Local Likelihood Density Estimation

Clive R. Loader

AT&T Bell Laboratories, Room 2C-279, 600 Mountain Avenue,

Murray Hill, NJ 07974, U.S.A.

©The Institute of Mathematical Statistics

To appear in: The Annals of Statistics, 1996.

Abstract

Local likelihood was introduced by Tibshirani and Hastie (1987) as a method of smoothing by local polynomials in nongaussian regression models. In this paper an extension of these methods to density estimation is discussed, and comparison with other methods of density estimation presented. The local likelihood method has particularly strong advantages over kernel methods when estimating tails of densities and in multivariate settings. Support constraints are incorporated in a simple manner. Asymptotic properties of the estimate are discussed. A method for computing the estimate is outlined.

C code to implement the estimation procedure described in this paper, together with S interfaces for graphical display of results, are available through WWW at

`http://netlib.att.com/netlib/att/stat/prog/index.html`

or anonymous ftp at `netlib.att.com`.

Some key words and phrases: Density estimation; Local likelihood; Local polynomials.

AMS 1991 subject classifications: Primary 62G07; Secondary 62G20, 62H12.

1 Introduction

Local regression is a popular form of nonparametric regression, combining excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. References include Stone (1977) and Cleveland (1979). Cleveland and Devlin (1988) discuss a multivariate setting. Recently, Fan (1992, 1993) has studied minimax properties of local linear regression. A detailed summary of the advantages of local regression compared to kernel fitting may be found in Hastie and Loader (1993).

Local regression may be viewed as a special case of the local likelihood procedure introduced by Tibshirani and Hastie (1987). This procedure is designed for nonparametric regression modeling in situations where a nongaussian likelihood is appropriate, such as logistic regression and proportional hazards models. A related paper is Staniswalis (1988).

The purpose of this paper is to extend local likelihood methods to the nonparametric density estimation setting. The estimate is introduced in section 2. The remainder of the paper is devoted to a study of theoretical and practical issues concerning the estimate. Computational methods are studied in section 3. Asymptotic theory for the estimate is developed in section 4. Section 5 contains some discussion of order choice and examples.

A previous application of local polynomial methods to density estimation is Lejeune and Sarda (1992). In this method the distribution function is estimated using a weighted quadratic penalty for the distribution formulation. Under the formulation we use, it turns out to be more natural to model the logarithm of the density, which is also more intuitively appealing in the tails. In independent work Hjort and Jones (1994) have also studied local likelihood procedures for density estimation. They concentrate mainly on the one dimensional case, but do not restrict their local models to a log-polynomial form. In particular, they find the important properties of the estimate are determined mainly by the number of parameters of the local model, rather than the precise form of the model. Hjort and Jones also provide more discussion of the relation of local likelihood to other methods that achieve some of the aims of local likelihood. Another interesting paper is Hjort (1994) where local likelihood is carried out in the closely related hazard rate setting.

Log-spline and penalized likelihood type methods have also been proposed in the density estimation literature; see for example Silverman (1982), Stone (1990) and references therein. These methods also have the advantages inherent in modeling $\log f(x)$, and performance should be competitive with the local likelihood method, although conceptually the methods are very different. Penalized likelihood methods are defined as solutions of global optimization problems, trading fidelity to data against roughness of the estimated curve. By contrast, the local likelihood method solves local optimization problems motivated by bias-variance considerations. Thus the methods seem difficult to compare directly; neither is likely to be uniformly better in practice.

2 The Local Likelihood Estimate

Suppose we have observations $X_1, \ldots, X_n$ lying in a subset $\mathcal{X}$ of R^d, having unknown density $f(x)$. The log-likelihood function is

$$\mathcal{L}(f) = \sum_{i=1}^{n} \log(f(X_i)) - n(\int_{\mathcal{X}} f(u)du - 1). \tag{1}$$

If f is a density, (1) coincides with the usual log-likelihood. The attractiveness of maximum likelihood estimation stems from the following property:

$$E_f \mathcal{L}(f_1) \leq E_f \mathcal{L}(f) \tag{2}$$

for all densities f_1, with equality only when $f_1 = f$ almost everywhere. Using the inequality

$$\log f_1(x) \leq \log f(x) + \frac{1}{f(x)}(f_1(x) - f(x))$$

shows (1) maintains this property for any non-negative integrable function f_1 defined on $\mathcal{X}$; we do not require f_1 to be a density. One consequence of this extension is that maximum likelihood estimation can be performed with multiplicative parameters. For example, fitting the family $f(x; C, \mu) = C\exp(-(x-\mu)^2/2)$ by maximum likelihood gives $\hat{C} = (2\pi)^{-1/2}$.

Frequently, we do not have sufficient information to specify a global family for f, but local smoothness assumptions may be reasonable. In this case, it is useful to consider a localized version of the log-likelihood:

$$\mathcal{L}(f,x) = \sum_{j=1}^{n} K(\frac{X_j - x}{h}) \log(f(X_j)) - n \int_{\mathcal{X}} K(\frac{u-x}{h}) f(u) du \tag{3}$$

where K is a suitable non-negative weight function and h a bandwidth. For the present we treat h as a constant, although in practice it is likely to be chosen by data-based methods. As with any local fitting procedure, performance can be enhanced by allowing h to vary as function of x or X_j (Jones, 1990). We note the property (2) extends to the local log-likelihood;

$$E_f \mathcal{L}(f_1, x) \leq E_f \mathcal{L}(f, x)$$

with equality when $f(u) = f_1(u)$ on the support of $K((u-x)/h)$. This suggests estimating $f(x)$ by maximizing (3) over a suitable class of functions.

