Fast bandwidth selection for generalized local linear models
Preliminary version

FREDERIC UDINA*

April 1, 2000

Abstract

In generalized linear models, the linear component is replaced by a local polynomial estimator. This requires a bandwidth to be selected. In the asymptotically optimal bandwidth expression some terms depend on the unknown regression function. The *rule of thumb* estimates these terms by fitting a global polynomial. We analize the performance and limitations of this rule in a Montecarlo study covering Bernoulli-logit and Poisson-Log models. To solve the problems that arise we devise a block method where the number of blocks is automatically selected by an information criterion.

1 Introduction

Local polynomial estimators are a de-facto standard in nonparametric curve estimators. In the simplest case the regression model

$$Y = m(X) + \epsilon$$

is considered where the regression function m is assumed to be smooth enough. To estimate $m(x_0)$ or any of its derivatives in a fixed point x_0, the polynomial

$$P_0(x) = \sum_{j=0..p} \beta_j (x - x_0)^j$$

*Departament d'Economía i Empresa, Universitat Pompeu Fabra. E-mail address: `udina@upf.es`.
Thanks to Jianqing Fan for the starting impulse. Some comments from Steve Marron, Matt Wand, Gabor Lugosi, and Albert Satorra have been also useful. Research work was supported by spanish DGES grant PB96.0300.

that minimizes the sum of squares

$$\sum_{i=1..n} K_h(X_i - x_0)\{Y_i - P_0(X_i)\}^2$$

is to be found, where K is a kernel function (usually a simetric density around zero) and $K_h(\cdot) = h^{-1}K(\cdot/h)$ denotes rescaling by the smoothing parameter h, also called the bandwidth. The order of the fitted polynomial, p, is usually chosen looking at the order of the derivative of m to be estimated. Typically, $p = 1$ is used to estimate $m(x_0)$ and the estimated function is called a local linear fitting. One of the advantages of this technique is that automatically selects the complexity of the model: a small bandwidth is equivalent to an interpolation of the data, while a big bandwidth gives a global fitting, a parametric model.

Local polynomial fitting is not a recent technique. Origins can be traced back to some moving average techniques in 1829 (CLEVELAND AND LOADER[1996]). Early development can be found in STONE[1977, 1980] and CLEVELAND[1979]. TIBSHIRANI AND HASTIE[1987] use local polynomials in generalized models calling it *local likelihood estimation*. Instead of using a kernel function, in this early papers the locality is given by the use of a fixed proportion of the sample: the data points closest to point where the estimation is to be computed. In the last decade a new approach appear, based on FAN[1992, 1993], that moves local polynomial fitting to the first range in nonparametric methods because of its conceptual simplicity, good mathematical properties and relatively good computational properties. Among many papers in the stream, we are interested in FAN, HECKMAN AND WAND[1995], a generalization of TIBSHIRANI AND HASTIE[1987]. In that paper, as well as in FAN AND GIJBELS[1996], in the context of generalized local linear models, the *rule-of-thumb* method to select a bandwidth is mentioned: it consists in fitting a global polynomial first, and then use it to estimate the unknown terms in the asymptotically optimal bandwidth formula. It was used by HÄRDLE AND MARRON[1995] in kernel regression models (locally constant) and by RUPPERT, SHEATHER AND WAND[1995] for the non generalized model. The main goal of this paper is to study the performance of the rule in several models and to propose some procedures to improve it, for it doesn't work successfully when the target function is far from a polynomial behavior.

In section 2, we define the general model we use and some of the particular models it includes that are important for applications. In section 3 we discuss bandwidth selection and the proposed *rule-of-thumb*. In section 4 the performance of the rule is studied in several particular models, and in section 5 we propose a block dividing method to avoid the problems that appear when the target regression function cannot be fitted by a global polynomial.

2 General model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. real random variables distributed as (X, Y) with conditional mean and variance given by

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

where V is a variance function, $m(x)$ is the regression function to be estimated and σ is an unknown dispersion parameter.

This kind of models appear often in practice when data do have conditional distributions $Y|X$ in the Bernoulli or Poisson families, for example, where $V(t) = t(1-t)$ and $V(t) = t$ respectively.

More specifically, we will consider generalized models as in MCCULLAGH AND NELDER[1989], where $Y|X$ has conditional density in the exponential family

$$f(y|x) = \exp\left[\frac{y\theta(x) - b(\theta(x))}{a(\phi)} + c(y, \phi)\right] \tag{1}$$

where $\theta(x)$ is the so-called *canonical parameter* and ϕ is called *dispersion parameter*. The functions a, b, c characterize the family.

Standard theory from McCullagh and Nelder introduces a *link function* g that links the mean regression $\mu(x)$ to a simpler space, usually a linear one:

$$\eta(x) = g(\mu(x)).$$

In particular, when $\eta(x) = \theta(x)$, the link function g is called canonical link and is given by $g = (b')^{-1}$, namely $b'(\theta(x)) = \mu(x)$. Note that the conditional density can be expressed as

$$\mu(x) = \mathbf{E}(Y|X = x) = b'(\theta(x)), \quad \text{var}(Y|X = x) = a(\phi)b''(\theta(x))$$

and that the conditional density can be written as

$$f(y|x) = \exp(\frac{y\, g_1(\eta) - b(g_1(\eta))}{a(\phi)} + c(y, \phi)), \quad g_1 := (g \circ b')^{-1} \tag{2}$$

In parametric modeling, $\eta(x)$ is often assumed to have a linear form. In non-parametric modeling, we relax this restrictive assumption and model the function $\eta(x)$ *locally* by a parametric family. The later assumption is valid for a wide class of functions admitting certain degree of smoothness. Let x_0 be a given point and assume that $\eta(x)$ has $(p+1)$th derivative. Then, locally around x_0 we have

$$\eta(x) = \sum_{0 \le j \le p} \frac{1}{j!}\eta^{(j)}(x_0)(x - x_0)^j + o((x - x_0)^{p+1})$$

From statistical modeling point of view, we write

$$\eta(x) = \eta_{\beta,p,x_0}(x) := \sum_{0 \le j \le p} \beta_j (x - x_0)^j,$$

where

$$\beta_j = \beta_j(x_0) := \frac{\eta^{(j)}(x_0)}{j!}, \qquad 0 \le j \le p$$

are parameters of our main interest.

So the locally weighted log-likelihood is given by (from (2))

$$l(\beta; x_0) = \sum_{1\leq i\leq n} K_h(X_i - x_0)\left\{\frac{Y_i g_1(\eta_{\beta,p,x_0}(X_i))}{a(\phi)} - b(g_1(\eta_{\beta,p,x_0}(X_i))) + c(Y_i, \phi)\right\}. \tag{3}$$

Let $\widehat{\beta}_0(x_0), \ldots, \widehat{\beta}_p(x_0)$ be the local maximum likelihood estimators, i.e. the maximizers of (3).

