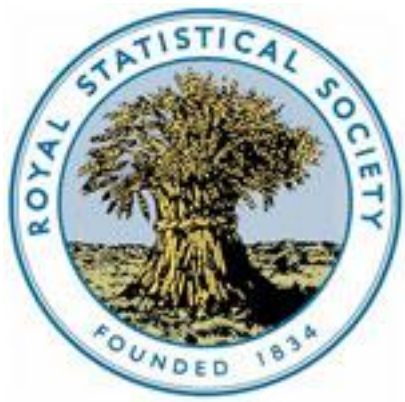

A Bandwidth Selector for Bivariate Kernel Regression
Author(s): Eva Herrmann, Joachim Engel, M. P. Wand and Theo Gasser
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1 (1995), pp. 171-180
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2346092
Accessed: 11-11-2016 07:05 UTC

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 171–180

A Bandwidth Selector for Bivariate Kernel Regression

By EVA HERRMANN†, *Technische Hochschule Darmstadt, Germany*
M. P. WAND, *University of New South Wales, Kensington, Australia*
JOACHIM ENGEL, *Universität Bonn, Germany*
and
THEO GASSER, *Universität Zürich, Switzerland*

[Received December 1990. Final revision November 1993]

SUMMARY

For two and higher dimensional kernel regression, currently available bandwidth selection procedures are based on cross-validation or related penalizing ideas. However, these techniques have been shown to suffer from high sample variability and, in addition, can sometimes be difficult to implement when a vector of bandwidths needs to be selected. In this paper we propose a selector based on an iterative plug-in approach for bivariate kernel regression. It is shown to give satisfactory results and can be quickly computed. Our ideas can be extended to higher dimensions.

Keywords: BANDWIDTH SELECTION; KERNEL ESTIMATOR; NONPARAMETRIC REGRESSION; TWO-DIMENSIONAL DATA SMOOTHING

1. INTRODUCTION

Nonparametric approaches to the smoothing of noisy regression data sets have been demonstrated to be effective for many applications. For detailed expositions see Eubank (1988), Müller (1988) and Härdle (1990).

An important component of all nonparametric regression estimators is the choice of the smoothing parameter. Marron (1988) gave an overview of proposals for data-driven smoothing parameter selection. So far most automatic smoothing parameter selectors are variants of cross-validation or asymptotically equivalent penalizing methods such as unbiased risk estimation (Rice, 1984). However, their performance has been seen to be unreliable both in practice and through theoretical analysis which indicates that they are subject to a high degree of sample variability (Härdle *et al.*, 1988). The most promising alternative methods belong to a type referred to as 'plug-in' rules since they involve plugging estimates of unknown functionals into asymptotic formulae for the optimal smoothing parameter. For kernel regression, where the smoothing parameter is usually called the 'bandwidth', Gasser *et al.* (1991) devised an iterative plug-in scheme with good theoretical properties and reliable performance in simulation. For kernel density estimation plug-in selectors have also been proposed and seen to be theoretically and practically superior to cross-validatory and penalizing rules (Sheather and Jones, 1991; Jones *et al.*, 1991).

Most of this work has been confined to the one-dimensional setting. There are important applications of nonparametric regression in higher dimensions and for

†*Address for correspondence*: Fachbereich Mathematik, Technische Hochschule Darmstadt, Schlossgartenstrasse 7, 64289 Darmstadt, Germany.
E-mail: eherrmann@mathematik.th-darmstadt.de

 0035–9246/95/57171

these situations it may be even more desirable to have an automatic smoothing parameter selection. In this paper we concentrate on bandwidth selection in bivariate regression. Other problems of bivariate kernel estimation, e.g. choice of the kernel or the form of the kernel estimator, are not addressed here. Because of its reliable performance in the one-dimensional case, we propose an extension of the plug-in rule of Gasser *et al.* (1991). Such an approach is computationally inexpensive and does not require numerical minimization which can be a hindrance for the implementation of some of the cross-validatory and penalizing rules, especially if a vector of bandwidths needs to be chosen.

In Section 2 we describe the bivariate regression model and the corresponding kernel estimator and we present the asymptotic theory required for plug-in bandwidth selection. A plug-in selector for selecting the bandwidth pair is presented in Section 3. Section 4 presents some simulations.

