Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation

David Ruppert *

November 8, 1995

Abstract

A data-based local bandwidth selector is proposed for nonparametric regression by local fitting of polynomials. The estimator, called the empirical-bias bandwidth selector (EBBS), is rather simple and easily allows multivariate predictor variables and estimation of any order derivative of the regression function. EBBS minimizes an estimate of mean square error consisting of a squared bias term plus a variance term. The variance term used is exact, not asymptotic, though it involves the conditonal variance of the response given the predictors that must be estimated. The bias term is estimated empirically, not from an asymptotic expression. Thus, EBBS is similar to the "double smoothing" approach of Härdle, Hall, and Marron, but is developed here for a far wider class of estimation problems than what those authors consider. EBBS is tested on simulated data and its performance seems quite satisfactory. Local polynomial smoothing of a histogram is a highly effective technique for density estimation, and several of the examples involve density estimation by EBBS applied to binned data.

Key words and phrases. Curve and surface fitting, derivative estimation, heteroscedasticity, local bandwidth, local regression, variance function estimation.

Short title. Empirical-bias bandwidths.

*David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853 (E-mail: davidr@orie.cornell.edu). This research was supported by NSF Grant DMS-9306196.

1 Introduction

In multivariate nonparametric regression, we have data $\{(\boldsymbol{X}_i, Y_i) : 1, \ldots, n\}$, where for each i, Y_i is a response and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^t$ is a d-dimensional vector of predictors. We assume the heteroscedastic model

$$Y_i = m(\boldsymbol{X}_i) + \sigma(\boldsymbol{X}_i)\epsilon_i, \ i = 1, \ldots, n, \tag{1}$$

where $m(\cdot)$ and $\sigma(\cdot)$ are smooth functions specifying the conditional mean and standard deviation of Y_i given $\boldsymbol{X}_i$, and ϵ_i has mean 0 and variance 1. We will assume that the ϵ_i's are mutually independent. The $\boldsymbol{X}_i$'s can be fixed or random, but in the latter case we will *not* make assumptions about their joint distribution. We will concentrate on estimation of $m(\cdot)$ and its derivatives, treating $\sigma(\cdot)$ as a nuisance parameter.

Fitting polynomials locally by weighted least squares is an appealing method of estimation, with good theoretical properties (Fan (1992)). The multivariate case ($d > 1$) was discussed by Cleveland and Devlin (1988) and Ruppert and Wand (1994), the latter concentrating on theory and also discussing estimation of derivatives in the univariate case. In this paper, we will treat the general multivariate case and estimation of derivatives of $m(\cdot)$ of any order. The interesting examples in Cleveland and Devlin (1988) show that multivariate local polynomial regression can be highly effective data-analytic tool, at least in two or three dimenions.

Suppose we want to estimate $m(\boldsymbol{x})$ for some given $\boldsymbol{x} = (x_1, \ldots, x_d)^t$. Here is how the local polynomial estimate of $m(\boldsymbol{x})$ is defined. Let $\boldsymbol{X} = (X_1, \ldots, X_d)^t$ be a d-dimensional variable, and consider a general degree-p polynomial in $\boldsymbol{X}$

$$P_p(\boldsymbol{X}; \boldsymbol{\beta}) = \sum_{K=0}^{p} \sum_{k_1+\cdots+k_d=K} \beta_{k_1,\ldots,k_d} \prod_{j=1}^{d} (X_j - x_j)^{k_j}, \tag{2}$$

where $\boldsymbol{\beta} = \{\beta_{k_1,\ldots,k_d} : k_1 + \cdots + k_d = K \text{ and } K = 0, \ldots, p\}$ is a vector of coefficients. Notice that $P_p(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_{0,\ldots,0}$. Therefore, to estimate $m(\boldsymbol{x})$, we approximate m locally by P_p, estimate $\boldsymbol{\beta}$ by weighted least squares, and then estimate $m(\boldsymbol{x})$ by $\widehat{\beta}_{0,\ldots,0}$. Given a bandwidth h, the weights used in the least squares estimation are $w_i(\boldsymbol{x}; h) = K_h(\boldsymbol{X}_i - \boldsymbol{x})$, where $K(\cdot)$ is a d-variate nonnegative function and $K_h(\cdot) = K(\cdot/h)/h^d$. When $d > 1$ we could consider a matrix of bandwidths, but, for simplicity, in this paper we restrict attention to h scalar. Let $\widehat{\boldsymbol{\beta}}(h)$ be the value of $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n} w_i(\boldsymbol{x}; h)\Big\{Y_i - P_p(\boldsymbol{X}_i; \boldsymbol{\beta})\Big\}^2. \tag{3}$$

Our estimate of $m(\boldsymbol{x})$ is $\widehat{m}(\boldsymbol{x}; h) = \widehat{\beta}_{0,\ldots,0}(h)$.

We can also estimate derivatives of $m(\cdot)$. Let $K \leq p$ be a nonnegative integer and let $\boldsymbol{k} = (k_1 \ldots, k_d)$ be a vector of nonnegative integers such that $k_1 + \cdots + k_d = K$. The mixed

Kth order derivative

$$m^{(\boldsymbol{k})}(\boldsymbol{x}) = \frac{\partial^K}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} m(\boldsymbol{x}) \tag{4}$$

is estimated by

$$\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}; h) = \frac{\partial^K}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} P_p(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}) = \left(\prod_{j=1}^{d} (k_j!) \right) \widehat{\beta}_{k_1,\ldots,k_d}(h). \tag{5}$$

For accurate estimation, the choice of the smoothing parameter h is crucial. When h is relatively small, $P_p(\cdot; h)$ is a good approximation to $m(\cdot)$ and $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}; h)$ has small bias, but then fewer data are used and $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}; h)$ is more variable. The best choice of h involves a tradeoff between bias and variance, which depend on the order of the derivative being estimated, the sample size, the empirical distribution of the $\boldsymbol{X}_i$'s, $\sigma(\cdot)$, and the values of the derivatives of $m(\cdot)$ beyond the pth order.

Several authors have considered automatic, that is data-based, selection of h. One basic issue is whether h should depend on $\boldsymbol{x}$ (local bandwidths) or not (global bandwidths). Ruppert, Sheather, and Wand (1995) consider a data-based global bandwidth, and in an innovative paper Fan and Gjibels (1995) proposed an automatic local bandwidth selector. Also the *loess* function in S+ (Cleveland, Grosse, and Shyu 1993) automatically varies a local bandwidth, but does require a user-specified (i.e., not automatic) global parameter called the span. The Ruppert, Sheather, and Wand (1995) proposal is restricted to $d = 1$ and to estimation of $m(\cdot)$, not its derivatives.

Fan and Gijbels (1995) do consider higher derivatives. They take an asymptotic expression for the bias, which involves derivatives of higher order than the degree, p, of the polynomial used in (5). These higher order derivatives are estimated and the estimates are plugged into the bias expression. Their procedure can become rather complicated in the multivariate case, especially when estimating derivatives. For example, estimating the gradient of a bivariate function is a problem not uncommon in engineering applications. In fact, the current research was partially motivated by collaboration with an engineer, Professor Stephen Pope at Cornell, who is faced with this problem when analyzing Monte Carlo data from the simulation of turbulence and combustion. If one uses local quadratic regression, then one must fit at least 10-parameter local cubics to estimate bias. In fact, following Fan and Gijbels suggestions one fits 15-parameter local quartics. In contrast, the procedure introduced in this manuscript only fits 6-parameter local quadratics.

In this paper, a new method of bandwidth estimation, called empirical-bias bandwidth selection (EBBS), is proposed. EBBS has the following features:

1. No asymptotic expressions for the bias are used. Rather, bias is estimated empirically by calculating $\hat{m}^{(\boldsymbol{k})}(\boldsymbol{x}; h)$ on a grid of h values, and then modeling the behavior of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}; h)$ as h varies. Thus, EBBS borrows the bootstrap philosophy of replacing

asymptotic approximations by computation. However, unlike most applications of the bootstrap, EBBS does not use simulation or resampling.

2. No asymptotic expressions for variance are used. Rather, an expression for the exact, finite sample variance is used.

3. There is no need to fit polynomials of degree higher than the degree p used in (5) to estimate $m^{(\boldsymbol{k})}$ itself.

4. Bandwidth selection for estimation of derivatives and when $\boldsymbol{x}$ is multivariate is simple.

5. The method can easily accommodate $p - K$ both odd and even.

Fan and Gijbels (1995) methodology also has features (2), and, in fact, our estimate of variance is the same as theirs. Also, to some extent, their methodology has feature (1). Specifically, they use an expression for the finite-sample bias that involves the unknown m, but then use a large-sample Taylor approximation to m. Another paper where bias is empirically modeled is by Härdle, Hall, and Marron (1992), who propose the technique of "double smoothing." Although, bias is modeled here somewhat differently than by those authors, the ideas here are similar to and inspired by their work. What is new here, besides our empirical modeling of bias, is that we consider local polynomial regression, multivariate $\boldsymbol{X}_i$'s, and estimation of derivatives. In contrast, Härdle, Hall, and Marron (1992) study only kernel estimation (equivalently $p = 0$), univariate $\boldsymbol{X}_i$'s, and only estimation of $m(\cdot)$.

The asymptotic distribution of $\widehat{m}^{(\boldsymbol{k})}(x; h)$ is of a rather different and more complex form when $p - K$ is even rather than odd. Therefore, Fan and Gijbels (1995) restrict their procedure to the odd case. This restriction is justified, to some extent, by an asymptotic minimax result showing that $p - K$ odd is optimal in a somewhat narrow, technical sense. In practice, however, $p - K$ equal to 2 can be appealing. For example, when estimating $m(\cdot)$ itself, quadratics work rather nicely and cubics seem less attractive, especially when $\boldsymbol{x}$ is multivariate.

2 The algorithm

We will estimate $m(\cdot)$ on a grid $G_x = \{\boldsymbol{x}_\ell : \ell \in \mathcal{L}\}$ where $\mathcal{L}$ is an index set. Typically G_x is a rectangular grid in R^d and $\mathcal{L}$ equals $\{1, \ldots, m_1\} \times \ldots \times \{1, \ldots, m_d\}$ for some integers $m_1, \ldots, m_d$. Notice the difference between a data point, $\boldsymbol{X}_i$, $i = 1, \ldots, n$, where a response has been observed, and a grid point $\boldsymbol{x}_\ell$, $\ell \in \mathcal{L}$, where m is estimated.

For some fixed $\boldsymbol{k}$ and p and for each $\boldsymbol{x}_\ell \in G_x$, we estimate $\mathrm{MSE}(\boldsymbol{x}_\ell; h)$, the mean square error of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h)$ as a function of h, by separately estimating the bias and variance functions. In this paper, all expectations are meant to be conditional on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ unless otherwise noted. Thus, bias, variance, and MSE will mean conditional on the $\boldsymbol{X}_i$'s. We

denote the estimated MSE function by $\widehat{\mathrm{MSE}}(\boldsymbol{x}_\ell; h)$ and then choose $\widehat{h}(\boldsymbol{x}_\ell)$ to minimize $\widehat{\mathrm{MSE}}(\boldsymbol{x}_\ell; h)$.

2.1 Estimating Bias

The bias of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h)$ is estimated as follows. The idea is to use estimates at several bandwidths to fit a model for the expectation of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h)$ as a function of h. Fix $\boldsymbol{x}_\ell$ and let h_0 be a point where we are to estimate the bias of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h_0)$. Let $J_b > 1$ be an integer and let $h_0^1, \ldots, h_0^{J_b}$ be in a neighborhood of h_0. Calculate $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h_0^j)$, $j = 1, \ldots, J_b$.

