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

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

Kernel smoothing methods for curve estimation are useful tools for data analysis. There is a huge literature on this subject. Good introduction comes from monographs, including Härdle (1990), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Bowman and Azzalini (1997) and Hart (1997). Consider the bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$, an i.i.d. sample from the model

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

where ε is random error with $E(\varepsilon|X) = 0$, $E(\varepsilon^2|X) = \sigma^2(X)$. Local polynomial regression can be employed to estimate the regression function. The choice of bandwidth h is of crucial importance in nonparametric kernel smoothing 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 survey of bandwidth selection procedures in kernel density estimation can be found in Jones, Marron and Sheather (1996a, b). 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 (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.

Chaudhuri and Marron (1999) proposed the idea of family smoothing, combined with statistical inference using the SiZer map, from the scale-space point of view. Scale-space ideas from computer vision, see, e.g., Lindeberg (1994) and ter Haar Romeny (2001), provide a different view-point on kernel smoothing. The "scale space surface," the family of all kernel smooths indexed by the bandwidth h, is a model used in computer vision. The essential idea is that large h models macroscopic (distant) vision where only large-scale features can be resolved, and small h models microscopic (zoomed in) resolution of small-scale features. This is very different from the classical statistical approach, where the focus is the underlying regression function m. See Chaudhuri and Marron (1999) for detailed discussion.

Chaudhuri and Marron (1999) have developed the SiZer map in least squares regression settings from the point of view of scale space. The SiZer map is powerful for determining which features in noisy data are strong enough to be distinguished from background noise. However this approach may be inefficient for discrete data when the likelihood function or the quasi-likelihood function of data can provide the basis for more powerful inference. This is illustrated in Figures 1 and 2.

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 indicates locations where the mean function is significantly increasing, red shows where it is significantly decreasing, and purple indicates where the 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.

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 $n = 500$. Figure 1 shows that the higher the mean is, the higher the variance is. Figures 2 and 3 compare the SiZer map proposed by Chaudhuri and Marron (1999) and the improved SiZer map, proposed in this paper, based on local likelihood for sample sizes 500 and 200, respectively. 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.

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. 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 function defined below for many distinct values of x in order to estimate the whole curve over all of the scale. The solution of such computational problem is a major contribution of this paper. 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 SiZer map based on local likelihood method. The SiZer maps for nonparametric Poisson regression and logistic regression are illustrated

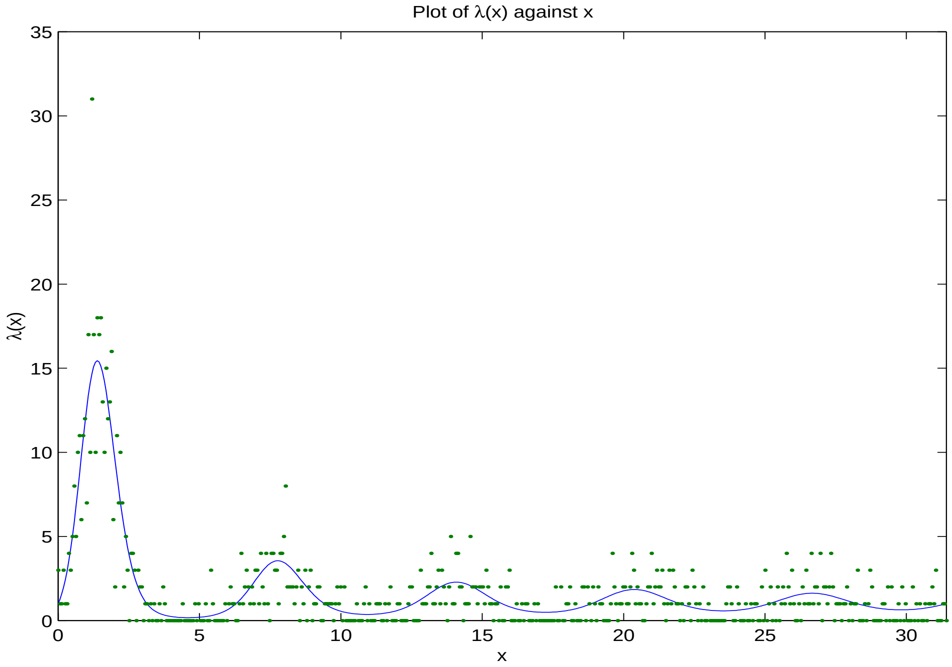


Figure 1: *True underlying signal in Example 1. Solid line stands for the mean function, and the dots are a realization of raw data with* $n = 500$.

