Working Paper 98-83
Statistics and Econometrics Series 38
November 1998

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax 34 - 91- 624.9849

INTEGRATION AND BACKFITTING METHODS IN ADDITIVE MODELS - FINITE SAMPLE PROPERTIES AND COMPARISON

Stefan Sperlich, Oliver B. Linton and Wolfgang Härdle*

Abstract

We examine and compare the finite sample performance of the competing backfitting and integration methods for estimating additive nonparametric regression using simulated data. Although, the asymptotic properties of the integration estimator, and to some extent the backfitting method too, are well understood, its small sample properties are not well investigated. Apart from some small experiments in the above cited papers, there is little hard evidence concerning the exact distribution of the estimates. It is our purpose to provide an extensive finite sample comparison between the backfitting procedure and the integration procedure using simulated data.

Key Words

Additive models; Curse of Dimensionality; Dimensionality reduction; Model choice; Nonparametric Regression.

*Sperlich, Department of Statistics y Econometrics, Universidad Carlos III de Madrid, e-mail: stefan@est-econ.uc3m.es; Linton, Department of Economics, Yale University, New Haven, CT 06511; Härdle, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, D-10178 Berlin. AMS classification: 62G07, 62G20, 62G35.

1 Introduction

Additive models are widely used in both theoretical economics and in econometric data analysis. The standard text of Deaton and Muellbauer (1980) provides many examples in microeconomics for which the additive structure provides interpretability and allows solution of choice problems. Additive structure is desirable from a purely statistical point of view because it circumvents the curse of dimensionality. There has been much theoretical and applied work in econometrics on semiparametric and nonparametric methods, see Härdle and Linton (1994), Newey (1990), and Powell (1994) for bibliography and discussion. Some recent work has shown that additivity has important implications for the rate at which certain components can be estimated. In this paper we consider the finite sample performance of two popular estimators for additive models: the backfitting estimators of Hastie and Tibshirani (1990) and the integration estimators of Linton and Nielsen (1995).

Let (X, Y) be a random variable with X of dimension d and Y a scalar. Consider the estimation of the regression function $m(x) = E(Y \mid X = x)$ based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$ from this population. Stone (1980, 1982) and Ibragimov and Hasminskii (1980) showed that the optimal rate for estimating m is $n^{-\frac{\ell}{2\ell+d}}$ with ℓ an index of smoothness of m. An additive structure for m is a regression function of the form

$$m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \tag{1}$$

where $x = (x_1, \ldots, x_d)^T$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions operating on each element of the vector or predictor variables with $E\{m_\alpha(X_\alpha)\} = 0$. Stone (1985, 1986) showed that for such regression curves the optimal rate for estimating m is the one-dimensional rate of convergence with $n^{-\frac{\ell}{2\ell+1}}$. Thus one speaks of dimensionality reduction through additive modelling.

In practice, the backfitting procedures proposed in Breiman and Friedman (1985) and Buja, Hastie and Tibshirani (1989) are widely used to estimate the additive components. The latter (equation (18)) consider the problem of finding the projection of m onto the space of additive functions representing the right hand side of (1). Replacing population by sample, this leads to a system of normal equations with $nd \times nd$ dimensions. To solve this in practice, the backfitting or Gauss-Seidel algorithm, is usually used, see Venables and Ripley (1994). This technique is iterative and depends on the starting values and convergence criterion. It converges very fast but has, in comparison with the direct solution of the large linear system, the slight disadvantage of a more complicated 'hat matrix', see Härdle and Hall (1993). These methods have been evaluated on numerous datasets and have been refined quite considerably since their introduction.

Recently, Linton and Nielsen (1995), Tjøstheim and Auestad (1994), and Newey (1994) have independently proposed an alternative procedure for estimating m_α based on integration of a standard kernel estimator. It exploits the following idea. Suppose that $m(x,z)$ is any bivariate function, and consider the quantities $\mu_1(x) = \int m(x,z)dQ_n(z)$ and $\mu_2(z) = \int m(x,z)dQ_n(x)$, where Q_n is a probability measure. If $m(x,z) = m_1(x) + m_2(z)$, then $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are $m_1(\cdot)$ and $m_2(\cdot)$, respectively, up to a constant. In practice one replaces m by an estimate and integrates with respect to some known measure. The procedure is explicitly defined and its asymptotic distribution is easily derived: it converges at the one-dimensional rate and satisfies a central limit theorem. This estimation procedure has been extended to a number of other contexts like estimating the derivatives (Severance-Lossin and Sperlich, 1997), to the generalized additive model (Linton and Härdle, 1996), to dependent variable transformation models (Linton, Chen, Wang, and Härdle, 1997), to econometric time series models (Masry and Tjøstheim, 1995,1997), to panel data models (Porter, 1996), and to hazard models with time varying covariates and right censoring (Nielsen, 1996). In this wide variety of sampling schemes and procedures the asymptotics have been derived because of the explicit form of the estimator. By contrast, backfitting or backfitting-like methods have until recently eluded theoretical analysis, until Opsomer and Ruppert (1997) provided conditional mean squared error expressions albeit under rather strong conditions on the smoothing matrices and design. More recently, Linton, Mammen, and Nielsen (1998) has established a central limit theorem for a modified form of backfitting which uses a bivariate integration step as well as the iterative updating of the other methods.

The purpose of this paper is to investigate the finite sample performance of the standard backfitting estimator and the integration estimator.

2 Methods and Theory

We suppose that

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where by definition $E(\varepsilon_i|X_i) = 0$; let also $\mathrm{var}(\varepsilon_i|X_i) = \sigma^2(X_i)$ be the conditional variance function. We denote the marginal density of the d-dimensional explanatory variable by $p(x)$ with marginals $p_\alpha(x_\alpha)$, $\alpha = 1, \ldots, d$. We shall sometimes partition $X_i = (X_{\alpha i}, X_{\underline{\alpha} i})^T$ and $x = (x_\alpha, x_{\underline{\alpha}})^T$ into scalar and $d-1$-dimensional subvectors respectively calling x_α the direction of interest and $x_{\underline{\alpha}}$ the direction not of interest; denote by $p_{\underline{\alpha}}(x_{\underline{\alpha}})$ the marginal density of the vector $X_{\underline{\alpha} i}$. In the following we assume the following additive form for the regression function

$$m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \ , \ x = (x_1, \ldots, x_d)^T \ , c \ \text{ constant.}$$

2.1 Integration

A commonly used estimate of $m(x)$ is provided by the multidimensional local polynomial product kernel estimator which solves the following minimization problem

$$(2) \qquad \begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix} = \min_{\theta_0,\theta_1} \sum_{i=1}^{n} \{Y_i - P_q(\theta_0,\theta_1; X_i - x)\}^2 \prod_{\alpha=1}^{d} K_\alpha \left(\frac{x_\alpha - X_{\alpha i}}{h_\alpha} \right),$$

where K_α and h_α, $\alpha = 1,\ldots,d$, are scalar kernels and bandwidths respectively, while $P_q(\theta_0,\theta_1;t)$ is a $(q-1)^{th}$ order polynomial in the vector t with coefficients θ_0,θ_1 for which $P_q(\theta_0,\theta_1;0) = \theta_0$ and e.g. $P_2(\theta_0,\theta_1;t) = \theta_0 + \theta_1 t$. Let $\widehat{m}(x) = \widehat{\theta}_0(x)$. Under regularity conditions, see Ruppert and Wand (1995) for example, the local polynomial estimator satisfies

$$(3) \qquad \widehat{m}_h(x) - m(x) \xrightarrow{\mathcal{L}} N\left\{ h^q \mu_q(K) b(x), \frac{1}{nh^d} \nu_q(K) v(x) \right\},$$

where $h = \left(\prod_{\alpha=1}^{d} h_\alpha\right)^{1/d}$ is the geometric average of the bandwidths, $\mu_q(K)$ and $\nu_q(K)$ are constants depending only on the kernels, while $v(x) = \sigma^2(x)/p(x)$ and $b(x)$ is the bias function depending on derivatives of m, and possibly p, up to and including order q. The (mean squared error) optimal bandwidth is of order $n^{-1/(2q+d)}$ for which the asymptotic mean squared error is of order $n^{-2q/(2q+d)}$, see Härdle and Linton (1994), which reflects the curse of dimensionality – as d increases, the rate of convergence decreases.