The local polynomial approximation supposes that $\log f(u)$ can be well approximated by a low degree polynomial in a neighbourhood of the fitting point x. That is,

$$\log f(u) \approx P(u-x) \tag{4}$$

with (in one dimension)

$$P(u-x) = a_0 + a_1(u-x) + \ldots + a_p(u-x)^p. \tag{5}$$

With this approximation, the local likelihood becomes

$$\mathcal{L}_p(f,x) = \sum_{j=1}^{n} K(\frac{X_j - x}{h}) P(X_j - x) - n \int_{\mathcal{X}} K(\frac{u-x}{h}) \exp(P(u-x)) du. \tag{6}$$

Let $\mathcal{C}$ be the parameter space,

$$\mathcal{C} = \{(a_0, \ldots, a_p) : \int_{\mathcal{X}} K(\frac{u-x}{h}) \exp(P(u-x)) du < \infty\}.$$

To avoid some technical problems, we assume that $\mathcal{C}$ is an open set. This holds if either $\mathcal{X}$ is bounded or K is continuous with bounded support, and $\mathcal{C} = R^d$. It also holds for the Gaussian kernel if $p \leq 2$. It fails if $K(u) = \exp(-|u|)/(1+|u|^3)$ and local linear fitting is used.

Definition 1 (Local likelihood density estimate) *For fixed $x \in \mathcal{X}$, let $(\hat{a}_0, \ldots, \hat{a}_p)$ be the maximizer of (6). The local likelihood density estimate is*

$$\hat{f}(x) = \exp(\hat{a}_0). \tag{7}$$

If no maximizer of (6) exists, or $x \notin \mathcal{X}$, then $\hat{f}(x) = 0$.

Example: For local constant fitting ($p = 0$), (7) gives

$$\hat{f}(x) = \exp(\hat{a}_0) = \frac{\sum_{j=1}^{n} K(\frac{X_j - x}{h})}{n \int_{\mathcal{X}} K(\frac{u-x}{h}) du}. \tag{8}$$

When $\mathcal{X}$ is unbounded, this is the kernel estimate introduced by Rosenblatt (1956) and Parzen (1962). This estimate has substantial bias problems in the tails of densities, and can also have bias near peaks. For bounded $\mathcal{X}$, the renormalization will usually reduce boundary bias, although problems remain if the density has substantial slope near the boundaries. Modifications exist to address many of these bias problems; see Scott (1992) for many recent references. In this paper we show the use of higher order fits also addresses these bias problems, but in a much more unified fashion.

If a maximizer of (6) exists, it must fall in the interior of the open set $\mathcal{C}$, and hence satisfies the system of local likelihood equations

$$\frac{1}{n} \sum_{j=1}^{n} A(\frac{X_j - x}{h}) K(\frac{X_j - x}{h}) = \int_{\mathcal{X}} A(\frac{u - x}{h}) K(\frac{u - x}{h}) \exp(P(u - x)) du \tag{9}$$

where $A(v) = (\, 1 \quad v \quad \ldots \quad v^p \,)^T$. These equations have a very simple and intuitive interpretation. The left hand side of (9) is simply a vector of localized sample moments up to order p, while the right hand side is localized population moments using the log-polynomial density approximation. The local likelihood estimate simply matches localized sample moments with localized population moments.

In the multidimensional case, the local polynomial expansion is extended to include all monomials of degree $\leq p$. This leads to a total of $k + 1$ terms, where $k = \binom{p+d}{d} - 1$. For convenience, we assume the constant term always comes first. The vector function $A(v)$ then includes the monomials in the same order. For example, in two dimensions with local quadratic fitting, $A(v) = (1 \quad v_1 \quad v_2 \quad v_1^2 \quad v_1v_2 \quad v_2^2)^T$. The local likelihood (6) continues to hold, with the outer sum now running through $i = 0, \ldots, k$. Definition 1 and the local likelihood equations (9) are unchanged.

In general, the system (9) has no closed form solution, leading to questions of existence and uniqueness. For fixed $X_1, \ldots, X_n$, the local likelihood surface is easily shown to be a concave function of $a_0, \ldots, a_p$; this implies at most one solution of (9) exists, and this solution will be a maximum. This ensures $\hat{f}(x)$ is well defined.

Existence of a solution of (9) is more difficult. Indeed, no solution exists if insufficient observations receive nonzero weights. A closely related problem is the existence of maximum likelihood estimates in exponential response models. Following the argument of Theorem 2 of Haberman (1977), a solution exists if the observed left hand side of (9) falls in the interior of its range. Assuming the kernel is continuous, this requires at least one observation to receive a non-zero weight for local linear fitting, and $d + 1$ observations (with a nonsingular sample covariance) to receive non-zero weights for local quadratic fitting.

3 Computation

Evaluation of the local likelihood density estimate can be split into two parts. First, one evaluates the summation vector on the left hand side of (9). The estimate is then found by solution of the equations. Except in special cases, there is no closed form for the solution and hence iterative methods must be used. The integrals must be evaluated numerically at each step.

Evaluation of the sums is a $O(nh)$ computation assuming a compactly supported kernel. The solution of the equations is independent of sample size, so for very large sample sizes the accumulation of sums is the most expensive part of the computation. However, for practical sample sizes, the iterative solution is the most expensive and difficult part of the solution.

The integral will usually be evaluated numerically. In multidimensional cases, substantial savings can be achieved through the use of a product kernel,

$$K(\frac{u-x}{h}) = \prod_{i=1}^{d} K_0(\frac{u_i - x_i}{h})$$

for a one dimensional kernel K_0, and choosing a local model so that the integrals factorize. If $\mathcal{X}$ is a rectangular subset of R^d and local linear fitting is used, then

$$\int_{\mathcal{X}} K(\frac{u-x}{h}) e^{a+b^T(u-x)} du = e^a \prod_{i=1}^{d} \int_{\mathcal{X}_i} K_0(\frac{u_i - x_i}{h}) e^{b_i(u_i - x_i)} du_i.$$

This enables all the integrals required in the iteration to be written as the product of one dimensional integrals.

This simplification does not work fitting a full local quadratic model. One possibility is to remove cross product terms from the model; with this modification the local likelihood again factorizes. The cost for excluding cross product terms will be some reduction in the ability to model curvature, particularly in tail regions.