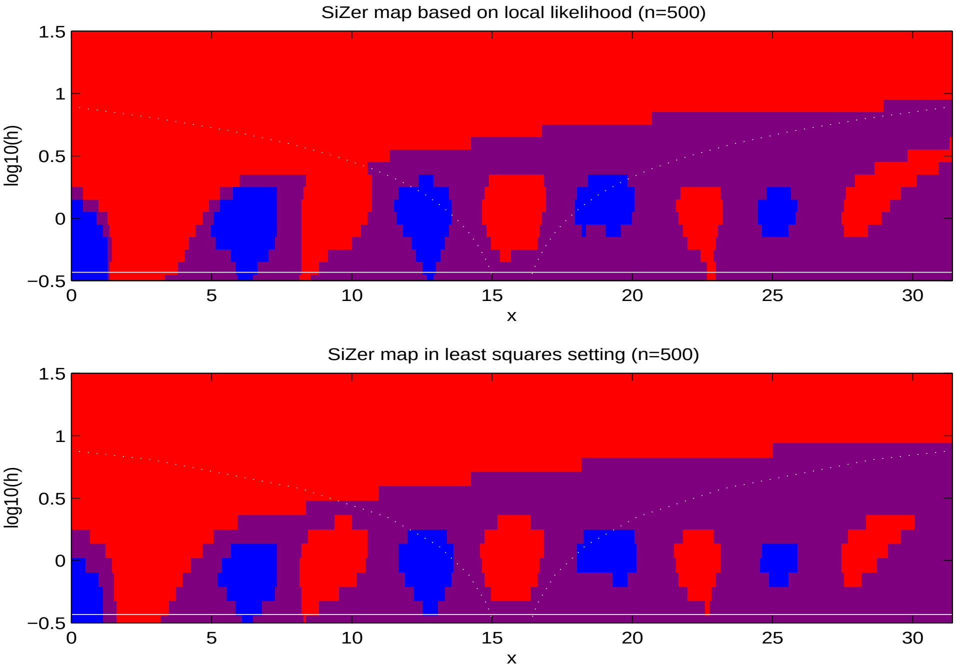


Figure 2: *SiZer maps for sample size* $n = 500$. *The top panel is the local likelihood SiZer map, and the bottom panel is the original SiZer map proposed by Chaudhuri and Marron (1999).*