When $m(\cdot)$ satisfies the additive model structure, we can estimate $m(x)$ with a better rate of convergence by imposing these restrictions. Let

$$(4) \qquad \widetilde{m}_\alpha(x_\alpha) = \int \widehat{m}_{h_1}(x_\alpha, x_{\underline{\alpha}}) dQ_n(x_{\underline{\alpha}}) - \widehat{c},$$

where $\widehat{c}$ is an estimate of c, while $Q_n(\cdot)$ is some easy to compute probability measure. The most convenient choice of $Q_n(\cdot)$ is the empirical measure of $\left\{X_{\underline{\alpha}i}\right\}_{i=1}^{n}$, which converges to the population distribution. It changes the integral in (4) to a sum over terms evaluated at $X_{\underline{\alpha}i}$ and implies for the constant $c = E(Y)$. The latter can be estimated root-n consistently by the sample mean $n^{-1}\sum_{i=1}^{n} Y_i$; an alternative estimate, which is not necessarily root- n consistent, is provided by $n^{-1}\sum_{i=1}^{n} \widetilde{m}_\alpha(X_{\alpha i})$. Whatever the estimates of c and $\widetilde{m}_\alpha(x_\alpha)$, we reestimate $m(x)$ by

$$(5) \qquad \widetilde{m}_{h_1}(x) = \widehat{c} + \sum_{\alpha=1}^{d} \widetilde{m}_\alpha(x_\alpha).$$

Linton and Härdle (1996) derived the pointwise asymptotic properties of the empirical integration versions of $\widetilde{m}_\alpha(x_\alpha)$ and $\widetilde{m}_{h_1}(x)$. To simplify matters, we set $h_\alpha = h_1$ and $K_\alpha = K$, while $\prod_{\beta\neq\alpha} K_\beta = L$ and $h_\beta = h_2$ for all $\beta \neq \alpha$. Under their regularity conditions,

$$(6) \qquad \widetilde{m}_{h_1}(x) - m(x) \xrightarrow{\mathcal{L}} N\left\{ \frac{h_1^q}{q!} \overline{\mu}_q(K) b_0(x), \frac{1}{nh_1} \overline{\nu}_q(K) v_0(x) \right\},$$

where $b_0(x) = \sum_\alpha b_{\alpha 0}(x_\alpha)$ and $v_0(x) = \sum_\alpha v_{\alpha 0}(x_\alpha)$ with $b_{\alpha 0}(x_\alpha) = \int b(x) p_{\underline{\alpha}}(x_{\underline{\alpha}}) dx_{\underline{\alpha}}$ and $v_{\alpha 0}(x_\alpha) = \int v(x) p^2_{\underline{\alpha}}(x_{\underline{\alpha}}) dx_{\underline{\alpha}}$. Here, $\overline{\mu}_q(K)$ and $\overline{\nu}_q(K)$ are constants depending only on the kernel K. By choosing $h_1 \propto n^{-1/(2q+1)}$ one can achieve the optimal rate of convergence i.e., mean squared error of order $n^{-2q/(2q+1)}$, which is independent of the dimensions d. See also Linton and Nielsen (1995) and Severance-Lossin and Sperlich (1997).

REMARK. The bandwidths $h_1, \ldots, h_d$ should be chosen differently as we discuss further in the simulation section. To achieve the optimal rate of convergence, we must impose some restrictions on the bandwidth sequences. This condition, which corresponds to (A7) in Linton and Härdle (1995) is needed for bias reduction of the nuisance components. In section 3 we will examine some bandwidth selection methods.

2.2 Backfitting

Hastie and Tibshirani (1990) motivate the backfitting method as follows. First consider the analogous population problem:

$$\min_m E\{Y - m(X)\}^2 \quad \text{s.t.} \quad m(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha),$$

which can be formulated inside a Hilbert space framework: let $[\mathcal{H}_{YX}, \langle \cdot, \cdot \rangle]$ be the Hilbert space of random variables which are functions of Y and X with $\langle a, b \rangle = E(ab)$, let also $[\mathcal{H}_X, \langle \cdot, \cdot \rangle]$, and $[\mathcal{H}_{X_\alpha}, \langle \cdot, \cdot \rangle]$, $\alpha = 1, \ldots, d$ be corresponding subspaces, where for example $\mathcal{H}_{X_\alpha}$ contains only functions of X_α. The above problem is equivalent to finding the element of the subspace $\mathcal{H}_{X_1} \oplus \cdots \oplus \mathcal{H}_{X_d}$ closest to a point $Y \in \mathcal{H}_{YX}$ or equivalently the point $m \in \mathcal{H}_X$. By the projection theorem, there exists a unique solution which is characterized by the following first order conditions

$$E[\{Y - m(X)\}|X_\alpha] = 0 \Leftrightarrow m_\alpha(X_\alpha) = E\left[\left\{Y - \sum_{\gamma \neq \alpha} m_\gamma(X_\gamma)\right\} | X_\alpha\right], \quad \alpha = 1, \ldots, d,$$

which leads to the formal representation:

$$\begin{pmatrix} I & P_1 & \cdots & P_1 \\ P_2 & I & \cdots & P_2 \\ \vdots & & \ddots & \vdots \\ P_d & \cdots & P_d & I \end{pmatrix} \begin{pmatrix} m_1(X_1) \\ m_2(X_2) \\ \vdots \\ m_d(X_d) \end{pmatrix} = \begin{pmatrix} P_1 Y \\ P_2 Y \\ \vdots \\ P_d Y \end{pmatrix},$$

where $P_\alpha(\cdot) = E(\cdot | X_\alpha)$. By analogy, let S_α $(n \times n)$ be the smoother matrix which when applied to the $n \times 1$ vector $y = (Y_1, \ldots, Y_n)^T$ yields an $n \times 1$ vector estimate $S_\alpha y$ of the

vector $\{E(Y_1|X_{\alpha 1}), \ldots, E(Y_n|X_{\alpha n})\}^T$. Substituting P_α by S_α we obtain the following

$$\underbrace{\begin{pmatrix} I & S_1 & \cdots & S_1 \\ S_2 & I & \cdots & S_2 \\ \vdots & & \ddots & \vdots \\ S_d & \cdots & S_d & I \end{pmatrix}}_{nd \times nd} \begin{pmatrix} \widehat{m}_1 \\ \widehat{m}_2 \\ \vdots \\ \widehat{m}_d \end{pmatrix} = \begin{pmatrix} S_1 y \\ S_2 y \\ \vdots \\ S_d y \end{pmatrix}.$$

This system can in principle be solved exactly for $\{\widehat{m}_\alpha(X_{\alpha 1}), \ldots, \widehat{m}_\alpha(X_{\alpha n})\}^T$, $\alpha = 1, \ldots, d$. However, when nd is large the required matrix inversion is not feasible. Further, often the matrix on the left is not regular in practice and thus this equation cannot be solved directly. In practice, the backfitting (Gauss-Seidel) algorithm is used to solve these equations: given starting values $\check{m}_\alpha^{(0)}$, $\alpha = 1, \ldots, d$, update the $n \times 1$ vectors as follows

$$\check{m}_\alpha^{(r)} = S_\alpha \left\{ y - \sum_{\gamma \neq \alpha} \check{m}_\gamma^{(r-1)} \right\}, \quad r = 1, \ldots$$

until some prespecified tolerance is reached. The estimator is linear in y, but the algorithm only converges under strong restrictions on the smoother matrices. Recent work by Opsomer and Ruppert (1997) discuss some improvements to this algorithm which are guaranteed to provide a unique solution. They also derive the conditional mean squared error of the resulting estimator under strong conditions: this has a similar expression to (6) in large samples.

3 Simulation Results

3.1 Introduction

In a number of different additive models, we determined the bias, variance and mean squared error for both estimation procedures. We considered designs with distributions: the uniform $U[-3, 3]^d$, the normal with mean 0, variance 1 and varying covariance $\rho = 0, 0.4, 0.8$, denoted as $N(\rho)$, for different numbers of observations and several dimensions. We drew all these designs once and kept them fixed for the investigation described in the following. The error term ε was always chosen as normal distributed with zero mean and variance $\sigma_\epsilon^2 = 0.5$. Since both estimators are linear, i.e.,

$$\widehat{m}_\alpha(x) = \sum_{i=1}^{n} w_{\alpha i}(x) Y_i$$

for some weights $\{w_{\alpha i}(x)\}$ we determined the conditional bias and variance as follows

$$\text{var}\{\widehat{m}_\alpha(x_\alpha)|X\} = \sigma_\epsilon^2 \sum_{i=1}^{n} w_{\alpha i}^2(x)$$

$$\text{bias}\{\widehat{m}_\alpha(x_\alpha)|X\} = \sum_{i=1}^{n} w_{\alpha i}(x) m(X_i) - m_\alpha(x_\alpha)$$

for the additive function estimators and by analogy for the regression estimator. In the following notation the MSE denotes the mean squared error and the MASE the averaged MSE. We focused on the following questions:

a) What is a reasonable bandwidth choice for an optimal fit ?

b) How sensitive are the estimators to the bandwidth ?

c) What are the MASE, MSE, bias and variance, boundary effects ?

d) We considered degrees of freedom, eigen analysis, singular values and eigen vectors.

e) We plotted the equivalent kernel weights of the estimates and

f) we investigated whether and when the asymptotics kick in.