To plot the density estimate, evaluation at a large number of points is frequently required. For this reason, the algorithm described by Cleveland and Gross (1991) for local regression is useful. In this method, a piecewise polynomial estimate is constructed over a partition of the space $\mathcal{X}$, with the polynomial pieces being represented in terms of function values and derivatives at the vertices of the partition. These vertex values are estimated using the local regression or likelihood algorithm, and the polynomial pieces are rapidly evaluated at any point in the domain. The specific partition used by Cleveland and Grosse is based on a k-d tree partition, and some other alternatives are suggested in Loader (1994).

Another possibility for computational savings is to use the Gaussian kernel, for which closed form evaluation of the integrals is possible, and at least in some cases, Hjort and Jones (1994) have derived closed form expressions for the estimate. A disadvantage with the Gaussian kernel is that for local quadratic fitting the parameters are constrained. This may limit the ability of the estimate to reproduce troughs in the data.

4 Asymptotic Theory

In this section we study some asymptotic properties for the local likelihood density estimates. In particular we study rates of convergence and obtain asymptotic distributions. Some of the ideas for the results are borrowed from Ruppert and Wand (1994).

Our asymptotic results will be stated for sequences of points $x = x_0 + hz$ where $x_0 \in \mathcal{X}$, $z \in \mathcal{R}^d$ is fixed and h, the bandwidth, converges to 0 as $n \to \infty$. For interior points we would usually consider $z = 0$ and $x = x_0$ for all h; by considering sequences, the results are also useful for studying the behavior of the estimate at points close to the boundary of $\mathcal{X}$.

We suppose throughout the kernel has compact support, although this could be weakened with some truncation arguments. Theorem 1 establishes rates of convergence for the coefficients $\hat{a}_j$, and shows these may be interpreted as estimates of the corresponding terms in the Taylor series expansion of $g(x)$. Theorem 2 gives an asymptotic representation for the estimate as a deterministic and random component. Theorem 3 provides an alternative asymptotic characterization of the estimate, where the bandwidth is held fixed as n increases.

Lemma 1 *Let*

$$Z = \frac{1}{n}\sum_{j=1}^{n} A(\frac{X_j - x}{h})K(\frac{X_j - x}{h}) - \int_{\mathcal{X}} A(\frac{u-x}{h})K(\frac{u-x}{h})f(u)du. \tag{10}$$

Let $n \to \infty$ and $h \to 0$ with $nh^d \to \infty$, and suppose $x = x_0 + hz$ with $x_0 \in \mathcal{X}$ and z a fixed vector in $\mathcal{R}^d$. Suppose also that f (restricted to $\mathcal{X}$) is continuous in an open neighborhood of x_0, and $f(x_0) > 0$. Suppose $\mathcal{X}(h) = \{v : v \in supp(K), x + hv \in \mathcal{X}\}$ has a limit $\mathcal{X}^*$ as $h \to 0$. Then

$$\sqrt{\frac{n}{h^d}}Z \xrightarrow{\mathcal{L}} N\left(0, f(x_0)\int_{\mathcal{X}^*} A(v)A(v)^T K(v)^2 dv\right). \tag{11}$$

Proof: It is easily shown $E(Z) = 0$ and

$$\begin{aligned} E(ZZ^T) \;\; = \;\; & \frac{1}{n}\int_{\mathcal{X}} A(\frac{u-x}{h})A(\frac{u-x}{h})^T K(\frac{u-x}{h})^2 f(u)du \\ & -\frac{1}{n}\int_{\mathcal{X}} A(\frac{u-x}{h})K(\frac{u-x}{h})f(u)du \int_{\mathcal{X}} A(\frac{u-x}{h})^T K(\frac{u-x}{h})f(u)du. \end{aligned}$$

A multivariate central limit argument completes the result; this can be easily established using moment generating functions. Note the first term of the covariance is $O(h^d/n)$; the second is $O(h^{2d}/n)$ and therefore is asymptotically negligible. Parzen (1962) gave this result for $p = 0$.

□

Before stating the asymptotic results we introduce some more notation. Let $g(u) = \log f(u)$. Suppose all derivatives of g of order $p+1$ exist and are continuous. Expand g in a Taylor series of order p around a point x, and let the coefficients, dependent on x, be $g_0, \ldots, g_k$ ($k = \binom{p+d}{d} - 1$). These should be arranged in an order corresponding to the components of $A(\cdot)$. For example, in one dimension, $A(v) = (1 \quad v \quad \ldots \quad v^p)^T$ and

$$g_j = \frac{1}{j!}\frac{d^j g(x)}{dx^j}.$$

For each j, let $m(j)$ denote the order of derivative represented by g_j; in the one dimensional example, $m(j) = j$.

Theorem 1 *Assume the conditions of Lemma 1 hold, and $\mathcal{X}^*$ has non-zero Lebesgue measure. Then*

$$h^{m(j)}(\hat{a}_j - g_j) = O(h^{p+1}) + O_p((nh^d)^{-1/2}).$$

Remarks:

i) The condition that $\mathcal{X}^*$ have positive Lebesgue measure excludes some badly behaved boundaries. Ruppert and Wand (1994) give the two dimensional example $\mathcal{X} = \{(x_1, x_2) : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq x_1^2\}$; Theorem 1 does not apply when $x_0 = (0, 0)$.

ii) Setting $h = n^{-1/(2p+2+d)}$ gives

$$\hat{a}_j - g_j = O_p(n^{-(p+1-m(j))/(2p+2+d)}); \tag{12}$$

Stone (1980) showed this rate to be optimal in a certain minimax sense. If two derivatives are assumed, and local linear fitting ($p = 1$) is used, we have the familiar rate $\hat{a}_0 - g(x) = O_p(n^{-2/(4+d)})$.

iii) Theorem 1 suggests $\hat{a}_j$ is a natural estimate of g_j. However, when $m(j) \geq 1$ the condition $nh^{d+2m(j)} \to \infty$ is required for consistency; this is stronger than the conditions previously stated.