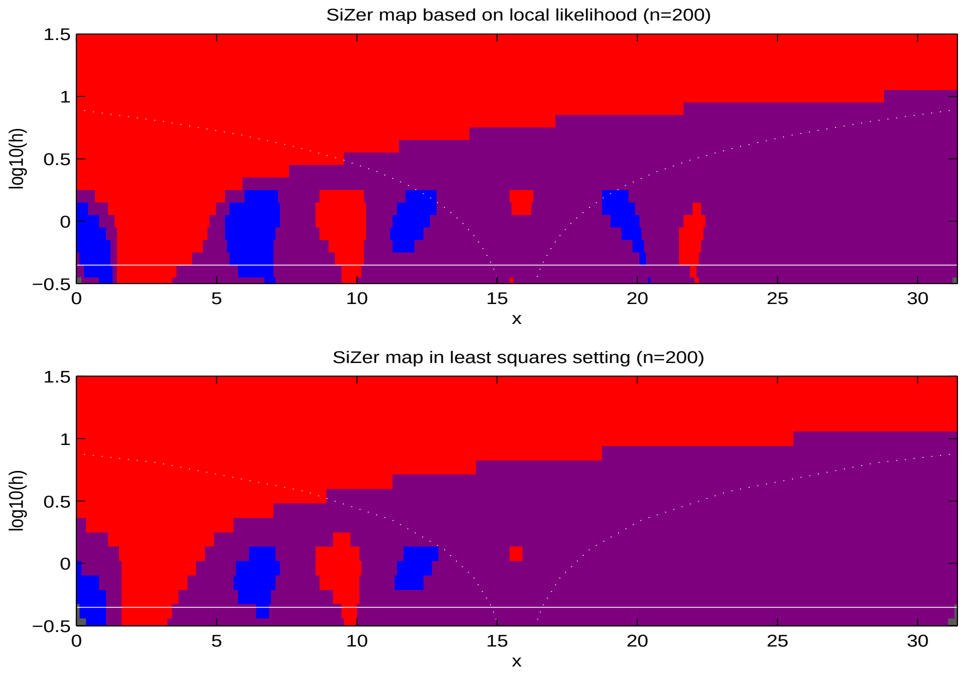


Figure 3: *SiZer maps for sample size* $n = 200$*. The top panel is the local likelihood SiZer map, and the bottom panel is the original SiZer map proposed by Chaudhuri and Marron (1999).*

via simulated examples and real data example in Section 4. Some discussions are given in Section 5.

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 iid 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. 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. 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 $\mathrm{Var}(Y|\mathbf{X} = \mathbf{x}) = a(\phi)b''\{\theta(\mathbf{x})\}$. Since our primary interest is to estimate the mean function, without loss of generality, the factors related to the dispersion parameter ϕ are omitted. This leads to the following conditional log-likelihood function

$$\ell\{\theta, y\} = \theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}.$$

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 $Q\{m(\mathbf{x}), y\}$. If the conditional variance is modeled as $\mathrm{Var}(Y|\mathbf{X} = \mathbf{x}) = V\{m(\mathbf{x})\}$ for some known positive function V, then the corresponding quasi-likelihood $Q(\mu, y)$ satisfies

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

(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 $V = b'' \circ (b')^{-1}$, 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. Tibshirani and Hastie (1987) proposed local likelihood estimation, 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. 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 β_0 and β_1 depend on x_0. Based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, the local quasi-likelihood is

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

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

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

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

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

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

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

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 in scale space $\eta_h'(x)$. The range of bandwidths will be discussed later on.

3.1 One-step local quasi-likelihood estimator

Repeated calculation of smoothers is required for such color maps. 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 computation expediency of least-squares local polynomial fitting.

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 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 prametric model, so 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, binned methods can save computational time by a factor of hundreds. The idea discussed by Fan and Marron (1994) can be directly extended to the setting of local quasi-likelihood. As aforementioned, this paper only considers 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

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

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

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

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

Using the "bin averages" $\{\bar{Y}_j\}$ and the "bin counts" $\{c_j\}$ given by

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

where the I_j are the index sets

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

Denote $\kappa_{l,j} = K_h(j\Delta)(j\Delta)^l$, where Δ is the binwidth and for $l = 0$, 1, 2 set

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

and

$$\begin{aligned} \mathbf{U}(x_{j'}) &= (U_1(x_{j'}), U_2(x_{j'}))^T \\ \mathbf{H}(x_{j'}) &= \begin{pmatrix} h_0(x_{j'}) & h_1(x_{j'}) \\ h_1(x_{j'}) & h_2(x_{j'}) \end{pmatrix}. \end{aligned}$$

Then

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

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

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

where

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

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. The construction of simultaneous confidence intervals relates to the topic 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 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. Here is a brief description how to construct 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.

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. In this paper we take q as an approximately simultaneous over x Gaussian quantiles, based on the "number of independent blocks". 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 calculated 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"

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

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

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

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. So regions where $\mathrm{ESS}(x,h) < n_0$ (we have followed the standard practice of $n_0 = 5$ at all points here) are shaded gray in our SiZer map, to rule out spurious features, and also to indicate regions where the smooth is essentially based on an inadequate amount of data. The above calculation of the block size $m(h)$ is modified to avoid problems with small ESS as

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

where D_h is the set of x locations where the data are dense

$$D_h = \{x : \mathrm{ESS}(x,h) \geq n_0\}.$$

Bandwidth selection is not an important issue for the SiZer map because it is based on the idea of family smoothing. However, a reference bandwidth may help one to interpret 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 use of an estimated optimal bandwidth for the least-squares local polynomial estimator 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 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 use 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 Poisson mean function $\lambda(x)$. The canonical link for Poisson regression is 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 4.1. In this example, the covariate X values are taken as equally spaced on [0, 10π], and the conditional distribution of Y given X is a Poisson distribution with mean function