We examined how well the estimation procedures performed in estimating one additive function. The parameters are $d = 2$ dimensions and $n = 100$ observations. We considered all combinations of the following additive functions for a two dimensional additive model:

$$m_1(x) = 2x \quad ; \quad m_2(x) = x^2 - E(x^2)$$
$$m_3(x) = \exp(x) - E\{\exp(x)\} \quad ; \quad m_4(x) = 0.5 \cdot \sin(-1.5x).$$

Our interest is mainly in the estimation of the marginal effect m_α . We first determined different optimal bandwidths for a given design distribution. In the second step we calculated for fixed designs bias, variance and mean average squared error (on the complete data set as well as on trimmed data) for both estimation procedures.

The advantages of using local polynomials are well known, especially with regard to the robustness against choice of bandwidth and the improvement in bias and consequently mean squared error if the requisite smoothness is present. In Severance-Lossin and Sperlich (1997) the consistency and asymptotic behavior of the integration estimator using local polynomial is shown. For these reasons we did the investigation for both, the Nadaraya Watson and the local linear estimator.

3.2 Bandwidth Choice

The choice of an appropriate smoothing parameter is always a critical point in nonparametric and semiparametric estimation. For the integration estimator we need even two bandwidths, h_1 and h_2, see section 2. There exist at least two rules for choosing them: the rule of thumb of Linton and Nielsen (1995) and the plug-in method suggested in Severance-Lossin and Sperlich (1997). Both methods give the MASE minimizing bandwidth, the first one approximately with the aid of parametric pre-estimators, the second

one by using nonparametric pre-estimators. We give here the formulas for the case of local linear smoothers. The rule of thumb is

$$h_1 = \left\{ \frac{\tilde{\sigma}^2 \nu(K)(max - min)}{\mu_2(K)(\sum_{j=1}^{d} \hat{\beta}_j)^2} \right\}^{1/5} n^{-1/5}$$

where $\nu(K) = \|K\|_2^2$, $\mu_2(K) = \int t^2 K(t) dt$ and max and min are the sample maximum and minimum of the direction of interest. We obtained $\tilde{\beta}_j$ as the coefficients of $x_j^2/2$ from a least squares regression of Y on a constant, x_j, $x_j^2/2$ and $x_j x_k$ for all $j, k = 1, \ldots, d$ $j < k$, while $\tilde{\sigma}^2$ was obtained from the residuals of this regression by taking the average of the squares.

The formula for the nonparametric plug-in method we used for calculating the asymptotically optimal bandwidth is

$$h_1 = \left\{ \frac{\nu(K) \int \sigma^2 \frac{p_{\underline{\alpha}}^2(x_{\underline{\alpha}}) p_\alpha(x_\alpha)}{p(x_\alpha, x_{\underline{\alpha}})} dx_{\underline{\alpha}} dx_\alpha}{4\{\frac{1}{2}\mu_2(K)\}^2 \int \{m_\alpha^{(2)}(x_\alpha)\}^2 p_\alpha(x_\alpha) dx_\alpha} \right\}^{1/5} n^{-1/5} .$$

Note that this formula is not valid for h_2, the bandwidth for the direction not of interest. We took the bandwidth h_2 that minimized the MASE in the particular finite sample model.

For a fair comparison of the optimal bandwidth and the corresponding MASE of both estimators we applied several procedures. We started with considering the minimal MASE of the overall regression function and the minimizing bandwidths. Then we looked for the bandwidths minimizing the MASE in each direction separately. For taking into account the influence of boundary effects we looked also for the optimal bandwidths on trimmed data.

For small samples of 100 observations we could not discover any information by comparing the numerically MASE-minimizing bandwidths. They differed a lot depending on the particularly drawn design. Therefore we focused on once drawn, in that sense fixed, designs for the whole paper and considered only analytically determined bandwidths h_1. Thus we compared the results for bandwidths calculated with the rule of thumb proposed by Linton and Nielsen and the analytically optimal one.

Selected numerical Results, using both, the Nadaraya Watson and the local linear Smoother Since the values of the MASE minimizing bandwidths that we found numerically for the particular designs in finite samples, were not particularly illuminting, we do not report them in the tables. In table 1 the bandwidths of the rule of thumb by Linton and Nielsen and the asymptotically optimal bandwidths for each estimation procedure are shown. Here we concentrated on bandwidths that minimize the MASE in

each direction separately. They are displayed for the additive components m_3, m_4 versus the particular model and design. The behavior for m_1, m_2 is the same, the results can be requested from the authors. One can see very well the strong influence of the distribution and the dependence of the additive function that has to be estimated. Furthermore, not only do the bandwidths determined by theory based rules differ a lot, we found them quite often far away from the MASE minimizing bandwidth value. This is also the case for the local linear smoothers. Mostly, the analytically chosen bandwidth was closer to the MASE minimizing one than the rule of thumb bandwidth, which, however, is much easier to calculate.

If the optimal value was infinity, we set it to 1 or in the case of a $N(0.8)$ distributed design to 2. In formulas where we had to integrate over a density from $-\infty$ to $+\infty$ we did this [for numerical reasons] over the interval $[-1.5, 1.5]$ for $N(0.8)$ and over $[-3, 3]$ else.

TABLE 1: ASYMPTOTICALLY OPTIMAL BANDWIDTHS WHEN USING NADARAYA WATSON

Additive Function:	$\hat{m}_3$				$\hat{m}_4$			
Distribution:	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model	$m = m_1 + m_3$				$m = m_1 + m_4$			
rule of thumb	0.222	0.246	0.243	0.242	0.308	0.294	0.316	0.376
backfitting	0.191	0.175	0.175	0.260	0.426	0.310	0.310	0.282
integration h_1	0.191	0.175	0.194	0.307	0.426	0.309	0.352	0.387
Model	$m = m_2 + m_3$				$m = m_2 + m_4$			
rule of thumb	0.194	0.210	0.209	0.209	0.279	0.251	0.260	0.276
backfitting	0.191	0.175	0.175	0.260	0.426	0.310	0.310	0.282
integration h_1	0.191	0.175	0.194	0.307	0.426	0.309	0.352	0.387
Model	$m = m_3 + m_4$				$m = m_3 + m_4$			
rule of thumb	0.185	0.230	0.234	0.243	0.185	0.230	0.234	0.243
backfitting	0.191	0.175	0.175	0.260	0.426	0.310	0.310	0.282
integration h_1	0.191	0.175	0.194	0.307	0.426	0.309	0.352	0.387

Table 2 gives the optimal bandwidths for different distributions, models, estimation routines and criteria when using local linear smoothing.

All findings from table 1 are replicated here. Furthermore, note that for uncorrelated regressors the bandwidths are almost the same for backfitting and integration method, which is in accordance with the theoretically similar MASE. As mentioned above we will now consider the choice of bandwidth for the local linear estimation procedure in a more detailed way.

TABLE 2: ASYMPTOTICALLY OPTIMAL BANDWIDTHS WHEN USING LOCAL LINEAR

Additive Function:	$\hat{m}_3$				$\hat{m}_4$			
Distribution:	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model	$m = m_1 + m_3$				$m = m_1 + m_4$			
rule of thumb	0.222	0.246	0.243	0.242	0.308	0.294	0.316	0.376
backfitting	0.191	0.267	0.267	0.284	0.426	0.423	0.423	0.374
integration h_1	0.191	0.267	0.324	0.571	0.426	0.423	0.513	0.752
Model	$m = m_2 + m_3$				$m = m_2 + m_4$			
rule of thumb	0.194	0.210	0.209	0.209	0.279	0.251	0.260	0.276
backfitting	0.191	0.267	0.267	0.284	0.426	0.423	0.423	0.374
integration h_1	0.191	0.267	0.324	0.571	0.426	0.423	0.513	0.752
Model	$m = m_3 + m_4$				$m = m_3 + m_4$			
rule of thumb	0.185	0.230	0.234	0.243	0.185	0.230	0.234	0.243
backfitting	0.191	0.267	0.267	0.284	0.426	0.423	0.423	0.374
integration h_1	0.191	0.267	0.324	0.571	0.426	0.423	0.513	0.752

3.3 Robustness with respect to the Choice of Bandwidth

To find out how sensitive the estimators are with respect to the choice of bandwidth for the direction of interest h_1 we plotted MASE and $\text{MSE}_{x=0}$ against bandwidth for the two models $m = m_2 + m_3$ and $m = m_2 + m_4$. The parameters were kept unchanged or were mentioned in the caption of the respective figures. We present our results first for the uniform design on $[-3, 3]^2$, then for designs with distribution $N(0.0)$ and $N(0.4)$, see figures 1 to 6.