Next for some $t \geq 1$ fit the curve

$$\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h) \approx bc_0(\boldsymbol{x}_\ell) + bc_{p+1-K}(\boldsymbol{x}_\ell)h^{p+1-K} + \cdots + bc_{p+t-K}h^{p+t-K} \tag{6}$$

to the "data" $\{(h_0^j, \widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h_0^j))\colon\ j = 1, \ldots, J_b\}$ by ordinary least squares, say. The notation "bc" means "bias coefficient." The terms after the first in (6) represents bias, so we estimate the bias of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}; h_0)$ by

$$\widehat{bc}_{p+1-K}(\boldsymbol{x}_\ell)h_0^{p-K+1} + \cdots + \widehat{bc}_{p+t-K}h_0^{p-K+t}. \tag{7}$$

The form of the bias function in (6) and (7) is suggested by asymptotics; see Ruppert and Wand (1994) for the case $t = 1$ and Huang (1995, section 3.2) for $t = 3$. Besides the use of asymptotics to suggest a model, we make no use of asymptotics in our estimation of the bias function.

When j is odd, then $bc_j = 0$ except in the "boundary region," that is at points within ν bandwidths of the boundary where the support of the kernel is $[-\nu, \nu]$ ($\nu = 1$ in our examples); see Ruppert and Wand (1994) for the case $t = 1$. Thus, for $p - K$ even $t \geq 2$ should be used since $bc_{p+1-K} = 0$ for interior values of x. In the current version of EBBS, bc_j is NOT set to 0 for j odd but rather is estimated from the data. One reason for this is that the boundary region cannot be determined until the bandwidths are selected. Also, often in practice a large portion of the observations are in the boundary region.

2.2 Estimating Variance

Since $\widehat{\boldsymbol{\beta}}$ is a weighted least-squares estimate, estimating the variance of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h)$ is straightforward. Define $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma^2(\boldsymbol{X}_1), \ldots, \sigma^2(\boldsymbol{X}_n)\}$ and $\boldsymbol{W}_\ell = \mathrm{diag}(w_1(\boldsymbol{x}_\ell; h), \ldots, w_n(\boldsymbol{x}_\ell; h))$. Let $\boldsymbol{\mathcal{X}}_\ell$ be the "X-matrix" for the the weighted least-squares problem of minimizing (3) with $\boldsymbol{x} = \boldsymbol{x}_\ell$. By standard least-squares theory, the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{\mathcal{X}}_\ell^t \boldsymbol{W}_\ell \boldsymbol{\mathcal{X}}_\ell)^{-1}(\boldsymbol{\mathcal{X}}_\ell^t \boldsymbol{W}_\ell \boldsymbol{\Sigma} \boldsymbol{W}_\ell \boldsymbol{\mathcal{X}}_\ell)(\boldsymbol{\mathcal{X}}_\ell^t \boldsymbol{W}_\ell \boldsymbol{\mathcal{X}}_\ell)^{-1}. \tag{8}$$

We now assume that $\sigma(\boldsymbol{X}_j) \approx \sigma(\boldsymbol{x}_\ell)$ for j such that $w_j(\boldsymbol{x}_\ell; h) \neq 0$; this approximation is increasingly accurate as h tend to zero, assuming that $\sigma(\cdot)$ is continuous. Fan and Gijbels

(1995) use the same approximation. Recall that $\boldsymbol{k}$, the order of the derivative we are estimating, is $(k_1, \ldots, k_d)$ where $k_1 + \ldots + k_d = K \leq p$. Thus, some column of $\boldsymbol{\mathcal{X}}_\ell$, say that rth, is of the form $(\prod_{j=1}^d (\boldsymbol{X}_{1j} - \boldsymbol{x}_{\ell j})^{k_j}, \ldots, \prod_{j=1}^d (\boldsymbol{X}_{nj} - \boldsymbol{x}_{\ell j})^{k_j})^t$. Then, using (5) and (8), the variance of $\widehat{m}^{(\boldsymbol{k})}(x; h)$ is approximately

$$\sigma^2(\boldsymbol{x}_\ell) \left(\prod_{j=1}^d (k_j!) \right)^2 \left[(\boldsymbol{\mathcal{X}}_\ell^t \boldsymbol{W}_\ell \boldsymbol{\mathcal{X}}_\ell)^{-1} (\boldsymbol{\mathcal{X}}_\ell^t \boldsymbol{W}_\ell^2 \boldsymbol{\mathcal{X}}_\ell) (\boldsymbol{\mathcal{X}}_\ell^t \boldsymbol{W}_\ell \boldsymbol{\mathcal{X}}_\ell)^{-1} \right]_{rr}, \tag{9}$$

since $\widehat{\beta}_{k_1,\ldots,k_d}$ is the rth component of $\widehat{\boldsymbol{\beta}}$. Define $\widehat{v}(\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h))$ to be equal to (9) with $\sigma^2(\boldsymbol{x}_\ell)$ omitted.

2.3 Selecting bandwidths

Now assume that we have an estimate $\widehat{\sigma}^2(\boldsymbol{x}_\ell)$ of $\sigma^2(\boldsymbol{x}_\ell)$; we will discuss this estimate in section 2.4. Using (7) to estimate bias and (9) to estimate variance, we obtain an estimate of the MSE of $\widehat{m}(\boldsymbol{x}_\ell; h_0)$:

$$\widehat{\mathrm{MSE}}(\boldsymbol{x}_\ell; h_0) = [\widehat{bc}_{p+1-K}(\boldsymbol{x}_\ell) h_0^{p+1-K} + \cdots + \widehat{bc}_{p+t-K} h_0^{p+t-K}]^2 + \widehat{\sigma}^2(\boldsymbol{x}_\ell) \widehat{v}(\widehat{m}(\boldsymbol{x}_\ell; h_0)). \tag{10}$$

Since (10) requires J_b fits to calculate $\widehat{\mathrm{MSE}}$ at a single h_0, it is helpful to reuse some of these fits when recalculating (10) at a new, but nearby, value of h_0. Here is the algorithm used in this paper. User supplied lower and upper bounds, h_a and h_b, on the bandwidths are needed. Then let $H_1 = \{h_1, \ldots, h_{M_1}\}$ be a grid of M_1 points from h_a to h_b, equally spaced on the log scale. We calculate $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell; h_j)$, $j = 1, \ldots, M_1$. Let J_1 and J_2 be integers, such that $J_1 + J_2 \geq t - 1$ and and J_2 is positive. Let $J_1^* = \max\{0, J_1\}$. We estimate bias at each of $h_{1+J_1^*}, \ldots, h_{M_1 - J_2}$, using $J_b = 1 + J_1 + J_2$ and using $h_{j-J_1}, \ldots, h_{j+J_2}$ as the neighboring values of h_j when fitting (6). The result is that we have an estimate of the MSE of $\widehat{m}^{(\boldsymbol{k})}(\boldsymbol{x}_\ell, h_j)$, $j = 1 + J_1^*, \ldots, M_1 - J_2$. We then use cubic interpolation to interpolate this MSE function onto a finer grid of $M_2 > M_1 - J_1^* - J_2$ values, $H_2 = \{h_1^*, \ldots, h_{M_2}^*\}$ where $h_1^* = h_{1+J_1^*}$ and $h_{M_2}^* = h_{M_1 - J_2}$. We will see that asymptotics suggest taking J_1 negative so that bias at h_0 is estimated using only bandwidths larger than h_0, though finite sample experimentation does not show a distinct advantage to negative values of J_1; see sections 3 and 4.

For any $h \in H_2$, $\widehat{\mathrm{MSE}}(\boldsymbol{x}_\ell; h)$ is typically quite rough as a function of $\boldsymbol{x}_\ell$. Therefore, we introduce a smoothed version of $\widehat{\mathrm{MSE}}(\boldsymbol{x}_\ell; h)$, called $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h, N_\ell)$, and defined to be $\widehat{\mathrm{MSE}}(\boldsymbol{x}; h)$ averaged over those $\boldsymbol{x}$ in N_ℓ, a given neighborhood of $\boldsymbol{x}_\ell$. The proposed local bandwidth is

$$\widehat{h}(\boldsymbol{x}_\ell; N_\ell) = \text{first local min}\{\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h, N_\ell) : h \in H_2\}, \tag{11}$$

i.e., $\widehat{h}(\boldsymbol{x}_\ell; N_\ell)$ is h_j^* where j is the smallest integer in $\{2, \ldots, M_2 - 1\}$ such that $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h_{j-1}, N_\ell) > \widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h_j, N_\ell)$ and $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h_j, N_\ell) < \widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h_{j+1}, N_\ell)$ and if there is

no such integer $\widehat{h}(\boldsymbol{x}_\ell; N_\ell) = h_1^*$ if $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; \cdot)$ is nondecreasing on H_2 and $\widehat{h}(\boldsymbol{x}_\ell; N_\ell) = h_{M_2}^*$ if $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; \cdot)$ is nonincreasing on H_2. It is crucial that this local minimum be used rather than the global minimum. Our method of estimating bias will greatly underestimate bias when h is so large that all features of m are smoothed away. Thus, the global minimum of $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h, N_\ell)$ is 0 and is reached as $h \to \infty$.

Even with the smoothing of the estimated MSE, $\widehat{h}(\boldsymbol{x}_\ell; N_\ell)$ is generally somewhat rough as a function of $\boldsymbol{x}_\ell$, and it is advisable to smooth $\widehat{h}(\boldsymbol{x}_\ell; N_\ell)$ to $\tilde{h}(\boldsymbol{x}_\ell; N_\ell)$, say by a moving average with a triangular weighting function.

If the sample size, n, is small, then one might use a global bandwidth since a locally varying bandwidth could prove unstable. To get a global bandwidth we simple let the local neighborhood, N_ℓ, of $\boldsymbol{x}_\ell$ always equal the entire grid, G_x and give all points equal weight when calculating $\widehat{\mathrm{SMSE}}$. Then, of course, the minimization in (11) needs to be done only once.

We will abbreviate $\tilde{h}(\boldsymbol{x}_\ell; N_\ell)$ to $\tilde{h}(\boldsymbol{x}_\ell)$ and we will use $\widehat{m}(\boldsymbol{x}_\ell)$ to denote $\widehat{m}(\boldsymbol{x}_\ell, \tilde{h}(\boldsymbol{x}_\ell))$.

2.4 Estimating the variance function

We estimate $\sigma^2(\cdot)$ by smoothing squared residuals as in Ruppert, Wand, Holst, and Hössjer (1995). Specifically, we start with the estimates $\{\widehat{m}(\boldsymbol{X}_i; h_m) : i = 1, \ldots, n\}$, where h_m is a user-supplied bandwidth for estimating the mean function. Next, we form squared residuals

$$e_i^2 = \{Y_i - \widehat{m}(\boldsymbol{X}_i; h_m)\}^2.$$

As in Ruppert, Wand, Holst, and Hössjer (1995), let V_i be the factor such that $\mathrm{Var}(e_i^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \sigma^2(X_i)V_i$ if $\sigma^2(\cdot)$ is constant in a neighborhood of X_i. Calculating V_i is a routine application of weighted least squares theory. Then we smooth the e_i^2's, $i = 1, \ldots, n$, by local polynomial regression using the same x-grid, G_x, as for estimation of $m(\cdot)$. The bandwidth for smoothing the squared residuals is also chosen by EBBS, but for this purpose we do *not* need a variance function for the squared residuals. Rather, we assume that the ϵ_i's in (1) come from a scale family so that for some $\kappa > 0$

$$\mathrm{Var}(\epsilon_i^2) = \kappa\{E(\epsilon_i^2)\}^2 = \kappa\sigma^4(\boldsymbol{X}_i).$$

To estimate κ we partition the data into blocks according to their $\boldsymbol{X}_i$ values. Within each block we calculate the ratio of the variance of the squared residuals to their squared mean. These ratios are averaged over blocks to form $\widehat{\kappa}$. When estimating the MSE of $\widehat{\sigma}^2(\boldsymbol{x}_\ell; h_0)$ at a given $\boldsymbol{x}_\ell$ and h_0, the variance function of the squared residuals is estimated by $\widehat{\kappa}\widehat{\sigma}^4(\boldsymbol{x}_\ell; h_0)$.

