# Local Likelihood SiZer Map


Runze Li
Department of Statistics
Pennsylvania State University
University Park, PA 16802-2111
Email: rli@stat.psu.edu

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
Email: marron@email.unc.edu


February 7, 2005


**Abstract**

The SiZer Map, proposed by Chaudhuri and Marron (1999), is a statistical tool for finding which features in noisy data are strong enough to be distinguished from background noise. In this paper, we propose the local likelihood SiZer map. Some simulation examples illustrate that the newly proposed SiZer map is more efficient in distinguishing features than the original one, because of the inferential advantage of the local likelihood approach. Some computational problems are addressed, with the result that the computational cost in constructing the local likelihood SiZer map is close to that of the original one.



**Key Words:** Confidence bands, generalized linear models, local polynomials, local likelihood, quasi-likelihood, significant features, SiZer map.

# 1 Introduction

Consider the bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$, an independent and identically distributed sample from the model

$$Y = m(X) + \varepsilon,$$

where $\varepsilon$ is random error with $E(\varepsilon|X) = 0$, $E(\varepsilon^2|X) = \sigma^2(X)$. Various nonparametric smoothing methods can be used to estimate $m(X)$. In this paper, we will use only local polynomial regression. The choice of bandwidth $h$ is of crucial importance in local polynomial regression. If $h$ is chosen too large then the resulting estimate misses fine features of the regression curve, while if $h$ is selected too small then spurious sharp structure becomes visible. There are a large number of proposals on bandwidth selection in the literature. A "best" data-driven bandwidth for local polynomial regression can be constructed using the ideas of cross-validation (see, for example, Härdle, 1990), nearest neighbor bandwidth (Fan and Gijbels, 1995), and plug-in method (Gasser, Kneip and Köhler, 1991, Sheather and Jones, 1991 and Ruppert, Sheather and Wand, 1995). All of these methods try to find a single best bandwidth under some criteria.

A different approach to smoothing can be found in scale space theory in the field of computer vision. Lindeberg (1994) and ter Haar Romeny (2001) provide good discussions of scale space modeling. From this point of view, choosing a bandwidth $h$ is like adjusting the focus on a camera. The set of all possible smooths, indexed by $h$, is called the scale space surface, and the goal of modeling is to provide an approximate map of this surface. A model with a large bandwidth gives a macroscopic view of the surface, showing only large-scale features, while a small $h$ gives a microscopic of "zoomed-in" view allowing resolution of small scale features. Chaudhuri and Marron (2000) discuss curve estimation from a scale space point of view, and Godtliebsen, Marron and Chaudhuri (2002) apply scale space theory to bivariate density estimation. Various interesting case studies and applications of scale space concepts are given in Ghosh, Chaudhuri and Murthy (2004), Ghosh, Chaudhuri and Sengupta (2004) and Rakesh, Chaudhuri and Murthy (2004).

Chaudhuri and Marron (1999) also applied scale space concepts to develop the SiZer map, a powerful new tool for use in least-squares smoothing. The SiZer map is helpful for determining which features in noisy data are strong enough to be distinguished from background noise. However, it may be inefficient for discrete data. Also, approaches based on the likelihood or quasi-likelihood function can provide much more powerful inferences in many cases, as in the following example.

**Example 1** (*Poisson regression*) In this example, the covariate $X$ values are taken to be equally

spaced on $[0,\ 10\pi]$, and the conditional distribution of $Y$ given $X$ is a Poisson distribution with mean function

$$\lambda(x) = \exp\left\{\frac{15\sin(x)}{x+4}\right\}.$$

Figure 1 shows the underlying mean function and scatter plot of a realization of the raw data with sample size $n = 500$. Note that the variance is higher when the mean is high. Figures 2 and 3 compare the SiZer map (as proposed by Chaudhuri and Marron, 1999) to the improved SiZer map, proposed in this paper, based on local likelihood for sample sizes 500 and 200, respectively. The color scheme and line type of the SiZer map used in this paper will follow those of Chaudhuri and Marron (1999). Specifically, behavior at an $(x, h)$ location is presented via the SiZer color map where blue (black in versions where only black and white are available) indicates locations where the true mean function is significantly increasing, red (white in black-white versions) shows where it is significantly decreasing, and purple (gray in black-white versions) indicates where the true mean function is not significantly different from zero. Moreover, a location is shaded gray when the "effective sample size in the window" is less than 5. The dotted curves in the SiZer maps show "effective window widths" of the smoothing windows, as intervals representing $\pm 2h$ (i.e. $\pm 2h$ standard deviations of the Gaussian smoothing kernel). A reference bandwidth is highlighted by the horizontal bar. In Figure 2, the two kinds of SiZer maps look similar because the sample size is large enough that all of the significant features are distinguished from background noise. However in Figure 3 there is less information in the data, and the local likelihood SiZer map is more efficient in distinguishing important features than the original SiZer using critical value given in (3.3).

This paper develops a local likelihood enhancement of the SiZer map. As pointed out by Chaudhuri and Marron (1999), the extension of the SiZer map to the context of local likelihood is conceptually straightforward. However, unlike the least-squares setting, the solution for the local likelihood score equations generally does not have a closed form. The Newton-Raphson or a similar iterative algorithm can be used in this case, but the computational cost can be very high, since the local likelihood or quasi-likelihood must be maximized for many different values of $x$ to estimate the curve over all of the scale. A major goal of this paper is to address this computational issue. Some applications for real data example are also given.

The reminder of the paper is organized as follows. In Section 2, we summarize the idea of local (quasi-) likelihood methodology. Section 3 develops the local likelihood SiZer map. SiZer maps for nonparametric Poisson regression and logistic regression are illustrated via simulated examples and a real data example in Section 4. A discussion is given in Section 5.

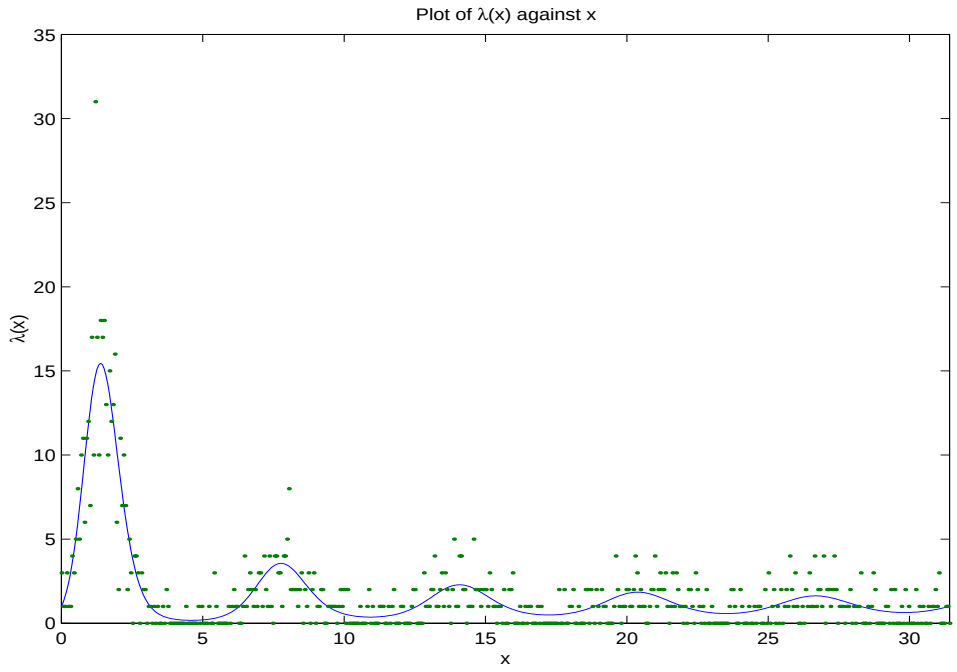


Figure 1: *True underlying signal in Example 1. Solid line stands for the mean function, and the dots are a realization of raw data with* $n = 500$.