Then, it is clear that, for a chosen $\nu \leq p$,

$$\widehat{\eta}^{(\nu)}(x_0) = \nu!\widehat{\beta}_\nu(x_0).$$

As x_0 ranges the appropriate interval, we obtain a curve estimate of $\eta^{(\nu)}$.

From standard nonparametric regression theory (see Fan and Gijbels, 1995), if primary interest is to estimate the νth derivative of the curve, p is taken to be $\nu+1$ or occasionally $\nu+3$. It has been argued in FAN[1992] and RUPPERT AND WAND[1994] that $p-\nu$ should be odd. See also FAN, HECKMAN AND WAND[1995] and FAN AND GIJBELS[1996]. From this work, one can conclude that, asymptotically, when $p-\nu$ is even, taking $p+1$ bias is reduced without any increasing in variability. We will assume from here on that $p-\nu$ is an odd integer.

Note that when $h = 0$ the resulting estimate is essentially an interpolation, and when $h = \infty$, the estimate corresponds to the parametric polynomial model. Thus, the model complexity is effectively controlled by h.

2.1 Some particular models

The models we will test are particular instances of the general model described in the previous section. Now we describe the normal model, a Bernoulli-logit model for binary data and a Poisson-log model.

Normal model, identity link

In the normal model, studied in FAN AND GIJBELS[1996], with identity as canonical link, we recover a standard local polynomial regression, as far as variance is supposed to be constant, say σ^2, independent from mean regression. In this case, we have

$$a(\phi) = \phi = \sigma^2, \qquad b(\theta) = \theta^2/2, \qquad c(y,\theta) = -\frac{1}{2}[y^2/\sigma^2 + \log(2\pi\sigma^2),$$

and (3) becomes

$$l(\beta; x_0) = -\frac{1}{\sigma^2}\sum_{1\leq i\leq n} K_h(X_i - x_0)\{Y_i - \eta_{\beta,p,x_0}(X_i)\}^2.$$

Bernoulli model, logit link

In the Bernoulli-logit model, Y_i is binary with values 0 i 1. The conditional density is

$$f(y|x) = \mu(x)^y(1-\mu(x))^{1-y} = \exp\left[y\log\frac{\mu(x)}{1-\mu(x)} + \log(1-\mu(x))\right],$$
$$\mu(x) = E(Y|X) = P(Y=1|X=x).$$

So the canonical link is

$$\theta(x) = \log\frac{\mu(x)}{1-\mu(x)}, \qquad \text{i.e.} \quad g(t) = \log\frac{t}{1-t},$$

and thus we have

$$\mu(x) = \frac{e^{\theta(x)}}{1+e^{\theta(x)}}.$$

This gives us $\log(1-\mu(x)) = -\log(1+e^{\theta(x)})$, and so

$$b(t) = \log(1+e^t).$$

With this canonical link function, the local likelihood is

$$l(\beta; x_0) = \sum_{1\le i\le n} K_h(X_i - x_0)\left\{Y_i\eta_{\beta,p,x_0}(X_i) - \log(1+\exp(\eta_{\beta,p,x_0}(X_i)))\right\}.$$

The estimator for $\mu(x)$ will be $g^{-1}(\widehat{\eta}(x))$.

Poisson model, Log link

Suppose that $y|x$ follows a Poisson distribution with conditional mean $\mu(x)$. Then, the conditional density of y given x is

$$f(y|x) = e^{-\mu(x)}\frac{\mu(x)^y}{y!} = \exp\left[y\log\mu(x) - \mu(x) - \log(y!)\right].$$

The canonical link is

$$\theta(x) = \log\mu(x), \qquad \text{i.e.} \quad g(t) = \log t$$

with

$$b(\theta) = \mu(x) = e^\theta.$$

The local likelihood is now

$$l(\beta; x_0) = \sum_{1\le i\le n} K_h(X_i - x_0)\left\{Y_i\eta_{\beta,p,x_0}(X_i) - \exp(\eta_{\beta,p,x_0}(X_i)) - \log(Y_i!)\right\}$$

The estimator for $\mu(x)$ will be $g^{-1}(\widehat{\eta}(x))$.

3 Bandwidth selection

The performance of the proposed estimators depend, as has been noted, on the bandwidth parameter h. In this section we will propose a simple rule to choose a bandwidth based on the data. For the generalized model (1), while it is possible to propose more sophisticated rules for bandwidth selection, (see FAN, HECKMAN AND WAND[1995] i FAN, FARMEN AND GIJBELS[1996]), a simple rule is particularly attractive since the computation of the local likelihood itself is very expensive.

Let

$$\mu_j := \int_{\mathcal{R}} x^j K(x)\,dx, \qquad \nu_j := \int_{\mathcal{R}} x^j K^2(x)\,dx.$$

Consider also the $(p+1)\times(p+1)$ matrices

$$S_p = \begin{pmatrix} \mu_0 & \mu_1 & \dots & \mu_p \\ \mu_1 & \mu_2 & \dots & \mu_{p+1} \\ \dots & & \dots & \\ \mu_p & \mu_{p+1} & \dots & \mu_{2p} \end{pmatrix} \quad \text{i} \quad S_p^* = \begin{pmatrix} \nu_0 & \nu_1 & \dots & \nu_p \\ \nu_1 & \nu_2 & \dots & \nu_{p+1} \\ \dots & & \dots & \\ \nu_p & \nu_{p+1} & \dots & \nu_{2p} \end{pmatrix}$$

Let $K_\nu^*(t) = K(t)\sum_{0\le l\le p} s^{\nu l} t^l$ be the effective kernel for estimating the νth derivative, where $\{s^{\nu l}\}$ is the ν-th row of S_p^{-1}. (see FAN AND GIJBELS[1996] and GASSER, MÜLLER AND MAMMITZSCH[1985]). Then it is shown by FAN, HECKMAN AND WAND[1995] that, for any fixed point x in the interior of $\text{supp}(f)$, as $n \longrightarrow \infty$, assuming $h = h_n \longrightarrow 0$ and $nh^{2\nu+1} \longrightarrow \infty$,

$$\frac{\hat{\eta}^{(\nu)}(x) - \eta^{(\nu)}(x) - \mathsf{B}(x)}{\mathsf{S}(x)} \xrightarrow{D} N(0,1) \tag{4}$$

where

$$\mathsf{B}(x) = \frac{\nu!\eta^{(p+1)}(x)}{(p+1)!} h^{p+1-\nu} \int_{\mathcal{R}} t^{p+1} K_\nu^*(t)\,dt \tag{5}$$

and

$$\mathsf{S}^2(x) = \frac{\nu!^2}{nh^{2\nu+1}} \text{var}(Y|X=x) g'(\mu(x))^2 f(x)^{-1} \int_{\mathcal{R}} K_\nu^*(t)\,dt \tag{6}$$

$f(x)$ being the design density of the predictor variable.