Next we smooth the V_i's in exactly the same manner, and then we define $\widehat{\sigma}^2(\boldsymbol{X}_i)$ as the ratio of the smooth of the squared residuals at $\boldsymbol{X}_i$ to the smooth of the V_i's at $\boldsymbol{X}_i$. In the examples, we smooth the residuals and the V_i's by local linear regression, but higher order polynomials could of course be used.

We recommend choosing h_m quite small to eliminate bias from the estimate of m. The large variance of $\widehat{m}(\boldsymbol{X}_i; h_m)$ caused by a small value of h_m is not a problem, since its effect on the estimate of $\sigma^2(\cdot)$ is offset by the use of the V_i's which is essentially a degrees of freedom correction. Monte Carlo experiments in section 4.2 show that the value of h_m has little, if any, effect on the final estimate of $m(\cdot)$ using the EBBS bandwidth.

If estimation of σ^2 is a primary interest, Ruppert, Wand, Holst, and Hössjer (1995) suggest a four step algorithm:

1. Smooth the Y_i's using h_m to obtain $\widehat{m}$.
2. Smooth the squared residuals from $\widehat{m}$ in step 1 using EBBS.
3. Re-estimate $m(\cdot)$ using EBBS and the variance function of the Y_i's obtained in step 2.
4. Re-estimate σ^2 (or estimate some derivative of σ^2) using EBBS applied to the residuals from $\widehat{m}$ in Step 3.

The first three steps are what we are suggesting here, and in step 3 we can estimate $m^{(\boldsymbol{k})}$ for and $\boldsymbol{k}$.

2.5 Density estimation by nonparametric regression

The traditional method of density estimation, kernel estimation with a global bandwidth, has several problems, including serious boundary bias and lack of spatial adaptivity. These problems can be corrected by a variety of special techniques, such as boundary kernels, transformations, and adaptive bandwidths; see Wand and Jones (1995) for discussion and further references. It is interesting, however, that local polynomial regression with local bandwidths, e.g., EBBS, can be a highly effective method of density estimation. Density estimation by local regression has been used by others, e.g., Cheng, Fan, and Marron (1993) and Cheng (1994). The idea is to use EBBS to smooth a histogram estimate of the density.

Suppose that $U_1, \ldots, U_M$ are iid from a density $m(u)$. For simplicity, we assume that the U_i's are univariate. Form a histogram with a very large number of equal-length bins, say 600, and normalized to have area 1. Then let n be the number of bins, let $Y_i, \ldots, Y_n$ be the bin heights, and let $X_1, \ldots, X_n$ be the bin centers. If C_i is the ith bin count and if L_x is the length of each bin, then

$$Y_i = \frac{C_i}{n\, L_x}.$$

Because n is very large, Y_i is a nearly unbiased, though highly variable, estimate of $m(X_i)$. Therefore, we can estimate m by smoothing the data $\{(X_i, Y_i)\}_{i=1}^n$ using EBBS.

Since the C_i will be nearly Poisson distributed,

$$\mathrm{Var}(Y_i) \approx \frac{1}{n\, L_x} E(Y_i).$$

As for smoothing with squared residuals as the "response" as in section 2.4, there is no need here to estimate the variance function of the normalized bin counts. Rather, in (10) we replace $\widehat{\sigma}^2(\boldsymbol{x}_\ell)$ by $\frac{1}{nL_x}\widehat{m}(\boldsymbol{x}_\ell; h_0)$.

In Section 5.2 it will be shown that $\widehat{m}(x;h)$ converges to a limit as $L_x \to 0$ so that $n \to \infty$. Thus, there is no upper limit to n except that imposed by time and memory considerations. However, $n \approx 500$ seems large enough for most practical applications.

2.6 Remarks

Except for converting density estimation into nonparametric regression, no use of binning is made in this implementation. Rather, following the design of *loess* (Cleveland, Grosse, and Shyu 1988), $\widehat{m}^{(\boldsymbol{k})}$ is computed on a fairly sparse grid (G_x) and then interpolated cubicly to obtain the estimate on a fine grid for plotting or other purposes. The extent to which binning will speed up EBBS, especially if $d > 1$ is unclear. Since the method of estimating $\sigma^2(\cdot)$ described in section 2.4 requires a fit at each data point X_i in order to calculate each V_i, binning would speed up this part of the EBBS algorithm.

3 Asymptotic Theory of Bias Estimation

In this section we investigate the asymptotic theory of empirical bias estimation in a special case, univariate local linear estimation of m ($d = 1$, $p = 1$, $K = 0$). The results here are only intended for guidance when choosing J_1, J_2, and other tuning parameters described in section 2. A more complete investigation including rates of convergence of $\widehat{h}(x)$ is planned for the future.

Let x be a fixed point where we wish to estimate m. Throughout this section we assume that m has five continuous derivatives in a neighborhood of x, that $X_1, \ldots, X_n$ are iid with density f, that f is twice continuously differentiable in a neighborhood of x, that $f(x) > 0$, and that σ^2 is continuous in a neighborhood of x.

We start with an expansion for the conditional bias that requires the smoothness assumptions on m and f that we have just imposed:

$$E(\widehat{m}(x;h)) = m(x) + bc_2h^2 + bc_3h^3 + bc_4h^4 + O_P(h^5 + n^{-1}h^{3/2}). \tag{12}$$

In (12) the expectation is conditional on $X_1, \ldots, X_n$, as are all expectations, variances, and covariances in this paper. Equation (12) is a refinement of earlier results (Fan; 1994) and is due to Huang (1995, Theorem 3.1). Huang only works with the case that x is an interior point and then $bc_3 = 0$ and $bc_2 = (m^{(2)}(x)/2)\mu_2(K)$ where $\mu_k(K) := \int u^k K(u)\,du$. Huang also shows that at interior points

$$bc_4 = \{\mu_2^2(K) - \mu_4(K)\}\frac{m^{(2)}(x)}{2}\left\{\frac{f'(x)}{f(x)}\right\}^2 - \frac{f^{(2)}(x)}{2f(x)} + \frac{m^{(4)}(x)}{4!}\mu_4(K). \tag{13}$$

We assume that

$$h_j = \alpha_j n^{-\gamma},\ j = 1, \ldots, J_b, \text{ for some } \gamma \in (0,1) \text{ and } \alpha_1, \ldots, \alpha_{J_b} > 0. \tag{14}$$

Define $\boldsymbol{S}$ to be the $J_b \times J_b$ matrix with entries

$$\boldsymbol{S}_{ij} = \frac{1}{\alpha_i \alpha_j} \int K\left(\frac{u}{\alpha_i}\right) K\left(\frac{u}{\alpha_j}\right) du.$$

The following result is a standard calculation:

Theorem 1 *Assume that (14) holds. Then*

$$\text{Cov} \begin{pmatrix} \widehat{m}(x; h_1) \\ \vdots \\ \widehat{m}(x; h_{J_b}) \end{pmatrix} = \sigma^2(x) \frac{1 + o_P(1)}{n^{1-\gamma} f(x)} \boldsymbol{S}. \tag{15}$$

Now suppose that we fit the model

$$\widehat{m}(x; h_j) \approx bc_0 + bc_2 h_j^2 + \cdots + bc_{t+1} h_j^{t+1},\ j = 1, \ldots, J_b, \tag{16}$$

for $t = 1$, 2, or 3 and $J_b \geq t + 1$. Let $\boldsymbol{D} = \text{diag}(1, n^{-\gamma}, \ldots, n^{-(t+1)\gamma})$,

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha_1^2 & \cdots & \alpha_1^{t+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_{J_b}^2 & \cdots & \alpha_{J_b}^{t+1} \end{pmatrix}, \text{ and } \widehat{\boldsymbol{m}} = \begin{pmatrix} \widehat{m}(x; h_1) \\ \vdots \\ \widehat{m}(x; h_{J_b}) \end{pmatrix}.$$

If (16) is fit by ordinary least squares, then

$$\begin{pmatrix} \widehat{bc}_0 \\ \widehat{bc}_2 \\ \vdots \\ \widehat{bc}_{t+1} \end{pmatrix} = [(\boldsymbol{AD})^t (\boldsymbol{AD})]^{-1} (\boldsymbol{AD})^t \widehat{\boldsymbol{m}} = \boldsymbol{D}^{-1} (\boldsymbol{A}^t \boldsymbol{A})^{-1} \boldsymbol{A}^t \widehat{\boldsymbol{m}}. \tag{17}$$

Using (15) we can readily prove:

Theorem 2 *Suppose that (12) and (14) hold and we use estimator (17). Then*

$$\text{Var} \begin{pmatrix} \widehat{bc}_0 \\ \widehat{bc}_2 \\ \vdots \\ \widehat{bc}_{t+1} \end{pmatrix} = \sigma^2(x) \boldsymbol{D}^{-1} (\boldsymbol{A}^t \boldsymbol{A})^{-1} (\boldsymbol{A}^t \boldsymbol{S} \boldsymbol{A}) (\boldsymbol{A}^t \boldsymbol{A})^{-1} \boldsymbol{D}^{-1} \left(\frac{1 + o_P(1)}{n^{1-\gamma} f(x)} \right).$$

and

$$E \begin{pmatrix} \widehat{bc}_0 \\ \widehat{bc}_2 \\ \vdots \\ \widehat{bc}_{t+1} \end{pmatrix} = \begin{pmatrix} bc_0 \\ bc_2 \\ \vdots \\ bc_{t+1} \end{pmatrix} + \boldsymbol{D}^{-1} R_n,$$

where R_n is a t-dimensional random vector such that $R_n = O_P(h^{t+2} + n^{-1} h^{3/2})$ for $t = 1$, 2, or 3.

If $t = 1$ and if x is an interior point, then $R_n = O_P(h^4 + n^{-1} h^{3/2})$.

We can use Theorem 2 for insight on how to choose $h_j = \alpha_j n^{-\gamma}$, $j = 1, \ldots, J_b$, to estimate the bias coefficients. Let $t^* = t$ except that $t^* = 2$ if $t = 1$ and x is an interior point. We will see that $n^{-1}h^{3/2} = o(h^{t+2})$ for the optimal value of γ. Thus, equating the orders of magnitude of variance and squared bias of $h^j \widehat{bc}_j$ we have $n^{\gamma-1} = h^{2(t^*+1)} = n^{-2\gamma(t^*+2)}$, so that $\gamma = 1/(2t^*+5)$ for $t^* = 1, 2,$ or 3. With this choice of γ, the squared bias and variance of $\widehat{bc}_j$ are both $O_P(n^{-(2t^*+4-2j)/(2t^*+5)})$. For example, the MSE of $\widehat{bc}_2$ is $O_P(n^{-4/9})$ for $t^* = 2$. Since the MSE optimal h for estimation of m is $O(n^{-1/5})$ and the optimal γ was just found to be less than $1/5$, we see that when estimating bias of $\widehat{m}$ at h_0, asymptotics suggest using only values of h that are larger than h_0, i.e., $J_1 < 0$ and $J_2 > 0$. However, finite samples results with 200 to 500 observations do not make a clear-cut case for such a policy. In fact, for functions with rapidly changing curvature, using h on both sides of h_0 generally outperforms using only h greater than h_0.