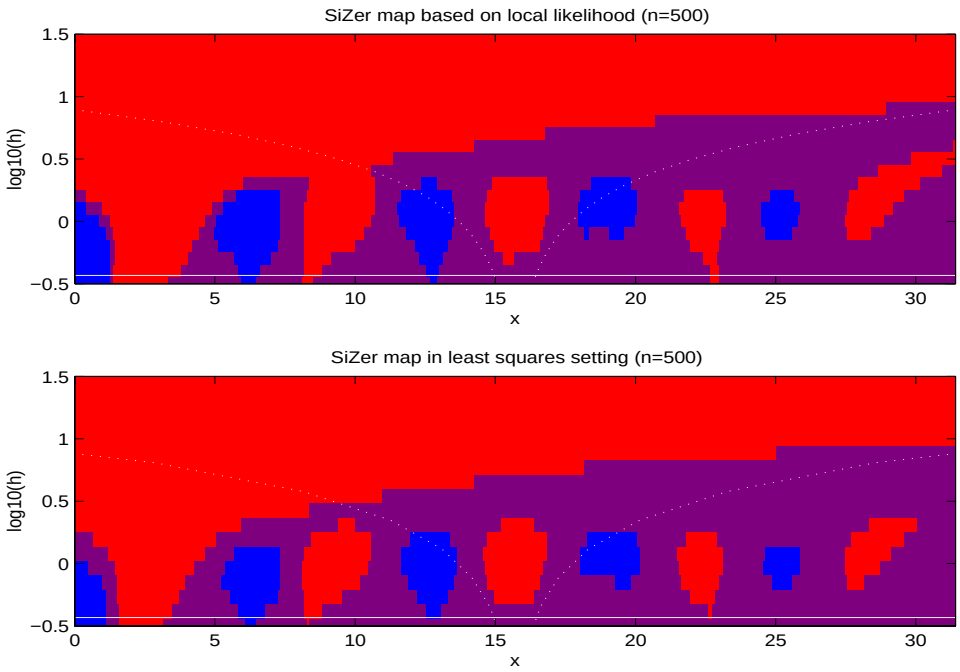


Figure 2: *SiZer maps for sample size* $n = 500$. *The top panel is the local likelihood SiZer map, and the bottom panel is the original SiZer map proposed by Chaudhuri and Marron (1999).*

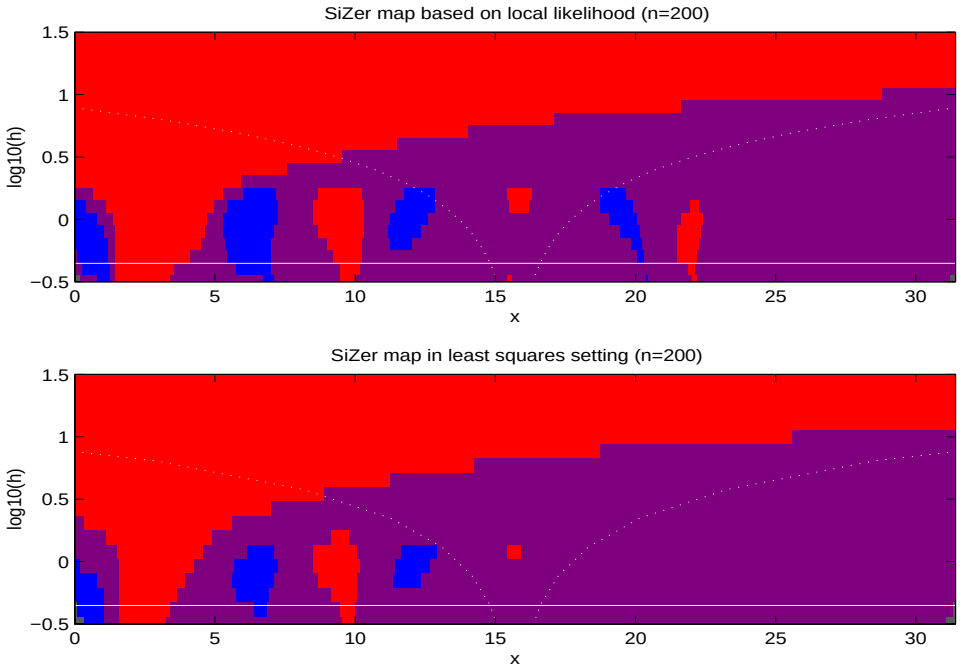


Figure 3: *SiZer maps for sample size* $n = 200$. *The top panel is the local likelihood SiZer map, and the bottom panel is the original SiZer map proposed by Chaudhuri and Marron (1999).*

# 2 Local Quasi-likelihood Approach

## 2.1 Generalized linear models and quasi-likelihood functions

Generalized linear models have been widely applied in various fields (see, for example, McCullagh and Nelder, 1989). Suppose that $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ are independent and identically distributed samples from the population $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a $d$-dimensional real vector of covariates, and $Y$ is a scalar response variable. Also suppose that the conditional density of $Y$ given covariate $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\left([\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)\right) \tag{2.1}$$

for known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. In parametric generalized linear models it is usual to model a transformation of a regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ as linear, that is

$$\eta(\mathbf{x}) = g\{m(\mathbf{x})\} = \mathbf{x}^T\boldsymbol{\beta},$$

and $g$ is a known *link* function. It is known for the exponential family that $\theta(\mathbf{x}) = (g \circ b')^{-1}\{\eta(\mathbf{x})\}$, where $\circ$ denote the composition. If $g = (b')^{-1}$, then $g$ is called the canonical link because $b'\{\theta(\mathbf{x})\} = m(\mathbf{x})$.

Under model (2.1), it can be easily shown that the conditional mean and conditional variance are given respectively by $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = b'\{\theta(\mathbf{x})\}$, and $\mathrm{Var}(Y|\mathbf{X} = \mathbf{x}) = a(\phi)b''\{\theta(\mathbf{x})\}$. There are many practical circumstances in which the conditional likelihood of $Y$ is unknown, but one is willing to assume a relationship between the mean function and the variance function. In this situation estimation of the mean function can be achieved by replacing the conditional log-likelihood $\log\{f_{Y|\mathbf{X}}(y|\mathbf{x})\}$ by a quasi-likelihood function $L(m(\mathbf{x}), y)$. If the conditional variance is modeled as $\mathrm{Var}(Y|\mathbf{X} = \mathbf{x}) = a(\phi)V\{m(\mathbf{x})\}$ for some known positive function $V$ and a unknown parameter $a(\phi)$, then the corresponding quasi-likelihood $L(\mu, y)$ satisfies

$$\frac{\partial}{\partial\mu}L(\mu, y) = \frac{y - \mu}{V(\mu)} \tag{2.2}$$

(due to Wedderburn, 1974). The quasi-score (2.2) possesses properties similar to those of the usual log-likelihood score function. Quasi-likelihood methods behave analogously to the usual likelihood methods and thus are reasonable substitutes when the likelihood function is not available. Note that the log-likelihood of the one-parameter exponential family is a special case of a quasi-likelihood function with $V = a(\phi)b'' \circ (b')^{-1}$, where $\circ$ denotes function composition.

## 2.2 Local Linear Estimation

Since the SiZer map for multi-dimensional covariates is beyond the scope of this paper, we only consider the $\mathbf{x}$ in (2.1) as one-dimensional. Because of the generality of the quasi-likelihood approach, we will formulate our results in these terms in this section. Results for the exponential family and for generalized linear models follow as a special case.