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

Figures 4 and 5 show SiZer maps which illustrate issues on several levels. First Figure 4 is for a sample size of $n = 500$, which shows SiZer behavior when the data are very informative about the structure of $\boldsymbol{\beta}(x)$, with in particular all of the peaks and valleys of the cos wave clearly visible. Figure 5 is for the smaller sample size of $n = 200$, which shows the situation where the data are much less informative, with the result that only some of the structure of the cos wave is flagged as statistically significant. Second, the middle rows of both Figures 4 and 5 show 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 shown in the top rows, because the red and blue shaded areas are less in the middle rows. Finally, the bottom rows 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.

To compare computation times of the fully iterated SiZer, the one-step SiZer using the least squares starting values, and our proposed one-step SiZer, 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, and “New” for our proposed one-step SiZer. 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.

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. $\text{logit}\{p(x)\} = p(x)/\{1 - p(x)\}$. Denote by $Q(\boldsymbol{\beta})$ the local likelihood based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, then

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

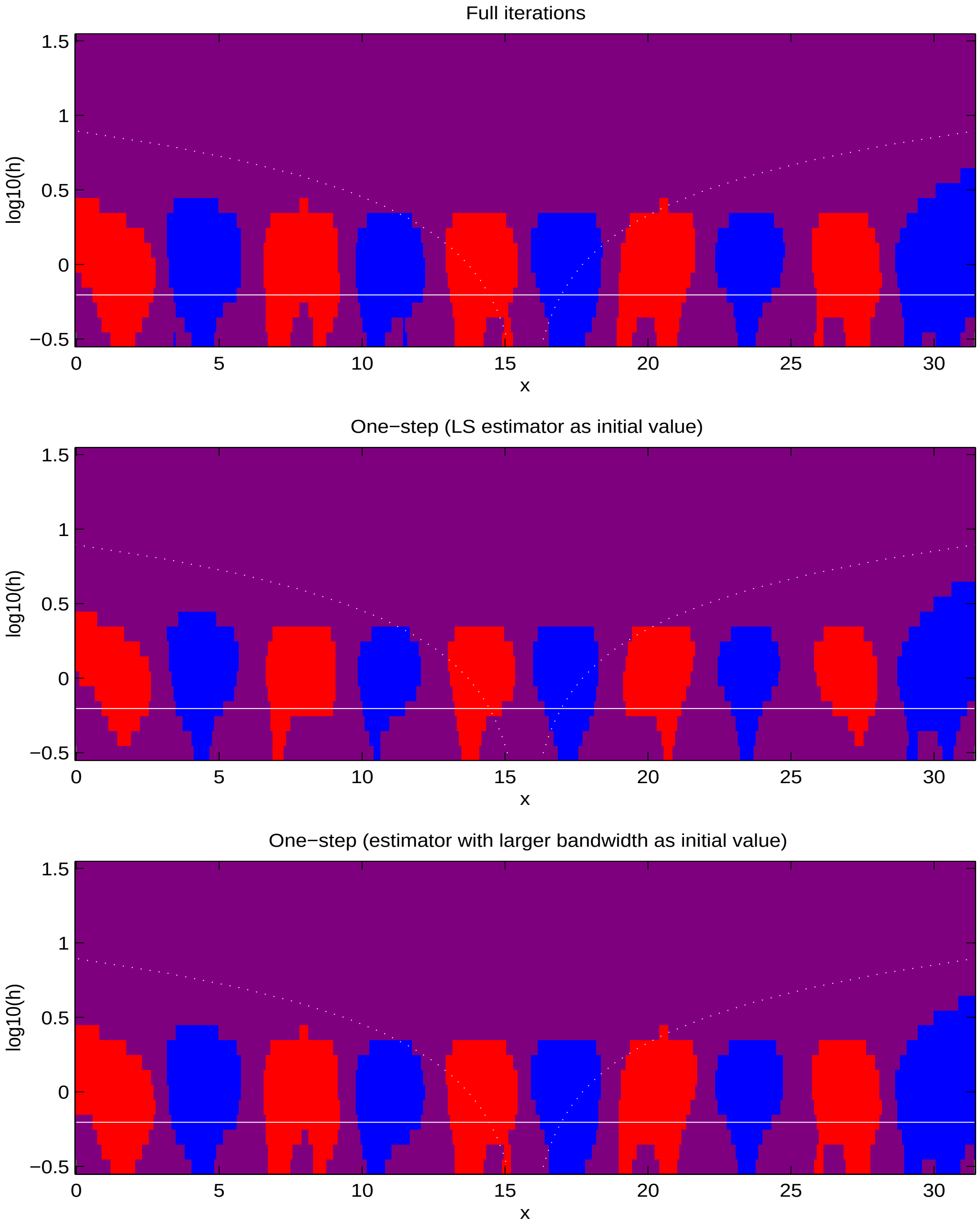
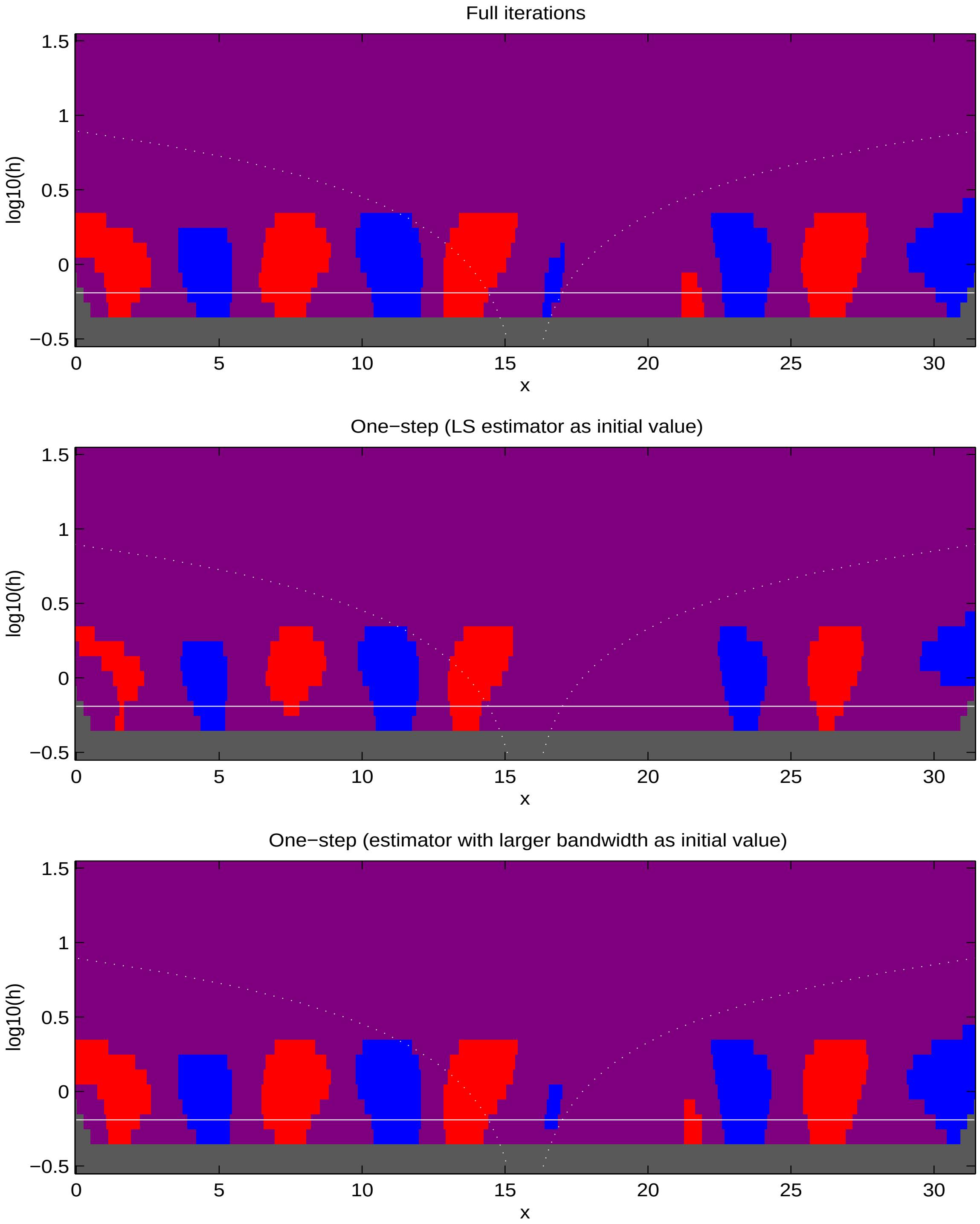