Because of correlations between the $\widehat{m}(x; h_j)$'s, spreading the α_j far apart is not necessary. To appreciate this, consider the interesting special case when $J_b = 2$ and $t = 1$. From Theorem 2, the asymptotic variance of $n^{(1-\gamma)/2}\widehat{bc}_2$ is finite for any fixed α_1 and α_2 with $\alpha_1 \neq \alpha_2$. Moreover, this asymptotic variance converges to a finite limit as $\alpha_2 \to \alpha_1$ since

$$\widehat{bc}_2 = \frac{\widehat{m}(x; h_2) - \widehat{m}(x; h_1)}{h_2^2 - h_1^2}.$$

and therefore

$$\begin{aligned} \mathrm{Var}\{\widehat{m}(x; h_2) - \widehat{m}(x; h_1)\} &\approx \frac{\sigma^2(x)}{n^{1-\gamma} f(x)} \int \left\{ \frac{1}{\alpha_1} K\left(\frac{u}{\alpha_1}\right) - \frac{1}{\alpha_2} K\left(\frac{u}{\alpha_2}\right) \right\}^2 du \\ &\to \frac{\sigma^2(x)}{n^{1-\gamma} f(x)} \frac{(\alpha_2 - \alpha_1)^2}{\alpha_1^3} \int \{zK'(x) + K(z)\}^2 dz \end{aligned}$$

as $\alpha_2 \to \alpha_1$. Therefore, for α_2 nearly equal to α_1,

$$\mathrm{Var}(\widehat{bc}_2) \approx \frac{\sigma^2(x)}{4nh_1^5 f(x)} \int [zK'(z) + K(z)]^2 dz. \tag{18}$$

Since $m''(x) = 2bc_2/\mu_2(K)$, define $\widetilde{m}''(x) = 2\widehat{bc}_2/\mu_2(K)$. It is interesting to see how $\widetilde{m}''(x)$ compares with the "usual" estimate of $m''(x)$ obtained as the second derivative of a locally fit cubic polynomial. In terms of variance, by (18) the "equivalent kernel" is

$$K^*(z) := \frac{\{zK'(z) + K(z)\}}{\mu_2(K)}.$$

If $K(z) = 3/4(1 - u^2)I(|u| \leq 1)$, i.e., the minimum MSE Epanechnikov kernel, then

$$K^*(z) = \frac{15}{4}(1 - u^2)I(|u| \leq 1),$$

which is a minimum variance kernel (Müller, 1988, p. 69)! Also, by (6) and since $bc_3 = 0$, the bias of $\widehat{bc}_2$ is $bc_4(h_2^4 - h_1^2)/(h_2^2 - h_1^2) \approx 2bc_4 h_1^2$. Thus, by (13), unless $f'(x) = f^{(2)}(x) = 0$

the bias of $2\widehat{bc}_2/\mu_2(K)$ as an estimate of $m''(x)$ is quite different and not comparable to that of the usual estimate based on differentiating a cubic fit. Note that this bias of $\widehat{bc}_2$, which is caused by the aliasing of higher order terms in model (6) with bc_2, might very well be helpful when using (6) without these terms to estimate the bias of $\widehat{m}$.

4 Examples

4.1 Introduction

The purpose of the following examples is to investigate how EBBS performs on simulated data and to indicate how the various tuning parameters should be chosen. Those reader interested only in advice about choosing the tuning parameters should go to section 4.6 for a summary. All computations were done in MATLAB on a SUN SPARC 20.

All the examples are univariate and G_x is an m-point grid of equally spaced points with $m = 40$ in all cases. Also, the endpoints of G_x are the endpoints of the support of the density of X, not the range of the sample X_i's. Thus, estimation is somewhat beyond the range of the observed data. The neighborhood N_ℓ discussed in section 2.3 depended on a tuning constant "*span*." Let Δ be the spacing between adjacent points on the grid G_x. Then N_ℓ consists of all grid points whose distance from x_ℓ is at most Δ times *span*. The weighting function is "triangular," i.e., its graph is an isosceles triangle so that the weights for producing $\widehat{\mathrm{SMSE}}(\boldsymbol{x}_\ell; h, N_\ell)$ are linearly decreasing with distance from $\boldsymbol{x}_\ell$.

The same neighborhoods and weighting function were used to smooth the bandwidths $\widehat{h}(\boldsymbol{x}_\ell)$ to produce $\tilde{h}(\boldsymbol{x}_\ell)$. However, because bandwidth selection can be unstable near boundaries we only calculate $\tilde{h}(\boldsymbol{x}_\ell)$ for $\ell = span+1, \ldots, m-span$. We then set $\tilde{h}(\boldsymbol{x}_\ell) = \tilde{h}(x_{span+1})$ for $l = 1, \ldots, span$ and $\tilde{h}(\boldsymbol{x}_\ell) = \tilde{h}(x_{m-span})$ for $l = m - span + 1, \ldots, m$. Thus, the bandwidth is constant on an interval of length $\Delta * span$ adjacent to each boundary.

4.2 Spatial Adaptation

A curve estimator is sometimes called "spatially adaptive" if it can adapt itself to regions of high and of low curvature. The function

$$m(x) = x + 2\exp(-16x^2) \tag{19}$$

has been used as a test case for spatial adaptivity by Fan and Gijbels (1995). Following them, in this example the X_i's are iid Uniform$(-2, 2)$ and the ϵ_i's are iid $N(0, (.4)^2)$.

EBBS was implemented according to the first three steps of the algorithm in section 2.4. In this example, we fixed $h_a = .3$ and $h_b = 2$ in step 3 (estimating $m(\cdot)$) and $h_a = .5$ and $h_b = 4$ in step 2 (estimating $\sigma^2(\cdot)$). This gives a wide choice for EBBS selected bandwidths and is intended to illustrate the use of EBBS when there is little prior knowledge of the best bandwidths. However, since h_a and h_b are larger in step 2 than step 3, there is some prior knowledge that $\sigma^2(\cdot)$ has little curvature; in fact, it has none since it is constant. The

EBBS algorithm has six other tuning parameters, t, M_1, J_1, $V := J_2 + J_1 + 1 - t$, *span*, and h_m; the first five are set at the same values at steps 2 and 3 and h_m in used only in step 1. V is the difference between the number of "observations" and the number of parameters when we fit model (6).

To measure the effects of these parameters on the accuracy of EBBS, a Monte Carlo experiments were performed. A factorial design was used with $M_1 = 10$, 14, 18 and 22; $J_1 = -1$ and 1; $V = 2$, 3 and 4; *span* = 0, 1, 3, 5, and 10; and $h_m = .3$ and .6. The design was a full factorial for a total of 720 independent runs. This experiment was repeated using $p = 1$, 2, and 3 for estimation of m; $p = 1$ was always used for estimating σ^2. The response was the average absolute deviation:

$$\text{AADE} = \frac{1}{40} \sum_{i=1}^{40} |\widehat{m}(x_j) - m(x_j)|,$$

where $x_1, \ldots, x_{40}$ is an equally spaced grid from -2 to 2.

It was clear from the results that $p = 2$ was the best choice of the degree of the local polynomials in this example; see below. We will analyze only the $p = 2$ data here. We used an analysis of variance model with all main effects and two-way interactions. Higher order interactions were pooled to estimate error. For this value of p, the factors J_1 and t had large main effects. The effect of h_m was not statistically significant and, by a confidence interval, apparently not of practical interest. There were large $M_1 * t$ and $J_1 * t$ interactions. Some other main effects and interactions were statistically significant but apparently not of large practical significance; with 720 observations this was to be expected.

The means of AADE for fixed levels of J_1, t, $M_1 * t$, and $J_1 * t$, i.e., averages across the levels of the other factors, are given in Table 1. One can see that AADE is minimized by using $t = 2$. For this choice of t, J_1 has little effect and can be either -1 or 1. If the inferior choice, $t = 1$ is used, then $J_1 = 1$ is much better than $J_1 = -1$. If $t = 2$, then M_1 equal to 14 or 18 is better than M_1 equal to 10 or 22. For $t = 1$, then $M_1 = 10$ should be avoided. These results agree with the asymptotics for bias estimation in section 3 in that $t > 1$ is superior to $t = 1$. However, they disagree with the asymptotics in that $t = 3$ is inferior to $t = 2$ and $J_1 = -1$ is no better (and sometimes worse) than $J_1 = 1$.

The same general conclusions about choosing the tuning parameters are reached when examining the results for $p = 1$ and $p = 3$; for brevity these results are not included.

Provided that one avoids the combination $t = 1$, $J_1 = 1$, and $M_1 = 10$ or 14, the performance of EBBS is not very sensitive to the choice of the tuning parameters, which is comforting. Why is $t = 1$, $J_1 = -1$, and $M_1 = 10$ or 14 a poor choice? The problem is that when $J_1 = -1$ then one is estimating bias at a particular bandwidth h_0 using only values of h larger than h_0, and only values substantially larger if M_1 is small. This apparently is not acceptable unless one uses a sufficiently accurate model for the bias, e.g., $t = 2$.

With tuning parameters chosen according to these guidelines and $p = 2$, an expected AADE of around .095 is achievable. In contrast, for $p = 1$ and $p = 3$ the best expected

AADE's are both about .11.

Figure 1a–e illustrates the performance of EBBS on a single sample. Local quadratic polynomials were used, so $p = 2$. The grid, G_x, consists of 40 equally spaced points from -2 to 2. We used two sets of tuning parameters: (I) $M_1 = 14$, $J_1 = 1$, $V = 2$. $h_m = .3$, $t = 2$, and $span = 3$; (II) same as (I) but $span = 0$.

Figure 1a shows the raw data. Figure 1b shows the true curve (dotted) and the estimated curves with tuning parameter sets (I) (solid) and (II) (dashed and dotted). The estimates were cubicly interpolated from a 40 point grid, G_x, to a 400 point grid before plotting.

Figure 1c shows $\tilde{h}(x_\ell; N_\ell)$ for the two parameter sets with line types as in (b)—the bandwidths were calculated only on the 40-point grid G_x and then *linearly* interpolated. The local quadratic regression is most biased where curvature changes most rapidly. Notice that the bandwidth adapts to rapidly changing curvature between $x = -.5$ and .5 by becoming smaller there than in the other regions. Although the two sets of tuning parameters produce bandwidth functions with rather different amounts of smoothness, they produce rather similar estimates of the regression function itself. The EBBS bandwidths are somewhat too small near the boundaries. Part of the problem is that the boundaries are of the support of the density of the $\boldsymbol{X}$'s, not the range of the observed $\boldsymbol{X}$'s. Thus, we are trying to estimate m somewhat beyond the observed data.

Figure 1d is the estimate of the conditional variance of Y cubicly interpolated onto a 400 point grid. The target of course is identically .16.

Figure 1e shows the errors, $\widehat{m}(x) - m(x)$, for the sample used in panels (a)–(e) (solid) and for three other independent samples using tuning parameter set (I). The bias in the region between $x = -.5$ and $x = .5$, especially around 0, is evident as is the increased variance at the boundaries. One can see that the estimate in panel (b) is typical of the four estimates, except that it has a bit more error at the boundaries.

In Figure 2 we compare the accuracy of local linear, quadratic, and cubic polynomials. At each grid point we define the mean absolute deviation error (MADE) as

$$\text{MADE}(x_\ell) = N^{-1} \sum_{i=1}^{N} |\widehat{m}(x_\ell) - m(x_\ell)|$$

where $N = 500$ is the number of Monte Carlo replications. Figure 2 is a plot of MADE for each degree polynomial. The local linear estimate is best in regions of low curvature and quadratic best in regions of high curvature. Local linear is quite biased near the bump—see Figure 2(d). Since the curvature changes from positive to negative and then back to positive as x increase from $-.5$ to .5, one might expect local cubic regression to outperform local quadratic, yet in the region of rapidly changing curvature local cubic regression is no better than local quadratic. The extra variability of local cubic regression near the boundaries is also evident.