Suppose that the second derivative of the $\eta(x)$ exists and is continuous. For each given point $x_0$, we approximate the function $\eta(x)$ locally by a linear function $\eta(x) \approx \beta_0 + \beta_1(x - x_0)$ for $x$ in a neighborhood of $x_0$. Note that $\beta_0$ and $\beta_1$ depend on $x_0$. Based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, the local quasi-likelihood is

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} L[g^{-1}\{\beta_0 + \beta_1(X_i - x_0)\}, Y_i]K_h(X_i - x_0), \tag{2.3}$$

where $K_h(\cdot) = K(./h)/h$ with $K(\cdot)$ being a kernel function, $h = h_n > 0$ is a bandwidth. Define the local quasi-likelihood estimator of $\boldsymbol{\beta}$ to be

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}\in R^2}Q(\boldsymbol{\beta}). \tag{2.4}$$

Thus, the local linear quasi-likelihood estimator of $\eta(x)$ is given by

$$\widehat{\eta}(x_0) = \widehat{\beta}_0,$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^T$. The conditional mean function $m(x)$ can then be estimated by applying the inverse of the link function to give

$$\widehat{m}(x) = g^{-1}\{\widehat{\eta}(x)\}.$$

Local likelihood estimation was proposed by Tibshirani and Hastie (1987), and Fan, Farmen and Gijbels (1998) studied statistical inference based on local likelihood. Fan, Heckman and Wand (1995) showed that the local polynomial quasi-likelihood method inherits the good statistical properties of the local polynomial least-squares approach to smoothing.

# 3 SiZer Map Based on Local Quasi-likelihood

The SiZer map developed here is similar to that of Chaudhuri and Marron (1999). It is based on estimation of first order derivatives of the estimated function. Because the link function $g(\cdot)$ usually is a strictly monotone function, we may study the statistical significance of features of $\eta(x)$ instead of the mean function $m(x) = g^{-1}\{\eta(x)\}$. Our approach to the visual assessment of significance of features such as peaks and valleys in a family of smoothers $\{\widehat{\eta}_h(x) : h \in [h_{\min}, h_{\max}]\}$ is based on confidence limits for the derivative $\eta'(x)$ in scale space. The range of bandwidths will be discussed later on. Under certain regularity conditions, $\widehat{\eta}_h$ is a consistent estimator of $\eta(x)$. Asymptotic bias and variance of $\widehat{\eta}_h$ can be found in Fan, Heckman and Wand (1995). Asymptotic normality of $\widehat{\eta}_h$ has been established in Cai, Fan and Li (2000).

## 3.1 One-step local quasi-likelihood estimator

Repeated calculation of smoothers is required for color maps, such as SiZer map. Unlike the least-squares setting, the solution for (2.4) generally does not have a closed form. To obtain the solution, one usually uses an iterative algorithm, such as Newton-Raphson. The computational cost of the iterative method can be very expensive as one needs to maximize the local quasi-likelihood (2.3) for many distinct values of $x$ in order to obtain the function $\widehat{\eta}'(\cdot)$. To reduce the computational cost, we suggest replacing the iterative local quasi-likelihood estimator by an explicit non-iterative estimator. An excellent candidate is the one-step Newton-Raphson scheme, which has been frequently used in parametric models (see for example, Bickel 1975) and extended to the setting of quasi-likelihood recently by Fan and Chen (1999). Since the local quasi-likelihood method involves finding hundreds of parametric maximum likelihood estimates, the computational gain of

one-step local quasi-likelihood estimates is much more significant than that for parametric models. It can be shown that the one-step local quasi-likelihood estimate does not lose any statistical efficiency provided that the initial estimator is good enough (see Fan and Chen (1999) for the univariate case, and Cai, Fan and Li (2000) for the multivariate case).

Now let us describe the one-step local quasi-likelihood estimator. Let $Q'(\boldsymbol{\beta})$ and $Q''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local quasi-likelihood $Q(\boldsymbol{\beta})$. Given an initial estimator $\widehat{\boldsymbol{\beta}}_0(x_0) = (\widehat{\beta}_0(x_0), \widehat{\beta}_1(x_0))^T$, the Newton-Raphson algorithm finds an updated estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} = \widehat{\boldsymbol{\beta}}_0(x_0) - [Q''\{\widehat{\boldsymbol{\beta}}_0(x_0)\}]^{-1} Q'\{\widehat{\boldsymbol{\beta}}_0(x_0)\}. \tag{3.1}$$

This one-step estimator inherits the computational expediency of least-squares local polynomial fitting.

A good choice of initial value is critical to the good performance of one step iteration methods. In usual applications, where only one smooth is computed, the methods of Fan and Chen (1999) or Cai, Fan and Li (2000) are recommended. But for the computation of the full family of smooths that underlies SiZer, the special structure of the computation allows a very simple and effective approach. For the largest scale member of the family, $h = h_{max}$, the smooth will be close to a parametric model, so start with a parametric initial value is used here. Then iterate downward through the bandwidths, with each initial value taken to be the previous smooth. Figures 5 and 6 below show that this type of initialization is much more effective than classical ones, and indeed gives performance quite close to the very slow fully iterated version.

### 3.2 Numerical implementation of binned methods

The one-step local quasi-likelihood estimator can save computational cost by a factor of tens without diminishing the performance of the fully iterative local quasi-likelihood estimator. In the least-squares regression setting, binning methods can save computational time by a factor of hundreds. Binning can also be directly extended to the setting of local quasi-likelihood.

In this paper, we always use the binning approach described in Fan and Marron (1994). For the equally spaced grid of points $\{x_j : j = 1, \cdots, g\}$, the sample $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ is summarized by the binned data

$$\{(x_j, \bar{Y}_j, c_j) : \ j = 1, \cdots, g\},$$

where $\{\bar{Y}_j\}$ are the "bin averages", and $\{c_j\}$ are the "bin counts". That is,

$$\bar{Y}_j \equiv \text{avg}\{Y_i : i \in I_j\}, \quad \text{and} \quad c_j = \#(I_j),$$

where the $I_j$ are the index sets

$$I_j \equiv \{i : X_i \to x_j\}, j = 1, \cdots, g.$$

Now we apply the binning approach to the local quasi-likelihood. Suppose that the link function $g$ is a canonical link. Extension to other link functions does not involve any difficulty except some additional tedious notation. Then

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - g^{-1}\{\beta_0 + \beta_1(X_i - x_0)\}] K_h(X_i - x_0) \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix},$$

and

$$Q''(\boldsymbol{\beta}) = -\sum_{i=1}^{n} V[g^{-1}\{\beta_0 + \beta_1(X_i - x_0)\}] K_h(X_i - x_0) \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} (1, X_i - x_0),$$

where $V\{g^{-1}(u)\} = dg^{-1}(u)/du$, the variance function of $V(Y|X)$, which is always positive. Therefore, the Hessian matrix $Q''(\boldsymbol{\beta})$ is always nonpositive definite, and $Q(\boldsymbol{\beta})$ is a concave function with respect to $\boldsymbol{\beta}$.

Denote $\kappa_{l,j} = K_h(j\Delta)(j\Delta)^l$, where $\Delta$ is the bin width. For $l = 0$, 1, 2 set

$$\begin{aligned} U_l(x_{j'}) &= \sum_{j=1}^{g} \kappa_{l,j-j'} c_j \bar{Y}_j - \sum_{j=1}^{g} \kappa_{l,j-j'} g^{-1}\{\beta_0 + \beta_1(j - j')\Delta\} c_j, \\ h_l(x_{j'}) &= \sum_{j=1}^{g} \kappa_{l,j-j'} c_j V[g^{-1}\{\beta_0 + \beta_1(j - j')\Delta\}] \end{aligned}$$

Thus, $Q'(\boldsymbol{\beta})$ and $Q''(\boldsymbol{\beta})$ can be approximated by

$$\mathbf{U}(x_{j'}) = (U_0(x_{j'}), U_1(x_{j'}))^T, \quad \text{and} \quad \mathbf{H}(x_{j'}) = -\begin{pmatrix} h_0(x_{j'}) & h_1(x_{j'}) \\ h_1(x_{j'}) & h_2(x_{j'}) \end{pmatrix},$$

respectively. Then

$$\widehat{\boldsymbol{\beta}}_{\text{OS}}(x_{j'}) = \widehat{\boldsymbol{\beta}}_0(x_{j'}) - [\mathbf{H}(x_{j'})]^{-1}\mathbf{U}(x_{j'}).$$