Our bandwidth choice will be the minimizer of

$$\text{AMSE} = \int_{\mathcal{R}} \left\{ \mathsf{B}^2(x) + \mathsf{S}^2(x) \right\} f(x) W(x)\,dx$$

that turns out to be

$$h_{\text{opt}} = \alpha_\nu(K) \left[\frac{\int \text{var}(Y|X=x) g'(\mu(x))^2 W(x)\,dx}{n \int \left[\eta^{(p+1)}(x)\right]^2 W(x) f(x)\,dx} \right]^{\frac{1}{2p+3}} \tag{7}$$

where

$$\alpha_\nu(K) = \left[\frac{(2\nu+1)(p+1)!^2 \int K_\nu^{*2}(t)\,dt}{(2p+2-2\nu)\left[\int t^{p+1} K_\nu^*(t)\,dt\right]^2} \right]^{\frac{1}{2p+3}} \tag{8}$$

The values of $\alpha_\nu(K)$ are tabulated by FAN AND GIJBELS[1996].

Remark 1. $W(x)$ is some weight function that can be used to lower the influence of noise in the ends of the working interval or the effect of points with high variance. In section4.3 a we analyze the influence on the performance of the estimator.

Remark 2. As noted above, $p-\nu$ is always odd, so the denominator in (8) does not vanish.

Remark 3. Note that the integrals in (8) are respectively the $(\nu+1)$ diagonal element of $S^{-1}S^*S^{-1}$ and the $\nu+1$ element of $S^{-1}(\mu_{p+1},\dots,\mu_{2p})$ so that they can be evaluated easily. When $p=1, \nu=0, K^*_\nu=K$.

Remark 4. If the kernel function is chosen with canonical rescaling (see Marron and Nolan (1988)) the constant α in (8) and h_{opt} will not depend on the kernel family in use.

There are some unknowns both on the numerator and on the denominator of (7) all depending essentially on $\eta(x)$ and its $(p+1)$-th derivative. To estimate these unknowns, we fit a global polynomial of degree $p+3$, namely,

$$\widehat{\eta}_q(x) = \sum_{0\le j\le p+3} \widehat{\gamma}_j x^j$$

and use it to obtain a *pilot estimate* of $\eta^{(p+1)}(x)$.

In this way, the bias factor in (7)

$$b_\nu := n\int_{\mathcal{R}} \left[\eta^{(p+1)}(x)\right]^2 W(x)f(x)\,dx$$

can be estimated by

$$n\int_{\mathcal{R}} \left[(p+1)!\widehat{\gamma}_{p+1} + (p+2)!\widehat{\gamma}_{p+2}\,x + (p+3)!\widehat{\gamma}_{p+3}\,x^2/2\right]^2 W(x)f(x)\,dx$$

The sample version of this is

$$\widehat{b}_\nu = (p+1)!^2 \sum_{1\le i\le n} \left[\widehat{\gamma}_{p+1} + (p+2)\widehat{\gamma}_{p+2}X_i + \frac{(p+3)(p+2)}{2}\widehat{\gamma}_{p+3}X_i^2\right]^2 W(X_i) \quad (9)$$

Under the class of models we are considering, $\text{var}(Y|X=x)$ and $g'(\mu(x))^2$ both depend on μ and can be estimated by plugging in the pilot estimate. This way, we also obtain an estimate of the numerator in (7), say, $\widehat{\nu}_\nu$. This leads to the simple rule

$$\widehat{h}_{\text{opt}} = \alpha_\nu(K)\left(\frac{\widehat{\nu}_\nu}{\widehat{b}_\nu}\right)^{\frac{1}{2p+3}} \quad (10)$$

It should be noted that fitting the polynomial η_q should be done by maximizing the likelihood in the generalized model context. The algorithm, based in the Fischer's scoring method is an adaptation of the one described for linear models in MCCULLAGH AND NELDER[1989], pàg 40.

1. Take an initial estimate for the parameters $\beta_j, \ j = 0 \dots p+3$ to form the initial polynomial (the strategy can vary according the particular model).
2. Compute the model prediction

$$\eta_i = \sum_{j=0..p+3} \beta_j \, x^j, \ \ i = 1 \dots n$$

and the corresponding values $\mu_i = g^{-1}(\eta_i)$.

3. Adjust the prediction with

$$z_i = \eta_i + (y_i - \mu_i) g'(\mu_i).$$

4. Now compute the weights

$$W_i = [g'(\mu_i)^2 V(\mu_i)]^{-1}$$

where $V(\mu_i)$ is the variance for μ_i.

5. The new values for $\beta_j, \ j = 0 \dots p+3$ are computed by regressing z_i on the vector $1, x_i, \dots, x_i^{p+3}$ using weighted least squares with weights W_i.
6. Iterate steps 1 to 5 unless the new coefficients are similar to the previous ones. Otherwise the algorithm is finished.

3.1 Some particular cases

We now examine the bandwidth rule (10) for the models described in section 2.1.

Normal model, identity link

In this case,

$$\hat{\nu}_v = \hat{\sigma}^2 \int_{\mathcal{R}} W(x) dx,$$

where $\hat{\sigma}^2$ is an estimator obtained from the $(p+3)$-degree fitting. The *rule-of-thumb* in this simple case has been analyzed in RUPPERT, SHEATHER AND WAND[1995] and FAN AND GIJBELS[1995b]. See also FAN AND GIJBELS[1996], pags. 110–113.

Bernoulli model, logit link

With this model we have $\text{var}(Y|X = x) g'(\mu(x))^2 = (\mu(x)(1 - \mu(x)))^{-1}$. So,

$$\hat{\nu}_v = \int_{\mathcal{R}} \frac{W(x)}{\hat{\mu}(x)(1 - \hat{\mu}(x))} dx$$

where $\hat{\mu}$ is obtained from the pilot parametric fit of degree $p + 3$.

Poisson model, Log link

Here, $\text{var}(Y|X=x)g'(\mu(x))^2 = \mu(x)^{-1}$. So,

$$\hat{\nu}_\nu = \int_{\mathcal{R}} \frac{W(x)}{\hat{\mu}(x)} dx$$

where also here $\hat{\mu}$ is obtained from the pilot parametric fit of degree $p+3$.

4 Simulations

In this section we describe some Monte Carlo simulations conducted to evaluate the performance of the described rule for bandwidth selection using small or medium sample sizes. We work with the models described above, Bernoulli-logit and Poisson-log.

