One-step Local Quasi-Likelihood Estimation

Jianqing Fan
Department of Statistics
University of North Carolina at Chapel Hill *

Jianwei Chen
Department of Statistics
Chinese University of Hong Kong

May 20, 1997

Abstract

Local quasi-likelihood estimation is a useful extension of the least-squares method, but its computation cost and algorithmic convergence problems make the procedure less appealing, particularly when it is iteratively used in the methods such as backfitting algorithm, cross-validation and bootstrapping. A one-step local quasi-likelihood estimator is introduced to overcome computational drawbacks of the local quasi-likelihood method. We demonstrate that as long as initial estimators are reasonably good, the one-step estimator has the same asymptotic behavior as the local quasi-likelihood method. Our intensive simulation shows that the one-step estimator performs at least as well as the local quasi-likelihood method for a wide range of choices of bandwidths. A data-driven bandwidth selector is proposed for the one-step estimator based on the pre-asymptotic substitution method of Fan and Gijbels (1995). It is then demonstrated via intensive simulation that the data-driven one-step local quasi-likelihood estimator performs as well as the maximum local quasi-likelihood estimator using the asymptotic optimal bandwidth. In other words, the one-step procedure reduces computation cost of the maximum local quasi-likelihood estimator without downgrading its performance.

Keywords and Phrases. Nonparametric regression, bandwidth selection, one-step estimation, quasi-likelihood, generalized linear models.
Abbreviated title: One-step local estimator.

AMS1991 subject classifications. Primary 62G07; secondary 62J12, 62E20, 62J02.

*The first author was supported by NSF grant DMS-9504414 and NSA Grant 96-1-0015. The authors thank Professor T.S. Lau for kindly providing us the data set used in Section 7.

1 Introduction

Generalized additive models (Hastie and Tibshirani, 1990) are very useful extensions of a popular family of models, Generalized Linear Models (GLIM, McCullagh and Nelder, 1989). They allow one to explore possible nonlinearity while avoid so-called "curse of dimensionality". A popular algorithm for fitting the generalized additive models is backfitting algorithm, which iteratively uses univariate nonparametric smoothing as its building blocks. Thus, it is very important to have data-driven univariate smoothers that possess computational expediency and high statistical efficiency. One-step local quasi-likelihood estimators introduced in this paper have these two advantages.

The basic smoothing method used in this paper is local polynomial fit (Fan and Gijbels, 1996). It possesses nice statistical properties. In particular, it has high minimax efficiency, copes very well with various design densities and corrects automatically edge effects. See Cleveland (1979), Cleveland *et al.* (1992), Fan (1992, 1993), Hastie and Loader (1993), Ruppert and Wand (1994), Wand and Jones (1995), Fan and Gijbels (1996), Jones (1997) and references therein. To deal with a wider class of stochastic models such as those in the GLIM, maximum local likelihood or its generalization local quasi-likelihood method was introduced to replace the least-squares method. See for example Tibshirani and Hastie (1987), Severini and Staniswalis (1994) and Hunsberger (1994), and Aragaki and Altman (1997). It was shown by Fan, Heckman and Wand (1995) that the local polynomial quasi-likelihood method inherits good statistical properties of the local polynomial least-squares approach. Carroll, Ruppert and Welsh (1996) and Fan, Farmen and Gijbels (1996) propose methods for selecting bandwidth and constructing confidence intervals for the local quasi-likelihood method.

Maximum local quasi-likelihood estimator is implicitly defined. Computing such an estimator requires an iterative algorithm such as the Newton-Raphson or Fisher's scoring method. This can be very time consuming and issues on the convergence of the algorithm arise. These make the maximum local quasi-likelihood method less appealing, particularly when the method is iteratively used such as in the backfitting algorithm, cross-validation or bootstrapping. A simple way out is to use the one-step local quasi-likelihood estimator introduced in this paper. The basic idea is simple. Given a good initial estimator such as those from the least-squares method, iterate the Newton-Raphson equation once and the resulting new estimator is the one-step local quasi-likelihood estimator. See for example Bickel (1975) for a similar idea. This new estimator clearly admits an explicit form and shares the same computational expediency as the least-squares polynomial fitting.

We will demonstrate that the one-step local quasi-likelihood estimator has the same asymptotic behavior as the maximum local quasi-likelihood estimator, as long as the initial estimators are reasonably good. In other words, the one-step estimator reduces computational burden of the maximum local quasi-likelihood estimator without down grading its performance. Thus, the one-step estimator truly inherits all good properties of the least-squares local polynomial fitting, in terms of both statistical efficiency and computation cost.

To verify the above asymptotic results at finite samples, we have conducted intensive simulations to compare the relative efficiency between the one-step method and the fully-iterative method, namely the maximum local quasi-likelihood estimators. For a wide range of bandwidths, we demonstrate that the one-step method performs at least as well as the fully-iterative method. Indeed, for small bandwidths, the one-step method tends to perform even better than the fully iterative method, and for moderate and large bandwidths both methods have about the same performance.

Choices of bandwidth are important to virtually all nonparametric smoothing problems. Various data-driven bandwidth selection techniques have been proposed, particularly in the density estimation setting. For a survey, see Jones, Marron and Sheather (1996). For local polynomial fitting, several useful bandwidth selection methods have been developed recently. They include the pre-asymptotic substitution method of Fan and Gijbels (1995), the plug-in bandwidth selector of Ruppert, Sheather and Wand (1995), the empirical-bias bandwidth selector of Ruppert (1995) and the generalized pre-asymptotic substitution method of Fan, Farmen and Gijbels (1996). In particular, Fan, Farmen and Gijbels (1996) outline a general method, based on a pre-asymptotic substitution idea, to assess the bias and variance of local maximum likelihood estimators. That general method can also be applied to the maximum local quasi-likelihood estimation discussed in this paper. However, the method is somewhat sophisticated, demanding intensive computation. In Section 6, we propose a less sophisticated bandwidth selection method that uses the bandwidth selection techniques in the least-squares setting as building blocks. One advantage is computational saving and another is that a wealth of least-squares bandwidth selection methods, such as those mentioned above, can be used.

The paper is organized as follows. Section 2 introduces the one-step local quasi-likelihood estimator and proposes good initial estimators. In Section 3, we study the asymptotic properties of the one-step estimator. Section 4 gives details on how to implement the one-step estimator for two important specific situations: Logistic and Poisson regression. Performance of the one-step and the fully-iterative estimator are compared at finite samples in Section 5. Section 6 gives a simple rule for bandwidth selection and for estimating standard errors. A few examples are given in Section

7 to illustrate the proposed method. Concluding remarks are made in Section 8. Technical proofs are given in the Appendix.

2 One-step local quasi-likelihood estimation

2.1 Quasi-likelihood

Suppose that the observed data $(X_1, Y_1), \cdots, (X_n, Y_n)$ can be regarded as a random sample from a population (X, Y) with the conditional mean and conditional variance given by

$$m(X) = E(Y|X = x), \qquad \text{var}(Y|X = x) = \sigma^2 V\{m(x)\}, \tag{2.1}$$

for a given function V and unknown scale parameter σ^2. In (2.1), the relationship between the conditional mean and the conditional variance are specified. Such a model structure appears quite often in statistical modeling. For example, Binomial distributions and Poisson distributions admit the relationship (2.1) with $V(t) = t(1-t)$ and $V(t) = t$, respectively. These are just two specific examples of the exponential family of distributions used in the generalized linear models (McCullagh and Nelder, 1989), under which the conditional density of Y given $X = x$ is assumed to belong to a canonical exponential family

$$f(y|x) = \exp([\theta(x)y - b\{\theta(x)\}]/a(\phi) + c(y, \phi)), \tag{2.2}$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. Here the parameter $\theta(\cdot)$ is called the canonical parameter and ϕ is called the dispersion parameter. Under this model, it can easily be shown (McCullagh and Nelder, 1989) that

$$m(x) = E(Y|X = x) = b'\{\theta(x)\}, \qquad \text{var}(Y|X = x) = a(\phi)b''\{\theta(x)\}. \tag{2.3}$$

Thus model (2.2) satisfies the assumption (2.1).

In parametric generalized linear models, the unknown regression function $m(x)$ is modeled linearly via a known link function $g(\cdot)$:

$$g\{m(x)\} = \alpha + \beta x. \tag{2.4}$$

The function g links the regression function to a linear space of the covariate. If $g = (b')^{-1}$, then g is called the canonical link function. The aim of this paper is to estimate the function $\eta(x) = g\{m(x)\}$ nonparametrically. There are several reasons for not estimating $m(\cdot)$ directly in the current context. As pointed out in Fan, Heckman and Wand (1995), the range of $\psi(x)$ is $(-\infty, +\infty)$ and hence the

estimate $\widehat{\psi}(\cdot)$ is range-preserving; the local log-likelihood (2.8) is concave in β and consequently computing $\widehat{\beta}(x)$ is much easier; the model reduces to the usual parametric model when h is large – this provides parsimonious models for the parametric generalized linear models.

Under the model assumption (2.1), the quasi-likelihood method is often used. The quasi-likelihood function $Q(\mu, y)$ is defined via

$$\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y - \mu}{V(\mu)}, \tag{2.5}$$

and the i^{th} data point contributes $Q\{m(X_i), Y_i\}$ to the quasi-likelihood. Thus, for the parametric model (2.4), one can estimate the parameters α and β via maximizing the quasi-likelihood

$$\sum_{i=1}^{n} Q\{g^{-1}(\alpha + \beta X_i), Y_i\}. \tag{2.6}$$

Note that for the exponential family of models (2.2), the quasi-likelihood (2.6) is just the conditional log-likelihood of $(Y_1, \cdots, Y_n)$ given $(X_1, \cdots, X_n)$. Thus, the quasi-likelihood approach is an extension of the likelihood method.