Proof of Theorem 1: The system (9) can be represented as

$$\begin{aligned} Z &= \int_{\mathcal{X}} A(\frac{u-x}{h})K(\frac{u-x}{h})(e^{\hat{P}(u-x)} - e^{g(u)})du \\ &= h^d \int_{\mathcal{X}(h)} A(v)K(v)(e^{\hat{P}(hv)} - e^{g(x+hv)})dv \end{aligned} \tag{13}$$

where $\hat{P}$ denotes P with coefficients $\hat{a}_0, \ldots, \hat{a}_p$, and Z is defined by (10).

Let $\tilde{P}$ denote the Taylor series expansion of g of degree p. Then

$$\begin{aligned} e^{\tilde{P}(hv)} - e^{g(x+hv)} &= e^{g(x+hv)}(e^{\tilde{P}(hv)-g(x+hv)} - 1) \\ &= f(x+hv)(\tilde{P}(hv) - g(x+hv) + O(h^{2p+2})) \end{aligned} \tag{14}$$

since by definition $\tilde{P}(hv) - g(x+hv) = O(h^{p+1})$. Using the $p+1$ order terms in the Taylor expansion of g around x, and noting $f(x+hv) = f(x) + O(h)$,

$$e^{\tilde{P}(hv)} - e^{g(u)} = -\frac{h^{p+1}f(x)}{(p+1)!}\sum_{i_1=1}^{d}\cdots\sum_{i_{p+1}=1}^{d}\left[\prod_{j=1}^{p+1} v_{i_j} g^{i_1,\ldots,i_{p+1}}(x)\right] + o(h^{p+1}) \tag{15}$$

where $g^{i_1,\ldots,i_{p+1}}(x)$ is the $(p+1)$st order partial derivative of g with respect to $x_{i_1}, \ldots, x_{i_{p+1}}$. Since the kernel has compact support, the error term holds uniformly on $\{v : K(v) > 0\}$, and

$$\begin{aligned} &\int_{\mathcal{X}(h)} A(v)K(v)(e^{\tilde{P}(hv)} - e^{g(x+hv)})dv \\ &= -\frac{h^{p+1}f(x)}{(p+1)!}\sum_{i_1,\ldots,i_{p+1}=1}^{d}\left[g^{i_1,\ldots,i_{p+1}}(x)\int_{\mathcal{X}^*}\prod_{j=1}^{p+1} v_{i_j} A(v)K(v)dv\right] + o(h^{p+1}). \end{aligned} \tag{16}$$

Substituting (16) and (11) into (13) and using $\mathcal{X}(h) \to \mathcal{X}^*$,

$$\begin{aligned} \int_{\mathcal{X}^*} A(v)K(v)(e^{\hat{P}(hv)} - e^{\tilde{P}(hv)})dv &= h^{-d}Z - \int_{\mathcal{X}^*} A(v)K(v)(e^{\tilde{P}(hv)} - e^{g(x+hv)})dv \\ &= O_p((nh^d)^{-1/2}) + O(h^{p+1}). \end{aligned}$$

Equivalently,

$$\int_{\mathcal{X}^*} A(v)K(v)e^{\hat{P}(hv)}dv = \int_{\mathcal{X}^*} A(v)K(v)e^{\tilde{P}(hv)}dv + O(h^{p+1}) + O_p((nh^d)^{-1/2}).$$

Treating

$$\int_{\mathcal{X}^*} A(v)K(v)e^{\tilde{P}(hv)}dv \tag{17}$$

as function of $(a_0, \ldots, h^{m(k)}a_k)$, the proof is completed by an application of the inverse function theorem in a neighborhood of the point $(g(x), 0, \ldots, 0)$ (Burkill and Burkill, 1970, page 223). Note the Jacobian matrix of (17),

$$\int_{\mathcal{X}^*} A(v)A(v)^T K(v)e^{P(hv)}dv$$

is strictly positive definite.

$\square$

Theorem 2 *Suppose $g(x) = \log f(x)$ has $p+1$ derivatives. Let*

$$M_1 = \int_{\mathcal{X}^*} A(v)A(v)^T K(v)dv.$$

Under the conditions of Lemma 1,

$$\begin{aligned}
&\begin{pmatrix} h^{m(0)}(\hat{a}_0 - g_0) \\ \vdots \\ h^{m(k)}(\hat{a}_k - g_k) \end{pmatrix} \\
&= M_1^{-1}\left[\frac{1}{f(x)h^d}Z + \frac{h^{p+1}}{(p+1)!}\sum_{i_1,\ldots,i_{p+1}=1}^{d}\left[g^{i_1,\ldots,i_{p+1}}(x)\int_{\mathcal{X}^*}\prod_{j=1}^{p+1} v_{i_j}A(v)K(v)dv\right]\right] \\
&\quad + o(h^{p+1}) + o_p((nh^d)^{-1/2}).
\end{aligned} \tag{18}$$

Suppose x_0 is an interior point of $\mathcal{X}$, p is even and $g(x)$ has $p+2$ derivatives. Also assume K is symmetric.

$$\begin{aligned}
\hat{a}_0 - g(x) &= e_0^T M_1^{-1}\left[\frac{1}{f(x)h^d}Z\right. \\
&\quad \left. + h^{p+2}\sum_{i_1,\ldots,i_{p+2}=1}^{d}\left[\frac{g^{i_1,\ldots,i_{p+2}}(x)}{(p+2)!} + \frac{g^{i_1,\ldots,i_{p+1}}(x)g^{i_{p+2}}(x)}{c(p)}\right]\int_{\mathcal{X}^*}\prod_{j=1}^{p+2} v_{i_j}A(v)K(v)dv\right] \\
&\quad + o(h^{p+2}) + o_p((nh^d)^{-1/2})
\end{aligned} \tag{19}$$

where $c(0) = 2$ *and* $c(p) = (p+1)!$ *for* $p \geq 2$.