Since m is a linear function plus a symmetric function, the expected MADE is symmetric

about 0. Any deviations from symmetry in Figure 2, e.g., differences between (b) and the mirror image of (c), are due to sampling variation.

Figure 3 is a plot of MADE when $p = 2$ and (i) $J_1 = 1$ and $t = 2$ and (ii) $J_1 = -1$ and $t = 1$. We see that the poor performance of (ii), which was noted before, is confined to the region of rapidly changing curvature at $x = 0$; elsewhere (ii) is noticeably superior to (i).

4.3 Normal mixture density

In this example we simulate samples of size 400 from a bimodal normal mixture, (2/3) $\mathrm{N}(0,1)$ + (1/3) $\mathrm{N}(1,(0.2)^2)$. The density has a sharp peak at the mean of the second component, so the spatial adaptibility of local bandwidths is helpful. Surprisingly, local linear regression proved slightly superior to local quadratic and local cubic regression, so we concentrate on the local linear and quadratic cases, especially the former.

As in the previous example, we experimented with the tuning parameters to find optimal values. We found that $p = 1$ was slightly superior to $p = 2$ and only the results for the former will be reported. We fixed $h_a = .08$, $h_b = 1$, and $span = 2$. The factors were $M_1 =$ 10, 14, 18, 22; $J_1 = -1$ and 1; $t = 1$, 2, and 3; and $V = 1$, 2, and 3. A full factorial design was replicated 6 times. The averages of AADE for fixed factor levels or factor-level combinations are given in Table 6. AADE is rather insensitive to the tuning parameters, except that $t = 2$ or 3 should be avoided unless M_1 is small. Also, $V = 1$ or 2 is better than $V = 3$.

Figure 4a–b show the behavior of EBBS for a somewhat typical sample. Figure 4a shows the binned data. In Figure 4b we see m and its estimate. The tuning constants were $M_1 = 10$, $J_1 = 1$, $t = 2$, $V = 2$, and $span = 0$ and 4. In panel (b) we see that $span = 4$ gives a noticeably smoother estimate of m than $span = 0$. The EBBS local bandwidths in Figure 4c show adaptation to the curvature of m near $x = 1$. Figure 4d contains the estimation error for the estimate in panel (b) and for three other samples, all using $span = 4$. One can see that the estimate in (b) estimates the peak somewhat better than is typical.

Figure 5 illustrates how EBBS is able to estimate bias. Panel (a) shows four estimates, with global bandwidths of .15, .25, .50, and .84. At the dotted vertical line through $x = -.72$, the four estimates differ little—bias is small for $x = -.72$. The dotted curve in panel (b) is $\widehat{m}(-.72; h)$ as a function of h.

The dashed and dotted vertical line in Figure 5 is at $x = 1.13$ where there is substantial negative bias. This is evident from the differences between the four estimates in panel (a) and in the dashed and dotted curve in panel (b) where $\widehat{m}(1.13; h)$ is plotted against h.

The solid vertical line in panel (a) is at $x = 1.74$ where there is moderate positive bias, as also can be seen in the solid curve in panel (b) which is a plot of $\widehat{m}(1.74; h)$.

4.4 A density with bounded support

Marron and Ruppert (1994) used the parabolic density $m(x) = 4/3 - 3(x - 1/3)^2$, $x \in [0, 1]$ to test several methods of correcting density estimators for boundary bias. Since the density has constant curvature, local quadratic EBBS works quite well, better than anything tested by Marron and Ruppert.

Again we used Monte Carlo experimentation to measure the effects of the tuning parameters. With $p = 1$ we fixed $h_a = .05$ and $h_b = .8$, and with $p = 2$ with set these parameters to .15 and 2. Then for each value of p we used a full factorial design replicated 6 times, where $M_1 = 12$, 16, 20, and 24; $J_1 = 0$ and 1, $J_2 = 0$, 1, and 2, and $span = 0$, 1, and 3.

As might be expected when m is exactly quadratic, $p = 2$ was superior to $p = 1$. When $p = 1$ the factors had little effect of AADE; we will not report the results.

For $p = 2$, there were large main effect for t but only small effects for the other factors and no significant two-way interactions. The results are shown in Table 3. Clearly, AADE increasing markedly with t. In this case, there is no bias at all, and the simpler the bias model the better.

Figure 6 shows the behavior of EBBS for four independent samples of size 500. For each sample, we plotted the local linear (dashed and dotted) and local quadratic (solid) fits, with the true density a dotted curve. We used $h_a = .05$ and $h_b = 1.5$ for local linear estimation and $h_a = .15$ and $h_b = 3$ for local quadratic estimation. In both cases $t = 2$, $V = 2$, and the relatively large value of $span = 5$ was used. Local linear does about as well as Marron and Ruppert's transformation methodology and Rice's boundary kernel approach, and local linear regression has the advantage of having automatic bandwidth selectors while the transformation and boundary kernel methods do not, at least as far as I am aware. Since the density is quadratic, the local quadratic estimates are quite good, better than local linear in all four samples. The EBBS bandwidths are plotted in Figure 6e ($p = 1$) and 6f ($p = 2$). EBBS choses large bandwidths for $p = 2$, as is appropriate.

To the best of my knowledge, no other method of density estimation in the literature performs nearly as well at this density as local quadratic estimation.

4.5 Discontinuous derivatives

Plug-in bandwidth selector such as in Fan and Gijbels (1995) assume the existence of derivatives higher than the order of the derivative being estimated, e.g., that $m^{(2)}(\cdot)$ exists when we are estimating $m(\cdot)$.

EBBS may be able to estimate rougher functions, and this is an important area for future research. To illustrate the potential of EBBS, we used the regression function $m(x) = |x|$. The X_i's were uniformly distributed on $(-3, 3)$ and the errors were normally distributed with $\sigma = .25$.

Figure 7 illustrates the EBBS estimator with $h_a = .2$ and $h_b = 3$ when estimating m and $h_a = 1$ and $h_b = 5$ when estimating σ^2, The other tuning parameters were $h_m = .15$,

$M_1 = 14$, $J_1 = 1$, $t = 1$, $V = 3$, and $span = 5$. Panel (a) shows one sample, and panel (b) shows the true function (dotted) and the estimate (solid). Panels (c) and (d) show the errors $(\widehat{m} - m)$ and bandwidths for the sample in (a)–(b) (solid) and three other samples.

Note that $\widehat{m}(x; h)$ is unbiased if $h \leq |x|$, but the bias increases rapidly for $h > |x|$. Thus, the optimal local bandwidth is approximately equal to $|x|$, except near $x = 0$ where the variance explodes if $h \leq |x|$.

We can see in panel (d) that EBBS underestimates the optimal bandwidth somewhat. This is to be expected since EBBS uses $h \geq h_0$ to estimate bias at h_0 and because using $span = 5$ smooths the MSE a fair amount. Nonetheless, the EBBS bandwidth grows linearly with $|x|$ as we would hope for, except at the boundaries where the bandwidth is constant by design as explained in section 4.1.

Thus, although EBBS is based on model (6) which does not apply here due to the discontinuous derivative, the model with $t = 1$ works well for bandwidth selection. However, the same model with $t = 2$ worked quite poorly for bandwidth selection, so poorly that we have not reported the results.

4.6 Summary

In the examples, we have seen that the performance of EBBS is not too sensitive to the choice of tuning parameters, and there is no choice that is uniformly best. Nonetheless, we can make some recommendations. First, $t = 2$ is generally better than $t = 1$, though $t = 1$ can be noticeably better in some cases and never seems poor. We recommend $M_1 = 10$ to 15, $J_1 = 1$ and $V = 1$.

Using an $m = 40$ point grid for G_x seems fine for routine use, and $span = 0$ to 4 is recommended for this value of m. Another possibility is to use a smaller value of m, say $m = 10$ or 20, with correspondingly smaller values of $span$. After $\tilde{h}(\boldsymbol{x}_\ell)$ is found on the grid G_x it can be extrapolated to a finer grid, say G'_x and $\widehat{m}$ can be computed on the finer grid; doing this speeds computation since computing $\tilde{h}$ is much more time consuming that computing $\widehat{m}(x; h)$ for a single h. The idea of extrapolating a local bandwidth from a coarse to a fine grid is taken from Hall, Marron, and Titterington (1995) and is referred to as a "partial local smoothing rule." Their paper appeared as this research was completed, so their interpolation idea was not tried on EBBS. Also, the smoothing of the estimated MSE that we propose is similar to their use of "integral averages."

5 Further Topics

5.1 Local ridge regression

Cleveland and Loader mention the interesting idea of "mixing" local polynomial estimators of degrees $p - 1$ and p. They state that the resulting estimator is also a "local ridge

regression" estimator of form

$$\widehat{m}(x;h,\delta) = e_1^t(\boldsymbol{\mathcal{X}}_x^t \boldsymbol{W}_x \boldsymbol{\mathcal{X}}_x + \mathrm{diag}(0,\ldots,0,\delta))^{-1}(\boldsymbol{\mathcal{X}}_x^t \boldsymbol{W}_x \boldsymbol{Y}),$$

where $e_1 = (1,0,,\ldots,0)^t$, $\boldsymbol{\mathcal{X}}_x$ has ijth element $(X_i - x)^{j-1}$ for $i = 1,\ldots,n$ and $j = 1,\ldots,p+1$, $\boldsymbol{W}_x = \mathrm{diag}(K_h(X_1 - x),\ldots,K_h(X_n - x))$, and $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^t$. If $\delta = 0$ then we have local p degree polynomial regression. As $\delta \to \infty$ the estimator converges to local $p-1$ degree regression.

With the ridge parameter δ fixed, EBBS can be applied directly to $\widehat{m}(x;h,\delta)$ to estimate bias, and estimation of the variance of $\widehat{m}(x;h,\delta)$ is a standard calculation. Thus, EBBS can be applied to $\widehat{m}(x;h,\delta)$ without change.

At the cost of more computational time, EBBS could estimate h, p, and δ. This seems to be a promising area for further research. Presumably, not all of h, p, and δ should be local.

5.2 Density estimation as the number of bins converges to infinity

Here we consider the behavior of the density estimator in Section 2.5 as the number of bins converges to ∞. For simplicity, consider the univariate case. Let S be the $(p+1)\times(p+1)$ matrix with $S_{ij} = \int_{\mathrm{supp}(f)} (y-x)^{i+j-2} K_h(y-x)dy$. Then it is easy to see that as $L_x \to 0$, $\widehat{m}(x;h)$ converges to $\widehat{\beta}_0$ where $(\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_p)^t$ solves

$$S\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \vdots\\ \widehat{\beta}_p\end{pmatrix} = \begin{pmatrix} M^{-1}\sum_{i=1}^M K_h(U_i - x)\\ M^{-1}\sum_{i=1}^M K_h(U_i - x)(U_i - x)^1\\ \vdots\\ M^{-1}\sum_{i=1}^M K_h(U_i - x)(U_i - x)^p\end{pmatrix}.$$

If $p = 1$ and x is an interior point, then S is diagonal and $\widehat{m}(x;h)$ is the usual kernel density estimator.

6 Conclusions

The empirical modeling of bias allows a simple, though somewhat computer intensive, method of local bandwidth selection (or partially local bandwidth selection in the terminology of Hall, Marron, and Titterington (1995)). The EBBS bandwidth selector worked well in the Monte Carlo studies that we performed, and was not particularly sensitive to the choice of tuning parameters.