2. BIVARIATE KERNEL REGRESSION

A bivariate regression model of the form $Y_i = r(t_i, u_i) + \epsilon_i$, $i = 1, \ldots, n$, is considered where r is an unknown real-valued function on a compact subset $A \subset \mathbf{R}^2$ and Y_i are the responses. We assume that the ϵ_i are independent with mean 0 and variance σ^2 and that the design follows a positive and continuous design density f on the set A. More explicitly, we assume that a partition $A_1, \ldots, A_n$ of A exists with $(t_i, u_i) \in A_i$, $\sup_i |\lambda(A_i) - n^{-1} f(t_i, u_i)^{-1}| = o(n^{-1})$ and $\sup_i \sup_{x,y \in A_i} \|x - y\| = O(n^{-1/2})$ where λ denotes the two-dimensional Lebesgue measure. The second condition on the design ensures that $\lambda(A_i)^{-1}$ is proportional to $f(t_i, u_i)$ and the last condition ensures that for each set A_i the expansions in both directions are of the same order. This corresponds closely to a fixed regular design that is often assumed in one dimension. For a random design we must assume that $\sup_i |\lambda(A_i) - n^{-1} f(t_i, u_i)^{-1}| = o_p(n^{-1})$ and $\sup_i \sup_{x,y \in A_i \cup \{(t_i,u_i)\}} \|x - y\| = O_p(n^{-1/2})$. In the following we assume a fixed design.

We consider product kernels. For integers ν and $k \geqslant \nu + 2$ such that $k - \nu$ is even we say that the function $W_{\nu,k}$ is a (ν, k)-kernel on $\mathbf{R}$ if

$$\int x^j W_{\nu,k}(x)\, dx = \begin{cases} 0 & 0 \leqslant j \leqslant k-1,\ j \neq \nu, \\ (-1)^\nu \nu! & j = \nu, \\ \mu_k(W_{\nu,k}) \neq 0 & j = k. \end{cases}$$

A (ν, k)-kernel is appropriate for estimation of νth-order components of partial derivatives of regression functions. Here we use optimal (ν, k)-kernels with respect to the mean-squared error (Gasser *et al.*, 1985). We shall use the convolution form of a kernel estimator for $r(t, u)$ given by

$$\hat{r}(t, u; b_t, b_u) = b_t^{-1} b_u^{-1} \sum_{i=1}^{n} \int_{A_i} K\{(t - v)/b_t\} K\{(u - w)/b_u\}\, d(v, w) Y_i,$$

where K is a (0, 2)-kernel with support on $[-1, 1]$ and (b_t, b_u) is a pair of bandwidths. In the special case of rectangular partition sets this estimator can be computed easily by using one-dimensional smoothing routines. In Appendix B we give a short description of how to reduce a regression model with arbitrary partition to an equivalent model with rectangular partition sets.

The starting point for plug-in rules is a formula for the asymptotically optimal bandwidths with respect to the mean integrated squared error (MISE), given by $\mathrm{MISE}(b_t, b_u) = E\{\mathrm{ISE}(b_t, b_u)\}$ where

$$\mathrm{ISE}(b_t, b_u) = \int_A \omega(t, u)\{\hat{r}(t, u; b_t, b_u) - r(t, u)\}^2 \, \mathrm{d}(t, u).$$

Here the weight function ω is introduced to restrict the integral on an inner part of A both for mathematical convenience and stability at the bandwidth selection stage. We assume that ω is a probability density with continuous second-order partial derivatives, support on a convex and compact subset B of A with $\lambda(B) > 0$ and that B has positive distance to the boundary of A. For the actual regression function estimation, boundary modifications (Gasser *et al.*, 1985) are used. The bandwidth pair which minimizes $\mathrm{ISE}(b_t, b_u)$ is denoted by $(b_{t,\mathrm{ISE}}, b_{u,\mathrm{ISE}})$ and the minimizer of $\mathrm{MISE}(b_t, b_u)$ by $(b_{t,\mathrm{MISE}}, b_{u,\mathrm{MISE}})$.

For the partial derivatives of a bivariate function g we define $g^{(i,j)}(t, u) = (\partial^{i+j}/\partial t^i \, \partial u^j) \, g(t, u)$ and for a kernel W with support on $[-1, 1]$ we let

$$R(W) = \int_{-1}^{1} W(x)^2 \, \mathrm{d}x.$$

Suppose that r has all second-order partial derivatives continuous, $r^{(2,0)}$ and $r^{(0,2)}$ are not completely vanishing, K is Lipschitz continuous and $b_t + b_u \to 0$, $nb_t b_u \to \infty$ as $n \to 0$. Then standard asymptotic theory (see for example Eubank (1988), section 4.3) leads to

$$\mathrm{MISE}(b_t, b_u) = \mathrm{AMISE}(b_t, b_u) + o(b_t^4 + b_u^4 + n^{-1}b_t^{-1}b_u^{-1}) + O(n^{-1})$$

where

$$\mathrm{AMISE}(b_t, b_u) = \tfrac{1}{4}\mu_2(K)^2(b_t^4 I_{tt} + 2b_t^2 b_u^2 I_{tu} + b_u^4 I_{uu}) + n^{-1}b_t^{-1}b_u^{-1}\sigma^2 R(K)^2 I_f$$