Remarks

i) Lemma 1 and Theorem 2 together imply asymptotic normality of the estimates. When h is larger than optimal, the deterministic error $o(h^{p+1})$ may dominate the random component $h^{-d}Z = O_p((nh^d)^{-1/2})$, and hence the proof is not constructive in this case. A better bias approximation can be achieved by retaining more terms in the Taylor series expansion (14), or avoiding the use of the Taylor series entirely, as we do in Theorem 3 below.

ii) Derivatives of f. Differentiating the relation $f(x) = \exp(g(x))$ yields

$$f'(x) = f(x)g'(x). \tag{20}$$

Substituting estimates of $\hat{f}(x) = e^{\hat{a}_0}$ and $\hat{g}'(x) = \hat{a}_1$ gives an estimate of $f'(x)$. Repeated differentiation of (20) gives natural estimates of derivatives of f up to order p.

iii) Simplification: The summands in (18) and (19) are invariant under permutation of $i_1, \ldots, i_{p+1}$, which leads to substantial simplification. Also, especially in the interior case, many of the integrals are 0 when K is symmetric. This yields simplifications similar to those obtained by Ruppert and Wand (1994) for local regression.

iv) Theorem 1 suggests

$$\int_{\mathcal{X}} \hat{f}(x)dx = 1 + O(h^{p+1}) + O_p((nh^d)^{-1/2}), \tag{21}$$

and hence renormalization of the estimate will not affect rates of convergence, but will affect constants. This argument is only heuristic and (21) fails in some specific cases.

Proof of Theorem 2: From Theorem 1 we obtain

$$e^{\hat{P}(hv)} - e^{\tilde{P}(hv)} = f(x)(\hat{P}(hv) - \tilde{P}(hv)) + o(h^{p+1}) + o_p((nh^d)^{-1/2}).$$

and hence

$$\int_{\mathcal{X}(h)} A(v)K(v)(e^{\hat{P}(hv)} - e^{\tilde{P}(hv)})dv$$
$$= f(x)\int_{\mathcal{X}(h)} A(v)A(v)^T K(v)dv \begin{pmatrix} h^{m(0)}(\hat{a}_0 - g_0) \\ \vdots \\ h^{m(k)}(\hat{a}_k - g_k) \end{pmatrix} + o(h^{p+1}) + o_p((nh^d)^{-1/2}). \quad (22)$$

Substituting (16) and (22) into (13) yields (18).

The result for even p follows by noting the bias component of $\hat{a}_0$ is 0 when p is even, and using the next term in the series expansion (15). The special case for $p = 0$ arises because the $O(h^{2p+2})$ term in (14) must be retained in this expansion.

□

An alternative asymptotic characterization proposed by Stoker (1993) is to keep the bandwidth fixed as $n \to \infty$. In a theoretical sense this scaling may be unsatisfactory; for example, the estimate is in general not even consistent. However, the motivation is that the result may be of more practical relevance. The 'optimal' bandwidths giving the rate (12) converges to 0 very slowly; the fixed bandwidth characterization may be much more closely related to the problems solved in practice.

Theorem 3 *Let $a^* = a^*(x)$ be the coefficients of $P(\cdot)$ satisfying*

$$\int_{\mathcal{X}} A(\frac{u-x}{h})K(\frac{u-x}{h})e^{P(u-x)}du = \int_{\mathcal{X}} A(\frac{u-x}{h})K(\frac{u-x}{h})f(u)du.$$

If $n \to \infty$ with x and h fixed,

$$\begin{pmatrix} h^{m(0)}(\hat{a}_0 - a_0^*) \\ \vdots \\ h^{m(k)}(\hat{a}_k - a_k^*) \end{pmatrix} = \left[\int_{\mathcal{X}} A(\frac{u-x}{h})A(\frac{u-x}{h})^T K(\frac{u-x}{h})e^{P^*(u-x)}du\right]^{-1} Z + o_p(n^{-1/2}). \quad (23)$$

where $P^*(u-x)$ has coefficients $a^*(x)$.

From Theorem 3 it is clear that $\hat{a}_0$ is really estimating a_0^*. If the true density is well approximated by $\exp(P^*(u-x))$ over the smoothing window the estimate will perform well; existence or otherwise of density derivatives is incidental.

This theorem is also useful for estimating standard errors for the estimate. Since $\hat{P} \to P^*$, the result continues to hold if P^* is replaced by $\hat{P}$. The Cholesky decomposition of matrix on the right side of (23) is available as a by-product of the optimization. One can estimate the covariance matrix of Z from the sample covariance of the $K((X_i - x)/h)A((X_i - x)/h)$; no further numerical integration is needed. Variances are then estimated by straightforward matrix multiplication.

5 Comparisons

In this section we discuss the important question of order choice in the local polynomial model, and also provide some brief comparisons with kernel methods. It is important to remember that no method will be universally best.

From the results in the last section, we have for large n and local linear fitting,

$$\hat{f}(x) - f(x) = e^{\hat{a}_0} - f(x) = \frac{h^2}{2} f(x) g''(x) \int v^2 K(v) dv + \frac{1}{h} Z_0 + o(h^2) + o_p((nh)^{-1/2})$$

where Z_0 is the first component of (10). For a local constant estimate,

$$\hat{f}(x) - f(x) = \frac{h^2}{2} f(x)(g''(x) + g'(x)^2) \int v^2 K(v) dv + \frac{1}{h} Z_0 + o(h^2) + o_p((nh)^{-1/2}).$$

The difference is in the bias; the local linear estimate is best when $|g''(x)| < |g''(x) + g'(x)^2|$. In the central part of the distribution neither estimate is uniformly better. The two bias terms are equal whenever $g'(x) = 0$, suggesting the estimates will have very similar ability to detect peaks and troughs in the density.

The main difference between the local constant and linear estimates is at boundary regions and in the tails. The difference at boundary regions can be seen from Theorem 2: The local linear fitting has $O(h^2)$ bias at the boundary; the local constant fit has $O(h)$ bias and may have bias induced by slope of the density. The local linear fit is more variable so the advantage is not always realized in practice.