A natural estimator of the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\text{OS}}$ is the corresponding sandwich formula. That is

$$\text{Cov}\{\widehat{\boldsymbol{\beta}}(x_{j'})\} = \mathbf{H}^{-1}(x_{j'})\mathbf{V}(x_{j'})\mathbf{H}^{-1}(x_{j'}),$$

where

$$\mathbf{V}(x_0) = \sum_{i=1}^{n} V[g^{-1}\{\beta_0 + \beta_1(X_i - x_0)\}] K_h^2(X_i - x_0) \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} (1, X_i - x_0).$$

Define $\tau_{l,j} = K_h^2(j\Delta)(j\Delta)^l V\{\beta_0 + \beta_1(j-j')\Delta\}$ for $l = 0, 1, 2$. Then the covariance matrix of the one-step estimator at grid point $x_{j'}$ is

$$\mathbf{V}(x_{j'}) = \begin{pmatrix} \tau_{0,j} & \tau_{1,j} \\ \tau_{1,j} & \tau_{2,j} \end{pmatrix},$$

Thus, we can calculate the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{OS}}$.

### 3.3 Local likelihood SiZer map

It is essential to the SiZer map to construct a good simultaneous confidence interval for the estimated derivative. Confidence limits for $\eta_h'(x)$ are of the form

$$\widehat{\eta}_h'(x) \pm q \cdot \widehat{sd}\{\widehat{\eta}_h'(x)\}, \tag{3.2}$$

where $q$ is an appropriate quantile, and the standard deviation is estimated as discussed in Section 3.2. An $(x, h)$ location (in scale space) is called significantly increasing, decreasing, or not significant, where zero is below, above or within these confidence limits respectively. The construction of simultaneous confidence intervals here is closely related to the problem of multiple comparisons in classical statistics. Chaudhuri and Marron (1999) proposed and compared several candidates for the quantile $q$. Intuitively, the pointwise Gaussian quantile $\Phi^{-1}(1-\alpha/2)$ is too small, meaning that the length of the confidence interval is too short, which leads to too many features being flagged as "significant". The calculation of the quantile $q$ based on bootstrap methods is time-consuming. Below is a brief description of the proposed method for constructing time-saving and appropriate confidence limits for $\eta_h'(x)$, whose behaviour is similar to that constructed via bootstrap. See Chaudhuri and Marron (1999) for details about variations on this approach.

In this paper we take $q$ as an approximately simultaneous over $x$ Gaussian quantiles, based on the "number of independent blocks" (Chaudhuri and Marron, 1999). The quantile $q$ is based on the fact that when $x$ and $x'$ are sufficiently far apart, the kernel windows centered at $x$ and $x'$ are essentially independent, but when $x$ and $x'$ are close together, the estimates are highly correlated. The simultaneous confidence limit problem is then approximated by $m$ independent confidence interval problems, where $m$ reflects the "number of independent blocks". We calculate $m$ through an "estimated effective sample size", defined for each $(x, h)$ as

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

Note that when $K(\cdot)$ is a uniform kernel, $\mathrm{ESS}(x, h)$ is the number of data points in the kernel window centered at $x$. For other kernel shapes, points are downweighted according to the height

of the kernel function, just as they are in the average represented by kernel estimators. Next we choose $m$ to be essentially the number of "independent blocks of average size available from our data set of size $n$"

$$m(h) = \frac{n}{\text{avg}_x \text{ESS}(x,h)}.$$

Now assuming independence of these $m(h)$ blocks of data the approximate simultaneous quantile is

$$q_\alpha(h) = \Phi^{-1}\left\{\frac{1+(1-\alpha)^{1/m(h)}}{2}\right\}. \tag{3.3}$$

The quantity ESS is also useful to highlight regions where the normal approximation implicit in (3.2) could be inadequate. This plays a role similar to $np$ in the Gaussian approximation to the binomial. Regions where $\text{ESS}(x,h)$ is less than some constant $n_0$ (we have followed the standard practice of $n_0 = 5$ at all points here) are shaded gray in the SiZer map, to rule out spurious features and to indicate regions where there is an inadequate amount of data. The above calculation of the block size $m(h)$ can be modified to avoid problems with small ESS as

$$m(h) = \frac{n}{\text{avg}_{x \in D_h} \text{ESS}(x,h)},$$

where $D_h$ is the set of $x$ locations where the data are dense, i.e. $D_h = \{x : \text{ESS}(x,h) \geq n_0\}$.

Bandwidth selection is not an important issue for the SiZer map, since it is based on the idea of family smoothing. However, a reference bandwidth may be helpful in interpreting the map. Fan and Chen (1999) have shown that optimal bandwidths for local least square smoothing have a simple relationship to optimal bandwidths for local quasi-likelihood smoothing. They suggested that an estimated optimal bandwidth for the least-squares local polynomial estimator be used with some modification as the bandwidth for the local one-step quasi-likelihood estimator. Thus the Ruppert-Sheather-Wand direct plug in bandwidth is taken as a reference bandwidth, highlighted as a horizontal bar in the SiZer map.

The bandwidth range $[h_{\min}, h_{\max}]$ can be chosen in several ways. We suggest a "wide range" approach, where $h_{\min}$ is taken to be the smallest bandwidth for which there is no substantial distortion in construction of the binned implementation of the smoother, $h_{\min} = 5 * (\text{binwidth})$, and $h_{\max}$ to be the range of the data. The choice of $h_{\min}$ is larger than that used by Chaudhuri and Marron (1999), but is recommended here because smaller values sometimes gave convergence difficulties due to kernel function discretization errors.

# 4 Simulation and Application

In this section, the proposed SiZer map is illustrated with both simulation examples and a real data example. It will be shown that the quickly computed SiZer map based on the proposed one-step estimator behaves as well as the one based on the much slower maximum local quasi-likelihood estimate with full iterations.

## 4.1 Poisson regression

For a Poisson regression model, the conditional distribution of $Y$ given $X$ is the Poisson mean function $\lambda(x)$. The canonical link for Poisson regression is the log-link. With the canonical link, the local (conditional) likelihood, based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, is

$$Q\{\boldsymbol{\beta}(x_0)\} = \sum_{i=1}^{n} \{Y_i \mathbf{X}_i^T \boldsymbol{\beta}(x_0) - \exp(\mathbf{X}_i^T \boldsymbol{\beta}(x_0)\} K_h(X_i - x_0),$$

where $\mathbf{X}_i = (1, X_i - x_0)^T$ and $\boldsymbol{\beta}(x_0) = (\beta_0(x_0), \beta_1(x_0))^T$. Therefore

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i \{Y_i - \exp(\mathbf{X}_i^T \boldsymbol{\beta})\} K_h(X_i - x_0),$$

and

$$Q''(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \exp(\mathbf{X}_i^T \boldsymbol{\beta}) K_h(X_i - x_0) \mathbf{X}_i \mathbf{X}_i^T.$$

Thus the one-step estimator for $\boldsymbol{\beta}(x_0)$ is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} = \widehat{\boldsymbol{\beta}}_0 - \left[Q''(\widehat{\boldsymbol{\beta}}_0)\right]^{-1} Q'(\widehat{\boldsymbol{\beta}}_0)$$

and the corresponding estimated covariance matrix for $\widehat{\boldsymbol{\beta}}_{\mathrm{OS}}$ is

$$\mathrm{Cov}\{\widehat{\boldsymbol{\beta}}_{\mathrm{OS}}\} = \left[Q''(\widehat{\boldsymbol{\beta}})\right]^{-1} \mathbf{V}(\widehat{\boldsymbol{\beta}}) \left[Q''(\widehat{\boldsymbol{\beta}})\right]^{-1},$$

where

$$\mathbf{V}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \exp(\mathbf{X}_i^T \boldsymbol{\beta}) K_h^2(X_i - x_0) \mathbf{X}_i \mathbf{X}_i^T.$$