is the asymptotic MISE. The functionals I_{tt}, I_{uu}, I_{tu} and I_f are given by

$$I_{tt} = \int_A \omega(t, u) \, r^{(2,0)}(t, u)^2 \, \mathrm{d}(t, u), \qquad I_{uu} = \int_A \omega(t, u) \, r^{(0,2)}(t, u)^2 \, \mathrm{d}(t, u),$$

$$I_{tu} = \int_A \omega(t, u) \, r^{(2,0)}(t, u) \, r^{(0,2)}(t, u) \, \mathrm{d}(t, u), \qquad I_f = \int_A \omega(t, u) \, f(t, u)^{-1} \, \mathrm{d}(t, u).$$

Let $(b_{t,\mathrm{AMISE}}, b_{u,\mathrm{AMISE}})$ denote the bandwidth pair which minimizes $\mathrm{AMISE}(b_t, b_u)$:

$$b_{t,\mathrm{AMISE}} = \left\{ \frac{\sigma^2 R(K)^2 I_{uu}^{3/4} I_f}{\mu_2(K)^2 I_{tt}^{3/4}(I_{tt}^{1/2} I_{uu}^{1/2} + I_{tu}) n} \right\}^{1/6} \tag{2.1}$$

and $b_{u,\mathrm{AMISE}} = (I_{tt}/I_{uu})^{1/4} b_{t,\mathrm{AMISE}}$. A plug-in approach requires an estimation of the residual variance σ^2 and the functionals I_{tt}, I_{uu} and I_{tu}. If the design density f is not known I_f can easily be estimated by

$$\hat{I}_f = n \sum_{i=1}^{n} \lambda(A_i)^2 \omega(t_i, u_i).$$

Since bias is usually the crucial point when estimating σ^2 we propose an estimator

that takes care especially of the bias and is unbiased for linear regression functions, similar to the one-dimensional estimator of Gasser *et al.* (1986) (Appendix A). Any other $O(n^{-1/2})$ consistent variance estimator can also be used; see for example Hall *et al.* (1990) for a one-dimensional variance estimator optimizing the asymptotic variance. The estimation of the functional is more delicate and is addressed below.

3. ITERATIVE PLUG-IN BANDWIDTH SELECTION

In this section we present our algorithm for plug-in selection of (b_t, b_u). We use the following kernel estimators of partial derivatives:

$$\hat{r}^{(2,0)}(t, u; \alpha_t, \alpha_u) = \frac{1}{\alpha_t^3 \alpha_u} \sum_{i=1}^{n} \int_{A_i} L_2\left(\frac{t-v}{\alpha_t}\right) M\left(\frac{u-w}{\alpha_u}\right) \mathrm{d}(v, w) Y_i$$

and

$$\hat{r}^{(0,2)}(t, u; \beta_t, \beta_u) = \frac{1}{\beta_t \beta_u^3} \sum_{i=1}^{n} \int_{A_i} M\left(\frac{t-v}{\beta_t}\right) L_2\left(\frac{u-w}{\beta_u}\right) \mathrm{d}(v, w) Y_i,$$

where L_2 is a (2, 4)-kernel, M is a (0, 2)-kernel and α_t, α_u, β_t and β_u are bandwidths which we refer to as 'pilot' bandwidths. Estimators of the functionals I_{tt}, I_{uu} and I_{tu} are obtained by replacing the partial derivatives by their kernel estimators and are denoted by $\hat{I}_{tt}$, $\hat{I}_{uu}$ and $\hat{I}_{tu}$.

We propose to choose the pilot bandwidths via an iterative algorithm (Gasser *et al.*, 1991). Let

$$\hat{G}(\alpha_t, \alpha_u, \beta_t, \beta_u) = (\hat{b}_{t, \widehat{\mathrm{AMISE}}}(\alpha_t, \alpha_u, \beta_t, \beta_u), \hat{b}_{u, \widehat{\mathrm{AMISE}}}(\alpha_t, \alpha_u, \beta_t, \beta_u))$$

where $\hat{b}_{t, \widehat{\mathrm{AMISE}}}(\alpha_t, \alpha_u, \beta_t, \beta_u)$ and $\hat{b}_{u, \widehat{\mathrm{AMISE}}}(\alpha_t, \alpha_u, \beta_t, \beta_u)$ are obtained from equation (2.1) by replacing σ^2 by $\hat{\sigma}^2$ (see Appendix A) and I_{tt}, I_{uu} and I_{tu} by the above estimates. Then the iterative algorithm for obtaining the estimates $(\hat{b}_t, \hat{b}_u)$ is

(a) set $\hat{b}_t^{(0)} = \hat{b}_u^{(0)} = \{\lambda(A)/n\}^{1/2}$,

(b) for $i = 1, 2, \ldots$ iterate using

$$(\hat{b}_t^{(i)}, \hat{b}_u^{(i)}) = \hat{G}(\hat{b}_t^{(i-1)} cn^{1/12}, \hat{b}_u^{(i-1)} dn^{1/12}, \hat{b}_t^{(i-1)} dn^{1/12}, \hat{b}_u^{(i-1)} cn^{1/12}), \qquad (3.1)$$

(c) stop after i^* iterations and set $(\hat{b}_t, \hat{b}_u) = (\hat{b}_t^{(i^*)}, \hat{b}_u^{(i^*)})$.