The relative efficiency of the local constant and local linear methods can be defined as the ratio of sample sizes to achieve the same MSE. In one dimension and ignoring boundary effects, the asymptotic relative efficiency is

$$R(x) = \left| \frac{g''(x)}{g''(x) + g'(x)^2} \right|^{1/2}.$$

Values of $R > 1$ indicate the local constant estimate is more efficient, while $R < 1$ indicates the local linear fit is more efficient. One can construct examples where either method wins; however, the local linear method is a clear winner in the tails for fairly broad classes of densities.

Suppose $d = 1$; the tails of $f(x)$ decay like $x^{-\alpha}$ ($\alpha > 1$) and the derivatives of f behave as expected. Examples include the Cauchy distribution, t distributions and F distributions. Then we can show

$$\lim_{x \to \infty} R(x) = \frac{1}{\sqrt{\alpha + 1}}.$$

That is, for a Cauchy distribution, the kernel method has a relative efficiency of about 58% in the tails; for t distributions with high degrees of freedom, the efficiency is even less.

Now consider families with tails $f(x) = \exp(-x^{\alpha} + o(x^{\alpha}))$ as $x \to \infty$, with $\alpha > 0$. This includes the normal distribution, Gamma distributions and Weibull distributions. In this case the asymptotic relative efficiency is

$$R(x) = |1 - \alpha^{-1}|^{1/2} x^{-\alpha/2} + o(x^{-\alpha/2})$$

as $x \to \infty$. That is, the relative efficiency of the kernel method is arbitrarily small in the tails.

A common way to correct kernel estimates for curvature bias is through the use of fourth order kernels; see Scott (1993, section 6.2.3). One construction of fourth order kernels is to multiply a symmetric second order kernel by a quadratic, with coefficients chosen to satisfy moment conditions. This fourth order kernel estimate has an asymptotic representation with the same random component as local quadratic and local cubic fitting, but the bias components of these three methods differ. We can again compute relative efficiencies of these methods, and compare the results in Table 1. The local quadratic fit dominates the 4th order kernel method,

	$f(x) = x^{-\alpha}(1 + o(1))$	$f(x) = \exp(-x^\alpha + o(x^\alpha))$
(1)	$(1+\alpha)^{-1/2}$	$\lvert 1-\alpha^{-1}\rvert^{1/2} x^{-\alpha/2}$
(2)	$\left(\frac{6+8\alpha}{(\alpha+1)(\alpha+2)(\alpha+3)}\right)^{1/4}$	$\left\lvert\frac{4(\alpha-1)(\alpha-2)}{\alpha^2}\right\rvert^{1/4} x^{-\alpha/2}$
(3)	$(\frac{3}{3+4\alpha})^{-1/4}$	$\lvert\frac{\alpha-3}{4\alpha}\rvert^{1/4} x^{-\alpha/4} (\alpha \neq 1, 2)$

Table 1: Tail behavior of $R(x)$, (1) for kernel vs local linear fitting; (2) for a fourth order kernel vs local quadratic fitting; and (3) for local quadratic vs local cubic fitting.

and the local cubic dominates both. It should be remembered that these comparisons are asymptotic; the difference at finite sample sizes may not be as substantial.

As a comparison of local constant and local quadratic fitting, we consider the Beta mixture

$$f(x) = \frac{2}{3}2(1-x) + \frac{1}{3}\frac{19!}{9!^2}x^9(1-x)^9$$

with support $\mathcal{X} = [0, 1]$; the local constant and local quadratic estimates were applied samples of size 200. In Figure 1 we show the pointwise 10 and 90 percentiles of the distribution of $\hat{f}(x)$, estimated using 100 replications. The local constant fit used a bandwidth $h = 0.15$, and the local quadratic fit used a bandwidth of $h = 0.30$. The weight function is $K(u) = (1 - |u|^3)^3 I_{[-1,1]}(u)$. Except at the left endpoint, these result in estimates with similar variability. However, the local constant method has increased bias in the peak region.

Figure 1 about here.

The slope of the density at the left boundary is fairly small, so the boundary bias of the local constant fit is only marginally evident here. The big feature is the larger variability of the local quadratic fit. This can be alleviated through the use of larger bandwidths in boundary regions. As a simple fix of similar variability problems in the case of local regression, Cleveland (1979) recommends the use of nearest neighbor based bandwidths, so the width of smoothing windows varies with the fitting point h, and always contains a fixed number of observations.

Both methods have problems at the right boundary, since the zero of the density becomes a singularity of the log density; this cannot be modeled by a local polynomial. One possible solution to this would be to consider boundary models including suitable non-polynomial terms. We do not pursue this here.

The difference between fits of various orders becomes much more substantial in more than one dimension. As a bivariate example consider a trimodal distribution used by Friedman, Stuetzle and Schroeder (1984). This consists of an equal mixture of three bivariate standard normal densities. A samples of size 225 is drawn.

Figure 2 shows the results of fitting a kernel estimate in two dimensions; the product tricube kernel is used, with a variable bandwidth covering 25% of the data. The small bandwidth results in a noisy estimate: one peak is quite sharp; another has a flat region near the peak. Moreover, the observed peak heights are 0.037, 0.032 and 0.044, substantially less than the true peak height of about 0.053. To reduce this bias would require an even smaller bandwidth.

Figure 2 about here.

Figure 3 applies a local quadratic fit (without the cross product term) to the same dataset. A larger bandwidth is used, here covering 40% of the data. The observed peak heights here are 0.047, 0.043 and 0.052, and the observed shape of the humps is much closer to that expected for the trimodal normal mixture.

Figure 3 about here.

REFERENCES

BURKILL, J. C. and BURKILL, H. (1970). *A second course in Mathematical Analysis.* Cambridge University Press.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assn.* **74**, 829-836.

CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assn.* **83**, 596-610.

CLEVELAND, W. S. and GROSSE, E. H. (1991). Computational methods for local regression. *Statist. and Computing* **1**, 47-62.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assn.* **87**, 998-1004.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

FRIEDMAN, J. H., STUETZLE, W. and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assn.* **79**, 599-608.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815-841.

HASTIE, T. and LOADER, C. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statist. Sci.* **8**, 120-143.