In all cases the target is the transformed regression function, $\eta(x)$, so $p=1$, and we take $\nu = 0$: we are using local linear estimation. Our goal is the asymptotically optimal bandwidth (7) and we measure the performance of a single result $\widehat{h}_{\text{opt}}$ by the relative error given by

$$\frac{\widehat{h}_{\text{opt}} - h_{\text{opt}}}{h_{\text{opt}}.}$$

We compute also the mean integrated squared error

$$\text{MISE} = \int_{\mathcal{R}} (\hat{\eta} - \eta)^2$$

We use the Bartlett-Epanechnikov kernel function

$$K(x) = \frac{3}{4}(1-x^2)_+.$$

We take as design density a uniform distribution in $[-2, 2]$, that is, the marginal distribution for X is $U(-2, 2)$. Sample sizes are $n = 250, 500, 1000$. Number of Monte Carlo samples is $B = 1000$. For each sample, we apply the algorithm described in section 3 and compute from the pilot estimate a bandwidth h_{opt}. We compute the relative error for this estimated bandwidth. The target function estimate is evaluated over a grid of points $\{g_i\}$, $i \in \{1, 2, \ldots, G\}$ and the MISE is approximated by

$$\frac{4}{G}\sum_{i=1}^{G} [\hat{\eta}(g_i) - \eta(g_i)]^2.$$

We use $G = 300$.

To summarize the obtained results, we present a density estimation both for the relative errors and the MISE values. Densities are estimated by a kernel method, with bandwidth obtained by the normal reference rule. This is justified by the asymptotically normal behavior shown in (4).

Example	$\eta(x)$
1	$0.5x^3$
2	$7\exp(-(x+1)^2) + \exp(-(x-1)^2) - 5.5$
3	$x + 2\exp(-8x^2)$

Figure 1: Example functions for the Bernoulli-Logit model.

4.1 Bernoulli-Logit models

The Bernoulli-logit model has been studied for the functions in figure 1.

The first function has been chosen to test the performance of the selector when the target function is polynomial against other cases where it is hard to fit a global polynomial. Example 2 is the same as in example 2 of FAN, FARMEN AND GIJBELS[1996] while example 3 is the same FAN AND GIJBELS[1995b] use in the normal case. The same function is used in HÄRDLE AND MARRON[1995] as an example of depart from a polynomial model.

In figure 2 we show typical results for the target function estimation with an intermediate sample size $n = 500$. We remark how the global polynomial fit fails to adjust the target function in example 3. This function really shows an atypical behavior for a polynomial.

In figure 3 we summarize results obtained in the simulations through the relative errors for the estimated optimal bandwidths and the estimated function integrated squared errors. Obtained errors are reasonable, and it can be clearly seen that in examples 2 and 3, where the target function is far from polynomial, increasing the sample size reduces the error variability but can't really reduce the bias because of the difficulty to fit an appropriate global polynomial. It can also be seen how the function curvature is underestimated in example 2, which function has relevant convexity changes. For example 3, the estimator sees a bigger curvature than real because of the difficulty for a polynomial to mimic the target function.

4.2 Poisson-log model

The Poisson model has been studied for the functions in figure 4, with $p = 1$ i $\nu = 0$ as before.

These function were studied in FAN AND CHEN[1997] and present different levels of difficulty to estimate, as shown by signal to noise ratios. We compute this indicator by

$$\frac{\mathrm{var}(\mu(x))}{\mathrm{var}(Y)}$$

where $\mu(x)$ is the regression function as described in the general model. We remark that the function in example 3 is specially difficult to estimate by the global polynomial we use as a pilot estimator.

Simulation parameters are as described in previous section. In figure 5 we show typical estimates obtained, as in the previous section. Note that the fore-mentioned

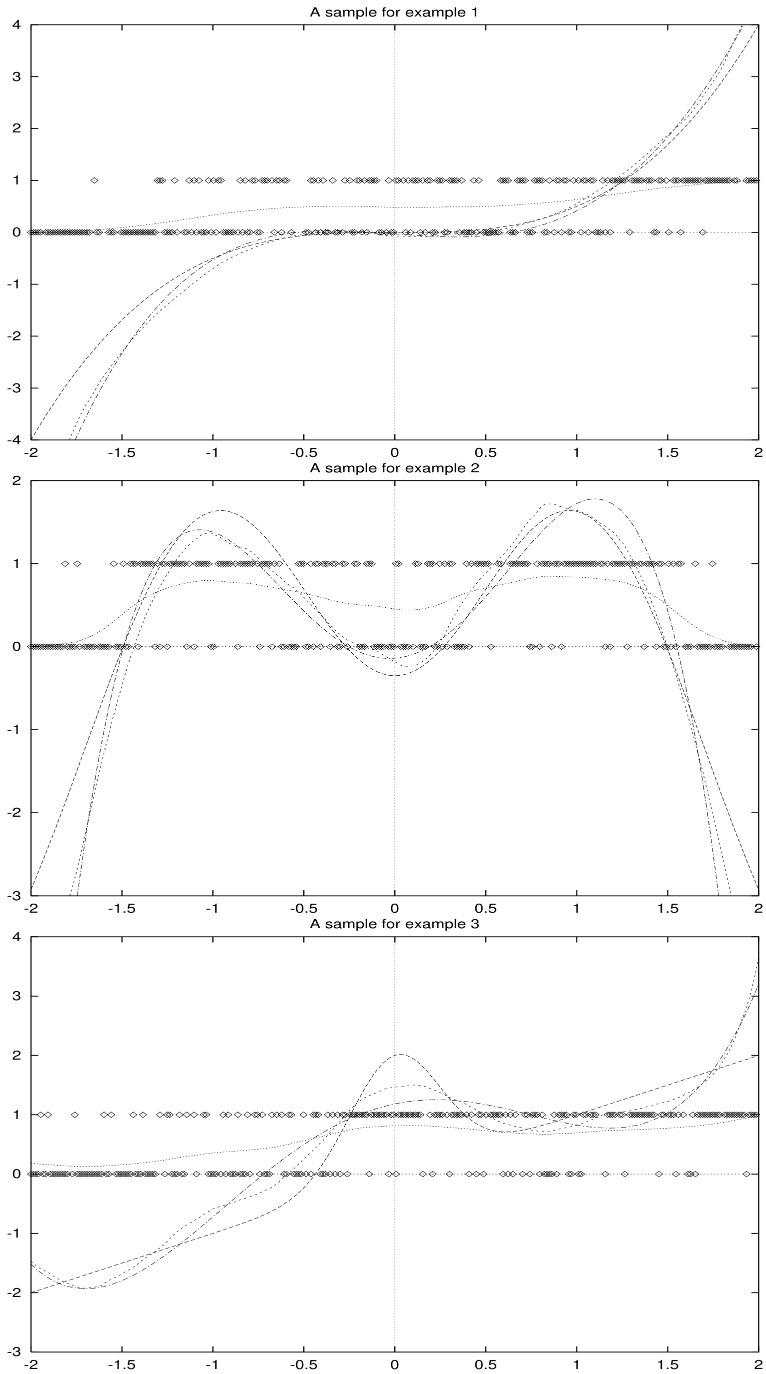


Figure 2: Typical simulations for the three examples of Bernoulli model described in the text. Here, $n = 500$. Only some sample points are shown for clarity. Target function (long dashed line), fourth order polynomial global fit (dotted-dashed line) and estimated function (short dashed line) are shown. The dotted line is the function $\widehat{\mu}(x)$ that estimates the probability $P[Y = 1]$.