Figure 4: *SiZer maps for Poisson regression with sample size* $n = 500$*. The top panel is the SiZer map based on a maximum local quasi-likelihood with full iterations, the middle panel is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan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



Full iterations

One−step (LS estimator as initial value)

One−step (estimator with larger bandwidth as initial value)

Figure 5: *SiZer maps for Poisson regression with sample size* $n = 200$. *The top panel is the SiZer map based on a maximum local quasi-likelihood with full iterations, the middle panel 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, and the bottom panel is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 3.1.*

Table 1: Computation Time (seconds)

Model	Sample Size	Full Iteration	Classic	New
Poisson	200	92.9362 (1.1298)	37.5611 (0.4982)	29.9463 (0.4598)
Poisson	500	89.4955 (1.3204)	38.9320 (0.8629)	30.9033 (0.7448)
Logistic	500	98.4047 (2.3083)	43.3013 (0.7939)	36.4970 (0.8044)
Logistic	1000	91.6759 (1.8998)	44.0240(0.6615)	36.2111 (0.0001)

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

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

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 4.2. 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 Bernoulli with probability function $p(x)$, where

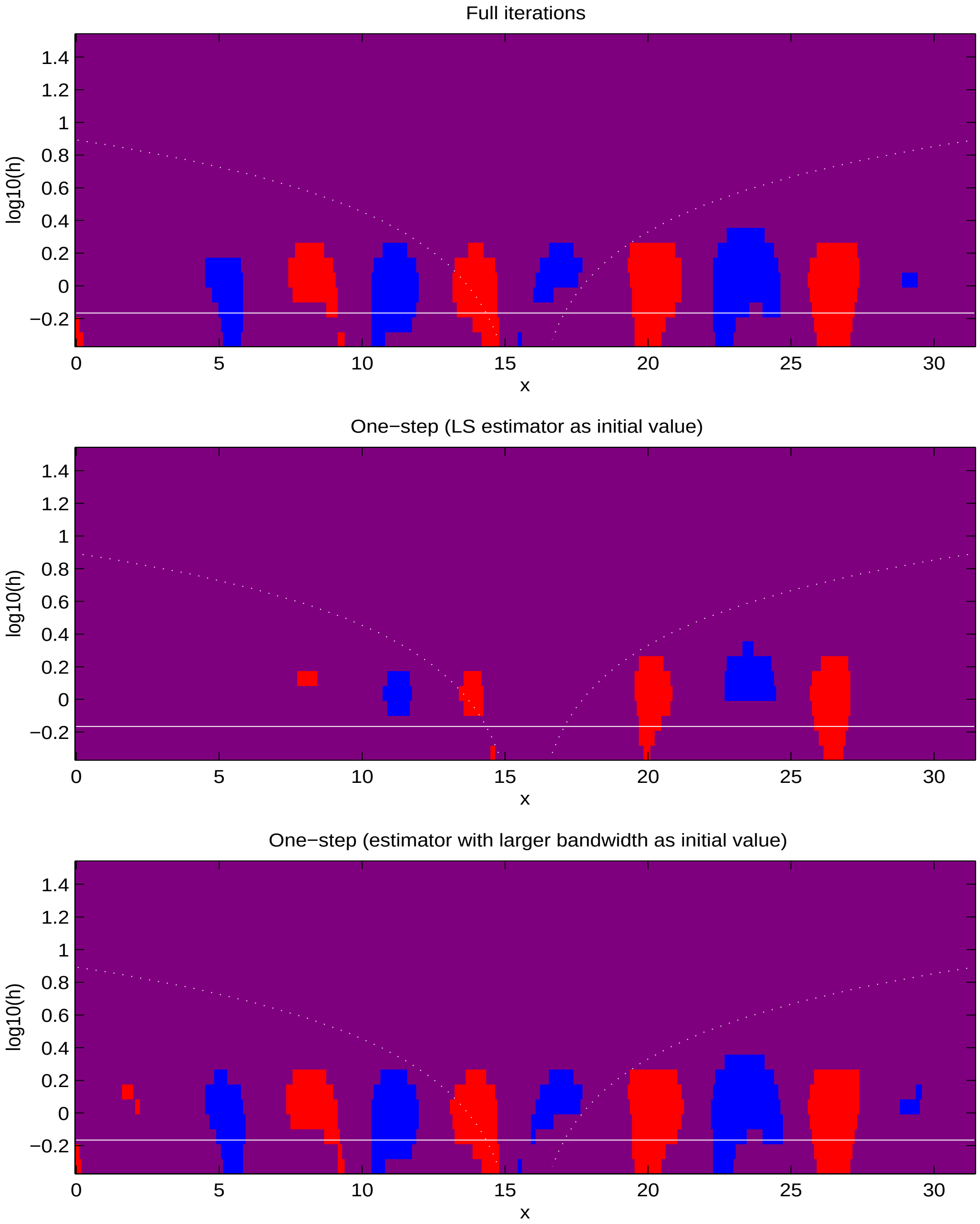