The results for MASE have been trimmed in the pictures, since otherwise they would have been dominated by boundary effects (compare with tables in the next section). The results for the integration estimator are drawn throughout the paper as solid lines, those for the backfitting algorithm as dashed lines.

Obviously, the backfitting estimator is very sensitive to the choice of bandwidth. To get a small MASE it is crucially important for the backfitting method to choose a good smoothing parameter. For correlated designs oversmoothing seems slightly preferable, otherwise there is no particular advantage to either oversmoothing or undersmoothing. The behavior of the estimates for the highly correlated design is slightly strange and hard to interpret. This is true for both estimation procedures. Therefore we skipped the figures for the $N(0.8)$ distributed design.

For the integration estimator the results differ depending on the model. In general this

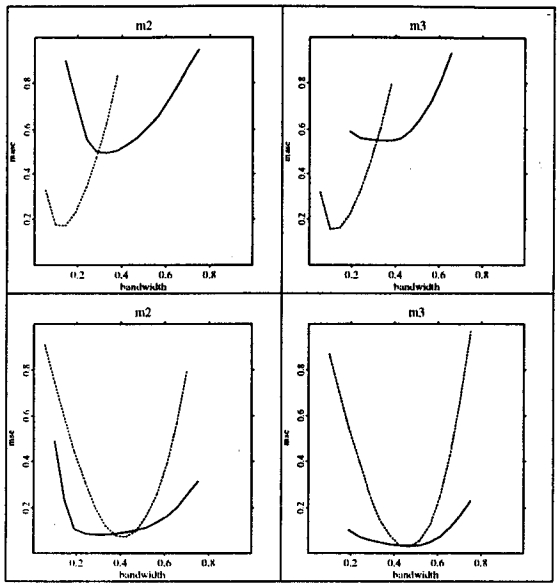

Figure 1: Performance by bandwidth h_1 of MASE(top) and $\text{MSE}_{x=0}$(bottom) in model $m = m_2 + m_3$, separately for m_2 (left), m_3(right). Design is $X \sim U[-3,3]^2$.

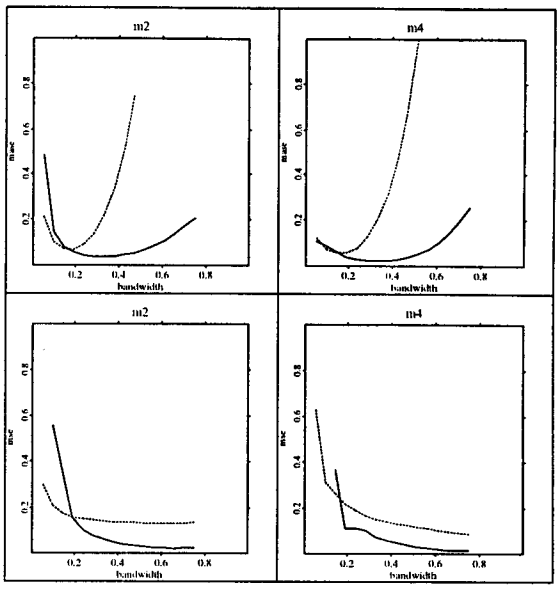

Figure 2: Performance by bandwidth h_1 of MASE(top) and $\text{MSE}_{x=0}$(bottom) in model $= m_2 + m_4$, separately for m_2 (left), m_4(right). Design is $X \sim U[-3,3]^2$.

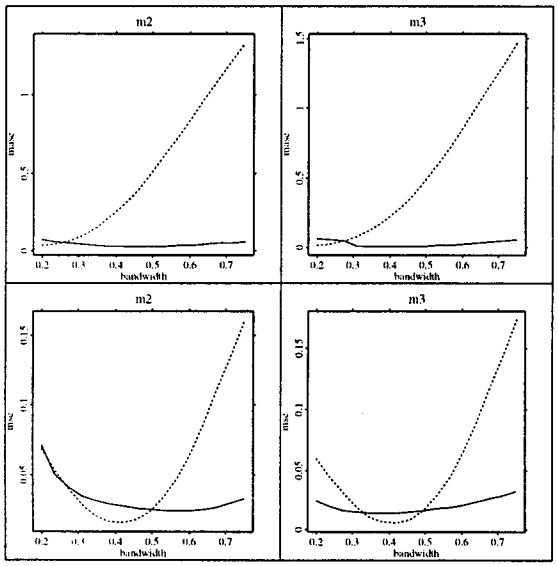

Figure 3: Performance by bandwidth h_1 of MASE(top) and $\text{MSE}_{x=0}$(bottom) in model $= m_2 + m_3$, separately for m_2 (left), m_3(right). Design is $X \sim N(0.0)$.

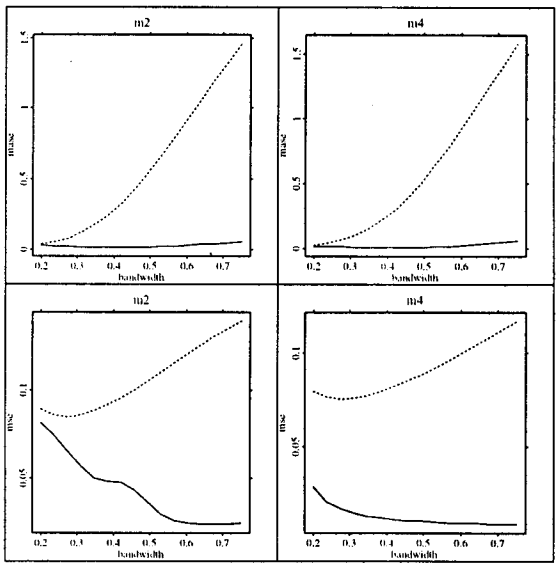

Figure 4: Performance by bandwidth h_1 of MASE(top) and $\text{MSE}_{x=0}$(bottom) in model $= m_2 + m_4$, separately for m_2 (left), m_4(right). Design is $X \sim N(0.0)$.

method is by far not as sensitive to the choice of bandwidth as the backfitting procedure is. If we focus on the $\text{MSE}_{x=0}$ we have similar results as for the MASE but weakened concerning the sensitivity. Here the results differ more depending on the data generating

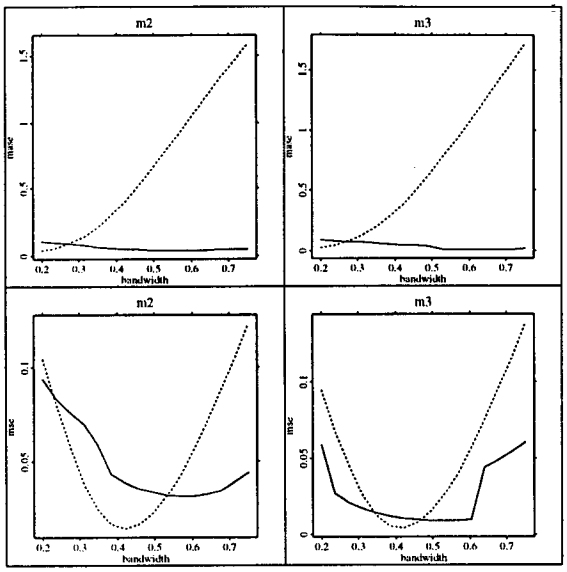

Figure 5: Performance by bandwidth h_1 of MASE(top) and $\text{MSE}_{x=0}$(bottom) in model $= m_2 + m_3$, separately for m_2 (left), m_3(right). Design is $X \sim N(0.4)$.

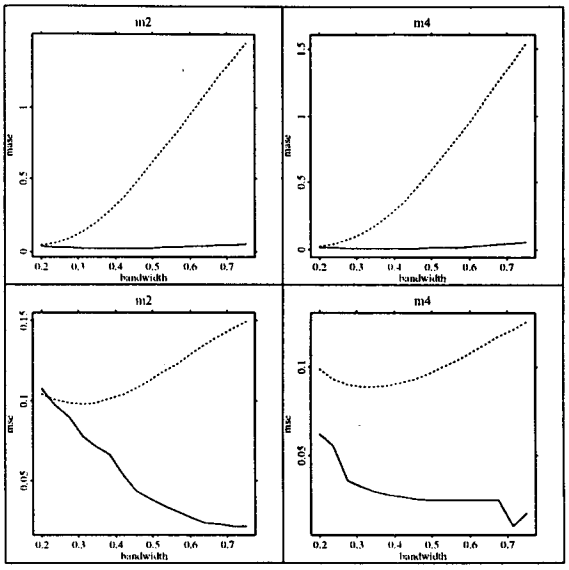

Figure 6: Performance by bandwidth h_1 of MASE(top) and $\text{MSE}_{x=0}$(bottom) in model $= m_2 + m_4$, separately for m_2 (left), m_4(right). Design is $X \sim N(0.4)$.

model.