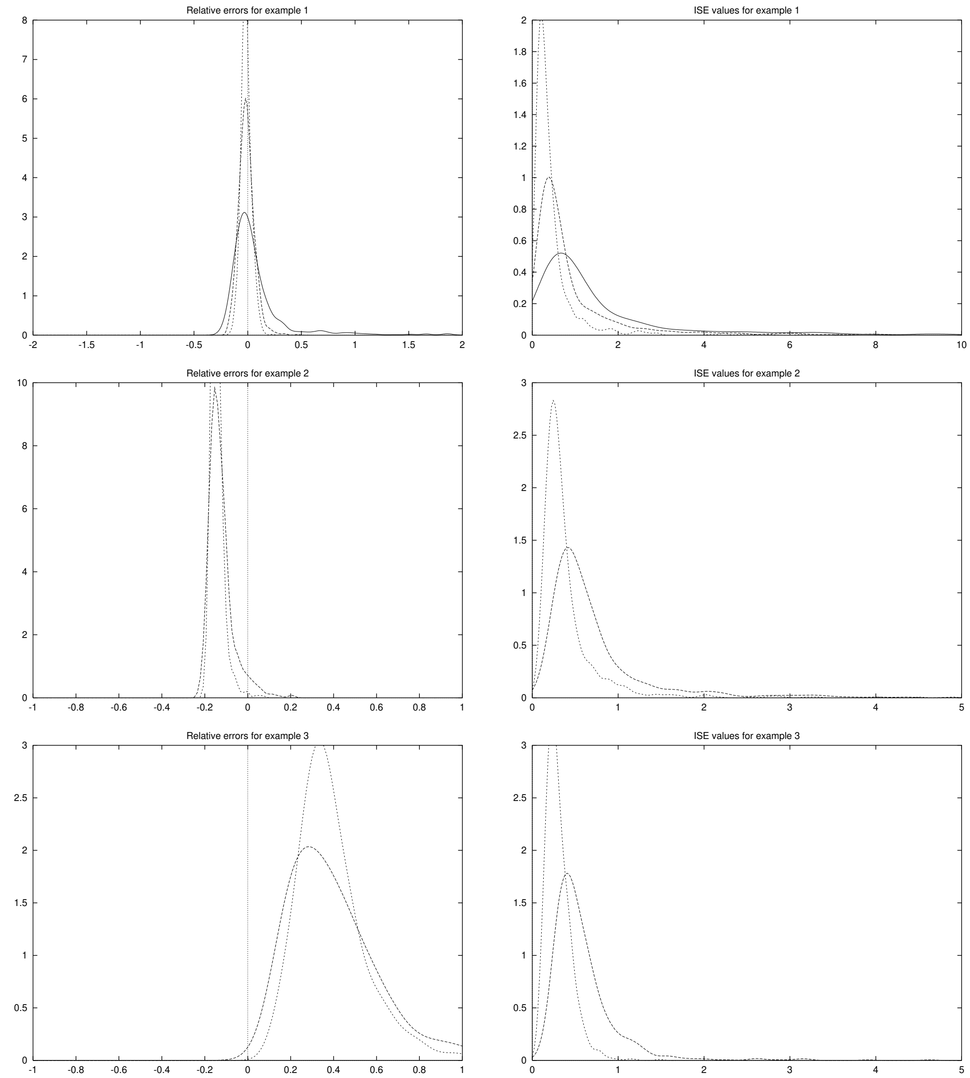


Figure 3: Result summaries for three examples of the Bernoulli-logit model as described in the text. Left: density estimates of relative errors in estimating h_{opt}. Right: density estimates of integrated squared errors for $\hat{\eta}(x)$ as estimator of the target function. Solid line is for sample size $n = 250$ in example 1, dashed line is for $n = 500$ and dotted line is for $n = 1000$.

Example	$\eta(x)$	Signal ratio
1	$1.25 + 1.5\sin(2x)$	$3/4$
2	$2 - .5x^2$	$1/2$
3	$3.5[\exp(-(x+1)^2) + \exp(-(x-1)^2)] - 1.5$	$1/2$

Figure 4: Example functions for Poisson models.

difficulty for example 3 is clearly revealed in this figure, as was in the example 3 for the Bernoulli-logit model. In the graphics, we omit some data point for better visual interpretation.

In figure 6 we show density estimates for the obtained relative errors in estimating the optimal bandwidth and MISE values. Results give similar conclusions as in previous section for the Bernoulli-Logit model. For the polynomial function example, quality of the estimate is quite blurred by the low signal level and the concavity of the target function. For examples 2 and 3, the bias reflects the estimation difficulties, giving under-smoothed estimates.

4.3 Role of the weighting function

Note first that the asymptotical validity of the optimal bandwidth in (7) can be *too asymptotical* in some models due to the variance. When the target function is not polynomial, the estimating procedure includes a non-correctable bias even for big sample sizes. The weighting function can help correcting these problems in two ways: compensating zones with high variance or giving less importance to the ends of the estimation interval, where it's more difficult.

To study the effect of boundary variability we use weighting functions that lower the influence of the global polynomial fit near the interval ends. Let π be the fraction of the interval $[a, b]$ to be affected by lower weights near each end. We define a function $w_\pi(x)$ so that it is constant 1 in the central part $[a+\pi(b-a), b-\pi(b-a)]$ and it behaves like the branches of the quartic $-x^4/(\pi(b-a))^4$ near the interval ends. This way we have a smooth function to fulfill our needs. A graph for $w_{.25}$ is shown in figure 7.

In the same figure, lower part, we show the variability reduction obtained by using this weighting functions w_k for $k = 0, 0.1, 0.25, 0.40$ for the Poisson-Log model, example 3. We show density estimates for the relative errors in estimating h_{opt}, taking into account that the target value is different for different weighting functions (see 7). Results clearly show that the variability induced by the boundary problem can be reduced but, as expected, there is no improvement in bias reduction.