EBBS allows one to use $p - K$ even, where p is the degree of the local polynomial model and K is the order of the derivative of m being estimated, e.g., one can use local quadratic regression to estimate m. This shows the potential to use EBBS in other situations where the use of asymptotic formulas to estimate bias seems cumbersome or impossible.

References

Cheng, M-Y. (1994). "A bandwidth selector for density estimation using local linear fit," Preprint.

Cheng, M-Y. (1994). "Minimax efficiency of local polynomial fit estimators at boundaries," Mimeo Series #2098, Department of Statistics, University of North Carolina, Chapel Hill, NC.

Cleveland, W.S. and Devlin, S. J. (1988), "Locally-weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, 83, 597–610.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1993), "Local regression models," in *Statistical Models in S*, edited by J.M. Chambers and T.J. Hastie, pp. 309–376, New York and London: Chapman & Hall.

Cleveland, W.S. and Loader, C. (1995). Smoothing by local regression: principles and methods. Preprint.

Fan, J. (1994). "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and Gijbels, I. (1995), "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation," *Journal of the Royal Statistics Society, Series B*, 57, 371–394.

Hall, P., Marron, J.S., and Titterington, D.M. (1995). "On partial local smoothing rules for curve estimation," *Biometrika*, 82, 575–588.

Härdle, W., Hall, P., and Marron, J.S., (1992), "Regression smoothing parameters that are not far from their optimum", *Journal of the American Statistical Association*, 87, 227-233.

Huang, L.-S. (1995). *On Nonparametric Estimation and Goodness-of-fit.* Ph.D. thesis, Department of Statistics, University of North Carolina at Chapel Hill.

Müller, H.-G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Berlin: Springer-Verlag.

Ruppert, D. and Wand, M.P. (1994), "Multivariate locally weighted least squares regression," *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D. and Wand, M.P., Holst, U., and Hössjer, O. (1995), "Local polynomial variance function estimation," manuscript.

Ruppert, D., Sheather, S.J., and Wand, M.P. (1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, to appear.

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

Table 1: Averages of AADE for levels of J_1 and t and factor-level combinations of $M_1 * t$ and $J_1 * t$ for the spatial adaptivity example and $p = 2$.

	Mean	Std. dev.
J_1		
-1	0.1094	0.0014
1	0.0993	0.0014
t		
1	0.1156	0.0018
2	0.0976	0.0018
3	0.0999	0.0018
$M_1 * t$		
10 1	0.1291	0.0035
10 2	0.1015	0.0035
10 3	0.0954	0.0035
14 1	0.1237	0.0035
14 2	0.0945	0.0035
14 3	0.0962	0.0035
18 1	0.1075	0.0035
18 2	0.0953	0.0035
18 3	0.1021	0.0035
22 1	0.1019	0.0035
22 2	0.0991	0.0035
22 3	0.1058	0.0035
$J_1 * t$		
-1 1	0.1292	0.0025
-1 2	0.0977	0.0025
-1 3	0.1015	0.0025
1 1	0.1019	0.0025
1 2	0.0976	0.0025
1 3	0.0983	0.0025

Table 2: Averages of AADE for factor levels and selected factor-level combinations when estimating a normal mixture density by local linear regression.

	Mean	Std. dev.
M_1		
10	0.0211	0.0004
14	0.0228	0.0004
18	0.0239	0.0004
22	0.0243	0.0004
J_1		
-1	0.0235	0.0003
1	0.0226	0.0003
t		
1	0.0222	0.0004
2	0.0226	0.0004
3	0.0243	0.0004
V		
1	0.0223	0.0004
2	0.0226	0.0004
3	0.0243	0.0004
$M_1 * t$		
10 1	0.0218	0.0008
10 2	0.0208	0.0008
10 3	0.0207	0.0008
14 1	0.0231	0.0008
14 2	0.0219	0.0008
14 3	0.0235	0.0008
18 1	0.0221	0.0008
18 2	0.0231	0.0008
18 3	0.0267	0.0008
22 1	0.0219	0.0008
22 2	0.0248	0.0008
22 3	0.0262	0.0008
$J_1 * t$		
-1 1	0.0237	0.0005
-1 2	0.0223	0.0005
-1 3	0.0245	0.0005
1 1	0.0208	0.0005
1 2	0.0229	0.0005
1 3	0.0240	0.0005

Table 3: Averages of AADE for factor levels when estimating a density with bounded support by local quadratic regression.

	Mean	Std. dev.
M_1		
10	0.0862	0.0031
14	0.0890	0.0031
18	0.0957	0.0031
22	0.0957	0.0031
J_1		
-1	0.0890	0.0022
1	0.0944	0.0022
t		
1	0.0679	0.0027
2	0.0933	0.0027
3	0.1138	0.0027
V		
1	0.0940	0.0027
2	0.0925	0.0027
3	0.0885	0.0027

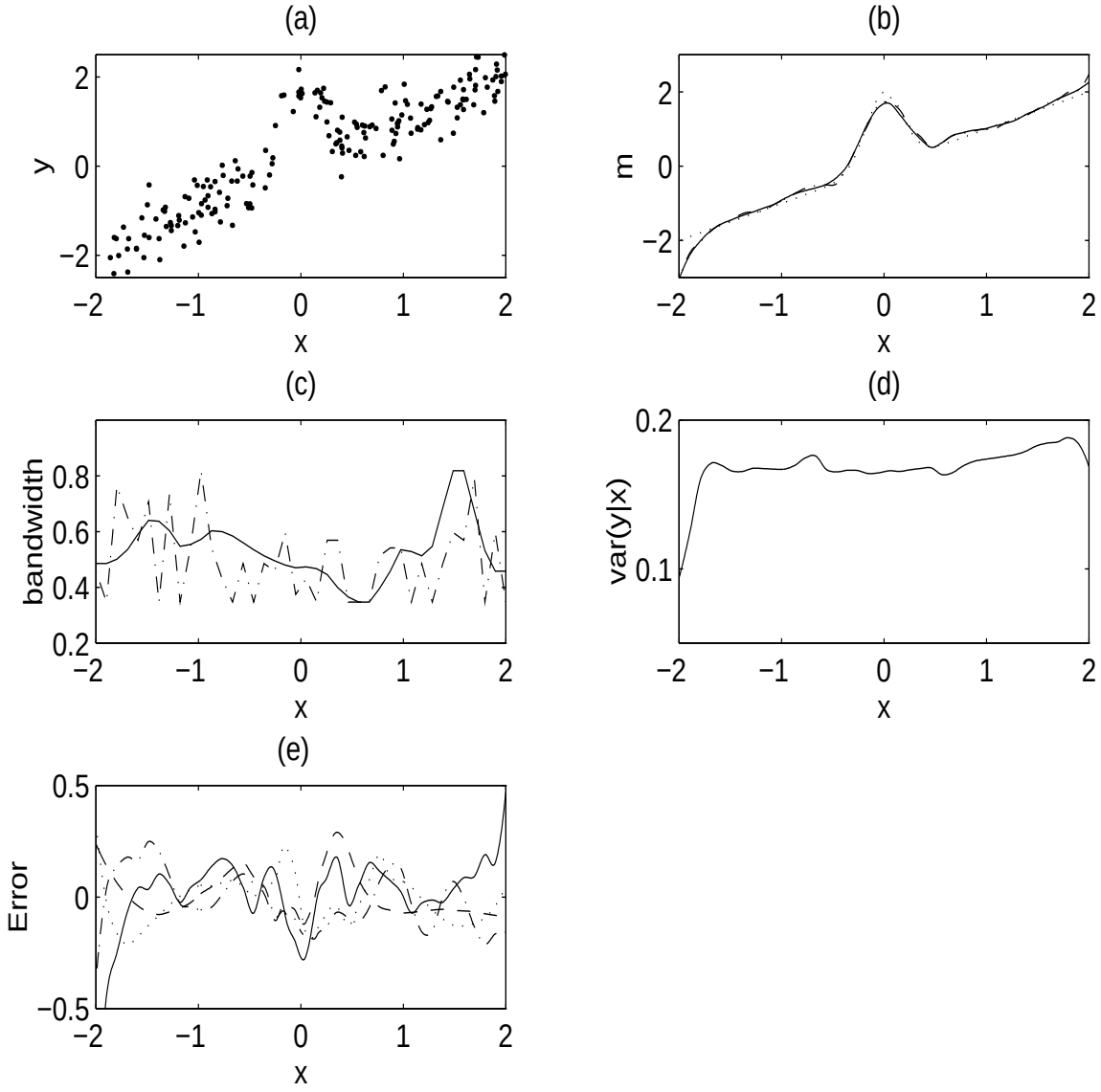

Figure 1: *Spatial Adaptation Example. Local quadratic regression. (a)–(e): Behavior of the EBBS method at a single Monte Carlo sample. The tuning constant used were (I)* $M_1 = 14$, $J_1 = 1$, $V = 2$, span $= 3$, $h_m = .3$, *and* $t = 2$*; (II) same as (I) but* span $= 0$. *(a) Raw data. (b) True regression function (dotted) and its estimates—(I) = solid and (II) = dashed and dotted. (c) Estimated optimal local bandwidth used to produce estimates in (b)—(I) = solid and (II) = dashed and dotted. (d) Estimate of the local variance. (e) Errors (=* $\widehat{m} - m$*) for the estimate in (b) and for three other independent samples using parameters (I).*