Since in a $[-3,3]^2$ rectangle $n = 100$ observations are fairly sparse and thus the behavior of the MASE or $\text{MSE}_{x=0}$ perhaps is not typical, we did the same investigation with $n = 100$ observations for the uniform design on $[-1.5,1.5]^2$. But, plotting the MASE and $\text{MSE}_{x=0}$ functions on the same scale as we did for the $U[-3,3]^2$ design, we detected that the general sensitivity is similar but certainly on a different range. Furthermore, the integration estimate improved a lot since it has been suffering more when data were sparse as e.g., in $[-3,3]^2$. All in all, our observations above are confirmed when data were not too sparse.

3.4 Simulation Results: Bias, Variance and MASE

Due to the excess of information we got out of doing our simulations, we concentrate on the results for the local linear case. Nevertheless we think it worthwhile to mention both, and, if there are differences in the results, to discuss them.

For the optimal bandwidths computed in section 3.2 and given fixed designs the following tables present MASE, squared bias and variance on the complete data set and on trimmed data. In table 3–5 the results are for the complete data set in the upper line, for the trimmed data in the lower line.

TABLE 3: MASE, USING THE LOCAL LINEAR SMOOTHER OVER ALL (UPPER) AND OVER TRIMMED (LOWER) DATA

Distrib.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_2$				$m = m_1 + m_3$				$m = m_1 + m_4$			
$\hat{m}_\alpha$	back	0.065	0.047	0.041	0.020	0.401	0.046	0.028	0.053	0.022	0.027	0.026	0.028
		0.060	0.038	0.031	0.014	0.393	0.037	0.018	0.033	0.018	0.018	0.016	0.019
	int	0.028	0.019	0.030	0.057	0.616	0.031	0.075	0.081	0.018	0.023	0.023	0.024
		0.023	0.013	0.017	0.047	0.555	0.024	0.059	0.071	0.014	0.016	0.015	0.017
$\hat{m}_\beta$	back	0.116	0.083	0.079	0.047	0.479	0.073	0.053	0.058	0.033	0.033	0.036	0.048
		0.113	0.071	0.060	0.024	0.470	0.058	0.032	0.028	0.027	0.022	0.020	0.026
	int	0.436	0.090	0.116	0.530	0.402	0.137	0.234	0.528	0.191	0.043	0.074	0.137
		0.427	0.028	0.029	0.205	0.411	0.027	0.031	0.149	0.180	0.020	0.023	0.093
$\hat{m}$	back	0.052	0.052	0.054	0.049	0.066	0.051	0.054	0.057	0.040	0.040	0.043	0.042
		0.045	0.032	0.031	0.028	0.057	0.030	0.029	0.035	0.333	0.024	0.024	0.024
	int	0.156	0.115	0.145	0.619	0.206	0.175	0.285	0.608	0.066	0.063	0.104	0.132
		0.143	0.041	0.041	0.252	0.159	0.043	0.053	0.194	0.051	0.031	0.041	0.089
Distrib.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_2 + m_3$				$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	back	0.228	0.075	0.109	0.344	0.142	0.124	0.135	0.128	0.107	0.068	0.081	0.111
		0.206	0.057	0.079	0.119	0.141	0.107	0.116	0.099	0.101	0.046	0.055	0.081
	int	0.645	0.065	0.105	3.490	0.048	0.047	0.048	0.089	0.078	0.053	0.049	0.056
		0.572	0.031	0.064	0.782	0.041	0.022	0.026	0.078	0.070	0.026	0.022	0.041
$\hat{m}_\beta$	back	0.271	0.060	0.086	0.238	0.132	0.112	0.121	0.110	0.077	0.051	0.062	0.096
		0.264	0.041	0.058	0.093	0.124	0.101	0.110	0.091	0.071	0.039	0.048	0.075
	int	0.310	0.070	0.191	1.499	0.061	0.048	0.480	1.322	0.177	0.057	0.603	2.413
		0.233	0.040	0.055	0.186	0.047	0.032	0.061	0.151	0.166	0.040	0.265	1.020
$\hat{m}$	back	0.087	0.072	0.074	0.124	0.061	0.061	0.063	0.068	0.079	0.065	0.066	0.064
		0.076	0.042	0.041	0.067	0.052	0.035	0.035	0.037	0.069	0.038	0.037	0.038
	int	0.407	0.204	0.443	6.283	0.102	0.118	0.561	1.368	0.145	0.085	0.670	2.238
		0.308	0.136	0.236	0.785	0.084	0.076	0.083	0.189	0.123	0.044	0.257	0.681

We found three main points:

1) It is not possible to declare one estimating procedure superior to the other one in general. The results are differing from model to model and for each additive component we want to estimate. We neither can say that one of the estimation procedures is in general outperforming the other one regarding the MASE nor that

TABLE 4: AVERAGED SQUARED BIAS, USING LOCAL LINEAR OVER ALL (UPPER) AND OVER TRIMMED (LOWER) DATA

Distrib.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_2$				$m = m_1 + m_3$				$m = m_1 + m_4$			
$\hat{m}_\alpha$	back	0.043	0.022	0.016	0.000	0.380	0.022	0.003	0.033	0.001	0.002	0.001	0.007
		0.039	0.020	0.014	0.000	0.338	0.019	0.003	0.017	0.000	0.002	0.001	0.004
	int	0.007	0.003	0.008	0.020	0.598	0.015	0.053	0.043	0.001	0.006	0.004	0.004
		0.005	0.002	0.003	0.017	0.481	0.010	0.036	0.036	0.000	0.005	0.003	0.002
$\hat{m}_\beta$	back	0.078	0.046	0.037	0.002	0.425	0.033	0.008	0.015	0.004	0.004	0.003	0.010
		0.068	0.040	0.031	0.002	0.364	0.027	0.006	0.006	0.002	0.003	0.002	0.007
	int	0.358	0.054	0.065	0.415	0.329	0.097	0.175	0.431	0.142	0.013	0.034	0.055
		0.266	0.003	0.006	0.078	0.227	0.002	0.004	0.072	0.119	0.003	0.003	0.031
$\hat{m}$	back	0.006	0.005	0.005	0.003	0.004	0.002	0.002	0.014	0.003	0.002	0.002	0.004
		0.005	0.004	0.004	0.003	0.001	0.001	0.001	0.007	0.002	0.001	0.001	0.002
	int	0.062	0.068	0.075	0.475	0.118	0.123	0.209	0.482	0.005	0.022	0.050	0.040
		0.032	0.008	0.010	0.103	0.062	0.009	0.015	0.099	0.003	0.006	0.012	0.024
Distrib.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_2 + m_3$				$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	back	0.187	0.033	0.066	0.289	0.102	0.082	0.091	0.072	0.050	0.024	0.034	0.059
		0.156	0.028	0.049	0.058	0.087	0.072	0.080	0.058	0.039	0.017	0.026	0.047
	int	0.590	0.022	0.058	3.465	0.006	0.004	0.013	0.064	0.017	0.005	0.011	0.035
		0.387	0.007	0.033	0.574	0.004	0.000	0.004	0.047	0.010	0.002	0.002	0.019
$\hat{m}_\beta$	back	0.219	0.023	0.051	0.198	0.105	0.086	0.095	0.074	0.050	0.025	0.036	0.060
		0.180	0.019	0.035	0.035	0.088	0.072	0.079	0.059	0.040	0.019	0.029	0.050
	int	0.238	0.030	0.132	1.473	0.018	0.018	0.455	1.300	0.132	0.027	0.577	2.391
		0.134	0.010	0.020	0.136	0.011	0.011	0.036	0.124	0.097	0.019	0.211	0.724
$\hat{m}$	back	0.006	0.003	0.003	0.055	0.005	0.003	0.003	0.004	0.007	0.005	0.003	0.003
		0.003	0.002	0.001	0.013	0.003	0.002	0.002	0.002	0.004	0.002	0.002	0.001
	int	0.280	0.124	0.340	6.220	0.020	0.050	0.500	1.310	0.040	0.012	0.600	2.182
		0.098	0.077	0.153	0.624	0.010	0.033	0.041	0.138	0.023	0.005	0.170	0.539

one of them is more biased or has less variance than the competing one.

2) The integration estimator is suffering more from boundary effects.