This iterative algorithm is motivated by searching for a fixed point similarly to the method of Sheather and Jones (1991) who use fixed point arguments directly for density bandwidth selection. We do not expect big differences between these two approaches, but the existence of finite fixed points is not necessary for the convergence of the above method. Further the asymptotic properties of the bandwidth estimator do not improve beyond a fixed number of iterations, here $i^* = 9$ as is shown below. Simulations proved that this is also true in practice.

Starting with bandwidths of rate $n^{-1/2}$ is motivated by our experience in the one-dimensional case. There, starting with a bandwidth near the minimal possible leads to better results in some difficult situations (Figs 1 and 2 of Gasser *et al.* (1991)) without affecting other cases.

The inflation factors $cn^{1/12}$ and $dn^{1/12}$ appearing in equation (3.1) are chosen

such that the algorithm returns asymptotically optimal bandwidths regardless of the value of c, $d > 0$. Since the bias terms are minimal for equal rates of all pilot bandwidths we restrict our attention to this case and denote their common rate in the ith iteration by $n^{-\xi^{(i)}}$. The bandwidths $\hat{b}_t^{(0)}$ and $\hat{b}_u^{(0)}$ are both of order $n^{-1/2}$. From equation (3.1) we obtain $\xi^{(1)} = 5/12$ so the application of result (a) in Appendix C shows that the rate $n^{-\xi^{(1)}}$ holds also for $\hat{b}_t^{(1)}$ and $\hat{b}_u^{(1)}$. Continuing for $i = 2, 3, 4$ we obtain $\hat{b}_t^{(i)}$ and $\hat{b}_u^{(i)}$ to have order $n^{-(6-i)/12}$. Applying result (c) in Appendix C just one more iteration is required to give $\hat{b}_t^{(5)} = b_{t,\mathrm{MISE}}\{1 + o_p(1)\}$ and $\hat{b}_u^{(5)} = b_{u,\mathrm{MISE}}\{1 + o_p(1)\}$ and these rates hold for all further iterations. This is guaranteed by the inflation factors which lead to $\xi^{(i)} = 1/12$ for $i \geqslant 5$. We have demonstrated that for any c, $d > 0$ and $i^* \geqslant 5$ the above algorithm leads to asymptotically optimal bandwidths in terms of rates and leading constants. Iterations beyond $i = 5$ yield an improvement of the variability of the estimated bandwidth reaching an order $O_p(n^{-1/2})$ for $i^* \geqslant 9$ iterations (see the results of Appendix C where the exponent β increases for each iteration between 5 and 9). We may think of the constants c and d as tuning parameters which can be adjusted to improve the finite sample performance as discussed in the next section.

4. IMPLEMENTATION AND EXAMPLES

For the examples presented we took K to be the Bartlett–Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)$, $|x| < 1$. The optimal boundary kernels of Gasser *et al.* (1985) were used in the boundary region. For the functional estimation we took

$$L_2(x) = \frac{105}{16}(-5x^4 + 6x^2 - 1),$$

$|x| < 1$ and $M = K$. The set A was the unit square and the function ω equals 1 on $[0.05, 0.95]^2$ and 0 elsewhere. The bandwidths are bounded from above by $\frac{1}{2}$ and are bounded from below so that

$$\bigcup_{i=1}^{n} [t_i - b_t,\ t_i + b_t] \times [u_i - b_u,\ u_i + b_u]$$

covers $[0.05, 0.95]^2$. We also calculated the bandwidths selected by the unbiased risk criterion of Rice (1984) and classical cross-validation bandwidths. For the iterative plug-in rule $i^* = 9$ iterations have been used as indicated from theory. We tried several values for c and d and found that $c = 1.5$, $d = 0.25$, gave a slightly superior performance so that these values were used throughout.