$$\text{logit}\{p(x)\} = \cos(x).$$

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 the top row. 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.

The averages and standard deviations of computation time for each simulation over 100 simulations are listed in Table 1, which shows that our proposed one-step SiZer map need least computation time among the three SiZer maps.

4.3 Application

The practical usefulness of the proposed SiZer map is illustrated in the context of analysis of an environmental data set. This data set consists of the number of daily hospital admissions



Full iterations

One−step (LS estimator as initial value)

One−step (estimator with larger bandwidth as initial value)

Figure 6: *SiZer maps for logistic regression with sample size* $n = 500$. *The top panel is the SiZer map based on a maximum local quasi-likelihood with full iterations, the middle panel 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, and the bottom panel is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 3.1.*

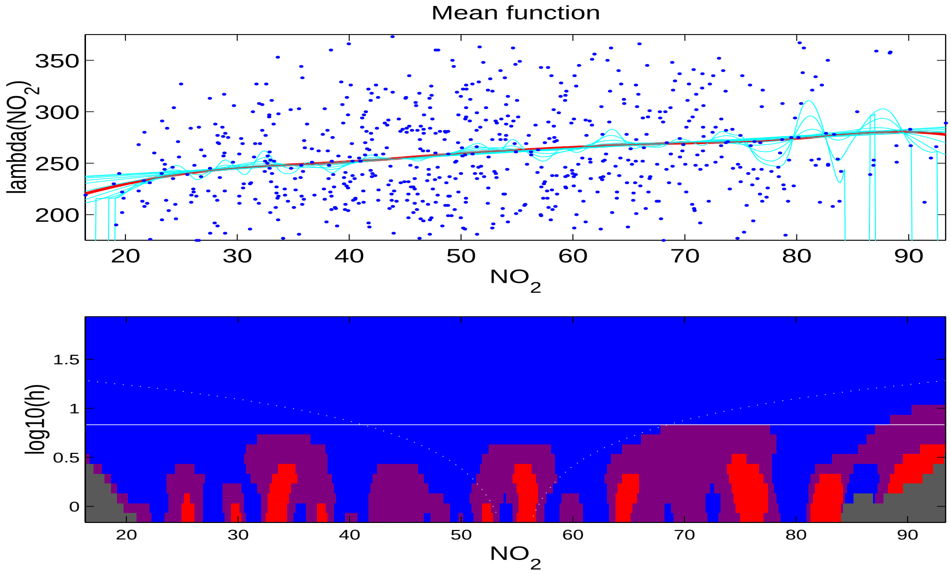


Figure 7: *SiZer map for the regression function of the number of admissions. The top panel is the plot of the family smoothing (See Chaudhuri and Marron (1999) for details). The bottom panel depicts the SiZer map for the regression function of the number of admissions given the level of NO_2.*