2.2 Local quasi likelihood

One can not directly use the quasi-likelihood (2.6) when function $\eta(\cdot)$ is not parameterized. Assuming that η is smooth with the $(p+1)^{th}$ derivative at a given point x, the following form holds approximately via Taylor's expansion:

$$\eta(z) \approx \beta_0 + \cdots + \beta_p (z - x)^p, \tag{2.7}$$

for z in a neighborhood of the point x. Following Fan, Heckman and Wand (1995), one can construct the local quasi-likelihood

$$\ell(\beta) \equiv \sum_{i=1}^{n} Q[g^{-1}\{\beta_0 + \ldots + \beta_p (X_i - x)^p\}, Y_i] K_h(X_i - x), \tag{2.8}$$

where $K_h = K(\cdot/h)/h$ with K being a kernel function and h a bandwidth. Let $\widehat{\beta}_0, \cdots, \widehat{\beta}_p$ maximize the local quasi-likelihood (2.8). Then, the maximum local quasi-likelihood estimator is

$$\widehat{\eta}_\nu(x) = \nu! \widehat{\beta}_\nu(x), \tag{2.9}$$

for $\eta^{(\nu)}(x), \nu = 0, \cdots, p$ with convention $\widehat{\eta}_0(x) = \widehat{\eta}(x)$. Note that the local quasi-likelihood (2.8) is just a weighted version of the quasi-likelihood (2.6) with weights $K_h(X_i - x)$ used to confine the polynomial model (2.7) being applied locally around the point x.

The quasi-likelihood (2.8) does not admit an explicit solution unless $p = 0$. The maximization is usually carried out via the Newton-Raphson method or its variation – Fisher's scoring method. For the case $p = 0$, the solution is nothing but a transform of the Nadaraya-Watson estimator:

$$\widehat{\eta}(x) = g\{\sum_{i=1}^{n} K_h(X_i - x)Y_i / \sum_{i=1}^{n} K_h(X_i - x)\}.$$

This method was studied by Staniswalis (1989) and Severini and Staniswalis (1994). Drawbacks of the kernel method are that it creates large bias, has serious boundary effect and can not cope well with non-uniform designs. For details, see Wand and Jones (1995), Simonoff (1995) and Fan and Gijbels (1996).

2.3 One-step local quasi-likelihood

The local quasi-likelihood estimation, while inherits many nice statistical properties from the least-squares local polynomial fitting (see Fan, Heckman and Wand, 1995), involves intensive computation via iteratively solving linear equations. This can be very time consuming and the issue on the convergence of the algorithm arises. These make the procedure less attractive. A simple way out is to employ one-step estimation scheme. See for example Bickel(1975). We now outline the procedure.

Denote by $q_\ell(x, y) = (\partial^\ell / \partial x^\ell) Q\{g^{-1}(x), y\}$. Let $\ell'(\beta)$ and $\ell''(\beta)$ be respectively the gradient vector and the Hessian matrix of the function $\ell(\beta)$. Namely

$$\ell'(\beta) = n^{-1} \sum_{i=1}^{n} q_1(\beta_0 + \ldots + \beta_p (X_i - x)^p, Y_i) K_h(X_i - x) \mathbf{X}_i \tag{2.10}$$

and

$$\ell''(\beta) = n^{-1} \sum_{i=1}^{n} q_2(\beta_0 + \ldots + \beta_p (X_i - x)^p, Y_i) K_h(X_i - x) \mathbf{X}_i \mathbf{X}_i^T, \tag{2.11}$$

where $\mathbf{X}_i = (1, X_i - x, \cdots, (X_i - x)^p)^T$.

Let $\widehat{\beta}_{MLE}$ be the maximum quasi-likelihood estimator that solves the quasi-likelihood equation $\ell'(\beta) = 0$, namely

$$\ell'(\widehat{\beta}_{MLE}) = 0.$$

Suppose that $\widehat{\beta}_0$ is a vector of initial estimators with reasonably good precision. Then by Taylor's expansion,

$$0 = \ell'(\widehat{\beta}_{MLE}) \approx \ell'(\widehat{\beta}_0) + \ell''(\widehat{\beta}_0)(\widehat{\beta}_{MLE} - \widehat{\beta}_0)$$

so that

$$\widehat{\beta}_{MLE} \approx \widehat{\beta}_0 - [\ell''(\widehat{\beta}_0)]^{-1} \ell'(\widehat{\beta}_0).$$

This leads us to defining the one-step local quasi-likelihood estimator as

$$\widehat{\beta}_{OS} = \widehat{\beta}_0 - [\ell''(\widehat{\beta}_0)]^{-1}\ell'(\widehat{\beta}_0). \tag{2.12}$$

The one-step estimator clearly inherits the computational expediency from the least-squares polynomial fitting.

We now briefly discuss the choice of the initial estimators. An intuitive and explicit method is based on the substitution of the least-squares local polynomial estimator. Let $\widehat{m}_\nu(x)$ $(\nu = 0, \cdots, p)$ be the local polynomial regression estimator of $m^{(\nu)}(x)$. Namely, they minimize

$$\sum_{i=1}^{n}\{Y_i - \beta_0 - \cdots - \beta_p(X_i - x)^p/p!\}^2 K_h(X_i - x).$$

Recall that $\eta(x) = g\{m(x)\}$ so that

$$\eta'(x) = g'(m(x))m'(x), \qquad \eta''(x) = g''(m(x))[m'(x)]^2 + g'(m(x))m''(x)$$

and so on. Substituting the least-squares estimator into the above relationship, we can easily obtain an initial estimator of $\eta^{(\nu)}(x)$ explicitly.

Note that the least-squares estimator is not necessarily range-preserving. Take the logistic regression in the Bernoulli model with $g(t) = \log\{t/(1-t)\}$ as an example. The mean regression function $m(x) = P(Y = 1|X = x)$ must fall in $[0, 1]$. However, the least-squares local polynomial regression estimator $\widehat{m}(x)$ is not necessarily in this range and hence $g(\widehat{m}(x))$ may not necessarily well defined. This gap can be bridged. See Section 4.1 for details of implementation.

3 Asymptotic properties

In this section, we will derive the asymptotic distribution of the one-step local quasi-likelihood estimator. We demonstrate that the one-step estimator performs as well as the maximum local quasi-likelihood (fully-iterative) estimator as long as the initial estimators $\widehat{\beta}_0$ in (2.12) are reasonably accurate. In other words, the one-step method reduces computational cost of the fully iterative estimator without downgrading its asymptotic performance.

Denote by $\mu_j = \int u^j K(u)du$ and $\nu_j = \int u^j K(u)^2 du$, $j = 0, 1, 2, \cdots$. Let

$$H = \mathrm{diag}(1, h, \cdots, h^p), \quad S = (\mu_{i+j-2})_{1\le i,j\le p+1}, \quad S^* = (\nu_{i+j-2})_{1\le i,j\le p+1} \tag{3.1}$$

be $(p+1)\times(p+1)$ matrices. Let

$$\beta_0(x) = \left(\eta(x), \eta'(x), \cdots, \eta^{(p)}(x)/p!\right)^T$$

be the true local parameters. To stress the dependency of $\widehat{\beta}_0$ and $\widehat{\beta}_{OS}$ on x in (2.12), we write them as $\widehat{\beta}_0(x)$ and $\widehat{\beta}_{OS}(x)$ in the next theorem. The following theorem describes the joint asymptotic normality of $\widehat{\beta}_{OS}(x)$. The marginal distribution for derivative estimators can easily be obtained.

Theorem 1. Under Conditions (1) – (4) in the appendix,

$$\sqrt{nh}\left(H\{\widehat{\beta}_{OS}(x)-\beta_0(x)\}-\frac{\eta^{(p+1)}(x)}{(p+1)!}S^{-1}\mu h^{p+1}+o_P(h^{p+1})\right)\to N\Big(0,v(x)S^{-1}S^*S^{-1}/f(x)\Big), \tag{3.2}$$

provided that the initial estimation vector satisfies

$$H\{\widehat{\beta}_0(x)-\beta_0(x)\}=O_p\{h^{p+1}+(nh)^{-1/2}\}, \tag{3.3}$$

where $\mu=(\mu_{p+1},\cdots,\mu_{2p+1})^T$ be a $(p+1)$-vector, and $v(x)=[g'\{m(x)\}]^2\mathrm{var}(Y|X=x)$.

Following the theory of local polynomial fitting (see Chapter 3 of Fan and Gijbels, 1996), the initial estimators discussed at the end of Section 2.3 satisfy condition (3.3).

Comparing with the result of Theorem 1 of Fan, Heckman and Wand (1995), for estimating $\eta^{(\nu)}$, one can easily see that the one-step estimator shares the same asymptotic bias and variance as the fully-iterative estimator when $p-\nu$ is odd. Fan, Heckman and Wand (1995) further pointed out that the local polynomial fitting with $p-\nu$ even is unappealing and not recommended. For this reason, we don't pursuit further the results in this direction. Theorem 1 improves the results of Fan, Heckman and Wand (1995) in two important directions. The quasi-likelihood function $Q(\mu,y)$ does not have to be concave with respect to the argument μ and the local quasi-likelihood estimator does not have to exist.

The above discussions reveal that the asymptotic optimal bandwidth, which minimizes the asymptotic weighted mean integrated squared error, should be the same for both the one-step and the fully-iterative estimator. Write $S^{-1}=(S^{ij})_{0\le i,j\le p}$ and let

$$K_\nu^*(t)=e_{\nu+1}^TS^{-1}(1,t,\cdots,t^p)^TK(t)=(\sum_{j=0}^p S^{\nu j}t^j)K(t)$$

be the equivalent kernel (see Fan and Gijbels 1996). Then, the asymptotic optimal bandwidth for estimating $\eta^{(\nu)}(\cdot)$ is given by

$$h_{opt}=C_{\nu,p}(K)\left[\frac{\int v(x)w(x)/f(x)dx}{\int\{\eta^{(p+1)}(x)\}^2w(x)dx}\right]^{1/(2p+3)}n^{-1/(2p+3)}, \tag{3.4}$$

where w is a weight function and

$$C_{\nu,p}(K)=\left[\frac{(p+1)!^2(2\nu+1)\int K_\nu^{*2}(t)dt}{2(p+1-\nu)\{\int t^{p+1}K_\nu^*(t)dt\}^2}\right]^{1/(2p+3)}.$$

An appealing property of the local polynomial fitting is that it copes well with the edge effect. This property is also inherited by the one-step estimator. To describe such a property, we follow the formulation given in Gasser and Müller (1979). Assume that the density has a bounded support [0, 1], say. Consider the one-step fitting scheme at the left-hand point $x = ch$. Let the matrices S_c and S_c^* be defined similarly to (3.1) except that the moments are now replaced respectively by

$$\mu_{i+j-2,c} = \int_{-c}^{+\infty} u^j K(u)du, \qquad \nu_{i+j-2,c} = \int_{-c}^{+\infty} u^j K^2(u)du, \quad \text{for } j = 0, 1, 2, \cdots.$$

Then we have the following result (a similar result holds for the right boundary point).