5 A block method

The main problem detected in the previous sections for the *rule of thumb* is that the global polynomial fit can't properly mimic the target function when it is not polynomial. It's clearly seen in example 3 for both models. Function #3 for Poisson examples is nearly linear at both sides of the estimation interval, while in the central part there is a dominating exponential component. HÄRDLE AND MARRON[1995] do face a similar problem in a simpler context: a simple nonparametric kernel regression using the Nadaraya-Watson estimator, equivalent to a locally constant fit ($p = 0$). Their approach is to divide the estimation interval in several blocks trying to make easier for a global polynomial to fit the target function in a smaller interval. A similar approach is used in RUPPERT, SHEATHER AND WAND[1995] in a local

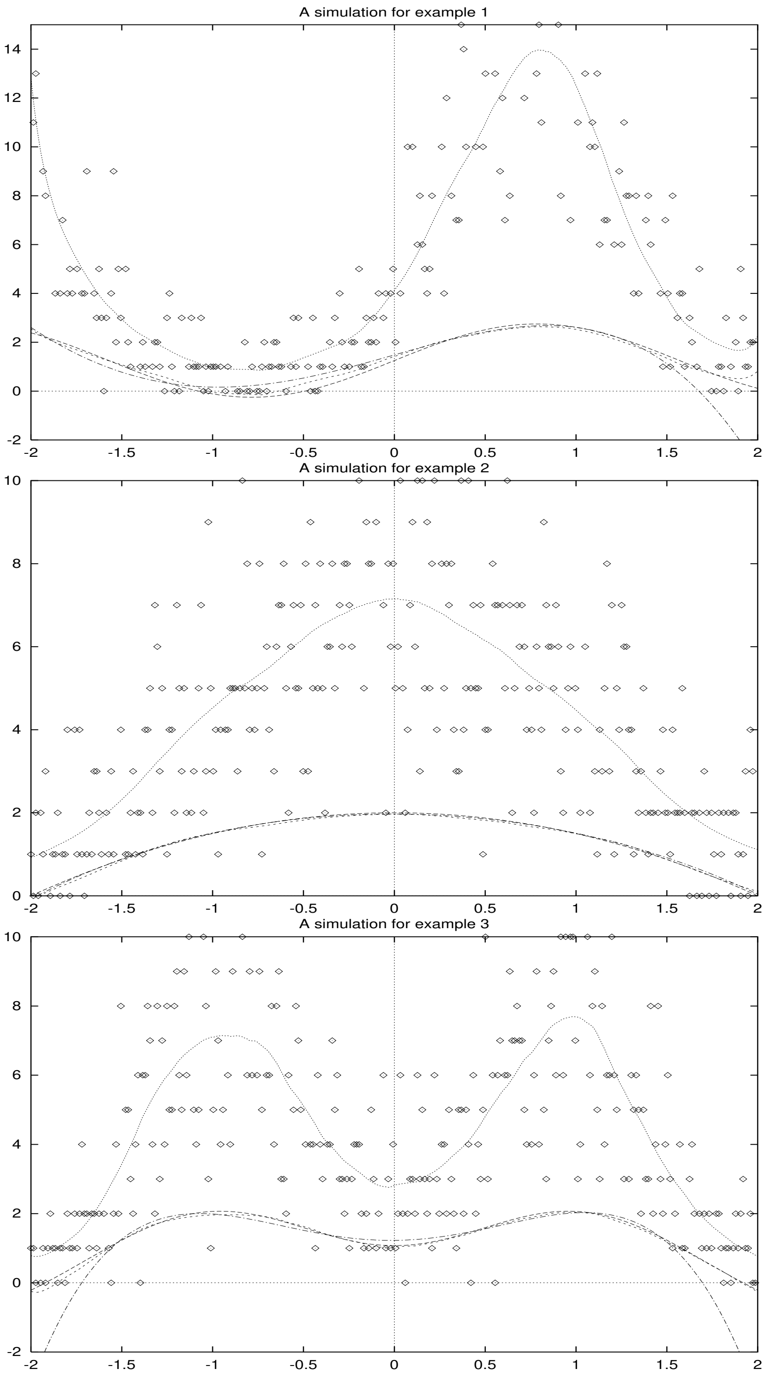


Figure 5: Typical simulations for the three examples of Poisson-Log model described in the text. Here, $n = 250$ for examples 1 and 2, $n = 500$ in example 3. Long dashed line is for the target function $\eta(x)$, dashed-dotted line is for the fourth order polynomial global fit, and short dashed line is for the estimated function. Dotted line is for the estimated $\hat{\mu}(x)$ that gives the conditional mean $\mathbf{E}(Y|X = x)$.

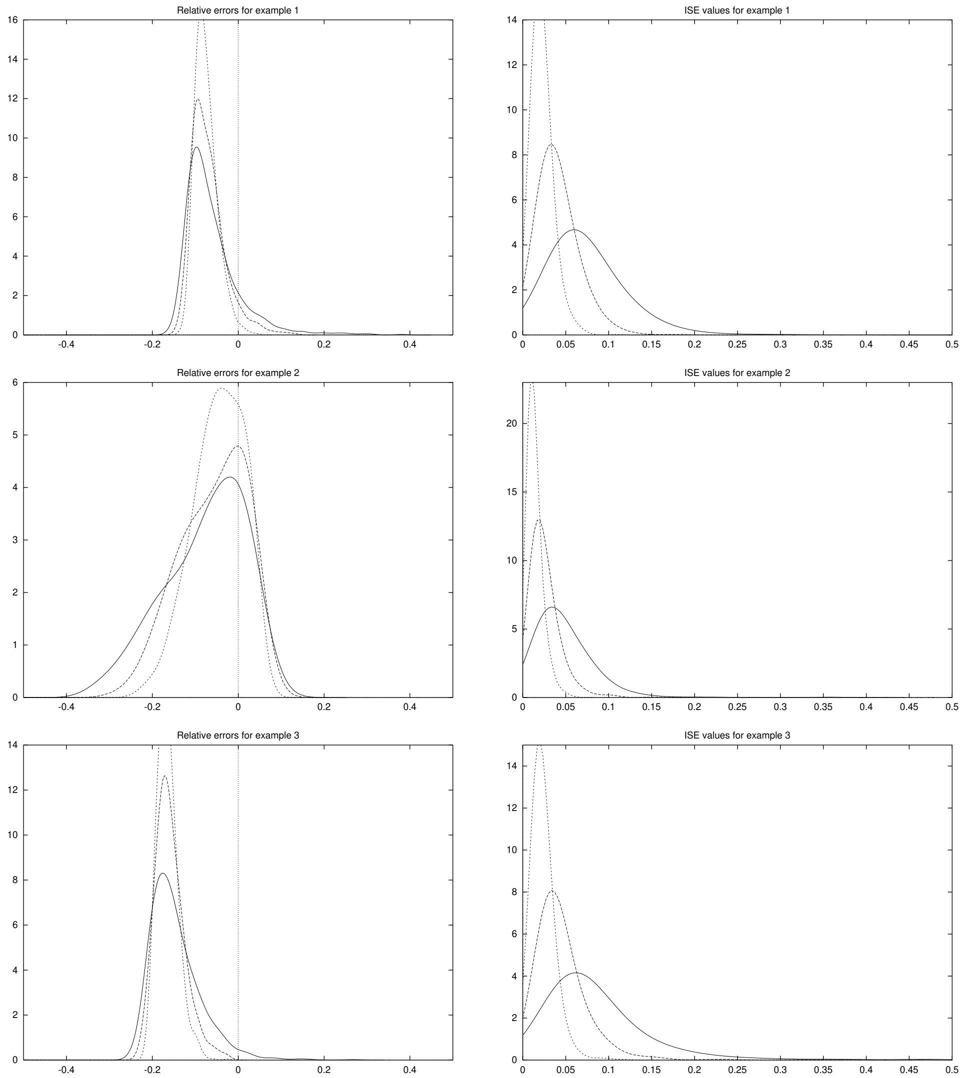
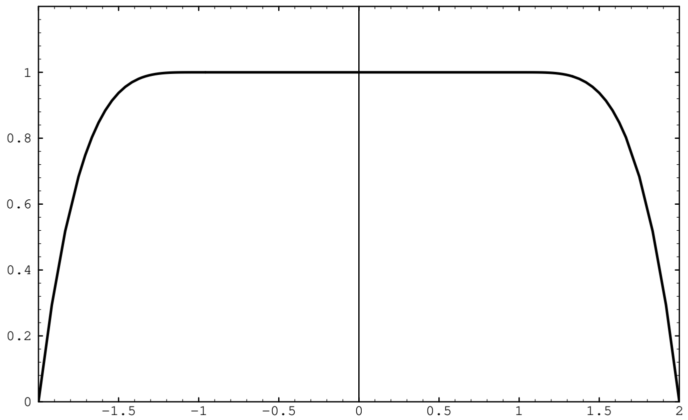


