^{-1}\boldsymbol{W}_n.$$

If x is away from the boundary, it follows from symmetry of the kernel that

$$\sum_{i=1}^{n} \Big(\frac{X_i - x}{h}\Big) K\Big(\frac{X_i - x}{h}\Big) \approx 0. \tag{9}$$

By (9) and $X_i = i\Delta$, A_n is a diagonal matrix having the jth diagonal

$$\frac{1}{nh}(g^{-1})'(\eta(x)) \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)\Big(\frac{X_i - x}{(j-1)!h^{j-1}}\Big)^{2j-2}, \quad j = 1, 2.$$

Then, the second element of $\hat{\boldsymbol{\beta}}^*$ is

$$\widehat{\beta}_1^* = \frac{1}{nh^2}\frac{1}{v(x)}\Big\{\frac{1}{nh}\sum_{i=1}^{n}\Big(\frac{X_i - x}{h}\Big)^2 K\Big(\frac{X_i - x}{h}\Big)\Big\}^{-1} \sum_{i=1}^{n}\Big(\frac{X_i - x}{h}\Big) K\Big(\frac{X_i - x}{h}\Big) Y_{i,x}^*, \tag{10}$$

which is an approximation form of $\widehat{\eta}_h'(x)$ as $n \to \infty$ when $p = 1$ and $X_i = i\Delta$. Therefore, by (9), $Y_{i,x}^*$ in (10) can be replaced by Y_i for $i = 1, \cdots, n$.

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 r be the number of pixels in each row, and $Q_1, \cdots, Q_r$ denote the test statistics of a row in the SiZer map. Then Q_j is proportional to the estimate of η' calculated for $x = j\tilde{\Delta} = jt\Delta$. In particular,

$$Q_j \approx \sum_{l=1}^{n} W_{jt-l}^{h} Y_l. \tag{11}$$

The exact form of the W_{jt-l}^{h} is given in (10). Note that W_{jt-l}^{h} is proportional to $-(jt - l)K_{h/\Delta}(jt - l)$, the derivative of the Gaussian kernel with standard deviation h/Δ.

If there is no signal, then Y_i's are iid random variables. Additionally, if Y'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. According to the Appendix in Hannig and Marron (2006) with $t\Delta = \tilde{\Delta}$,

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

Therefore, using Theorem 1 in Hannig and Marron (2006), for each fixed r we can get j step correlation $\rho_{j,r}$ such as

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

by setting $\tilde{\Delta}/h = C/\sqrt{\log r}$ and

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

Following the similar arguments in the paragraphs after Theorem 1 of Hannig and Marron (2006), we conclude that

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

where the cluster index θ

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

References

Berndt, E. R. (1991). *The Practice of Econometrics: Classic and Contemporary.* Addison-Wesley, Reading, MA.

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *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. *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., Heckman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90:141–150.

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

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.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

Hannig, J. and Lee, T. (2006). Robust SiZer for exploration of regression structures and outlier detection. *Journal of Computational & 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.

Huh, J. (2010). Detection of a change point based on local-likelihood. *Journal of Multivariate Analysis*, 101:1681–1700.

Huh, J. and Park, B. U. (2002). Likelihood-based local polynomial fitting for single-index models. *Journal of Multivariate Analysis*, 80:302–321.

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.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, 2nd ed.* Chapman & Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384.

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

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.

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., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge, New York.

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.