**Example 2**. In this example, the covariate $X$ values are taken as equally spaced on [0, $10\pi$], and the conditional distribution of $Y$ given $X$ is a Poisson distribution with mean function

$$\lambda(x) = \exp\{\cos(x)\}.$$

Figure 4 shows a SiZer map for a sample size of $n = 500$, which shows SiZer behavior when the data are very informative about the structure of $\eta(x)$, with in particular all of the peaks and valleys

of the cosine wave clearly visible. Figure 5 shows a SiZer map for $n = 200$, in which case the data are much less informative. Therefore, only some of the structure of the cos wave is flagged as statistically significant. Comparison of the full iteration SiZer maps and the one-step SiZer maps using the least squares starting values in Figure 4 and and 5 shows that, in both situations, the one-step SiZer maps using the least squares starting values result in an inadequate representation of the full iteration versions (the red and blue shaded areas are smaller in the middle rows). Figures 4(c) and 5(c) show that our proposed one-step SiZer gives a good quality representation of the fully iterated SiZer, with essentially the same red and blue shaded regions. As a comparison, the original SiZer maps using critical value $q$ in (3.3), are depicted in Figures 4(d) and 5(d). They are similar to the one-step using the least squares starting values, and is less efficient in distinguishing important features than the full iteration version SiZer map and our proposed one-step SiZer map.

To compare computation times of the fully iterated SiZer, the one-step SiZer using the least squares starting values, our proposed one-step SiZer and the original SiZer map, we conducted 100 Monte Carlo simulations with $n = 200$ and 500 on SUN Ultra 5 workstation with 400 MHz. The average and standard deviation of computation time (in second) for each simulation are displayed in Table 1, in which "Classic" stands for the one-step SiZer using the least squares starting values, "New" for our proposed one-step SiZer, and "LS" for the original SiZer map using critical value $q$ given in (3.3). From Table 1, both one-step SiZer maps dramatically reduce computation time relative to the fully iterated SiZer, and our proposed one-step SiZer needs less computation time than the one with the least squares starting values. The computation times do not significantly vary with the sample size because of the binned method is used in the numerical implementation. As expected, the original SiZer map with critical value $q$ given in (3.3) needs least computational time because it only needs to compute hundreds of least squares estimates rather than maximum likelihood estimates using iterative algorithm. However, it is less efficient in distinguishing important features than our proposed one-step SiZer map.

Table 1: Computation Time (seconds)

| Model | Sample Size | Full Iteration | Classic | New | LS |
|---|---|---|---|---|---|
| Poisson | 200 | 91.423 (1.103) | 37.223 (0.432) | 29.472 (0.580) | 11.183 (0.357) |
| Poisson | 500 | 87.095 (1.438) | 37.790 (0.608) | 29.979 (0.589) | 12.785 (0.553) |
| Logistic | 500 | 94.196 (2.298) | 40.919 (0.799) | 34.678 (0.693) | 12.760 (0.598) |
| Logistic | 1000 | 88.062 (1.927) | 41.840 (0.533) | 34.373 (0.693) | 13.549 (0.493) |

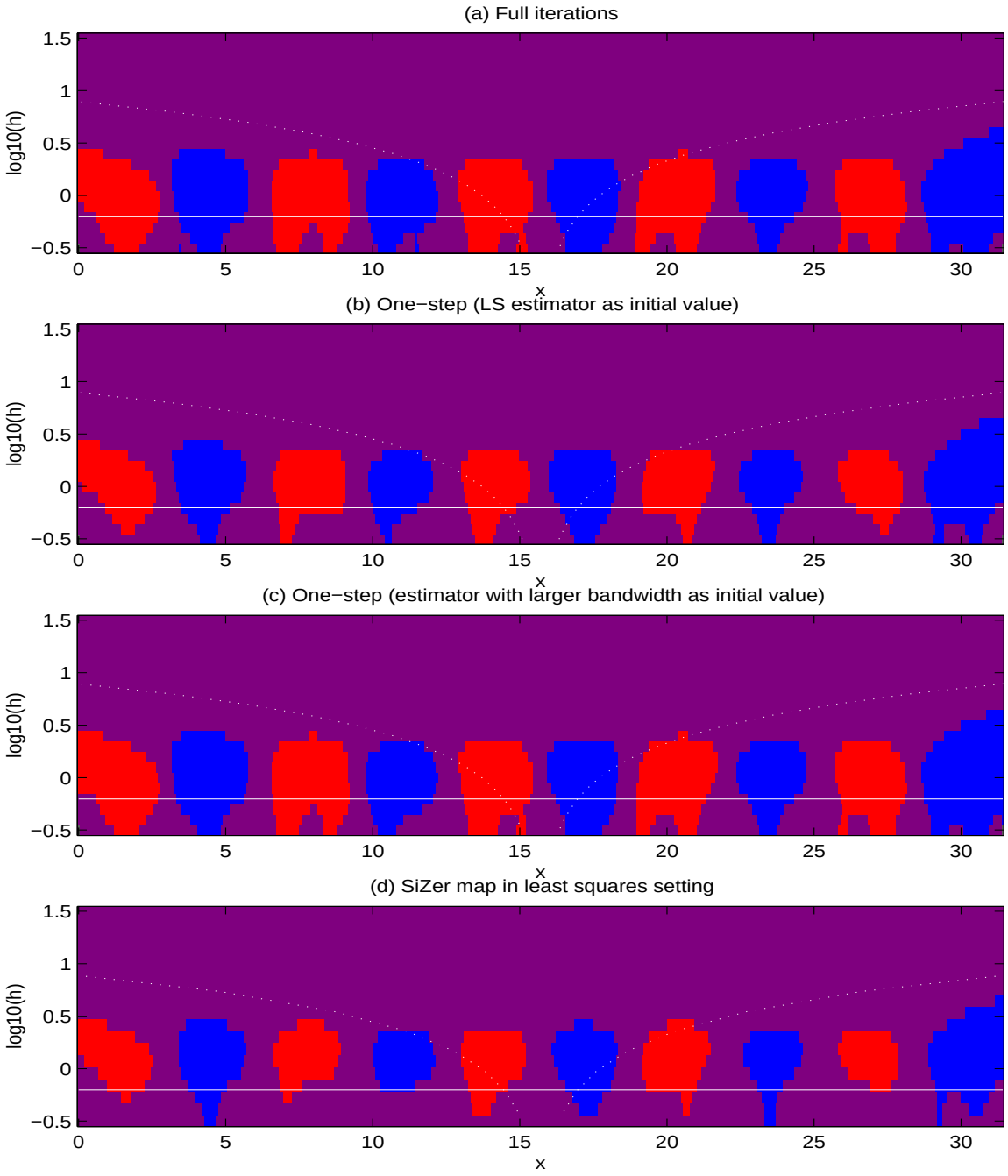


Figure 4: *SiZer maps for Poisson regression with sample size* $n = 500$. *(a) is the SiZer map based on a maximum local quasi-likelihood with full iterations, (b) is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan and Chen (1999) as initial values, (c) is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 3.1, and (d) is the original SiZer map proposed by Chaudhuri and Marron (1999).*