The algorithm for the plug-in estimator was stated in quite a general form and can also be used for non-rectangular design. In our simulations we used design densities of product form, i.e. $f(t, u) = f_t(t) f_u(u)$ and rectangular design. The case of non-rectangular design can be traced back to the rectangular design case (Appendix B). Thus, design points (t_i, u_j), for $i = 1, \ldots, n_t$, $j = 1, \ldots, n_u$, are given by $t_i = F_t^{-1}\{(i - 0.5)/n_t\}$ and $u_j = F_u^{-1}\{(j - 0.5)/n_u\}$ where f_t and f_u are linear, normal and normal mixtures. The corresponding partition sets are given by

$$A_{i,j} = [(t_{i-1} + t_i)/2,\ (t_i + t_{i+1})/2] \times [(u_{j-1} + u_j)/2,\ (u_j + u_{j+1})/2]$$

for $i = 2, \ldots, n_t - 1$, $j = 2, \ldots, n_u - 1$ and analogously at the boundary.

In our simulations we considered several regression surfaces corresponding to Gaussian mixtures because of their familiarity and flexibility. Let $N(\mu, \Sigma)$ denote the normal density with mean vector μ and 2×2 covariance matrix Σ. We used a regression function with a single Gaussian peak,

$$r_1\colon N\left\{\begin{pmatrix}0.5\\0.5\end{pmatrix}, \begin{pmatrix}0.05 & 0\\0 & 0.05\end{pmatrix}\right\},$$

TABLE 1
Simulation results for various regression functions and sample size†

	Quantile	*Function r_2, two peaks, $n_t = n_u = 15$, $\sigma^2 = 0.05$*				*Function r_3, peak and trough, $n_t = n_u = 25$, $\sigma^2 = 0.05$*				*Function r_4, tunnel, $n_t = n_u = 25$, $\sigma^2 = 0.05$*			
		PI	*UR*	*CV*	*OPT*	*PI*	*UR*	*CV*	*OPT*	*PI*	*UR*	*CV*	*OPT*
b_t	10%	0.114	0.088	0.048	0.082	0.073	0.056	0.029	0.033	0.089	0.029	0.050	0.078
	50%	0.124	0.135	0.087	0.103	0.077	0.065	0.041	0.060	0.096	0.068	0.084	0.084
	90%	0.138	0.168	0.154	0.144	0.081	0.081	0.067	0.071	0.106	0.101	0.111	0.112
b_u	10%	0.132	0.096	0.058	0.109	0.088	0.057	0.029	0.057	0.139	0.132	0.319	0.472
	50%	0.143	0.158	0.139	0.160	0.093	0.079	0.061	0.073	0.159	0.306	0.480	0.489
	90%	0.154	0.198	0.181	0.183	0.098	0.098	0.091	0.091	0.187	0.359	0.500	0.500
ISE	50%	0.014	0.014	0.017	0.013	0.012	0.012	0.015	0.011	0.004	0.005	0.003	0.002
	75%	0.015	0.017	0.022	0.015	0.013	0.013	0.018	0.012	0.005	0.007	0.004	0.003
	90%	0.018	0.019	0.026	0.018	0.014	0.015	0.020	0.013	0.006	0.009	0.005	0.003
QISE	90%	1.07	1.35	1.90	1.00	1.14	1.17	1.75	1.00	2.57	5.41	1.54	1.00

†Characteristic sample quantiles are compared for the plug-in (PI), unbiased risk (UR), cross-validation (CV) and the ISE optimal (OPT) bandwidths. QISE denotes ISE/inf(ISE).

TABLE 2
Simulation results for regression function r_1 with one Gaussian peak, $n_t = 20$, $n_u = 10$, $\sigma^2 = 0.1$ and three different design densities†

	Quantile	*Design f_1*			*Design f_2*			*Design f_3*		
		PI	*UR*	*OPT*	*PI*	*UR*	*OPT*	*PI*	*UR*	*OPT*
b_t	10%	0.132	0.033	0.145	0.170	0.103	0.172	0.180	0.095	0.175
	50%	0.155	0.134	0.177	0.201	0.148	0.246	0.209	0.162	0.273
	90%	0.178	0.212	0.228	0.228	0.191	0.301	0.231	0.205	0.350
b_u	10%	0.180	0.064	0.130	0.214	0.226	0.214	0.234	0.238	0.234
	50%	0.195	0.159	0.198	0.214	0.232	0.214	0.234	0.252	0.234
	90%	0.209	0.261	0.257	0.214	0.251	0.221	0.234	0.276	0.235
ISE	50%	0.012	0.017	0.011	0.025	0.033	0.023	0.040	0.051	0.035
	75%	0.015	0.046	0.014	0.039	0.048	0.031	0.050	0.071	0.042
	90%	0:017	0.054	0.016	0.050	0.062	0.042	0.065	0.085	0.052
QISE	90%	1.276	5.464	1.000	1.400	2.111	1.000	1.404	2.134	1.000