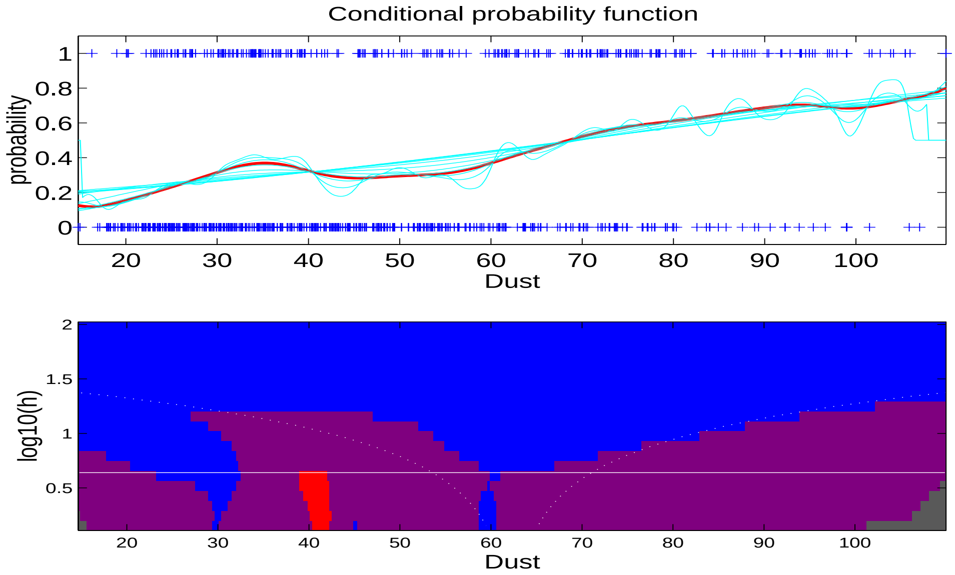


Figure 8: *SiZer map for conditional probability of the number of admissions. The top panel is the plot of family smoothing. The bottom panel depicts SiZer map for conditional probability of high level Sulphur Dioxide given the level of dust.*

for circulation and respiration problems and daily measurements of air pollutants, and has been analyzed in papers of Fan and Chen (1999) and Cai, Fan and Li (2000). As an illustration, we consider how the number of hospital admissions is associated with levels of Nitrogen Dioxide NO_2 and how conditional probability of high level Sulphur Dioxide SO_2 (with values $> 20\mu g/m^3$) depends on the levels of dust. First the number of hospital admissions is taken as the response variable, and the levels of NO_2 as covariate. Assume that the conditional distribution given the levels of NO_2 is a Poisson distribution. The newly proposed one-step SiZer map is depicted in the bottom row of Figure 7. The top row of Figure 7 is the family plot (Chaudhuri and Marron, 1999), in which a thick red curve is the estimated regression function using the reference bandwidth. From Figure 7, the number of admissions globally increases as the levels of NO_2 increases. However, there are some small wiggles when the smoothing parameters are small. This suggests that the number of admissions also depends on other pollutants. Figure 8 illustrates the proposed one-step SiZer map for the conditional probability of high level Sulphur Dioxide given the level of dust. The conditional probability function globally increases as the levels of dust increase.

5 Discussion

In this paper, we propose the local likelihood SiZer map. It has been shown that the local likelihood SiZer map is more efficient, compared with direct application of the original SiZer map, when a quasi-likelihood function is available.

Acknowledgement: This work was part of the Ph.D. dissertation of the first author, under the supervision of the second. The work of J.S. Marron was partially supported by NSF Grant DMS-9971649.

References

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

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis, The Kernel Approach with S-Plus Illustrations.* Oxford Science Publications, Oxford.

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

Chauduhuri, 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 olcal 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.

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

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests.* Springer, New York.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a). A brief survey of bandwidth selection for density esitmation, *J. Amer. Statist. Assoc.*, **83**, 941-953.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b). Progress in data-based bandwidth selection for kernel density estiamtion, *Computational Statistics*, **11**, 337-381.

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.

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.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, London: Chapman and Hall.

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