HJORT, N. L. (1994). Dynamic likelihood hazard rate estimation. *Biometrika* **82**, to appear.

HJORT, N. L. and JONES, M. C. (1994). Locally Parametric Nonparametric Density Estimation. Statistical Research Report, Institute of Mathematics, University of Oslo.

JONES, M. C. (1990). Variable kernel density estimates and variable kernel density estimates. *Austral. J. Statist.* **32**, 361-371.

LEJEUNE, M. and SARDA, P. (1992). Smooth estimators of distribution and density functions. *Comp. Statist. & Data Anal.* **14**, 457-471.

LOADER, C. (1994). Computation of nonparametric function estimates. *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface*, 356-361.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065-1076.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832-837.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

SCOTT, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization.* Wiley, New York.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795-810.

STANISWALIS, J. (1988). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assn.* **84**, 276-283.

STOKER, T. M. (1993). Smoothing bias in density derivative estimation. *J. Amer. Statist. Assn.* **88**, 855-863.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5**, 595-620.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

STONE, C. J. (1990). Large sample inference for log-spline models. *Ann. Statist.* **18**, 717-741.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assn.* **82**, 559-567.

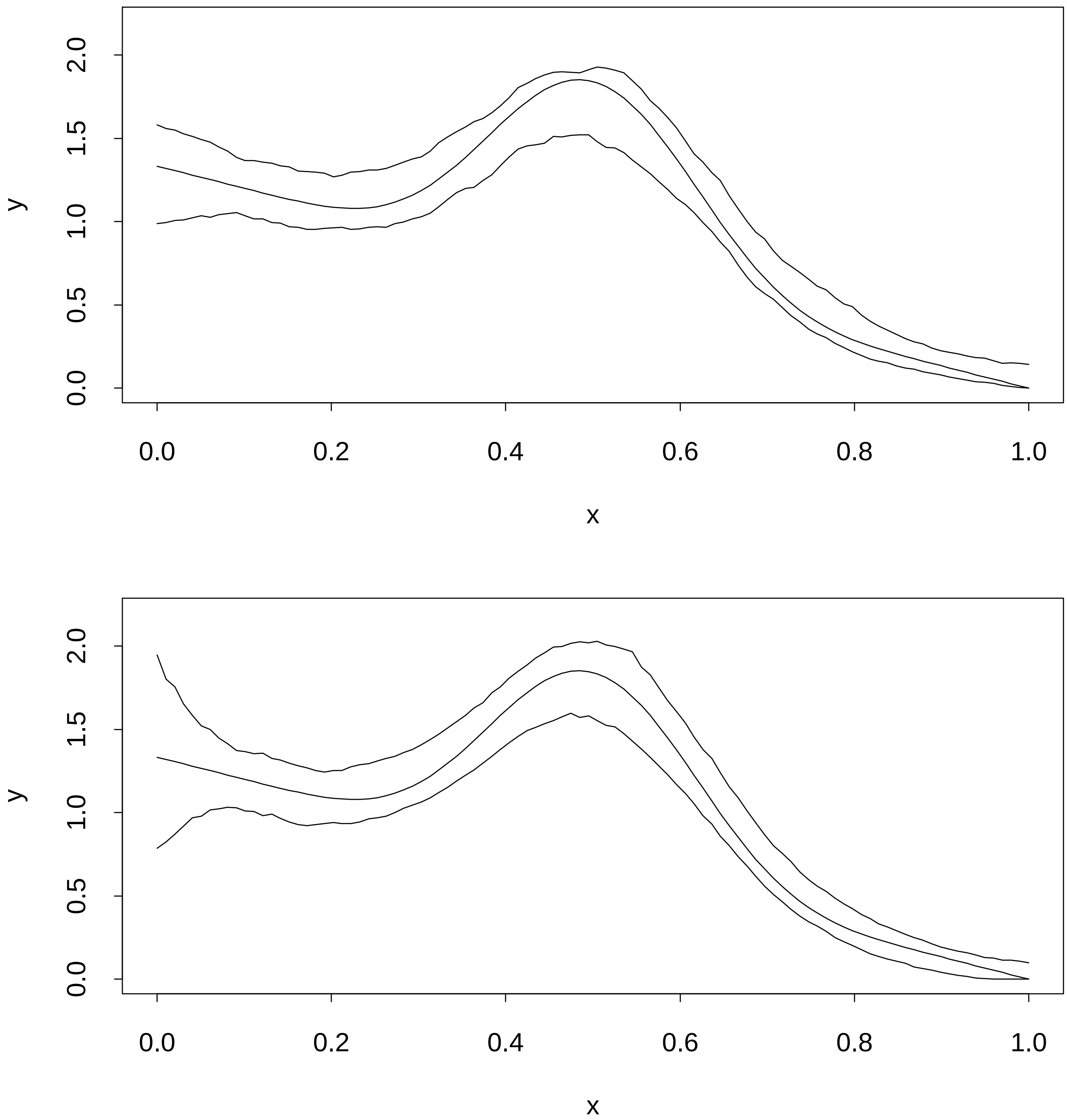

Figure 1: Estimating a $\frac{2}{3}$beta$(1,2) + \frac{1}{3}$beta$(10,10)$ density, using local constant fitting (top) and local quadratic fitting (bottom). The 10 and 90 percentiles of the distribution of $\hat{f}(x)$, for samples of size 200, are estimated from 100 Monte Carlo replications.