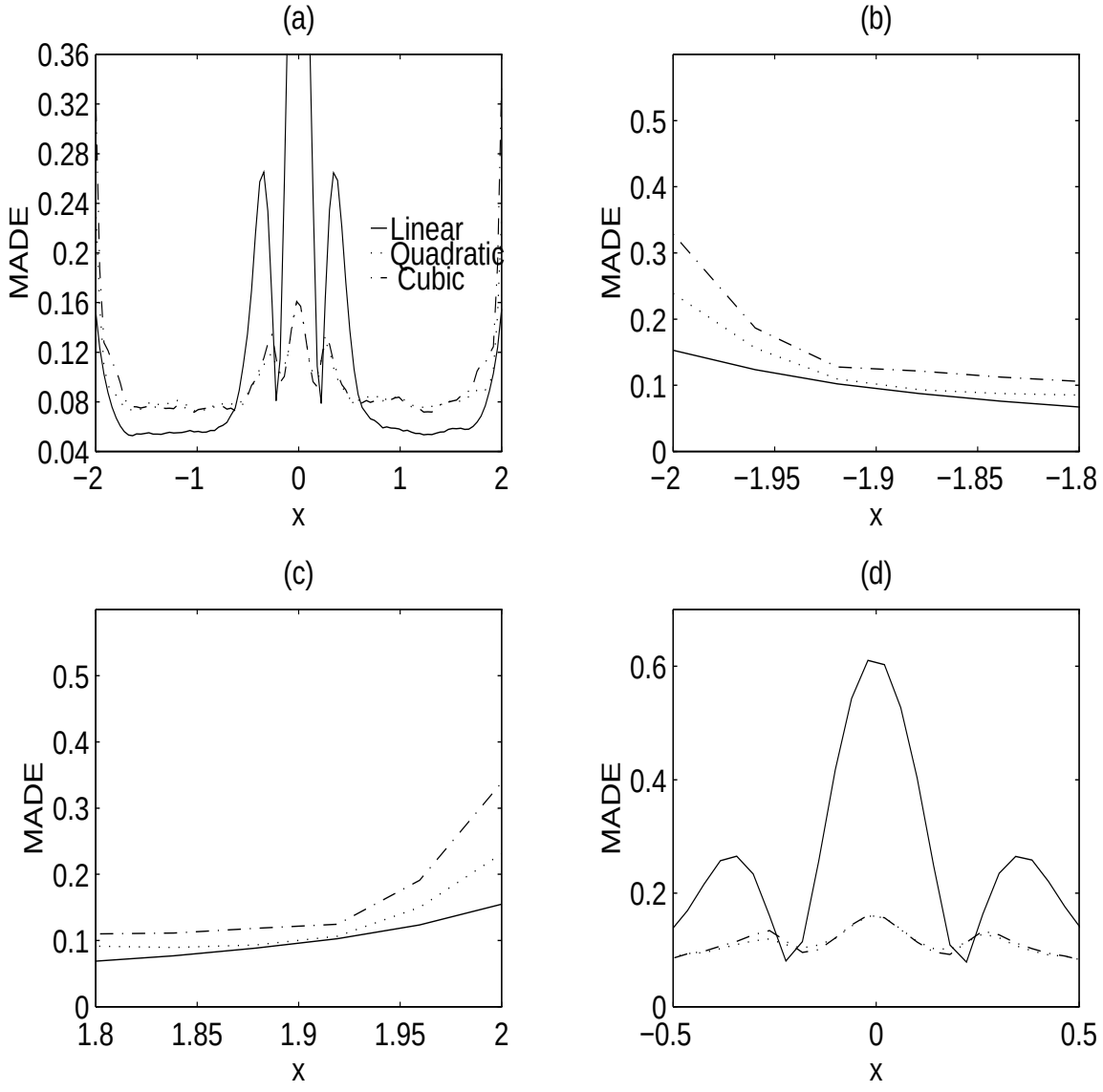

Figure 2: *Spatial Adaptation Example. Comparison of Mean Absolute Deviation Error (MADE) for linear, quadratic, and cubic local polynomials. Based on 500 Monte Carlo trials. The same tuning constants* $h_m = .3$, $M_1 = 14$, $J_1 = 1$, $V = 3$, $t = 2$, *and* span $= 0$ *were used for all three estimators. (a) MADE for all* x. *(b) Detailed view of MADE at left boundary. (c) MADE at right boundary. (d) MADE at bump centered at 0.*

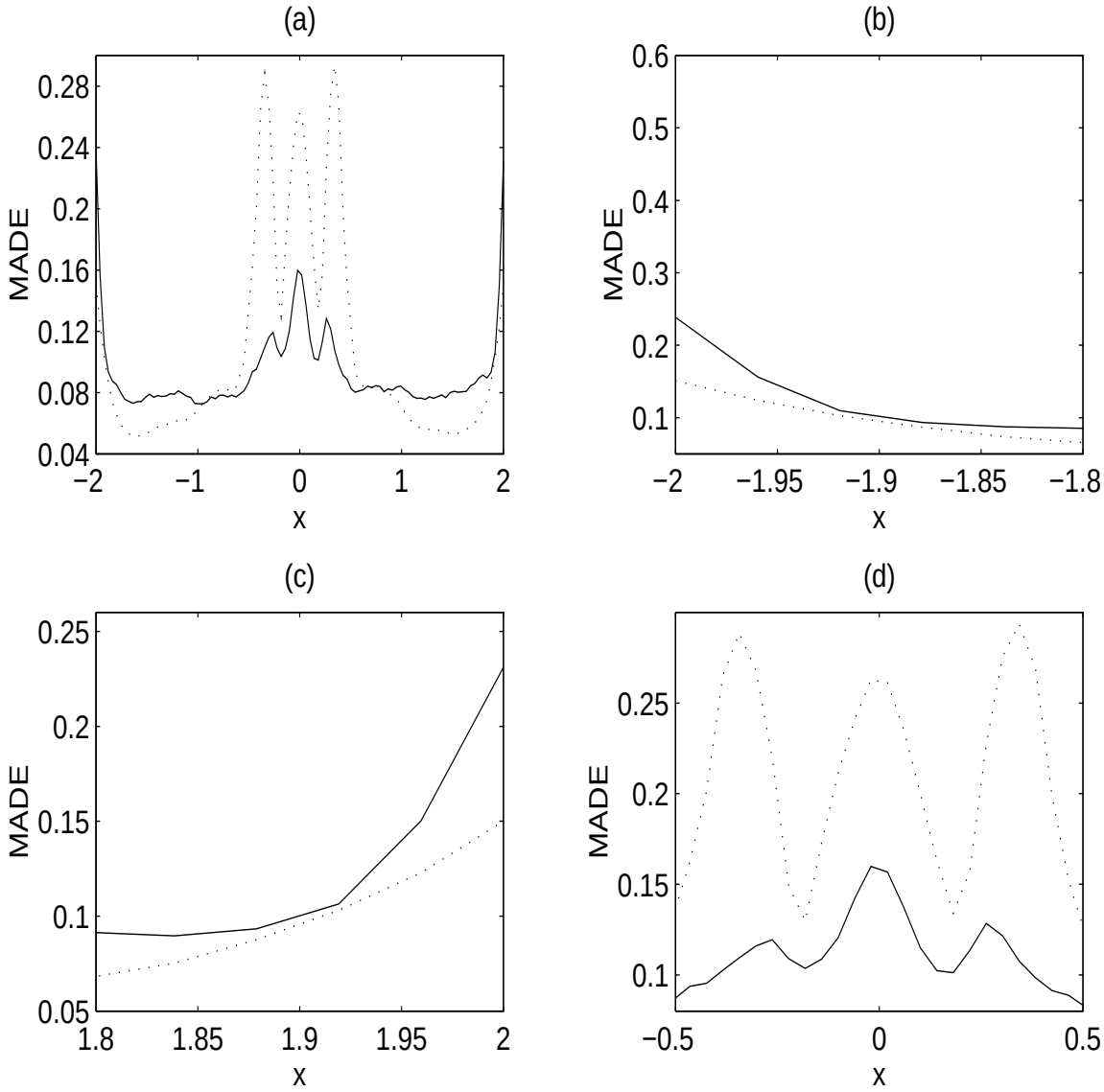

Figure 3: *Spatial Adaptation Example. Comparison of Mean Absolute Deviation Error (MADE) for two local quadratic estimators: (i)* $J_1 = 1$ *and* $t = 2$ *(solid), (ii)* $J_1 = -1$ *and* $t = 1$ *(dotted). Based on 500 Monte Carlo trials. The other tuning constants were* $h_m = .3$*,* $M_1 = 14$*,* $V = 3$*, and span = 0 for both estimators. (a) MADE for all* x*. (b) Detailed view of MADE at left boundary. (c) MADE at right boundary. (d) MADE at bump centered at 0.*

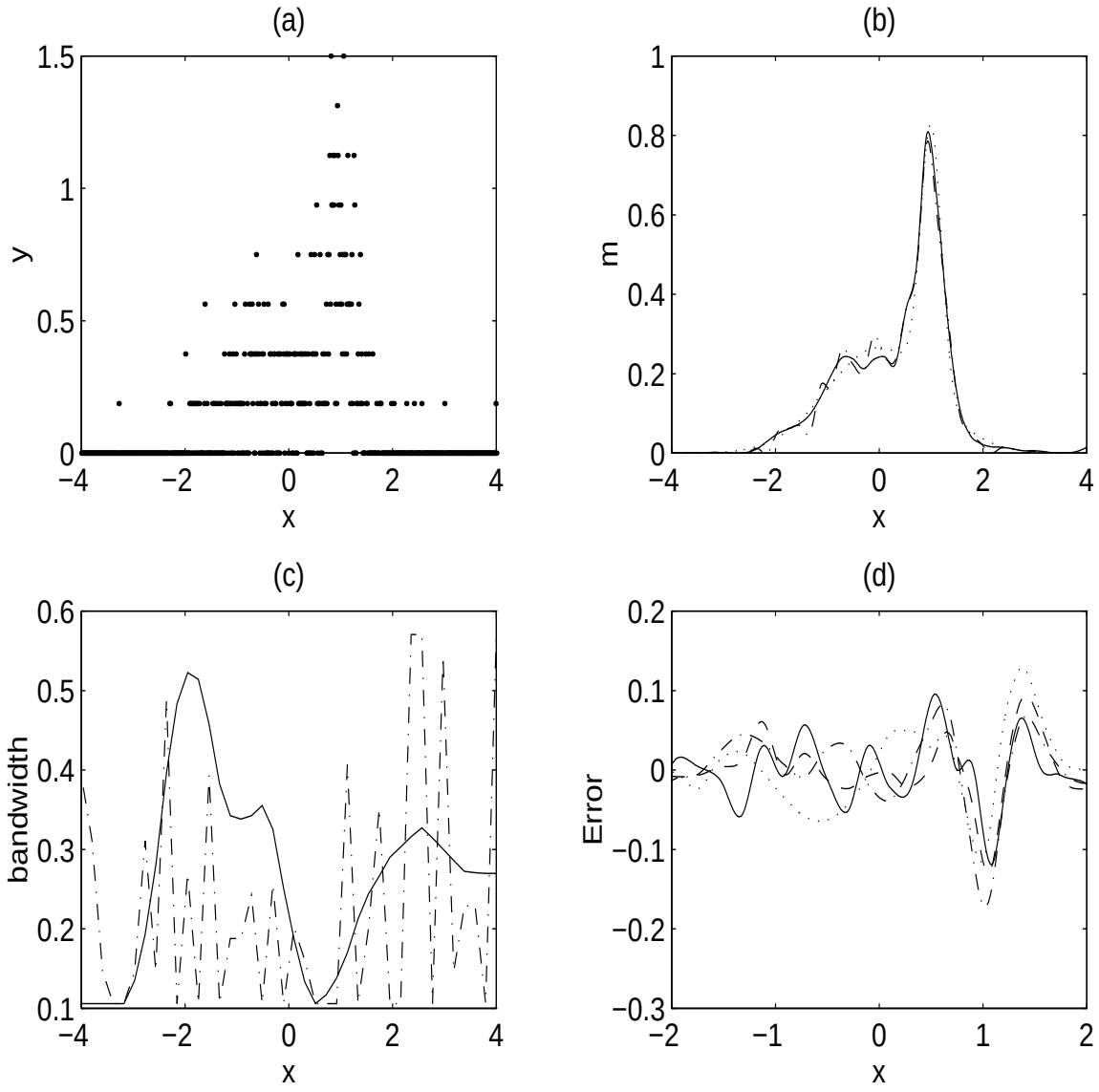

Figure 4: *Estimation of a normal mixture density. The sample size is 400. (a)–(d) Behavior of EBBS at one Monte Carlo sample. (a) Bin centers versus bin heights for a 600 bin histogram estimate of the density. (b) True density and its estimate for* span $= 4$ *(solid) and* span $= 0$ *(dashed and dotted). (c) Estimated optimal local bandwidth. (d) Estimated MSE. (e) Error* $(= \widehat{m} - x)$ *for the sample in (a)–(d) (solid) and for three other independent samples*