†Here the plug-in (PI), unbiased risk (UR) and the ISE optimal (OPT) bandwidths are compared. 0.214 and 0.234 are the minimal possible bandwidths for b_u and designs f_2 and f_3 respectively. QISE denotes ISE/inf(ISE).

one with two Gaussian peaks,

$$r_2\colon\ 0.385N\left\{\begin{pmatrix}0.25\\0.35\end{pmatrix},\ \begin{pmatrix}0.025 & 0\\0 & 0.025\end{pmatrix}\right\} + 0.667N\left\{\begin{pmatrix}0.75\\0.75\end{pmatrix},\ \begin{pmatrix}0.025 & 0\\0 & 0.025\end{pmatrix}\right\},$$

one with a peak and a trough,

$$r_3\colon\ 3600N\left\{\begin{pmatrix}0.3\\0.3\end{pmatrix},\ 0.0055\begin{pmatrix}1 & -0.3\\-0.3 & 1\end{pmatrix}\right\} - 185N\left\{\begin{pmatrix}0.7\\0.6\end{pmatrix},\ 0.083\begin{pmatrix}1 & -1.414\\-1.414 & 2\end{pmatrix}\right\}$$

and additionally a tunnel function,

$$r_4(t,\ u) = \frac{5}{\pi}\exp\left\{-10\left(t-\frac{1}{2}\right)^2\right\}$$

where the asymptotic formula (2.1) does not hold.

The simulation study involved 100 replications of each situation. We calculated some quantiles of the bandwidths selected, of ISE and of the performance ratio $\mathrm{ISE}(\hat{b}_t,\ \hat{b}_u)/\inf_{b_t, b_u}\{\mathrm{ISE}(b_t,\ b_u)\}$.

Table 1 shows some quantiles for different regression functions and equidistant design.

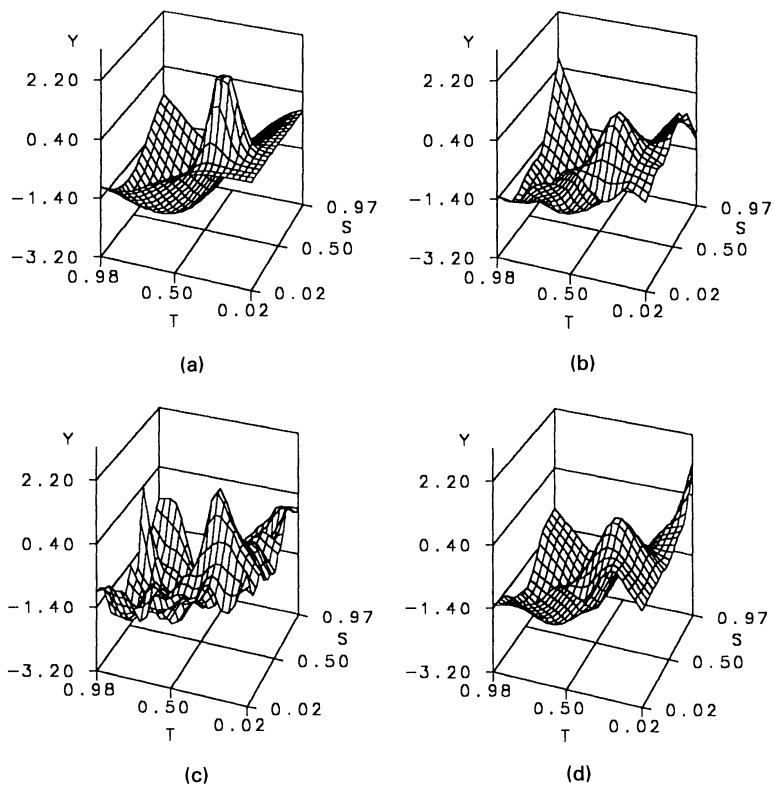

Fig. 1. (a) Surface of the regression curve r_3, (b) kernel estimate based on ISE optimal bandwidths, (c) kernel estimate based on the unbiased risk bandwidth selector and (d) kernel estimate based on the iterative plug-in bandwidth selector ((b)–(d) use realizations with the 95% quantile in ISE)

Various design functions are compared in Table 2, which gives some quantiles for function r_1, $n_t = 20$, $n_u = 10$, $\sigma^2 = 0.1$ and for equidistant design f_1, for a restricted Gauss design density

$$f_2\colon\ c_2 N\left\{\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix},\ \begin{pmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix}\right\}$$

and a linear design density

$$f_3(t, u) = (t + \tfrac{1}{2}) \times \left(\frac{19}{10}u + \frac{1}{20}\right).$$

These resuls are quite typical for several regression functions and design densities. We can see that especially for relatively small sample size or high residual variance the plug-in rule beats both others. Besides this its calculation is simpler and much faster.