Figure 6: Result summaries for three examples of the Poisson-Log model as described in the text. Left: density estimates of relative errors in estimating h_{opt}. Right: density estimates of integrated squared errors for $\widehat{\eta}(x)$ as estimator of the target function. Solid lines are for sample size $n = 250$, dashed line is for $n = 500$ and dotted line is for $n = 1000$.

(a)



(b)

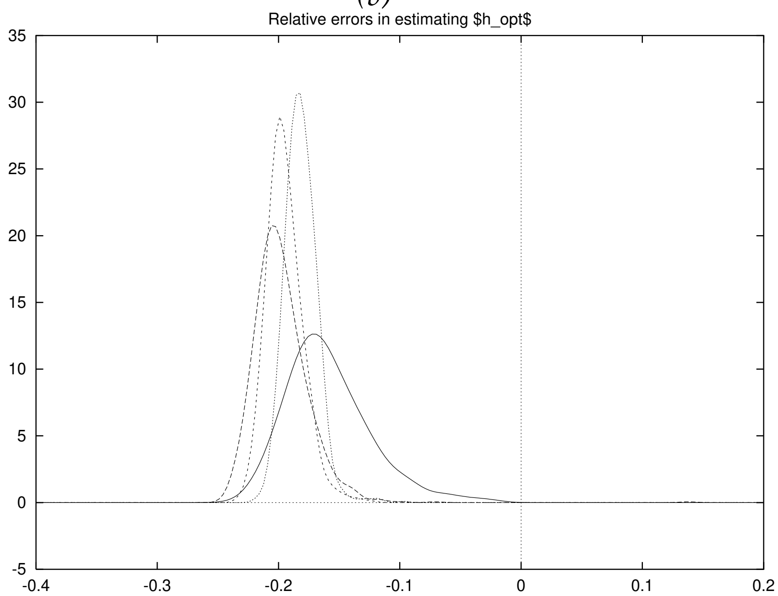


Figure 7: (a) graph of the weighting function $w_{.25}$. (b) density estimates for the relative errors obtained in estimating respective optimal bandwidths when using w_k weighting functions for $k = 0$ (solid line), $k = 0.1$ (long dashed line), $k = 0.25$ (short dashed line) i $k = 0.4$ (dotted line). Example function is #3, $n = 500$.

polynomial regression model setting (any p, normal model as in section 3.1). Here we take some of this ideas to our more general setting.

There are two issues to discuss: the number of blocks (see next section) and how to divide the interval into blocks. Let $[a, b]$ the estimation interval, usually equal to the support of the design density, and N the number of blocks. One possibility is to take equal width blocks,

$$\mathcal{B}_j = [a + \frac{(j-1)(b-a)}{N}, a + \frac{j(b-a)}{N}], \qquad 1 \leq j \leq N.$$

This simple rule can produce blocks with too few data points if the design density is not uniform. We prefer to take disjoint and adjacent blocks containing equal number of data points. If N divides n then

$$\mathcal{B}_j = \left\{ X_{((j-1)\frac{n}{N}+1)}, \dots, X_{(j\frac{n}{N})} \right\}.$$

If N does not divide n, the remaining values can be allocated to the first and last blocks.

In any case, for a given block $\mathcal{B}_j$, we apply the methods described in section 3, fitting a global polynomial of order $p + 3$ using only the data points (X_i, Y_i) with $X_i \in \mathcal{B}_j$, and compute, according to (7), (9), and (10),

$$\widehat{\nu}_{\nu,j}, \quad \widehat{b}_{\nu,j}, \qquad 1 \leq j \leq N.$$

This can give either a single global bandwidth

$$\widehat{h}_{\text{opt},B} = \alpha_\nu(K) \left(\frac{\sum_j \widehat{\nu}_{\nu,j}}{\sum_j \widehat{b}_{\nu,j}} \right)^{\frac{1}{2p+3}} \tag{11}$$

or a variable bandwidth $\widehat{h}_{\text{opt},V}$ given by a step function taking in each block the value

$$\widehat{h}_{\text{opt},j} = \alpha_\nu(K) \left(\frac{\widehat{\nu}_{\nu,j}}{\widehat{b}_{\nu,j}} \right)^{\frac{1}{2p+3}}. \tag{12}$$

This step function can be smoothed using the block radius as bandwidth as suggested in HÄRDLE AND MARRON[1995].

5.1 Number of blocks

HÄRDLE AND MARRON[1995] use a simple rule to choose the number of blocks and provide some graphical clues for a visual selection of the number of blocks that give better fitting to the data cloud. They report N between 3 and 5 as usual values. RUPPERT, SHEATHER AND WAND[1995] use an automatic method based on C_p ideas from MALLOWS[1973]. We move these ideas to an automatic method to select N based in Akaike information criteria (AIC), see MCCULLAGH AND NELDER[1989] and TIBSHIRANI AND HASTIE[1987].

We define

$$\mathrm{AIC}(N) = D(N) + 2(p+4)N/n$$

where $D(N)$ is the deviance of the model consisting in fitting a $p+3$ order polynomial in each of the N blocks. This deviance is defined as

$$2l(\mathbf{Y},\mathbf{Y}) - 2l(\mathbf{Y},\hat{\mu}),$$

that is, the lack of fit with respect to a completely fitted model, where l denotes the log-likelihood function

$$l(y,\theta) = \sum_i f(y,\theta)$$

In practical implementation, a maximum number of blocks needs to be fixed. Fixing it as a sub-linear function of n, e.g. $n^{1/2}$ gives guarantees of consistency of the estimator of h_{opt}. In our examples, with medium sample sizes, we take 5 as maximum value for N, obtaining good results.