Theorem 2. Assume Conditions (1) – (4) in the appendix hold for the point $x = 0+$. Then, at the left-hand boundary point $x = ch$, we have

$$\sqrt{nh}\left(H\{\widehat{\beta}_{OS}(ch) - \beta_0(ch)\} - \frac{\eta^{(p+1)}(0+)}{(p+1)!}S_c^{-1}\mu_c h^{p+1} + o_P(h^{p+1})\right) \to N\Big(0, v(0+)S_c^{-1}S_c^*S_c^{-1}/f(0+)\Big),$$

provided that the initial estimation vector satisfies

$$H\{\widehat{\beta}_0(ch) - \beta_0(ch)\} = O_p\{h^{p+1} + (nh)^{-1/2}\}, \tag{3.5}$$

where $\mu_c = (\mu_{p+1,c}, \cdots, \mu_{2p+1,c})^T$.

4 Applications to specific models

In this section, we will discuss how to implement one-step local quasi-likelihood estimators and local quasi-likelihood estimators for two important members of the exponential family (2.2), namely, the Bernoulli model and Poisson model. For simplicity, we use local linear fits throughout this section; other orders can be implemented analogously.

4.1 Logistic regression

Given a random sample $\{(X_i, Y_i), i = 1, \cdots, n\}$ from a population (X, Y) whose conditional distribution is a Bernoulli distribution with $P(Y = 1|X = x) = p(x)$, we are interested in estimating the logistic regression function

$$\eta(x) = \log \frac{p(x)}{1 - p(x)}.$$

As noted in §2.1, the quasi-likelihood in this case becomes the log-likelihood given by

$$Q(p(x), y) = \log[p(x)^y\{1 - p(x)\}^{1-y}] = y\eta(x) - \log[1 + \exp\{\eta(x)\}].$$

Thus, the local quasi-likelihood (2.8) now becomes

$$\ell(a,b) = \sum_{i=1}^{n} K_h(X_i - x)\Big(Y_i\{a + b(X_i - x)\} - \log[1 + \exp\{a + b(X_i - x)\}]\Big). \tag{4.1}$$

For the initial estimator $(\widehat{a}_0, \widehat{b}_0)$ of $(\eta(x), \eta'(x))$, the one-step estimator is given by

$$\begin{pmatrix} \widehat{a}_{OS} \\ \widehat{b}_{OS} \end{pmatrix} = \begin{pmatrix} \widehat{a}_0 \\ \widehat{b}_0 \end{pmatrix} + \begin{pmatrix} u_{n,0} & u_{n,1} \\ u_{n,1} & u_{n,2} \end{pmatrix}^{-1} \begin{pmatrix} v_{n,0} \\ v_{n,1} \end{pmatrix}, \tag{4.2}$$

where

$$u_{n,j} = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j \widehat{p}_{i0}(1 - \widehat{p}_{i0}), \qquad v_{n,j} = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j (Y_i - \widehat{p}_{i0})$$

with $\widehat{p}_{i0} = \exp\{\widehat{a}_0 + \widehat{b}_0(X_i - x)\}[1 + \exp\{\widehat{a}_0 + \widehat{b}_0(X_i - x)\}]^{-1}$. The maximum local likelihood estimation is simply iteratively using equation (4.2).

In practical implementation, the matrix in (4.2) can be singular or nearly singular when the local data are sparse. A commonly-used technique to deal with this problem is the ridge regression technique (see e.g. Seift and Gasser, 1996). Then an issue arises how large the ridge parameter should be used. Note that if $h_n \to 0$ and $nh_n \to \infty$, we have the asymptotic approximation,

$$u_{n,j} \approx \widehat{p}_0(1 - \widehat{p}_0)h^{j-1} \int u^j K(u)du N \quad \text{with} \quad \widehat{p}_0 = \frac{\exp(\widehat{a}_0)}{1 + \exp(\widehat{a}_0)},$$

where $N = \{nhf(x)\}$ with f being the marginal density of X can intuitively be understood as the effective number of local data points. Replacing $u_{n,j}$ by $u_{n,j} + \widehat{p}_0(1 - \widehat{p}_0)h^{j-1} \int u^j K(u)$ for $j = 0$ and $j = 2$ in (4.2) will not alter the asymptotic behavior and will prevent the matrix from nearly singular when N is small. In other words, we suggest to use ridge parameters

$$\widehat{p}_0(1 - \widehat{p}_0)h^{j-1} \int u^j K(u), \quad \text{for } j = 0, 2. \tag{4.3}$$

These ridge parameters will be used in implementation of the one-step local likelihood estimation. They will also be used in the implementation of maximum local likelihood estimation, which iteratively uses (4.2). Note that the resulting local ridge regression estimator does not alter asymptotic properties of the one-step estimator and the local likelihood estimator because the ridge parameters are of smaller order.

We now turn to discussing the use of initial estimator. Following Fan (1992), the least-squares local linear estimator of the regression function $p(x)$ and its derivative $p'(x)$ are given by

$$\begin{pmatrix} \widehat{p}_{LS}(x) \\ \widehat{p'}_{LS}(x) \end{pmatrix} = \begin{pmatrix} s_{n,0} & s_{n,1} \\ s_{n,1} & s_{n,2} \end{pmatrix}^{-1} \begin{pmatrix} t_{n,0} \\ t_{n,1} \end{pmatrix}, \tag{4.4}$$

where

$$s_{n,j} = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j, \qquad t_{n,j} = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j Y_i.$$

Again, we use the ridge regression to guard against the possibility of singularity of the matrix in (4.4). Following the same heuristic as in the last paragraph, the ridge parameters

$$h^{j-1} \int u^j K(u), \quad \text{for } j = 0, 2. \tag{4.5}$$

are used, namely replacing $s_{n,j}$ by $s_{n,j} + h^{j-1} \int u^j K(u)$ for $j = 0, 2$ for the matrix in (4.4).

The estimator $\widehat{p}_{LS}(x)$ is not necessarily between 0 and 1. This drawback can easily be repaired via using the estimator

$$\widehat{p}_{LS,1}(x) = \min\{\max(\widehat{p}_{LS}(x), 0), 1\}.$$

This is still not satisfactory since $\widehat{p}^*_{LS}(x)$ can take value zero or one so that its logit transform does not exist. To overcome this difficulty, we borrow the idea of Bayesian estimation of proportion with the uniform prior and suggest to use

$$\widehat{p}_{LS,2} = \frac{N}{N+2} \widehat{p}_{LS,1}(x) + \frac{2}{N+2} \times 0.5,$$

where N is the effective number of local data points defined by

$$N = \frac{2nh}{X_{\max} - X_{\min}} \left(\frac{\int u^2 K(u) du}{\int K^2(u) du} \right)^{1/5}, \tag{4.6}$$

where $X_{\max}$ and $X_{\min}$ are respectively the maximum and minimum order statistic of design points. The first factor is the effective number of data points under the uniform design with the standard uniform kernel and the second factor is used to standardize the kernel (see Marron and Nolan, 1988 for the idea of canonical kernel). Note that we define N to be independent of x [instead of $N = \sum_{i=1}^{n} K_h(X_i - x_0)/K_h(0)$] so that all estimates are pulled the same amount towards 0.5. Otherwise, the sparse region will be pulled more and this can create some artifacts (An artificial dip can occur for large x in Figure 7 if we use the non-constant N).

With this modification, following the idea outlined at the end of Section 2, we define the initial estimator as

$$\widehat{a}_0(x) = \log \frac{\widehat{p}_{LS,2}(x)}{1 - \widehat{p}_{LS,2}(x)}, \qquad \widehat{b}_0(x) = \frac{\widehat{p}'_{LS,2}(x)}{\widehat{p}_{LS,2}(x)\{1 - \widehat{p}_{LS,2}(x)\}} \ . \tag{4.7}$$

4.2 Poisson regression

We now consider the Poisson regression model, with conditional probability as

$$P(Y = k | X = x) = \exp\{-\lambda(x)\}\lambda(x)^k / k!, \qquad \text{for } k = 0, 1, \cdots$$

The canonical link function is $g_1(t) = \log t$ and the canonical function $\eta(x) = \log\{\lambda(x)\}$ is of interest. The quasi-likelihood in this case is the log-likelihood and is given by

$$\ell(a,b) = \sum_{i=1}^{n} K_h(X_i - x)[Y_i\{a + b(X_i - x)\} - \exp\{a + b(X_i - x)\}].$$

The one-step estimator is given similarly to (4.2) except now that

$$u_{n,j} = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j \widehat{\lambda}_{i0} \qquad v_{n,j} = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j (Y_i - \widehat{\lambda}_{i0}),$$

where $\widehat{\lambda}_{i0} = \exp\{\widehat{a}_0 + \widehat{b}_0(X_i - x)\}$ with $(\widehat{a}_0, \widehat{b}_0)$ being an initial estimator. Using the same arguments as in §4.1, the ridge parameters

$$\widehat{\lambda}_0 h^{j-1} \int u^j K(u) \quad \text{with} \quad \widehat{\lambda}_0 = \exp(\widehat{a}_0), \quad \text{for } j = 0, 2 \tag{4.8}$$

are used to guard against the singularity of the matrix in (4.2). As noted in §4.1, the maximum local likelihood estimation is simply iteratively using equation (4.2) and hence a similar ridge regression strategy is employed.

The least-squares local linear estimator of the regression function $\lambda(x)$ and its derivative $\lambda'(x)$ are still given by the right hand side of (4.4). Again, the ridge parameters (4.5) are used. We denote the resulting estimators by $\widehat{\lambda}_{LS}(x)$ and $\widehat{\lambda}'_{LS}(x)$. The estimator $\widehat{\lambda}_{LS}(x)$ is not necessarily positive. A simple modification is to use

$$\widehat{\lambda}_{LS,1}(x) = \max\{\widehat{\lambda}_{LS}(x), 0\} + 0.2N^{-1}$$

where N is given by (4.6). Clearly this modification does not alter the asymptotic property of $\widehat{\lambda}_{LS}(x)$. Following the recipe given at the end of §2.3, initial estimators are defined as

$$\widehat{a}_0(x) = \log\{\widehat{\lambda}_{LS,1}(x)\}, \qquad \widehat{b}_0(x) = \widehat{\lambda}'_{LS}(x)/\widehat{\lambda}_{LS,1}(x). \tag{4.9}$$

With this initial estimator, one can calculate the one-step estimator (4.2).