Fig. 1 shows the estimation of the regression function r_3, $n_t = 25$, $n_u = 20$ and $\sigma^2 = 0.5$ for equidistant design. Besides the true regression function the resulting estimators when using the plug-in bandwidths, the unbiased risk bandwidths and the ISE optimal bandwidths are shown, each with a realization where the 95% quantile in ISE is reached. It is quite typical that if the ISE is large for the unbiased risk bandwidths they are chosen to be much too small whereas the plug-in bandwidths are typically too large if their ISE is large.

ACKNOWLEDGEMENTS

Part of this research was carried out while M. P. Wand was visiting the statistics group of Sonderforschungbereich 123 (project B1), Universität Heidelberg, while affiliated to the Department of Statistics, Texas A&M University.

APPENDIX A

We introduce a nonparametric method for estimating σ^2 with small bias by using the edges of all possible Delaunay triangulations (see for example Ripley (1981)). For a design point (t_i, u_i) we consider all triplets of design points with the following properties. Firstly, each point is connected with (t_i, u_i) by a single edge. Secondly, there is a path of two edges connecting the three design points forming a triplet and, thirdly, the three design points are not located on a straight line. An interior point of a rectangular design, for example, has four triplets fulfilling all three requirements. For a given design point (t_i, u_i) we define a pseudoresidual $\tilde{\epsilon}_i$ as the mean difference between Y_i and all planes through three such neighbouring pionts at (t_i, u_i). Let us denote $a_i = \sigma^2/\mathrm{var}(\tilde{\epsilon}_i)$. Our variance estimator $\hat{\sigma}^2$ is defined as the mean of all $a_i\tilde{\epsilon}_i^2$. For a rectangular design we have explicit formulae

$$\tilde{\epsilon}_{ij} = Y_{ij} - \tfrac{1}{2}(Y_{i-1,j} + Y_{i+1,j} + Y_{i,j-1} + Y_{i,j+1}) + \tfrac{1}{4}(Y_{i-1,j-1} + Y_{i-1,j+1} + Y_{i+1,j-1} + Y_{i+1,j+1})$$

and $a_{ij} = 4/9$ for $2 \leqslant i \leqslant n_t - 1$, $2 \leqslant j \leqslant n_u - 1$, and analogously at the boundary.

Under the assumptions on the partition and assuming that all second-order partial derivatives of r are continuous, the bias and variance of our variance estimator are of the order $O(n^{-1})$. For a rectangular design the order of the bias can be improved to $O(n^{-2})$.

APPENDIX B

Here we sketch one possibility on how to transfer the analysis for a non-rectangular design to the case of a rectangular design, after estimating the variance. Note that the simplicity of smoothing for the rectangular design stems from the rectangular partition sets $A_1, \ldots, A_n$ only. For non-rectangular design points we can always find a partition $A_1, \ldots, A_n$ with elements built up of several small rectangles such that the conditions of Section 2 are fulfilled. For example, let A be the unit square and $t_{(1)}, \ldots, t_{(n)}$ and $u_{(1)}, \ldots, u_{(n)}$ be the ordered one-dimensional design points. Firstly, we suppose that they are strictly ordered. Let $\tau_0 = 0$, $\tau_j = (t_{(j)} + t_{(j+1)})/2$ for $j = 1, \ldots, n-1$, $\tau_n = 1$ and $v_0 = 0$, $v_k = (u_{(k)} + u_{(k+1)})/2$ for $k = 1, \ldots, n-1$, $v_n = 1$. Then we have n^2 small rectangles $\tilde{A}_{jk} = (\tau_{j-1}, \tau_j] \times (v_{k-1}, v_k]$. The following procedure unambiguously assigns each rectangle $\tilde{A}_{jk}$ to an element of the partition $A_1, \ldots, A_n$. For given $\tilde{A}_{jk}$ we can associate an original design point located nearby, e.g.

$$\|(t_{(j)}, u_{(k)}) - (t_{i*}, u_{i*})\| = \min_i \|(t_{(j)}, u_{(k)}) - (t_i, u_i)\|.$$

For fixed (t_{i*}, u_{i*}) we can compose the partition set A_{i*} by all rectangles $\tilde{A}_{jk}$ which are associated with (t_{i*}, u_{i*}) in the above sense. Kernel smoothing with this partition $A_1, \ldots, A_n$ gives the same surface as smoothing in the enlarged rectangular model with partition $\tilde{A}_{jk}$ and sample points $\tilde{Y}_{jk}$ where $\tilde{Y}_{jk} = Y_{i*}$ and i^* is defined as above as the index of the associated design point for the set $\tilde{A}_{jk}$ for $j, k = 1, \ldots, n$. This can easily be seen by definition of the kernel estimator. If the original one-dimensional design points cannot be strictly ordered the same procedure can be done with the remaining rectangles with positive Lebesgue measure. For a rectangular design this procedure leads to the partition used in our simulations.