We have some promising results from application of this blocked methodology, but they are not already processed to present any conclusion.

6 A real data application

As an example of application to real data, we estimate the proportion of sex in a group of students as a function of the scores (up to 25) in an exam. Data, with $n = 164$, has a women proportion of 53.7%, and for them the average score is 11.0 with standard deviation 4.4. For men, it is 12.1 with standard deviation 5.0. The mean difference is not significant, but it's interesting to study the scores distribution for the two subpopulation. This can be shown by adjusting a Bernoulli-logit model using the sex as binary response ($Y = 0$ for men, $Y = 1$ for women) and the scores as predictive variable, see figure 8a. Sample size is quite small and so the behavior of the fitted function close to the interval ends is not very reliable, but it shows clearly that woman do concentrate mainly in the central part with two maxima around $x = 6$ and $x = 16$. The global quartic fit would indicate a wrong position $x = 20$ for the second maximum. A complementary description for this data can be obtained as in the figure 8b by comparing the scores distributions for the two subsamples, but it doesn't show so clearly the *response* of sex to scores. Density functions have been estimated by kernel method, using a bandwidth of $h \approx 2.5$ obtained by normal reference.

7 Conclusion

The rule of thumb, consisting of using a global polynomial fit as a pilot estimator to compute an estimated asymptotically optimal bandwidth, can be used with confidence in generalized locally linear models. It works fine with small samples when the target function is polynomial, and has important bias problems when the target function is difficult to match by the global polynomial fit.

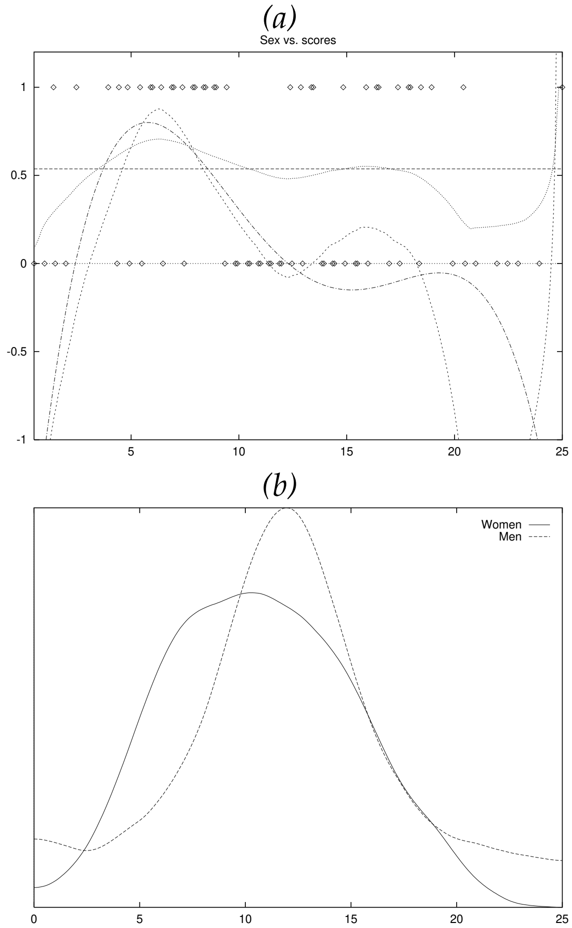


Figure 8: (a) A Bernoulli-logit model adjusted to scores-sex data for $n = 164$ students. $X \in [0, 25]$ is scores in a math exam, $Y = 1$ for women, $Y = 0$ for men. Dotted line is for the estimated conditional probability $P[Y = 1|X = x]$, short dashed line is for the logit of it, and dashed dotted line is for the quartic global fit. The horizontal long dashed line is for the global proportion of women, 0.537 (b) Density functions for the scores of respective subsamples by sex (solid line for women), computed by kernel method with bandwidth $h \approx 2.5$ obtained by the normal reference rule.

Use of weighting functions to lower the influence of the interval ends can reduce the variability of the estimator but with some increase in bias.

The block dividing method with number of blocks selected by AIC, provide an automatic way to select a good data driven bandwidth without losing the simplicity of the rule of thumb.

References

CLEVELAND, W. (1979) Robust locally weigthed regression and smoothing scatterplots. *Journal of the American Statistical Association*. **74**, 829–36.

CLEVELAND, W. S.; LOADER, C. R. (1996) Smoothing by local regression: principles and methods. In W. Härdle and M. G. Schimek, editors *Statistical Theory and Computational Aspects of Smoothing*. Physica Verlag, Heidelberg, 10–49.

FAN, J. (1992) Design adaptive nonparametric regression. *Journal of the American Statistical Association*. **87**, 998–1004.

FAN, J. (1993) Local linear regression smoothers and their minimax efficiencies. *Ann. Statist*. **20**, 155–69.

FAN, J.; CHEN, J. (1997) One-step Local Quasi-Likekihood Estimation. *Unpublished manuscript*. Available in
`ftp://ftp.stat.unc.edu/pub/fan`.

FAN, J.; FARMEN, M.; GIJBELS, I. (1996) A blueprint of local maximum likelihood estimation. *Journal of the Royal Statistics Society, series B*. to appear. Discussion paper #9604, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

FAN, J.; GIJBELS, I. (1995B) Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*. **4**, 213–227.

FAN, J.; GIJBELS, I. (1996) *Local Polynomial Modelling and Its Application – Theory and Methodologies*.. Chapman and Hall, New York.

FAN, J.; HECKMAN, N. E.; WAND, M. (1995) Local Polynomial Kernel Regresison for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association*. **90**, 141–150.

GASSER, T.; MÜLLER, H-G.; MAMMITZSCH (1985) Kernels for Nonparametric Curve Estimation. *Journal of the Royal Statistics Society, series B*. **47**, No. 2, 238–252.

HÄRDLE, W.; MARRON, J. S. (1995) Fast and simple scatterplot smoothing. *Computational Statistics and Data Analysis*. **20**, 1–18.

MALLOWS, C. L. (1973) Some Comments on C_p. *Technometrics*. **15**, 661–675.

MCCULLAGH, P.; NELDER, J. A. (1989) *Generalized linear models*. Second Edition. London: Chapman and Hall.

RUPPERT, D.; SHEATHER, S. J.; WAND, M. (1995) An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*. **90**, 1257–1270.

RUPPERT, D.; WAND, M. (1994) Multivariate locally weighted least squares regression. *Annals of Statistics*. **22**, 1346–1370.

STONE, C. J. (1977) Consistent nonparametric regression (with discussion). *Annals of Statistics*. **5**, 595–645.

STONE, C. J. (1980) Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*. **8**, 1348–1360.

TIBSHIRANI, R.; HASTIE, T.J. (1987) Local Likelihood Estimation. *Journal of the American Statistical Association*. **82**, 559–567.