5 Comparisons with local quasi-likelihood method

In this section, we compare finite sample performance of the one-step local quasi-likelihood estimators with the local quasi-likelihood estimators via simulations. The purpose is to examine if the two types of estimators perform comparably at finite sample with a wide range of choices of bandwidths. Logistic regression and Poisson regression models will be used.

In the implementation, we employed the local linear fit with the Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$. Let h_{opt} be the asymptotic optimal bandwidth given by (3.4) with the uniform weight. The bandwidths $h = h_{opt}/2, h_{opt}$ and $2h_{opt}$ are used. This range is wide enough to cover most of practical applications.

The performance of each given estimator $\widehat{\eta}(x)$ is assessed via the square-Root of Average Square Errors (RASE):

$$RASE = \left(n_{grid}^{-1} \sum_{j=1}^{n_{grid}} \{\widehat{\eta}(x_j) - \eta(x_j)\}^2 \right)^{1/2}, \tag{5.1}$$

where $\{x_j, j = 1, \cdots, n_{grid}\}$ are the grid points at which the function η is estimated. For completeness, we also compute the RASE of the initial estimator based on the least-squares method. We will call such an initial estimator of η as a (modified) least-squares estimator. For the logistic regression and Poisson regression models, the least-squares estimator of $\eta(x)$ is $\widehat{a}_0(x)$ given respectively by (4.7) and (4.9).

5.1 Logistic regression

We now use the simulation models from Fan, Farmen and Gijbels (1997) as testing examples. The design density is the uniform distribution on $[-2, 2]$ and logit regression function is given by

Example 1. $\eta(x) = 3\sin(2x)$

Example 2. $\eta(x) = 7[\exp\{-(x+1)^2\} + \exp\{-(x-1)^2\}] - 5.5$

Example 3. $\eta(x) = 2 - x^2$.

These curves appear as the solid line in part (c) of Figures 1-3 and the asymptotic optimal bandwidths h_{opt} for $n = 250, 500, 1000$ were given in Table 1 of Fan, Farmen and Gijbels (1997).

For each of the above examples, we conducted 400 simulations with sample size $n = 250, 500, 1000$. The results for $n = 500$ and $n = 1000$ are basically the same as those for $n = 250$ and hence are omitted except for Example 1. For each given sample, we computed the ratio of the RASE of $\widehat{\eta}_{LS}(x)$ (the least-square method) to that of $\widehat{\eta}_{MLE}(x)$ (the maximum local likelihood method), and the ratio of RASE of $\widehat{\eta}_{OS}(x)$ (the one-step local likelihood estimator) to that of $\widehat{\eta}_{MLE}(x)$.

In part (a) of Figures 1–3, we summarize the marginal distributions of the ratios for three different choices of bandwidths. From these figures, it is clear that for $h = h_{opt}/2$ the one-step local quasi-likelihood estimator is the best and even the least-squares method performs somewhat better the maximum local quasi-likelihood approach (which will also be called as "fully-iterative approach"). For bandwidths $h = h_{opt}$ and $2h_{opt}$, both the one-step and the fully-iteratively method

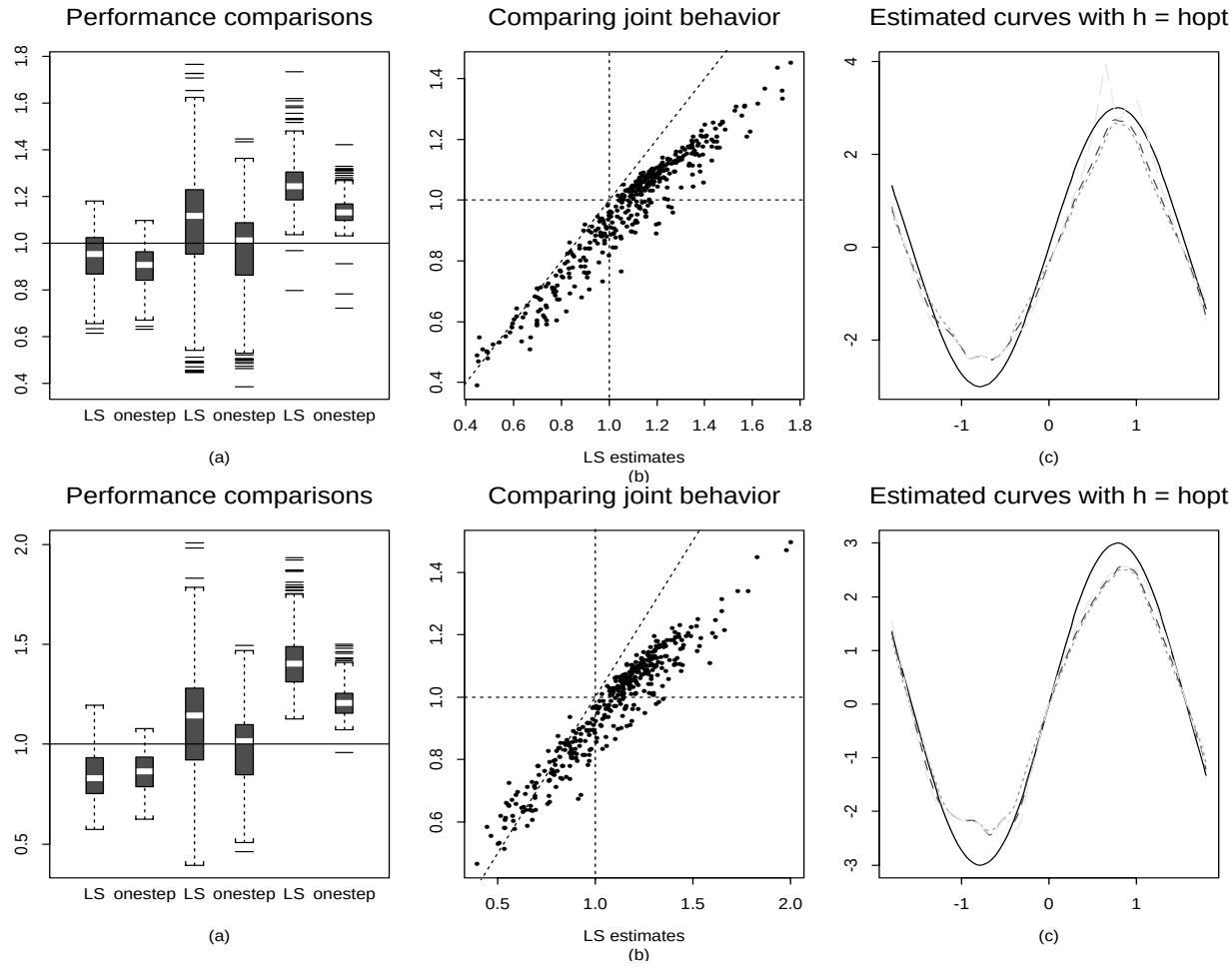

Figure 1: *Simulation results for Example 1. The top panel is for* $n = 250$ *and the bottom panel is for* $n = 500$. *(a) Boxplot for the ratios of RASE of the least-squares method and one-step local likelihood approach to that of the local likelihood estimator, using bandwidths (from left to right)* $h_{opt}/2, h_{opt}, 2h_{opt}$ *(b) The scatter plot the ratios of RASE of the least-squares method versus those of the one-step local quasi-likelihood approach. (c) A typical estimate with bandwidth* h_{opt}. *Solid curve — true function. Dash curves (from shortest to longest dash) are the least-squares, one-step and local likelihood estimate.*

perform comparably and both methods outperform the least-squares method. This is consistent with the asymptotic theory in Section 3.

In part (b) of Figures $1-3$, we depict the joint behavior of the three estimation methods using $h = h_{opt}$. For each given sample, the ratio of RASE of $\widehat{\eta}_{LS}(x)$ to that of $\widehat{\eta}_{MLE}(x)$ is plotted against the ratio of RASE of $\widehat{\eta}_{OS}(x)$ to that of $\widehat{\eta}_{MLE}(x)$. The horizontal dash line marks the positions where $\widehat{\eta}_{OS}(x)$ and $\widehat{\eta}_{MLE}(x)$ perform equally well and the slope dashed line marks the position where $\widehat{\eta}_{LS}(x)$ and $\widehat{\eta}_{OS}(x)$ have the same performance. A similar remark can be made for the horizontal dashed lines. For example, Figure (1b) indicates that $\widehat{\eta}_{OS}$ almost always outperforms $\widehat{\eta}_{LS}$. From

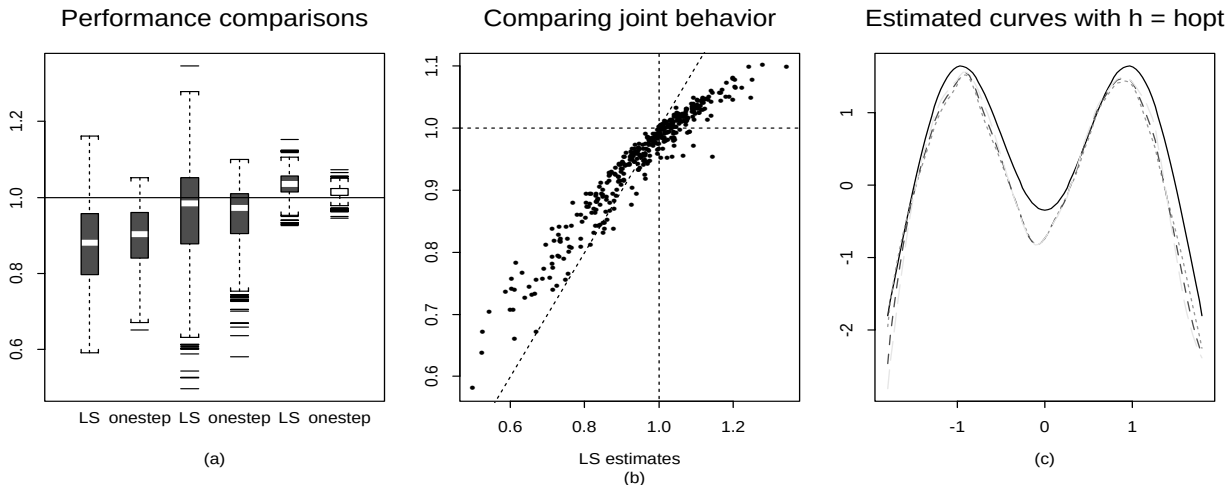

Figure 2: *Simulation results for Example 2 with* $n = 250$. *Similar captions to Figure 1 are used.*