APPENDIX C

We present the main technical result to justify the iterative step defined by equation (3.1).

Suppose that r has all fourth-order partial derivatives continuous and that L_2 and M are Lipschitz continuous with support $[-1, 1]$. Assume that the bandwidths α_t, α_u, β_t and β_u satisfy

$$\begin{aligned}
\alpha_t &= c_t n^{-\xi}\{1 + o(1) + o_p(n^{-\beta})\},\\
\alpha_u &= c_u n^{-\zeta}\{1 + o(1) + o_p(n^{-\beta})\},\\
\beta_t &= d_t n^{-\zeta}\{1 + o(1) + o_p(n^{-\beta})\},\\
\beta_u &= d_u n^{-\xi}\{1 + o(1) + o_p(n^{-\beta})\},
\end{aligned}$$

where $c_t, c_u, d_t, d_u, \xi, \zeta > 0$, $\beta \geqslant 0$ and $\xi \leqslant \zeta$. Then:

(a) for $5\xi + \zeta > 1$,

$$\begin{aligned}
\hat{I}_{tt}(\alpha_t, \alpha_u) &= Jn^{-1+5\xi+\zeta} c_t^{-5} c_u^{-1}\{1 + o_p(1)\},\\
\hat{I}_{uu}(\beta_t, \beta_u) &= Jn^{-1+5\xi+\zeta} d_t^{-1} d_u^{-5}\{1 + o_p(1)\}
\end{aligned}$$

and

$$\hat{I}_{tt}^{1/2}\hat{I}_{uu}^{1/2} + \hat{I}_{tu} = 2Jn^{-1+5\xi+\zeta} c_t^{-5/2} c_u^{-1/2} d_t^{-1/2} d_u^{-5/2}\{1 + o_p(1)\},$$

where

$$J = \sigma^2 R(L_2) R(M) I_f \int_0^1 \int_0^1 \omega(t, u)\, dt\, du;$$

(b) for $5\xi + \zeta = 1$,

$$\hat{I}_{tt}(\alpha_t, \alpha_u) = I_{tt} + Jc_t^{-5}c_u^{-1} + o_p(1),$$
$$\hat{I}_{uu}(\beta_t, \beta_u) = I_{uu} + Jd_t^{-1}d_u^{-5} + o_p(1)$$

and

$$\hat{I}_{tt}^{1/2}\hat{I}_{uu}^{1/2} + \hat{I}_{tu} = I_{tt}^{1/2}I_{uu}^{1/2} + I_{tu} + 2Jc_t^{-5/2}c_u^{-1/2}d_t^{-1/2}d_u^{-5/2} + o_p(1);$$

(c) for $5\xi + \zeta = \frac{1}{2}$,

$$\hat{I}_{tt}(\alpha_t, \alpha_u) = I_{tt} + O(n^{-2\xi}) + O_p(n^{-1/2}) + o_p(n^{-2\xi-\beta}),$$
$$\hat{I}_{uu}(\beta_t, \beta_u) = I_{uu} + O(n^{-2\xi}) + O_p(n^{-1/2}) + o_p(n^{-2\xi-\beta})$$

and

$$\hat{I}_{tt}^{1/2}\hat{I}_{uu}^{1/2} + \hat{I}_{tu} = I_{tt}^{1/2}I_{uu}^{1/2} + I_{tu} + o_p(n^{-2\xi-\beta}) + O(n^{-2\xi}) + O_p(n^{-1/2}).$$

These results can be verified by applying techniques similar to the proof of theorem 2 of Gasser *et al.* (1991).

REFERENCES

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression.* New York: Dekker.

Gasser, Th., Kneip, A. and Köhler, W. (1991) A flexible and fast method for automatic smoothing. *J. Am. Statist. Ass.*, **86**, 643–652.

Gasser, Th., Müller, H.-G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *J. R. Statist. Soc.* B, **47**, 238–252.

Gasser, Th., Sroka, L. and Jennen-Steinmetz, C. (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

Hall, P., Kay, J. W. and Titterington, D. M. (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.

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

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? *J. Am. Statist. Ass.*, **83**, 86–95.

Jones, M. C., Marron, J. S. and Park, B. U. (1991) A simple root-*n* bandwidth selector. *Ann. Statist.*, **19**, 1919–1932.

Marron, J. S. (1988) Automatic smoothing parameter selection: a survey. *Emp. Econ.*, **13**, 187–208.

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

Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.

Ripley, B. D. (1981) *Spatial Statistics.* New York: Wiley.

Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc.* B, **53**, 683–690.