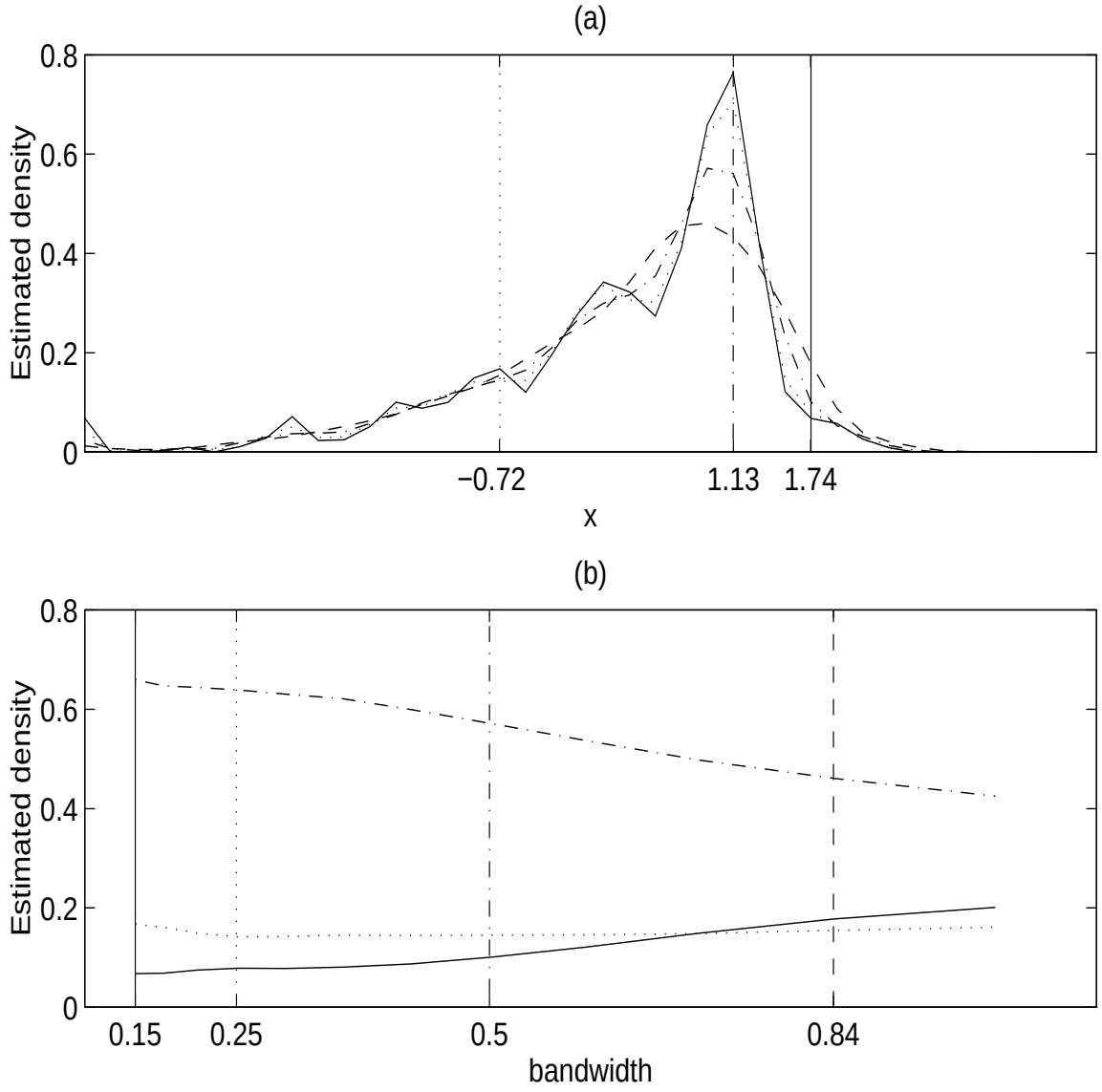

Figure 5: *Estimation of a normal mixture density. Effects of the bandwidth on the estimate. The sample size is 400. In (a) we see estimates with four different global bandwidths, all at the same sample. The line types correspond to those of the vertical lines in (b) through the bandwidths used in (a), e.g., the solid curve in (a) used a bandwidth equal to .15 so the solid vertical line in (b) goes through* $h = .15$*. In (b) we plot the estimated density at a function of the bandwidth at three fixed values of* x*. The line types of the estimates correspond to those of the vertical lines in (a) through the* x *value of the estimates in (b), e.g., the solid curve in (b) is a plot of* $\widehat{m}(1.75; h)$ *versus* h *and the solid vertical line in (a) goes through* $x = 1.75$*.*

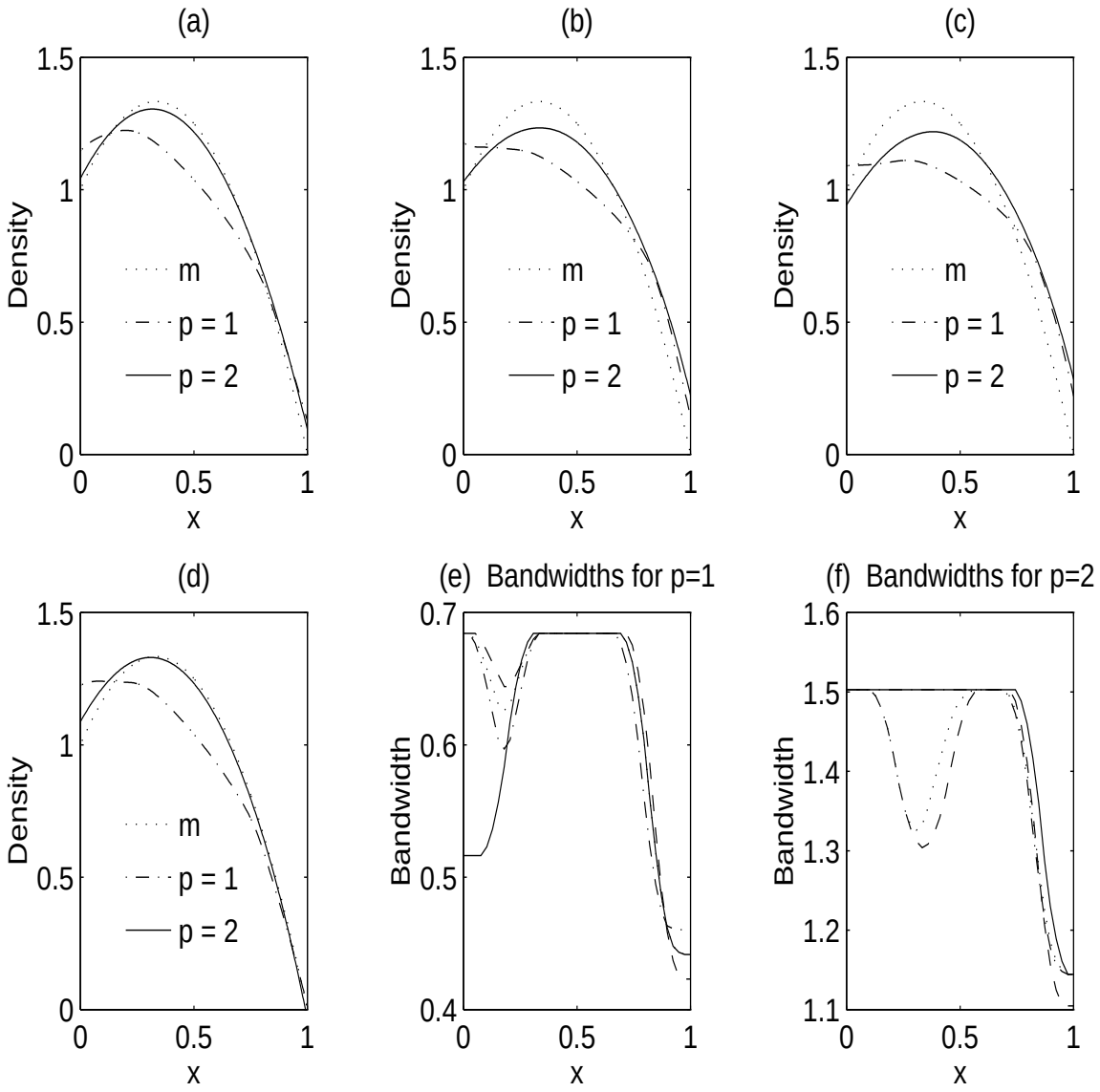

Figure 6: *Estimation of a density with bounded support.(a)–(d) Behavior of EBBS for both* $p = 1$ *and* $p = 2$ *at four independent Monte Carlo samples. (e) and (f) Bandwidths for the four samples. The sample size is 500.*

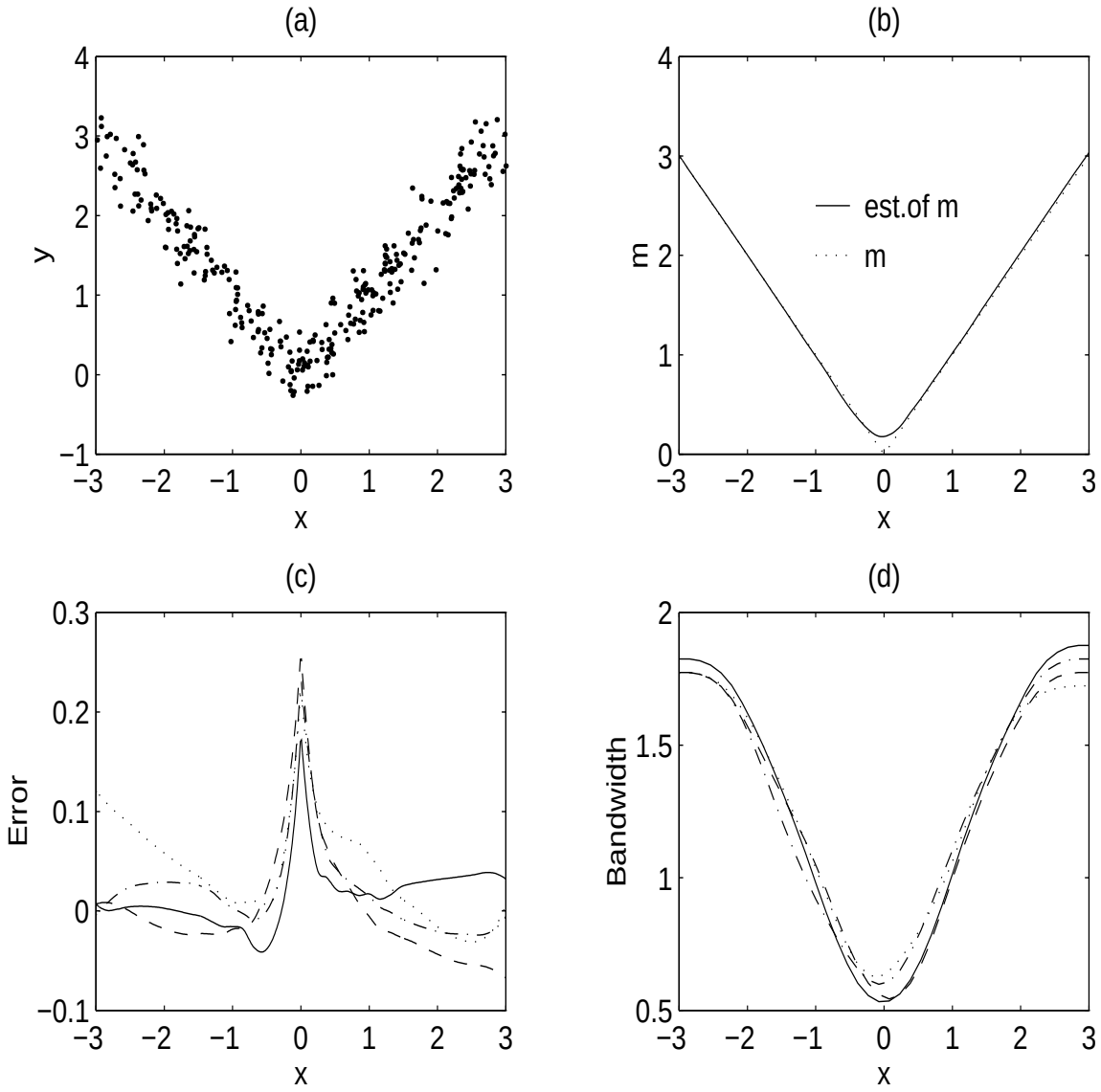

Figure 7: *Absolute value function. Behavior of EBBS at one Monte Carlo sample. The sample size is 300.*