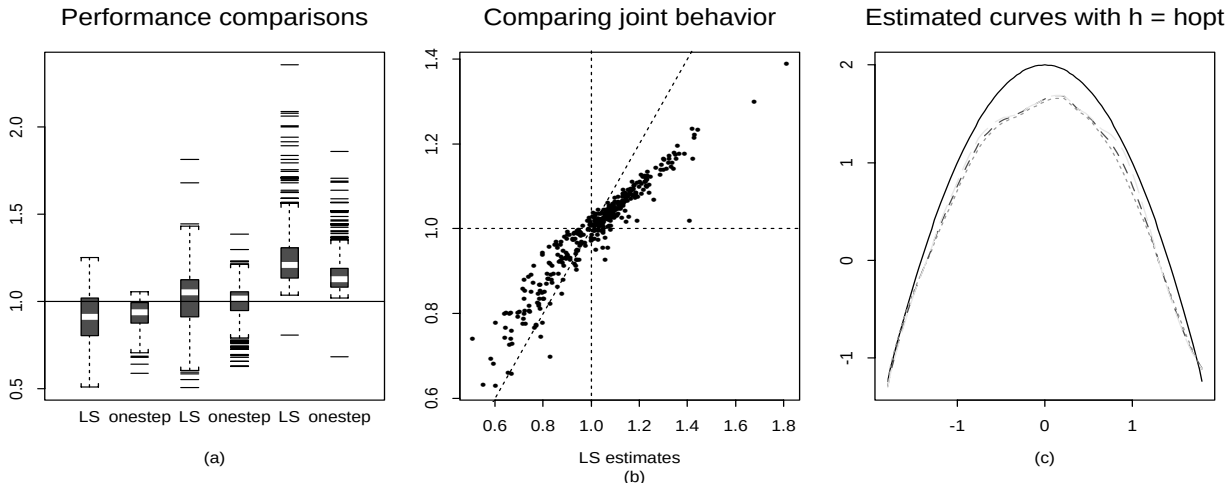

Figure 3: *Simulation results for Example 3 with* $n = 250$. *Similar captions to Figure 1 are used.*

these figures, one can see that the local one-step estimator performs better than the least-squares method, and is even somewhat better than the fully-iterative method in the sense the ratios have heavier tail on the small side than the large side. Take the lower panel of Figure 1(b) as an example. There are quite some points with ratios smaller than 0.8 but there are only a few points with the ratio higher than $1/0.8 = 1.25$.

In part (c) of Figures $1 - 3$, we took a random sample and computed the estimates for each of the three methods. This gives us visual impression how well each method performs.

5.2 Poisson regression

We now test our asymptotic results in the Poisson regression setting. The design density is

n	250	500	1000
Example 4	.369	.321	.279
Example 5	.349	.301	.264
Example 6	.599	.521	.453

Table 1: Asymptotic optimal bandwidth (3.4) for Poisson models with uniform weight function

again the uniform distribution on $[-2, 2]$ and the function $\eta(x) = \log\{\lambda(x)\}$ is given by

$$\begin{aligned}
&\text{Example 4.} \quad \eta(x) = 1.5\sin(2x) + 1.25 \\
&\text{Example 5.} \quad \eta(x) = 3.5[\exp\{-(x+1)^2\} + \exp\{-(x-1)^2)\}] - 1.5 \\
&\text{Example 6.} \quad \eta(x) = 2.0 - 0.5x^2.
\end{aligned}$$

These functions are respectively a rescaling (linear transform) of the functions given in Examples $1-3$ so that the signal to noise ratios are between 2 and 2.5 under the Poisson distributions. For these Poisson models, the asymptotic optimal bandwidths h_{opt} with the uniform weight function on $[-2, 2]$ are summarized in Table 1 for $n = 250, 500$ and $1,000$.

Figures 4(a)–(c) compare the performance of the least-squares, the one-step, and the local likelihood estimator using bandwidths $h = h_{opt}/2, h_{opt}$ and $2h_{opt}$ under the Poisson models. Only the results for $n = 250$ are presented here; others are similar and hence are omitted. Figures 4 (d)–(f) present a typical estimated curve using $h = h_{opt}$ respectively for Examples $4-6$.

Figure 4 reveals the fact that for small bandwidths, the one-step estimator and the least-squares estimator tend to perform better than the local likelihood estimator and for large bandwidths the one-step estimator and the local likelihood estimator perform about equally well. This is consistent with our asymptotic theory.

6 Bandwidth selection and estimation of standard error

The one-step estimator and the local quasi-likelihood estimator share the same asymptotic bandwidth. Hence, one can apply the sophisticated bandwidth selection rule proposed in Fan, Farmen and Gijbels (1996) to the one-step estimator. However, this will increase significantly computation cost. In this section, we make a simple connection for the bandwidth selection problem between the local least-squares method and the local quasi-likelihood method. The benefit of this is that a wealth of bandwidth selectors for the local least-squares method can be applied to the local maximum quasi-likelihood estimation and the local one-step quasi-likelihood estimation.

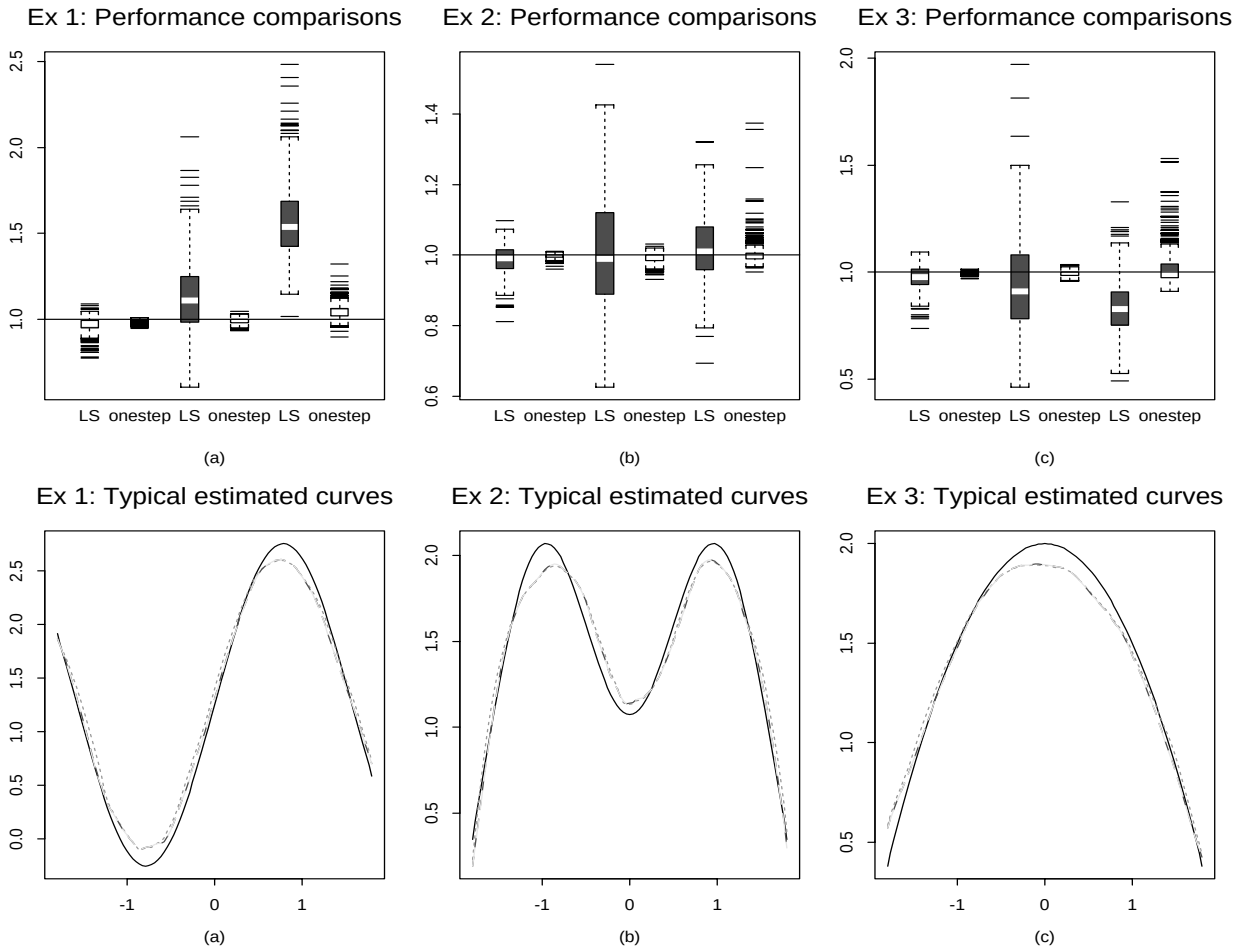

Figure 4: *Simulation results for Poisson models with* $n = 250$. *Top panel compares the performance for the least-squares estimator, the one-step estimator and the local likelihood estimator. Similar captions to Figure 1 (a) are used. The bottom summarizes typical performance of the estimators using bandwidth* h_{opt}. *Similar captions to Figure 1(c) are used.*

For a given estimator $\widehat{m}$, let $\widehat{\eta}(x) = g\{\widehat{m}(x)\}$. Then by the mean-value theorem, we have

$$\widehat{\eta}(x) - \widehat{\eta}(x) \approx g'\{\widehat{m}(x)\}\{\widehat{m}(x) - m(x)\}.$$

Therefore, the Integrated Square Error (ISE) for $\widehat{\eta}(\cdot)$ is approximately the same as weighted ISE for $\widehat{m}(\cdot)$. More generally,

$$\int \big\{\widehat{\eta}(x) - \widehat{\eta}(x)\}^2 w_0(x)dx \approx \int \{\widehat{m}(x) - \widehat{m}(x)\}^2 g'\{\widehat{m}(x)\}^2 w_0(x)dx, \tag{6.1}$$

where w_0 is a given weight function.