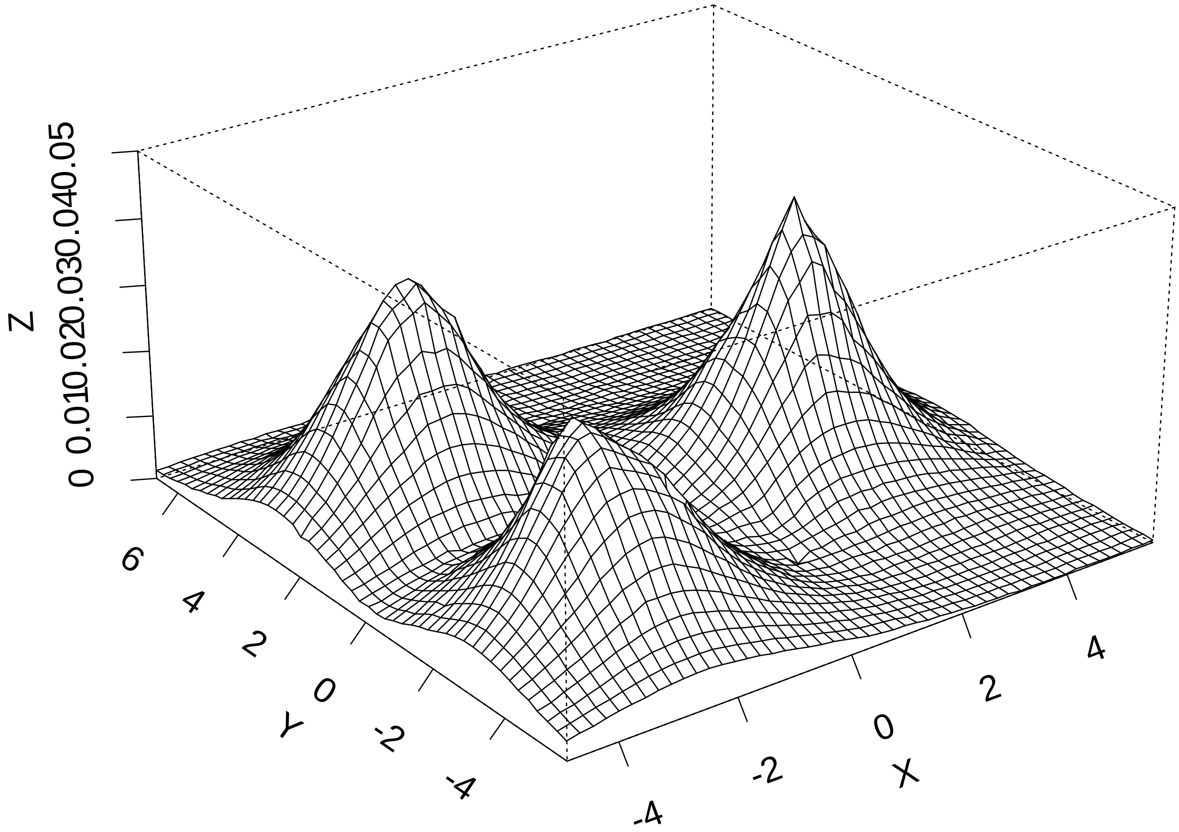

Figure 2: Two dimensional density: 225 observations from a trimodal distribution. A local constant fit with a variable bandwidth covering 25% of the data.

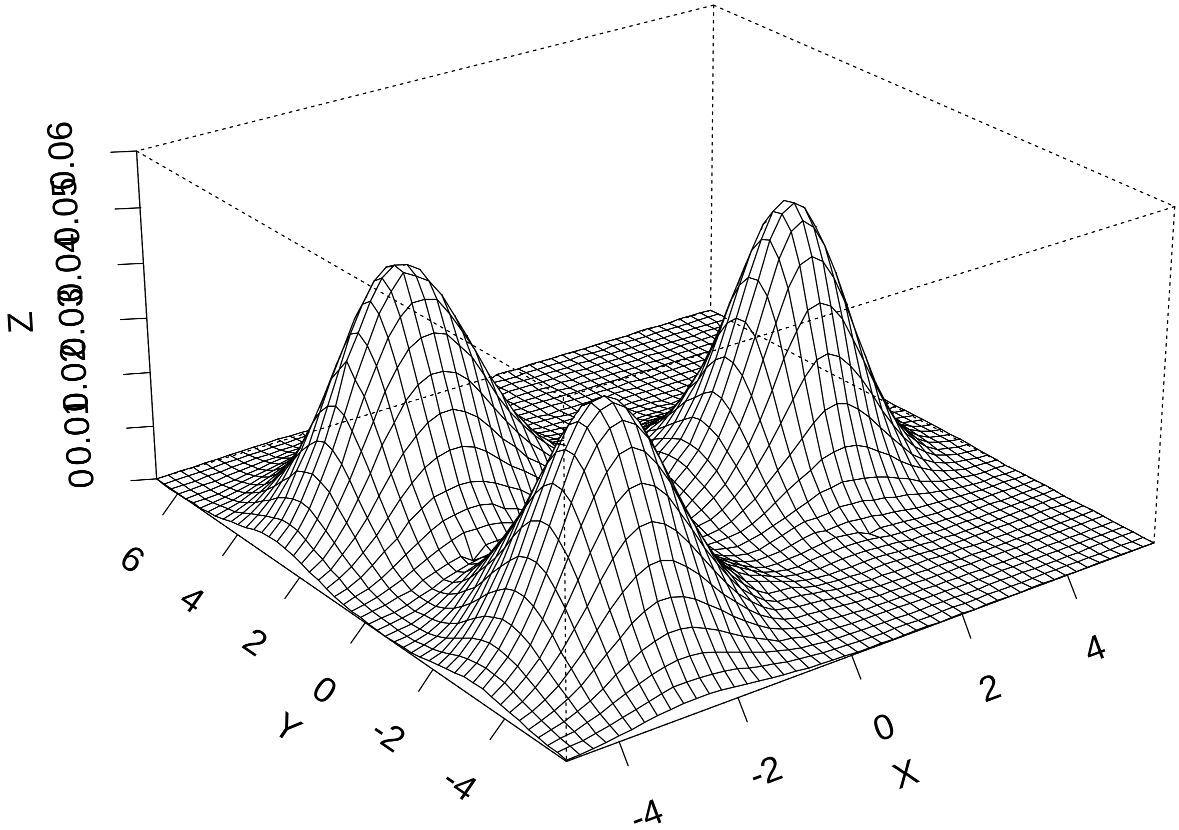

Figure 3: Two dimensional density: Local marginal quadratic fit with bandwidth covering 40% of the data.