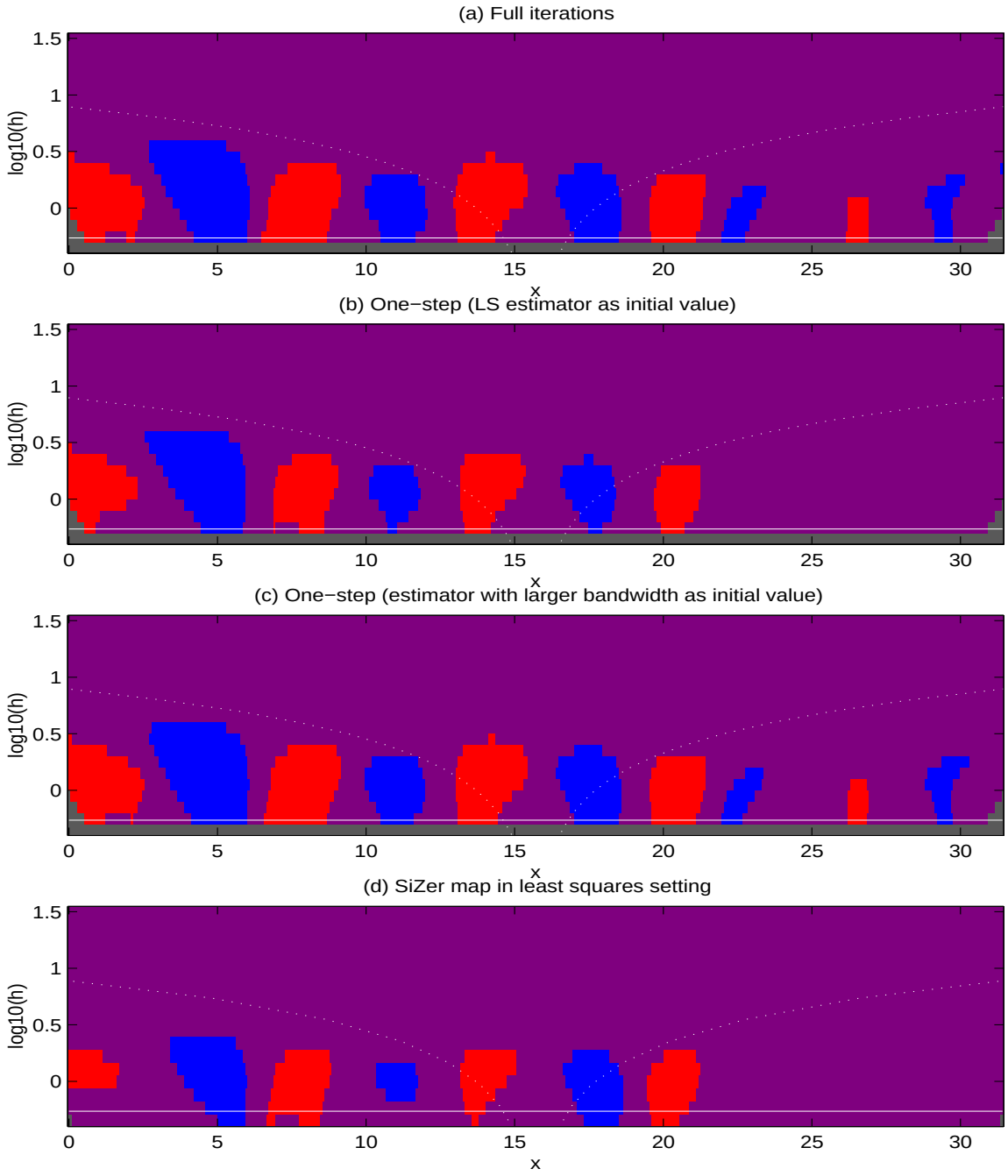


Figure 5: *SiZer maps for Poisson regression with sample size* $n = 200$. *(a) is the SiZer map based on a maximum local quasi-likelihood with full iterations, (b) is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan and Chen (1999) as initial values, (c) is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 3.1, and (d) is the original SiZer map proposed by Chaudhuri and Marron (1999).*

### 4.2 Logistic regression

For a Bernoulli distribution, the mean function is the probability function $p(x) = P(Y = 1|X = x)$, the variance function is $p(x)\{1-p(x)\}$ and the canonical link is logit, i.e. $\text{logit}\{p(x)\} = \log[p(x)/\{1-p(x)\}]$. Denote by $Q(\boldsymbol{\beta})$ the local likelihood based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, then

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i \left\{ Y_i - \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})} \right\} K_h(X_i - x_0),$$

and

$$Q''(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{\{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})\}^2} K_h(X_i - x_0)\mathbf{X}_i\mathbf{X}_i^T.$$

Thus the one-step estimator for $\boldsymbol{\beta}(x_0)$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{OS}} = \widehat{\boldsymbol{\beta}}_0 - \left[Q''(\widehat{\boldsymbol{\beta}}_0)\right]^{-1} Q'(\widehat{\boldsymbol{\beta}}_0)$$

and the corresponding estimated covariance matrix for $\widehat{\boldsymbol{\beta}}_{\text{OS}}$ is

$$\text{Cov}\{\widehat{\boldsymbol{\beta}}_{\text{OS}}\} = \left[Q''(\widehat{\boldsymbol{\beta}}_{OS})\right]^{-1} \mathbf{V}(\widehat{\boldsymbol{\beta}}_{OS}) \left[Q''(\widehat{\boldsymbol{\beta}}_{OS})\right]^{-1},$$

where

$$\mathbf{V}\{\boldsymbol{\beta}\} = \sum_{i=1}^{n} \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{\{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})\}^2} K_h^2(X_i - x_0)\mathbf{X}_i\mathbf{X}_i^T.$$

**Example 3**. In this example, the covariate $X$ values are taken as equally spaced on $[0, 10\pi]$, and the conditional distribution of $Y$ given $X$ is Bernoulli with probability function $p(x)$, where

$$\text{logit}\{p(x)\} = \cos(x).$$

Figure 6 shows the SiZer maps with the sample size $n = 500$, and indicates that the one-step SiZer map using the least squares starting values yields an inadequate representation of the full iteration version shown in Figure 6 (a). Furthermore, our proposed one-step SiZer gives a good quality representation of the fully iterated SiZer, with essentially the same red and blue shaded areas. From Figure 6, the full iteration SiZer and our proposed one-step SiZer are more efficient in distinguishing important features than the original SiZer using critical value $q$ in (3.3).

The averages and standard deviations of computation time for each simulation over 100 simulations are listed in Table 1. Our proposed one-step SiZer requires less computation time than the other SiZer maps based on local likelihood. Although the original SiZer need less computational time than our proposed one-step SiZer, but it is less efficient.