The relation (6.1) suggests a simple, rule of thumb, bandwidth selection rule. Use an estimated optimal bandwidth for the least-squares local polynomial estimator $\widehat{m}_{LS}$ with weight $g'\{\widehat{m}\}^2 w_0$ as the bandwidth for the local one-step estimator $\widehat{\eta}_{OS}$ with weight w_0.

Theorem 1 gives a simple formula for estimating the standard errors of the one-step estimator $\widehat{\beta}(x)$. Let $\widehat{v}(x)$ be a consistent estimate of $v(x)$ defined in Theorem 1. Then the covariance matrix of the one-step estimator $\widehat{\beta}_{OS}(x)$ can simply be estimated by

$$\widehat{v}(x)\left(\sum_{i=1}^{n} K_h(X_i-x)\mathbf{X}_i\mathbf{X}_i^T\right)^{-1}\left(\sum_{i=1}^{n} K_h^2(X_i-x)\mathbf{X}_i\mathbf{X}_i^T\right)\left(\sum_{i=1}^{n} K_h(X_i-x)\mathbf{X}_i\mathbf{X}_i^T\right)^{-1} \tag{6.2}$$

or by

$$\begin{aligned}[\widehat{v}(x)]^{-1}&\left(\sum_{i=1}^{n} q_2(\mathbf{X}_i^T\widehat{\beta}_0,Y_i)K_h(X_i-x)\mathbf{X}_i\mathbf{X}_i^T\right)^{-1}\left(\sum_{i=1}^{n} K_h^2(X_i-x)\mathbf{X}_i\mathbf{X}_i^T\right)\\ &\times\left(\sum_{i=1}^{n} q_2(\mathbf{X}_i^T\widehat{\beta}_0,Y_i)K_h(X_i-x)\mathbf{X}_i\mathbf{X}_i^T\right)^{-1},\end{aligned} \tag{6.3}$$

for some consistent estimate $\widehat{\beta}_0$. Note that the first and the last matrix in (6.3) is the same as that given in the definition of the one-step estimator (2.12). In the implementation, we use expression (6.3) to save computation. For the logistic regression and Poisson regression, $v(x)^{-1} = p(x)q(x)$ and $v(x)^{-1} = \lambda(x)$, respectively.

The bias vector can also be assessed via the ideas outlined in Fan, Farmen and Gijbels (1996). We omit the detail here.

7 Examples

We now illustrate our bandwidth selection rules via both simulations and real data examples.

7.1 Simulations

To examine the efficacy of the above bandwidth selection rule, we revisit Examples $1-6$ with sample size $n = 250$ (The results for $n = 500, 1000$ are similar and hence omitted). We take the weight function w_0 to be the indicator function on the interval $[-2, 2]$ where the curve to be estimated and employ the pre-asymptotic substitution bandwidth selection method of Fan and Gijbels (1995).

For a given sample, we compute the RASE defined by (5.1) for the one-step estimator $\widehat{\eta}_{OS}$ using the bandwidth selection rule given in Section 6 and for the maximum local likelihood estimator $\widehat{\eta}_{MLE}$ using the asymptotic optimal bandwidth h_{opt}. The latter estimator can be regarded as an ideal estimator and serves as a benchmark. The ratios of RASE of $\widehat{\eta}_{OS}$ to that of $\widehat{\eta}_{MLE}$ are depicted in Figure 5 (a). It demonstrates clearly that our data-driven one-step estimator performs as well as the ideal estimator. Figures 5 (b) – (c) show a typical estimated curve by using the data-driven one-step estimator and the ideal estimator: The sample is chosen such that the resulting data-driven

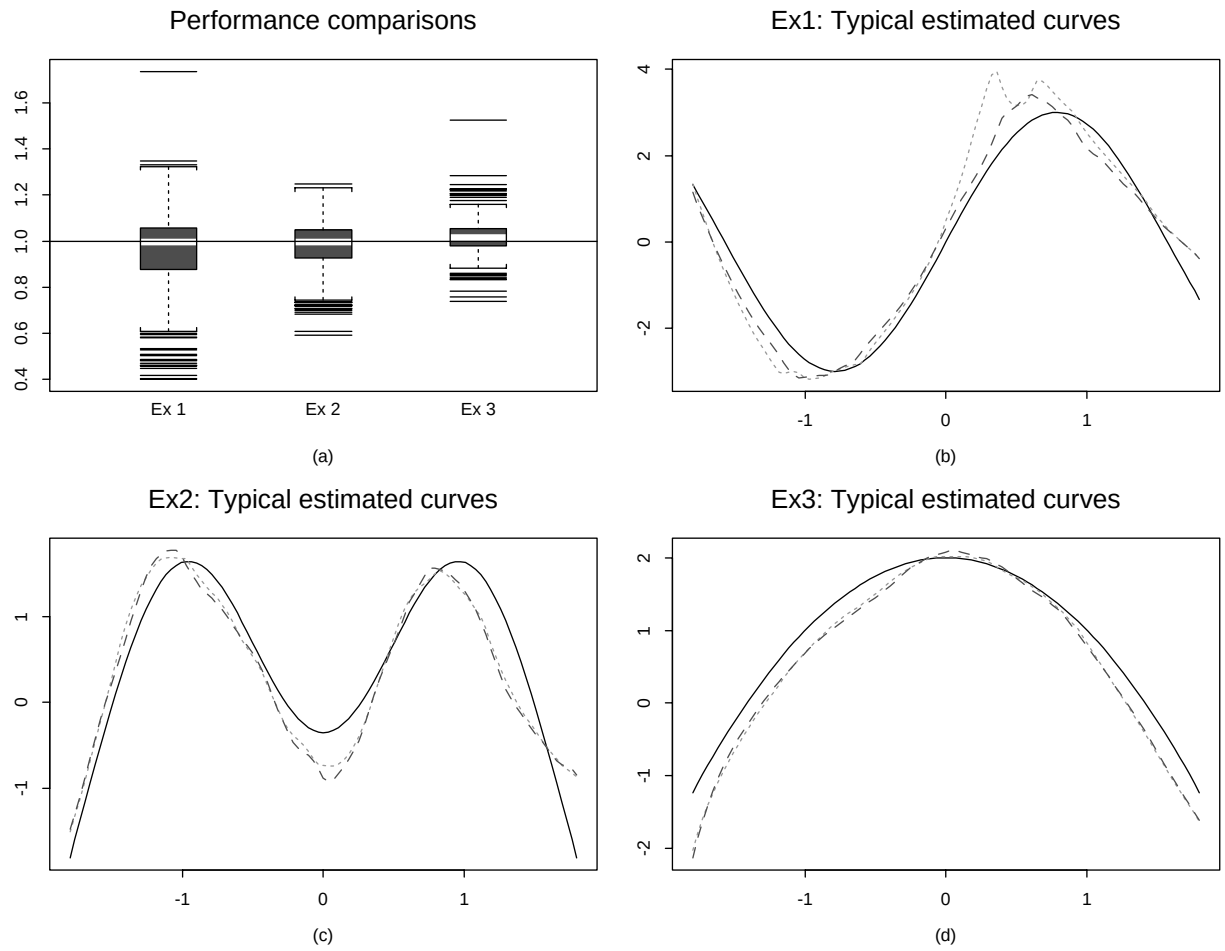

Figure 5: *Comparisons of the data-driven one-step estimator with the ideal estimator for logistic regression model. (a) Boxplot of the ratios of the RASE of the data-driven one-step estimator to that of the ideal estimator. (b), (c) and (d). A typical estimate curve by using the data-driven one-step estimator (long-dashed curve) and the ideal estimator (short-dashed curve).*

one-step estimator has the median performance among 400 samples. Once the sample is selected, the data-driven one-step estimator and the ideal estimator are applied to the same sampled data and the resulting curves are depicted in Figure 5.

Figure 6 summarizes similar results for the Poisson models given in Examples $4-6$ with sample size $n = 250$. For larger sample sizes ($n = 500$ and $n = 1000$), the results are very similar and hence are omitted. Figure 6 shows that the one-step estimator with the proposed bandwidth selection rule performs comparably with the ideal estimator.