3) For increasing correlation both estimators get problems but much more the integration estimator. This is in line with the theory saying that the integration

TABLE 5: AVERAGED VARIANCE, USING LOCAL LINEAR OVER ALL (UPPER) AND OVER TRIMMED (LOWER) DATA

Distrib.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_1 + m_2$				$m = m_1 + m_3$				$m = m_1 + m_4$			
$\hat{m}_\alpha$	back	0.022	0.024	0.025	0.020	0.022	0.024	0.025	0.020	0.022	0.025	0.025	0.020
		0.015	0.014	0.014	0.012	0.015	0.014	0.014	0.012	0.015	0.014	0.014	0.013
	int	0.021	0.017	0.023	0.037	0.018	0.017	0.023	0.039	0.017	0.017	0.019	0.020
		0.015	0.009	0.011	0.025	0.013	0.010	0.011	0.026	0.012	0.009	0.010	0.013
$\hat{m}_\beta$	back	0.038	0.037	0.042	0.045	0.054	0.040	0.045	0.043	0.029	0.029	0.034	0.038
		0.028	0.020	0.019	0.020	0.042	0.021	0.020	0.018	0.021	0.016	0.015	0.015
	int	0.078	0.036	0.051	0.115	0.073	0.040	0.059	0.097	0.049	0.029	0.040	0.083
		0.051	0.020	0.019	0.044	0.055	0.021	0.021	0.040	0.033	0.015	0.016	0.046
$\hat{m}$	back	0.046	0.047	0.050	0.046	0.061	0.050	0.052	0.043	0.037	0.038	0.041	0.038
		0.036	0.025	0.026	0.024	0.049	0.027	0.027	0.022	0.028	0.021	0.021	0.019
	int	0.094	0.048	0.069	0.144	0.088	0.052	0.077	0.126	0.061	0.041	0.054	0.092
		0.066	0.027	0.030	0.067	0.067	0.029	0.033	0.064	0.044	0.023	0.025	0.053
Distrib.		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
Model		$m = m_2 + m_3$				$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	back	0.041	0.041	0.043	0.056	0.040	0.042	0.044	0.056	0.057	0.044	0.047	0.052
		0.030	0.020	0.020	0.025	0.030	0.021	0.020	0.025	0.043	0.022	0.021	0.023
	int	0.054	0.043	0.047	0.025	0.042	0.043	0.035	0.025	0.061	0.047	0.038	0.024
		0.039	0.020	0.022	0.012	0.031	0.019	0.016	0.012	0.046	0.021	0.017	0.012
$\hat{m}_\beta$	back	0.053	0.037	0.035	0.041	0.027	0.026	0.026	0.035	0.027	0.026	0.026	0.036
		0.040	0.017	0.016	0.017	0.018	0.012	0.011	0.014	0.018	0.012	0.011	0.014
	int	0.073	0.040	0.059	0.026	0.043	0.030	0.026	0.022	0.045	0.030	0.026	0.022
		0.055	0.021	0.021	0.011	0.030	0.015	0.012	0.010	0.031	0.015	0.012	0.010
$\hat{m}$	back	0.081	0.069	0.071	0.069	0.057	0.058	0.061	0.064	0.073	0.061	0.063	0.061
		0.065	0.038	0.039	0.036	0.044	0.032	0.032	0.033	0.058	0.034	0.033	0.031
	int	0.127	0.080	0.102	0.063	0.082	0.069	0.061	0.058	0.105	0.073	0.070	0.056
		0.098	0.043	0.048	0.026	0.064	0.035	0.029	0.024	0.082	0.037	0.030	0.023

estimator is inefficient for correlated designs, see Linton (1997). He suggested an estimator for additive models constructed as a mixture of marginal integration and one-iteration-backfit and proved that for correlated designs this procedure dominates asymptotically the integration method in its variance part.

We want to emphasize our statements by looking closer to the behavior of squared bias, variance and MSE over the range.

The main difference from the results for using the Nadaraya Watson smoother is that the local linear smoother improves the integration estimator more than the backfitting estimator. The effects concerning the distribution of X and model structure are, as we expected, quite similar.

The following figures illustrate the behavior and the trade-off of variance and bias for both estimators in each additive direction. They are plotted on the range of the support. The boundaries of the data are cut off at a level of 5% each side, since otherwise their effects would dominate the pictures. The figures 7–11 reinforce clearly our observations and remarks concerning the tables 3–5.

The *integration estimator* suffers more from sparseness of observations than the backfitting estimator does. In what follows the boundary effects are worse in the integration estimator and it does better for the normal distribution than for the uniform considering the MASE. At the mass of observations this estimator mostly has lower squared bias and variance for the estimators of the additive functions. Finally, we observe that an increasing ρ (covariance of the explanatory vector) affects strongly its MASE in a negative sense.

The *backfitting estimator* is less affected by boundary effects or correlation of the explanatory variables. For the regression estimator it fits the regression in general better than the integration estimator, at least the MASE is almost always smaller. But it pays for a low MSE (or MASE) for the regression with high MSE (MASE respectively) in the additive function estimation. Here we see the main difference of these estimators; the integration estimator is estimating the additive function by integrating out the directions not of interest, which means it is measuring the marginal influence of the considered input, whereas the backfitting estimator is looking in the space of additive models for the best fit of the response Y vs. X. For a more detailed discussion about their different interpretation, see Nielsen and Linton (1997). An increase in the correlation ρ of the design again leads to a worse estimate.

Making Use of the Bandwidth Matrix For the purpose of correcting for correlation between the components of X we furthermore refined the estimation procedure by replacing the bandwidth vector h by its multivariate counterpart, a nonsingular bandwidth matrix H. This leads to the following multivariate kernel function:

$$K_H(u) = \frac{1}{|H|} K(H^{-1}u)$$

Motivated by the bandwidth matrix selection in the book of Wand and Jones (1995), the matrix H is constructed in the following way: Its diagonal elements are equivalent to the

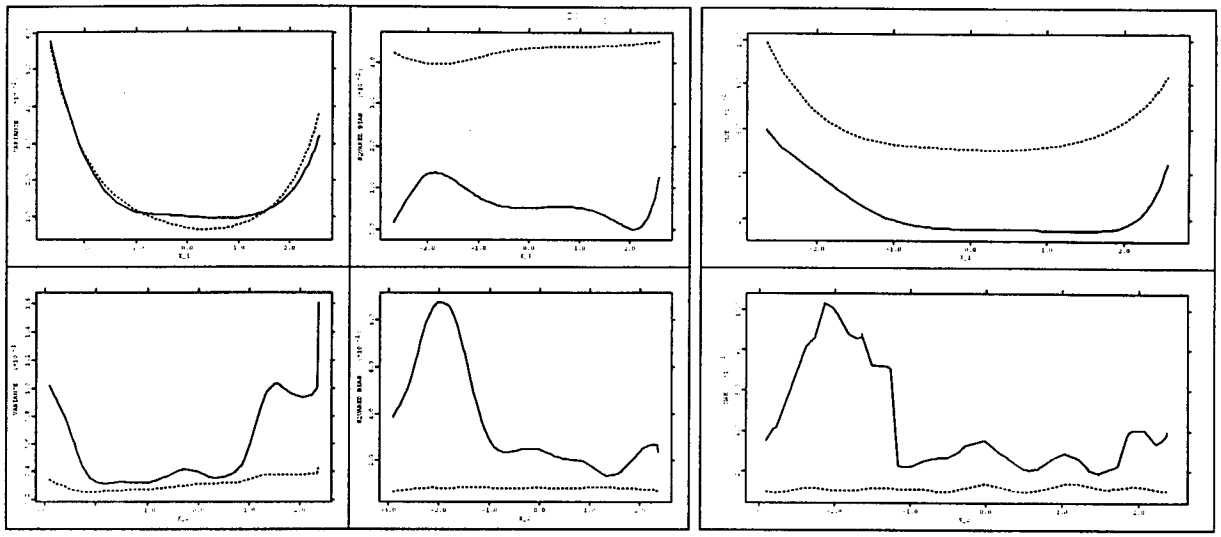

Figure 7: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_1 + m_2$ for m_1 (top), m_2(bottom) separately. Uniform design on $[-3,3]^2$, using local linear smoother.