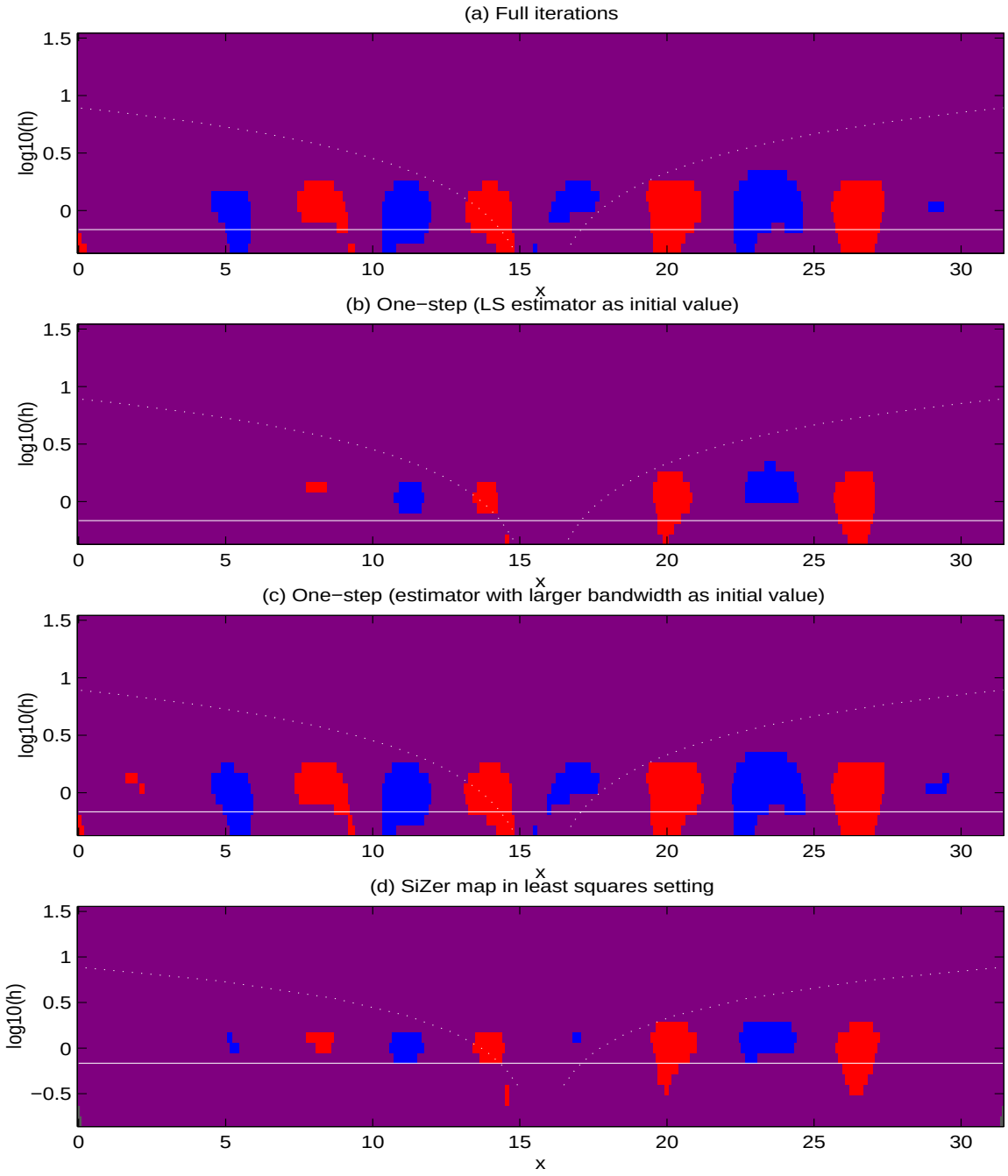


Figure 6: *SiZer maps for logistic regression with sample size* $n = 500$. *(a) is the SiZer map based on a maximum local quasi-likelihood with full iterations, (b) is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan and Chen (1999) as initial values, (c) is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 3.1, and (d) is the original SiZer map proposed by Chaudhuri and Marron (1999).*

### 4.3 Application

The practical usefulness of the proposed SiZer map is illustrated by the following analysis of an environmental data set. This data set was collected from January 1, 1994 to December 31, 1995 and consists of daily measurements of air pollutants along with the daily number of hospital admissions for circulatory or respiratory problems.

**Seasonal Trends of Hospital Admissions**

We first demonstrate how to use the local likelihood SiZer map to analyze seasonal trend of the number of hospital admissions. The number of hospital admissions is taken as response variable, and the day (or date on which data were collected) as a covariate. Since the response is a variable of count data, we consider Poisson regression with mean function:

$$E(Y|\text{day}) = \lambda(\text{day}).$$

We further apply the local likelihood SiZer map of Poisson regression to the data collected in 1994 and 1995 separately. The resulting SiZer maps are depicted in the bottom panel of Figures 7 and 8. The top row in Figures 7 and 8 is an extension of the family smoothing plot (Chaudhuri and Marron, 1999) to Poisson regression, in which a thick red curve is the estimated regression function using the reference bandwidth. From Figure 7, we can see that overall seasonal trend for Year 1994 is increasing, while the trend for Year 1995 is decreasing from Figure 8. Both Figures 7 and 8 indicate that there exists a peak in March. This implies that seasonal change from winter to spring may be a risk factor for circulatory and respiratory problems. Furthermore, we can see from Figures 7 and 8, there is an overall increasing trend from Day 1 to Day 75 and from Day 250 to the end of the years. This indicates that people are more likely to have circulatory and respiratory problems during the winter. Some small wiggles in Figure 9 and 10 suggest that a model taking into account of other risk factors, such as air pollutants may be helpful.

**The Effects of Air Pollutants**

It is believed that air pollutants may cause circulatory and respiratory problems. As an illustration, we take the number of hospital admissions as a response variable $Y$, and either the level of Nitrogen Dioxide $NO_2$ or the level of *dust* as a covariate. Again, we consider a Poisson regression model with mean

$$E(Y|X = x) = \lambda(x),$$

where $x$ is either the level of $NO_2$ or the level of *dust*. Figure 9 and 10 depict the family plots and the local likelihood SiZer maps for $NO_2$ and *dust*, respectively. From Figure 9 and 10, we can

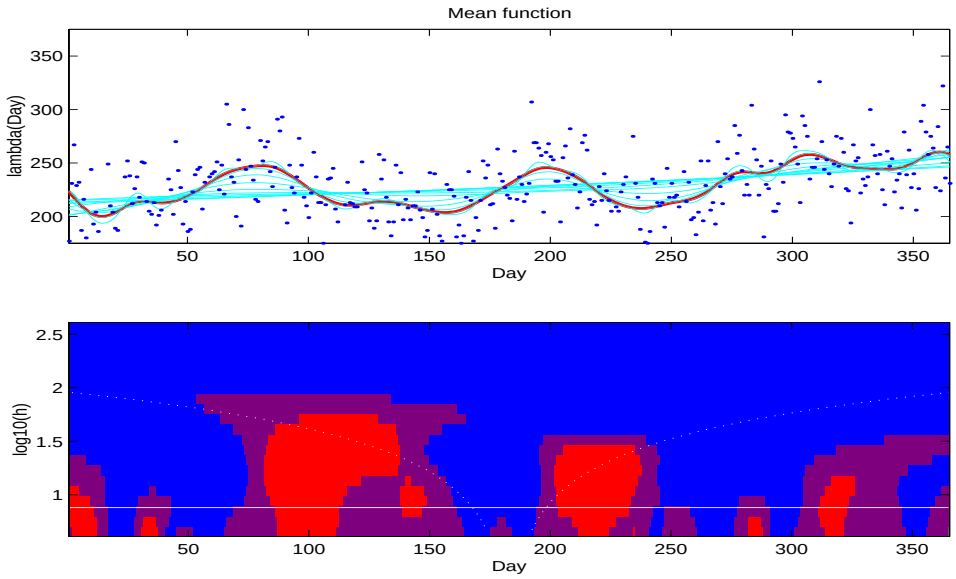


Figure 7: *SiZer map for the regression function of the number of daily hospital admissions during Year 1994. The top panel is the plot of the family smoothing (see Chaudhuri and Marron (1999) for details). The bottom panel depicts the SiZer map for the Poisson regression function.*

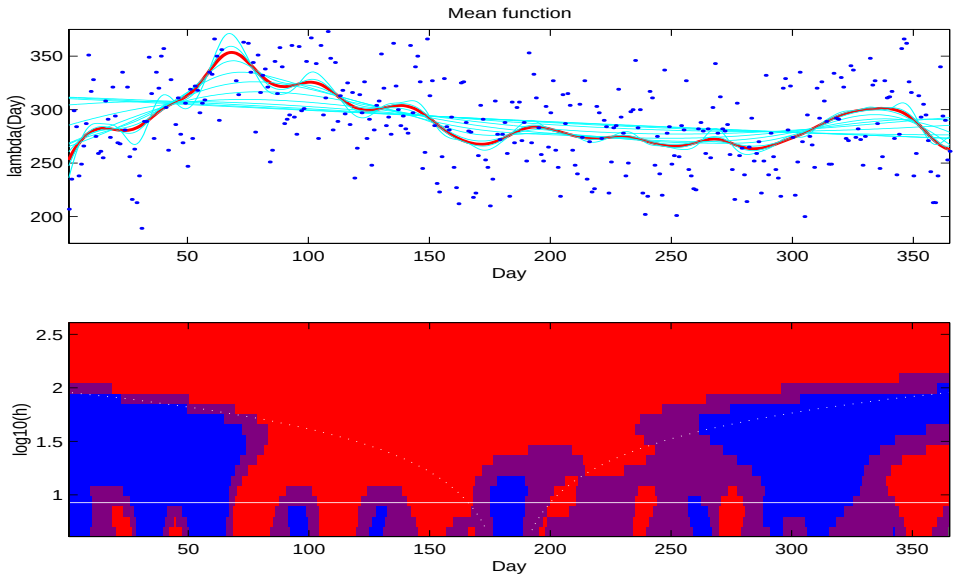


Figure 8: *SiZer map for the regression function of the number of daily hospital admissions during Year 1995. Caption is the same as Figure 7.*

see that the number of hospital admissions increases as the level of $NO_2$ and/or dust increases. Furthermore, both Figure 9 and 10 present some small wiggles when the smoothing parameters are small. This suggests that one should consider more complicated models, such as a generalized