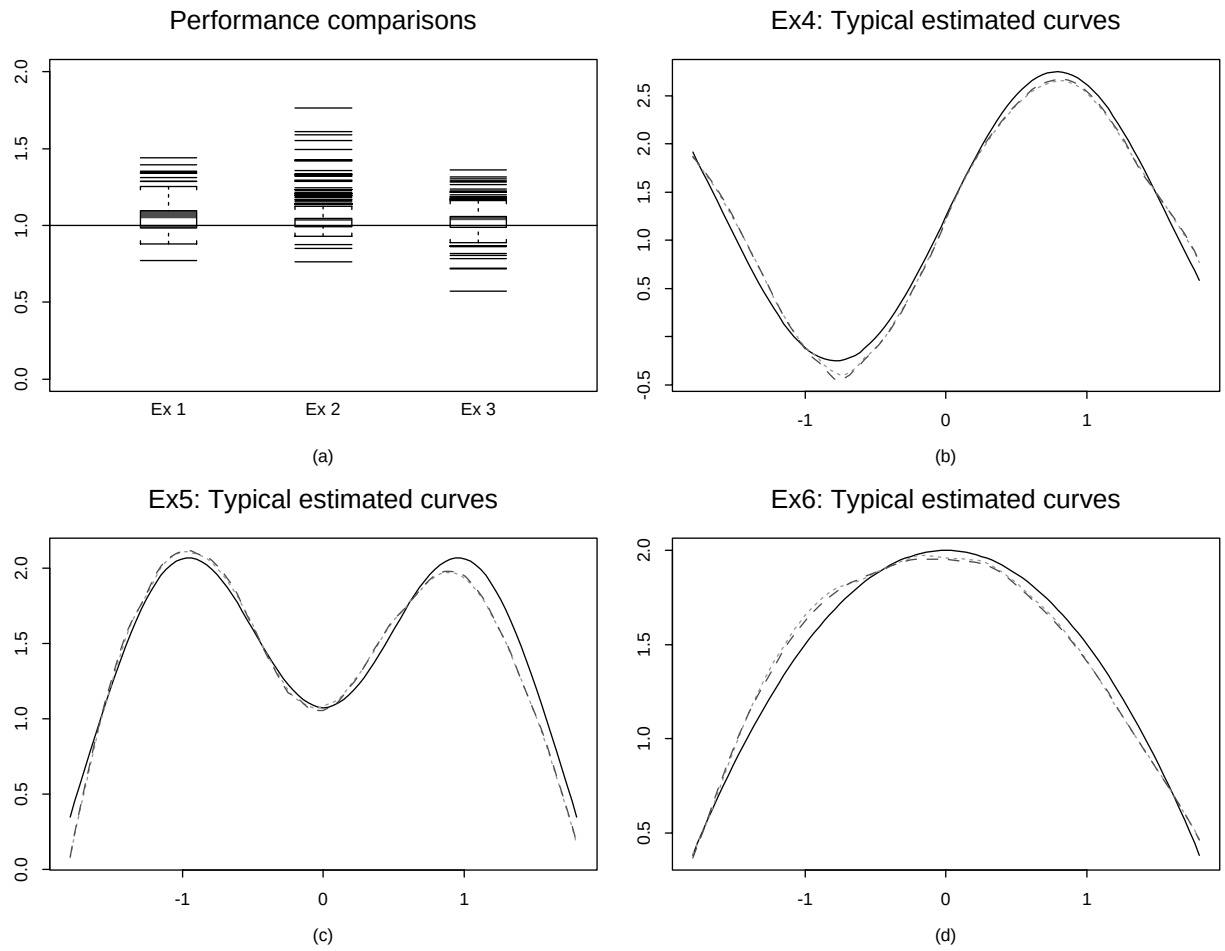

Figure 6: *Comparisons of the data-driven one-step estimator with the ideal estimator for Poisson regression model. Similar captions to Figure 5 are used.*

7.2 Applications to an environmental dataset

The data set used here consists of a collection of daily measurements of pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995. Of particular interest is to study the association between levels of pollutants and number of daily hospital admissions for circulation and respiration. The data were kindly provided by Professor T.S. Lau of the Chinese University of Hong Kong. It is known that measurements on pollutants are highly correlated. As an example, we consider how the probability of high level Sulphur Dioxide SO_2 (with values $> 20\mu g/m^3$) is associated with level of pollutant Nitrogen Dioxide NO_2 (in $\mu g/m^3$). Figure 7 gives the estimated conditional probability and its logit-transform which is monotone in the most part of the region except a dip occurs around 45. The bandwidth $h = 10.1$ was chosen by our software. The dashed lines are the estimated function plus (or minus) two estimated standard

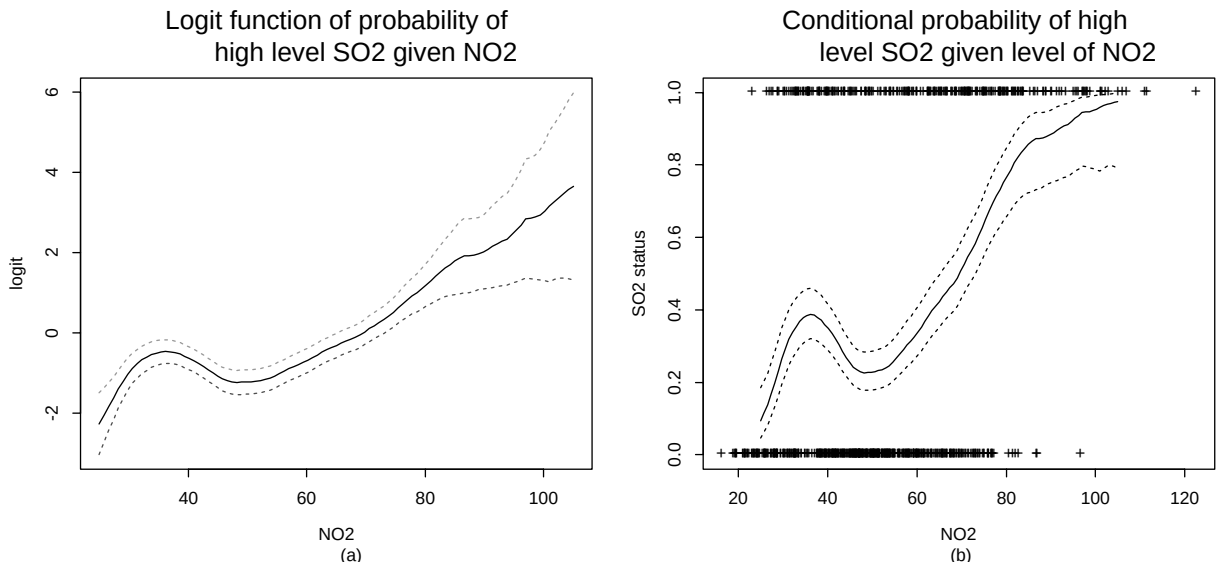

Figure 7: *Logistic regression of the status of high level SO_2 on the level of NO_2. (a) Estimated logit function of the conditional probability. (b) Estimated probability function and observed data.*

errors at each given points. They give us rough ideas of the estimation error. The estimates are obtained by using the one-step local quasi-likelihood method with bandwidth selected automatically by data.

We now study how the number of hospital admissions depend on time and the level of NO_2. The model used here is the Poisson distribution. Figures 8 (a) and (b) depict how the number of hospital admissions changes over time. Seasonal patterns can be observed. The bandwidth $h = 29.5$ was selected. We have also plotted the level of NO_2 against the time variable and found a similar seasonal pattern. This motivates us to examine the relationship between the number of hospital admissions and the level of NO_2. The results are depicted in Figure 8 with bandwidth $h = 17.9$. The figures show an increasing trend and indicates the extent to which the level of pollutant NO_2 affects the number of hospital admission. It is reasonable to think that the daily number of admissions is correlated and that it takes some time for pollutants to affect the circulatory and respiratory system. These aspects have not yet addressed in the above preliminary analyses.

8 Concluding remarks

Local maximum quasi-likelihood estimation is a useful technique for nonparametric fitting of generalized linear models. The computation cost and issues of algorithmic convergence make the procedure less appealing, particularly when the procedure is iteratively used such as in the backfitting

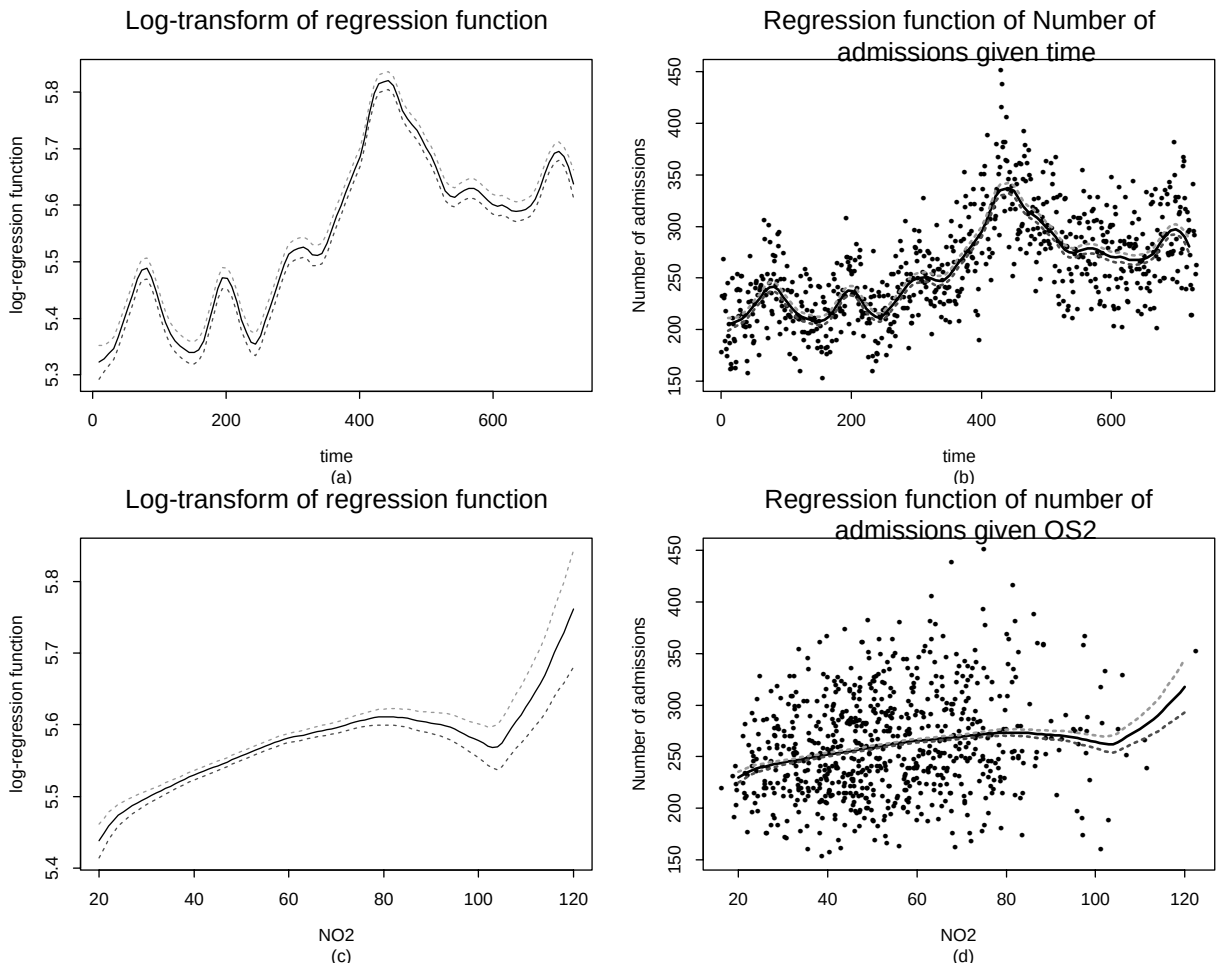

Figure 8: *Regression of number of hospital admissions on time and level of NO_2. (a) Estimated log-regression function of number of admissions given time. (b) Estimated regression function of number of admissions given time. The points are the raw data. (c) and (d). Similar captions to (a) and (b) are used. The dashed lines are the estimated function plus (or minus) twice estimated standard error at each given points.*

algorithm, cross-validation and bootstrapping. One simple approach to overcoming these problems is the one-step local quasi-likelihood estimation. In light of the asymptotic theory and numerical examples given in Sections 5 and 6, one can conclude that the one-step procedure performs at least as well as the fully iterative local maximum quasi-likelihood method. In other words, the one-step procedure reduces the computation burden of the fully-iterative method without deteriorating its performance. The idea outlined in this paper should be applicable in other context.