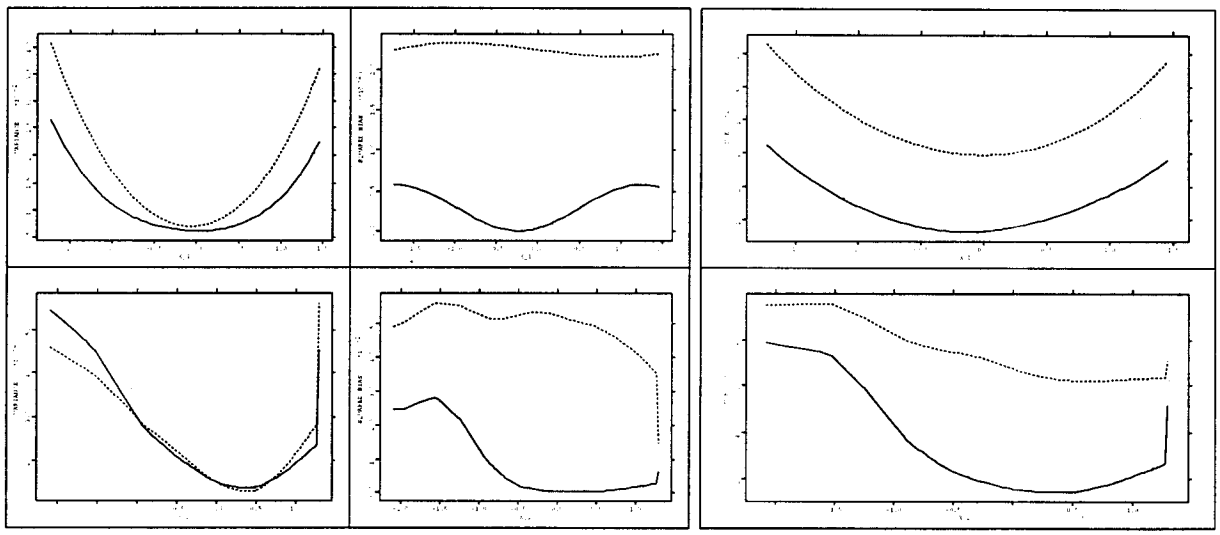

Figure 8: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_1 + m_2$, for m_1(top), m_2(bottom) separately. Standard normal design with cov= 0.0, using local linear smoother.

elements of the bandwidth vector h and its off-diagonal elements can be derived from the covariance matrix. In addition we included a factor δ which allows us to control the influence of the off-diagonal elements. Hence, for the two dimensional model one gets:

$$H = \begin{pmatrix} h_1 & \delta\rho \\ \delta\rho & h_2 \end{pmatrix}$$

Note that for $\delta = 0$ we would get the results of the previously applied estimation procedure. This can be checked from the tables of section 3.4. Defining the bandwidth matrix

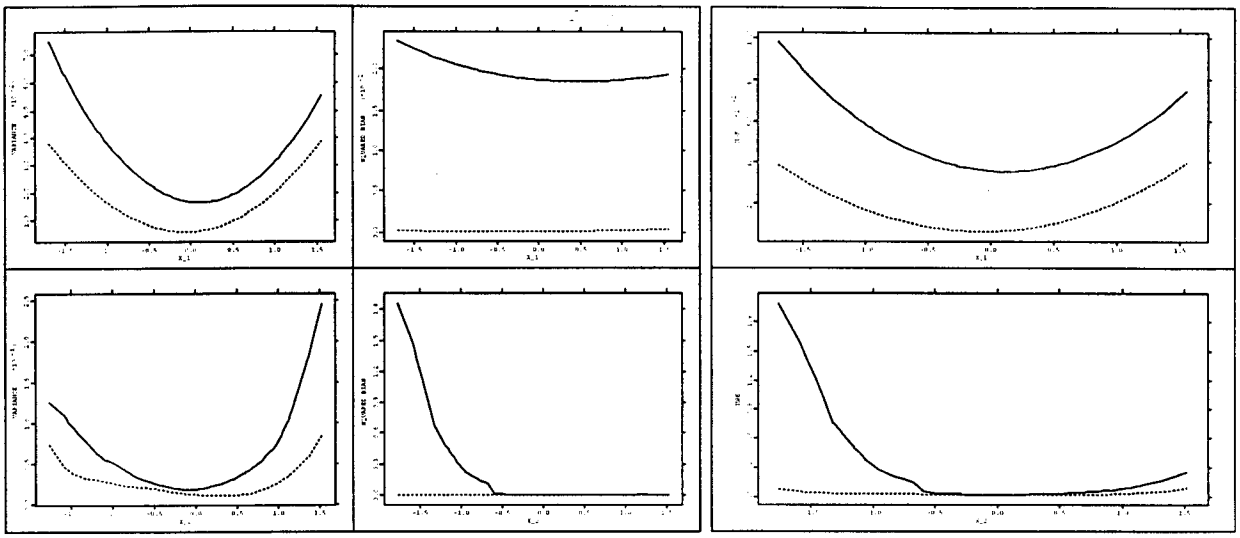

Figure 9: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_1 + m_2$, for m_1(top), m_2(bottom) separately. Standard normal design with cov= 0.8 , using local linear smoother.

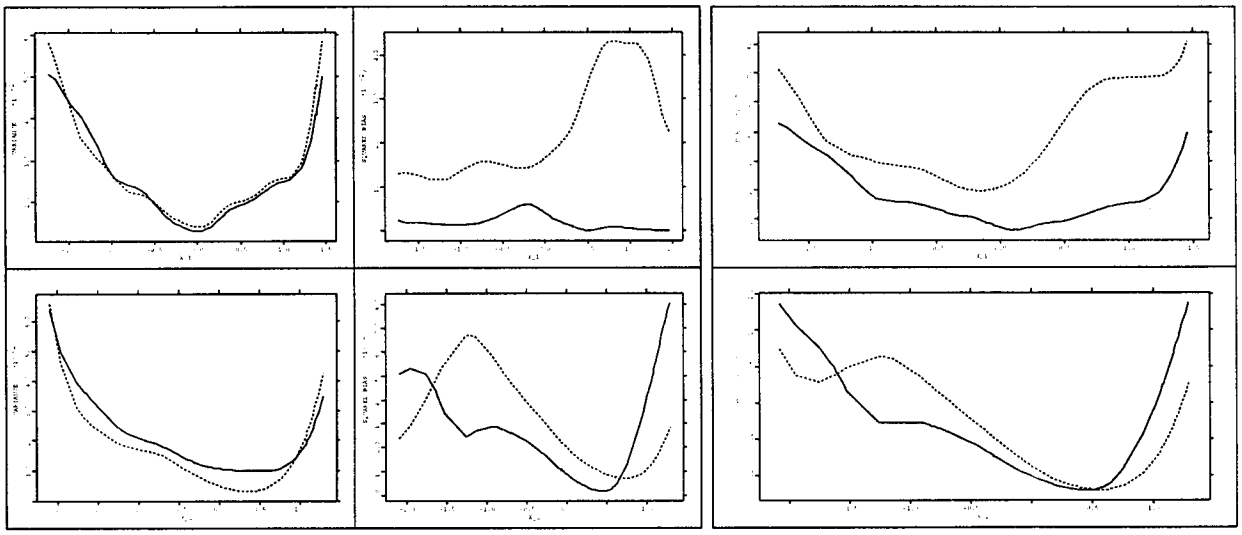

Figure 10: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_3 + m_4$, for m_3(top), m_4(bottom) separately. Standard normal design with cov= 0.0 , using local linear smoother.

in this way we were able to run the estimation on a grid for δ. The table 6 presents our results for the standard normal design for $cov(x_1, x_2) = 0.4$ and $cov(x_1, x_2) = 0.8$. They show MASE together with the value of δ, by which it is minimized, for trimmed data in brackets. To compare the results with the former one we present them in the following table together with the MASE which we got for diagonal bandwidth matrices in the integration procedure. Obviously the fit can be improved significantly if we use a proper off–diagonal element in the bandwidth matrices. Since interpretation using a nondiagonal bandwidth matrix is different for the backfitting, due to its iterative charac-