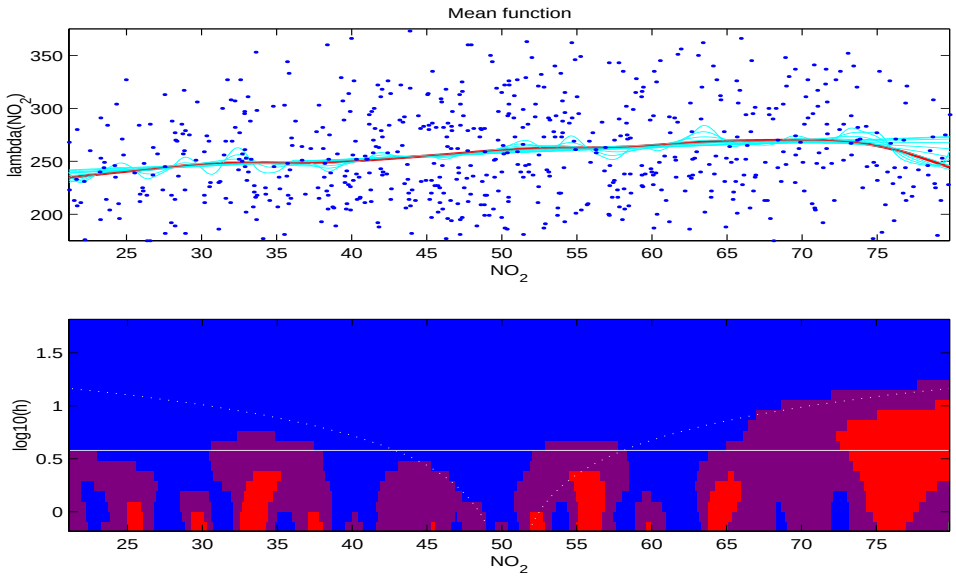


Figure 9: *SiZer map for the regression function of the number of daily hospital admissions on the level of $NO_2$. Caption is the same as Figure 7.*

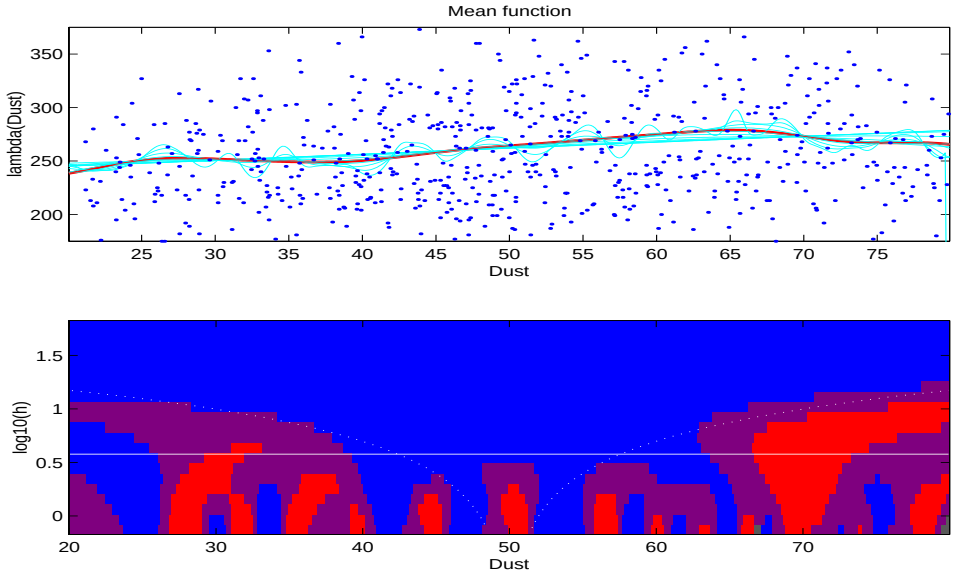


Figure 10: *SiZer map for the regression function of the number of the daily hospital admissions on the level of* dust. *Caption is the same as Figure 7.*

varying-coefficient model (Cai, Fan and Li, 2000):

$$\log\{E(Y|U,\mathbf{x})\} = \beta_0(U) + \beta_1(U)X_1 + \cdots + \beta_p(U)X_p,$$

where $U$ is the day, $X_1, \cdots, X_p$ are the levels of air pollutants. Further research is needed to develop

the SiZer map for such a model.

## 5 Discussion

In this paper, we propose the local likelihood SiZer map, and address computational issues for such a map. We demonstrate that the local likelihood SiZer map is more efficient than direct application of the original SiZer map, when a quasi-likelihood function is available. It may be useful to extend the SiZer map for generalized varying coefficient model (Cai, Fan and Li, 2000), in which one may have to use different bandwidths for different coefficients since different coefficients may have different smoothness. Further research is needed.

As a referee pointed out, there are several versions of SiZer maps. The main difference among different versions of SiZer map is the critical value $q$ obtained by different approaches. In this paper, we only compared our local likelihood SiZer map with one of them, using the same critical value $q$ given in (3.3). it is expected that, using the same approach to constructing the critical value for simultaneous confidence intervals for both local likelihood SiZer map and the original SiZer map, local likelihood SiZer map performs better when a quasi-likelihood function is available because quasi-likelihood estimate is more efficient than least squares estimate in the presence of heteroscedastic errors, such as Poisson regression and logistic regression.

**Acknowledgement**: This work was part of the Ph.D. dissertation of the first author, under the supervision of the second. Li's research was supported by a NSF grant DMS-0102505, and Marron's research was partially supported by a NSF Grant DMS-9971649.

## References

Bickel, P.J. (1975). One-step Huber estimates in linear models. *J. Amer. Statist. Assoc.*, **70**, 428-433.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.*. **95**, 888–902.

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.*, **94**, 807–823.

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation, *J. Royal Statist. Soc. B*, **61**, 927-943.

Fan, J., Farmen, M. and Gijbels, I. (1998). Local Maximum Likelihood Estimation and Inference, *J. Royal Statist. Soc. B*, **60**, 591-608.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *I. Royal. Statist. Soc. B*, **57**, 371-394.

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

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.*, **90**, 141–150.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators, *J. Comput. and Graph. Statist.*, **3**, 35-56.

Gasser, T., Kneip, A. and Köhler, W., (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.*, **86**, 643-652.

Ghosh, A. K., Chaudhuri, P., and Murthy, C.A. (2004). On visualization and aggregation of nearest neighbor. Manuscript.

Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2004). Classification using kernel density estimates: multi-scale analysis and visualization. *Technometrics*. Tentatively accepted.

Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *J. Comput. Graph. Statist.*, **11**, 1-22.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Boston.

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

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

Rakesh, R. R., Chaudhuri, P. and Murthy, C.A. (2004). Thresholding in edge detection: a statistical approach. *IEEE Transaction on Image Processing*. To appear.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.,* **90**, 1257-1270.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Statist. Soc. B*, **53**, 683-690.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics,* Springer, New York.

ter Haar Romeny, B. M. (2001). *Front-End Vision and Multiscale Image Analysis*, Kluwer, Dordrecht.

Tibshirani, R. and Hastie, T.J. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.*, **82**, 559–567.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method, *Biometrika*, **61**, 439-447.