In the implementation of local linear regression smoother, ridge regression techniques are often employed to avoid singularities of design matrices. In light of examples in Sections 5 and 6, we conclude that our choice of ridge parameters is simple and useful.

References

Aragaki, A. and Altman, N. (1997), "Local polynomial regression for binary response". *Manuscript.*

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

Cleveland, W.S. (1979). "Robust locally weighted regression and smoothing scatterplots" *J. Amer. Statist. Assoc.*, 74, 829–836.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992), "Local regression models", In *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309–376, Wadsworth & Brooks, California.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1996), "Nonparametric estimation via local estimating equations", *Unpublished manuscript.*

Fan, J. (1992), "Design-adaptive nonparametric regression", *J. Amer. Statist. Assoc.*, 87, 998–1004.

Fan, J. (1993), "Local linear regression smoothers and their minimax efficiency", *Ann. Statist.*, **21**, 196–216.

Fan, J., Farmen, M. and Gijbels, I. (1996), "A blueprint of local maximum likelihood estimation", Discussion Paper #9604, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

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

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling 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.

Gasser, T. and Müller, H.-G. (1979), "Kernel estimation of regression functions", In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, **757**, 23–68. Springer-Verlag, New York.

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

Hastie, T.J. and Loader, C. (1993), "Local regression: automatic kernel carpentry (with discussion)", *Statist. Sci.*, 8, 120–143.

Hastie, T.J. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Hunsberger, S. (1994), "Semiparametric regression in likelihood-based models", *J. Amer. Statist. Assoc.*, 89, 1354–1365.

Jones, M.C. (1997), "A variation on local linear regression", *Statistica Sinica*, to appear.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996), "A brief survey of bandwidth selection for density estimation", *J. Amer. Statist. Assoc.*, **91**, 401-407.

Ruppert, D. (1995), "Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation", *Technical Report* #1137, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York.

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.

Ruppert, D. and Wand, M.P. (1994), "Multivariate weighted least squares regression", *Ann. Statist.*, 22, 1346–1370.

Severini, T.A. and Staniswalis, J.G. (1994), "Quasi-likelihood estimation in semiparametric models", *J. Amer. Statist. Assoc.*, **89**, 501–511.

Seift, B. and Gasser, Th. (1996), "Finite-sample variance of local polynomials: Analysis and solutions", *J. Amer. Statist. Assoc.*, **91**, 267-275.

Simonoff, J.S. (1995), *Smoothing methods in Statistics*, Springer, New York.

Staniswalis, J.G. (1989), "The kernel estimate of a regression function in likelihood-based models", *J. Amer. Statist. Assoc.*, **84**, 276–283.

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*, Chapman and Hall, London.

Appendix: Proofs of Theorems 1 and 2

Theorem 2 can be proved similarly to as Theorem 1. For brevity, we only outline the proof of Theorem 1.

For each given point x, the following conditions are needed:

(1) The functions $f(.), \eta^{(p+1)}(.), V'(m(.))$, $g'(m(.))$ exist and are continuous at the point x and $f(x) > 0$;

(2) $E(Y^{2+\varepsilon}|X = .)$ for some $\varepsilon > 0$ is bounded in a neighborhood of the point x;

(3) K has a bounded support;

(4) $h \to 0$ and $nh \to \infty$.

Note that the set of conditions is weaker than that given in Fan, Heckman and Wand (1995). In particular, we do not require $Q(x, y)$ be concave in x and do not impose conditions on V'', g''' and $E(Y^4|X = .)$. Condition (3) is imposed for technical convenience and can be removed at lengthier proofs via imposing some tail conditions on K and functions η, f, V and g.

Proof of Theorem 1. For simplicity, we will drop the dependency of $\beta_0(x), \widehat{\beta}_0(x)$ and $\widehat{\beta}_{OS}(x)$ on x. In the sequel, we will show that

$$H^{-1}\ell''(\widehat{\beta}^*)H^{-1} = -a(x)S + o_p(1) \tag{A.1}$$

for any $\widehat{\beta}^*$ satisfying $H(\widehat{\beta}^* - \beta_0) = o_p(1)$, where

$$a(x) = \frac{f(x)}{V(m(x))g'(m(x))^2};$$

and that

$$\sqrt{nh}H^{-1}\{\ell'(\beta_0) - E\ell'(\beta_0)\} \to N(0, b(x)S^*) \tag{A.2}$$

with

$$b(x) = \frac{\mathrm{var}(Y|X=x)f(x)}{V(m(x))^2 g'(m(x))^2},$$

and that

$$H^{-1}E\ell'(\beta_0) = c(x)\mu h^{p+1} + o(h^{p+1}) \tag{A.3}$$

where

$$c(x) = \frac{\eta^{p+1}(x)f(x)}{V(m(x))g'(m(x))^2(p+1)!}.$$

Assume that (1)-(3) hold. Then, by Taylor's expansion

$$\ell'(\widehat{\beta}_0) = \ell'(\beta_0) + \ell''(\widehat{\beta}^*)(\widehat{\beta}_0 - \beta_0) \tag{A.4}$$

where $\widehat{\beta}^*$ lies between β_0 and $\widehat{\beta}_0$ and hence satisfies

$$H(\widehat{\beta}^* - \beta_0) = O_p\{h^{p+1} + (nh)^{-\frac{1}{2}}\} \tag{A.5}$$

according to (3.3). Combination of (A.1) and (A.4) leads to

$$\widehat{\beta}_{OS} - \beta_0 = a(x)^{-1}H^{-1}\{S^{-1} + o_p(1)\}H^{-1}\ell'(\beta_0) + o_p(\widehat{\beta}_0 - \beta_0).$$

Using this and (A.2), we obtain

$$H(\widehat{\beta}_{OS} - \beta_0) = a(x)^{-1}S^{-1}H^{-1}\ell'(\beta_0) + o_p\{h^{p+1} + (nh)^{-\frac{1}{2}}\}.$$

Therefore, by (A.3) we have

$$\begin{aligned} & H(\widehat{\beta}_{OS} - \beta_0) - a(x)^{-1}c(x)S^{-1}\mu h^{p+1} \\ = \;& a(x)^{-1}S^{-1}H^{-1}\{\ell'(\beta_0) - E\ell'(\beta_0)\} + o_p\{h^{p+1} + (nh)^{-\frac{1}{2}}\}. \end{aligned} \tag{A.6}$$

The conclusion follows from (A.2) and (A.6).

We now establish results (A.1)-(A.3). Note that a typical element in the matrix $H^{-1}\ell''(\widehat{\beta}^*)H^{-1}$ is

$$S^*_{n,k} = n^{-1}\sum_{i=1}^{n} q_2(\widehat{\beta}^*_0 + \cdots + \widehat{\beta}^*_p(X_i - x)^p, Y_i)K_h(X_i - x)\left(\frac{X_i - x}{h}\right)^k$$

To prove (A.1), we need only to show that

$$S^*_{n,k} = -a(x)\mu_k + o_p(1), \quad \text{for} \quad k = 0, 1, \ldots, 2p. \tag{A.7}$$

By Conditions (1) –(4), one can easily show that

$$S^*_{n,k} = n^{-1}\sum_{i=1}^{n} q_2(\eta(x), Y_i)K_h(X_i - x)\left(\frac{X_i - x}{h}\right)^k + o_p(1).$$

Let $d(x) = [g^{-1}(x)/V\{g^{-1}(x)\}]'$. Then, by calculating the mean and variance of the leading term in the above expression, one can obtain

$$\begin{aligned} S^*_{n,k} &= Eq_2(\eta(x), Y)K_h(X - x)\left(\frac{X - x}{h}\right)^k + o_p(1). \\ &= E\left\{[m(X) - m(x)]d(m(x)) - \frac{1}{V(m(x))g'(m(x))^2}\right\}K_h(X - x)\left(\frac{X - x}{h}\right)^k + o_p(1) \\ &= -a(x)\mu_k + o_p(1). \end{aligned}$$

This proves (A.7) and hence (A.1).

The result (A.2) follows from the multivariate central limit theorem and is established in Lemma 2 of Fan, Heckman and Wand (1995). To prove (A.3), we need only to show that

$$\begin{aligned} & En^{-1}\sum_{i=1}^{n} q_1(\bar{\eta}(X_i), Y_i)K_h(X_i - x)(\frac{X_i - x}{h})^k \\ = \;& c(x)\mu_{k+p+1}h^{p+1} + o(h^{p+1}) \quad \text{for} \quad k = 0, 1, \ldots, p, \end{aligned} \tag{A.8}$$

where $\bar{\eta}(X_i) = \beta_0 + \cdots + \beta_p(X_i - x)^p$. Indeed, the right hand side of (A.8) is given by

$$E\frac{g^{-1}(\eta(X)) - g^{-1}(\bar{\eta}(X))}{V(g^{-1}(\bar{\eta}(X)))}\frac{1}{g'(g^{-1}(\bar{\eta}(X)))}K_h(X - x)\left(\frac{X - x}{h}\right)^k$$

and by the mean-value theorem, it is equal to

$$E\frac{\eta(X) - \bar{\eta}(X)}{g'(g^{-1}(\xi))V(g^{-1}(\bar{\eta}(X)))}\frac{1}{g'(g^{-1}(\bar{\eta}(X)))}K_h(X - x)(\frac{X - x}{h})^k.$$

where ξ lies between $\eta(X)$ and $\bar{\eta}(X)$. By using Taylor's expansion and continuity assumptions, one can easily show that the last expectation is given by

$$c(x)\mu_{k+p+1}h^{p+1} + o(h^{p+1}).$$

This proves (A.8) and completes the proof of Theorem 1.