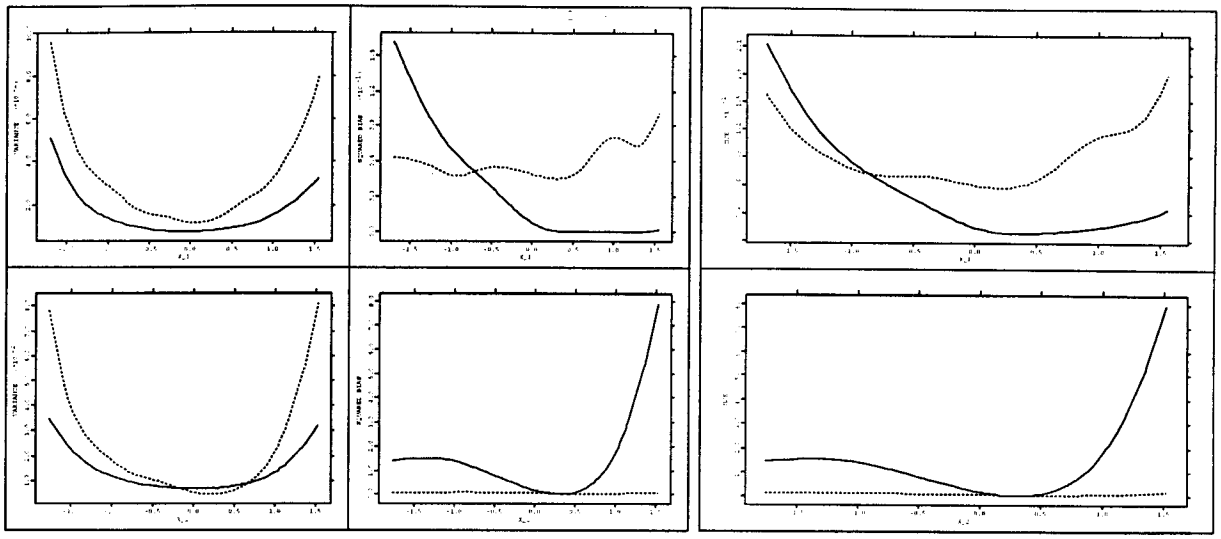

Figure 11: Variance (left), bias^2 (middle) and MASE (right) by bandwidth h_1 in model $m = m_3 + m_4$, for m_3(top), m_4(bottom) separately. Standard normal design with cov= 0.0 , using local linear smoother.

ter and smoothing always in univariate subspaces, we skipped this investigation for the backfitting. However, such an investigation in theory and practice would be interesting for that method, too.

Using the local linear smoother, the optimal bandwidth matrix to estimate the linear function m_1 should be huge or even infinite on the first diagonal, but we cannot ameliorate the results by changing the off–diagonals. For that reason we skipped models that include m_1.

3.5 Singular and Eigenvalue Analysis

Eigendecomposition of the smoother matrix of an estimator can be used to describe the behavior of the smoother, especially when this matrix is symmetric and thus the eigenvalues are real. In that case this is much like the use of a *transfer function* to describe a linear filter for time series. This connection is made precise in Hastie and Tibshirani (1990). If the smoother matrix is not symmetric we have to turn to the singular value analysis since a eigendecomposition often would lead to complex eigenvalues.

In the following we present the first respectively the biggest singular values of the weight matrices. These smoothing matrices are symmetric for the backfitting, using local polynomial kernel smoothing but they are not symmetric for the integration estimator. Thus we did a singular value analysis for the integration procedure.

In figure 12 to 15 we give the calculated values. Again in all figures the lines for the integration estimator are solid, for the backfitting estimator dotted. For each design

TABLE 6: PERFORMANCE (MASE) WITH VS WITHOUT OFF–DIAGONALS δ_{min} IN THE BANDWIDTH MATRIX USING LOC. LIN. SMOOTHER

$cov(x_1, x_2)$	Model	m_j	δ_{min}		$MASE^{int.}_{with}$		$MASE^{int.}_{old}$	
	$m = m_2 + m_3$	m_2	0.9	(0.7)	0.090	(0.040)	0.105	(0.064)
		m_3	0.5	(0.6)	0.144	(0.030)	0.191	(0.055)
0.4	$m = m_2 + m_4$	m_2	1.5	(0.6)	0.045	(0.022)	0.048	(0.026)
		m_4	0.0	(0.1)	0.480	(0.061)	0.480	(0.061)
	$m = m_3 + m_4$	m_3	0.2	(0.2)	0.049	(0.022)	0.049	(0.022)
		m_4	1.5	(1.5)	0.590	(0.217)	0.603	(0.265)
	$m = m_2 + m_3$	m_2	0.6	(0.5)	0.244	(0.167)	3.490	(0.782)
		m_3	0.5	(1.5)	0.246	(0.038)	1.499	(0.186)
0.8	$m = m_2 + m_4$	m_2	0.4	(1.1)	0.079	(0.045)	0.089	(0.078)
		m_4	1.5	(0.8)	1.463	(0.194)	1.322	(0.151)
	$m = m_3 + m_4$	m_3	0.1	(0.4)	0.074	(0.052)	0.056	(0.041)
		m_4	1.5	(1.5)	2.385	(1.008)	2.413	(1.020)

distribution presented we give the results for two randomly drawn samples.

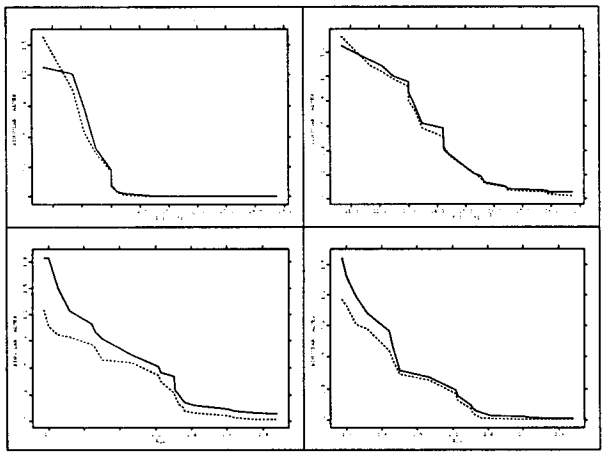

Figure 12: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1(top), x_2(bottom) vs eigen/singular values for two uniformly distributed samples (left, right).

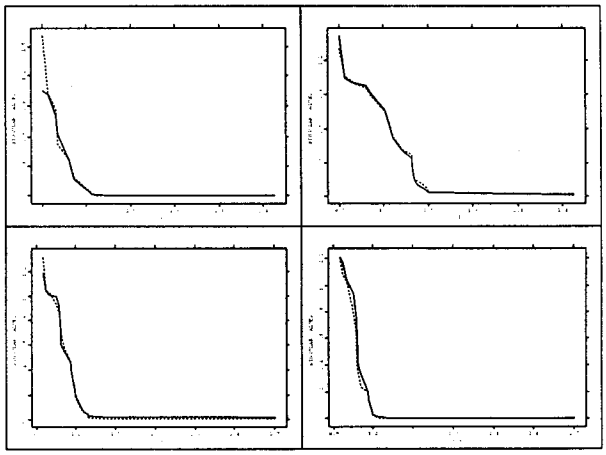

Figure 13: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1(top), x_2(bottom) vs eigen/singular values for two normal (cov= 0.0) distributed samples (left, right).

The slope of the eigen or singular values, see figure 12 to 15, gives us an idea of the smoothness of the specific estimator. They almost always cross, often the backfitting eigenvalue is a little bit steeper, what depends on the bandwidth choice, but there seems to be no remarkable difference between the integration and the backfitting method regarding

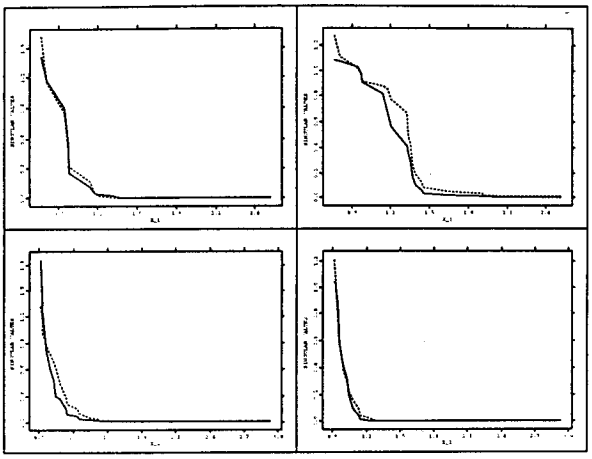

Figure 14: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1(top), x_2(bottom) vs eigen/singular values for two normal (cov= 0.4) distributed samples (left, right).

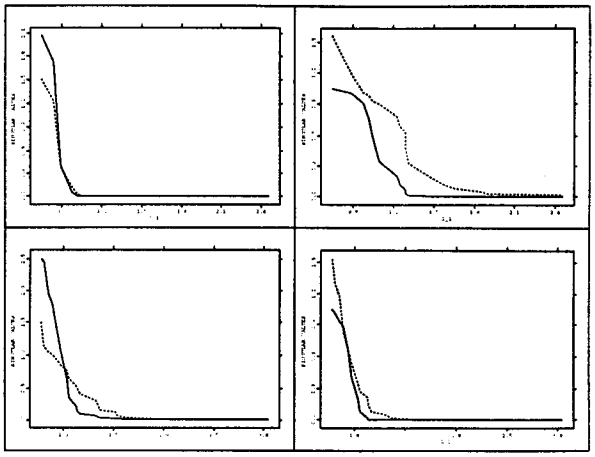

Figure 15: Eigen-/Singular value analysis using local linear smoother. Plotted are x_1(top), x_2(bottom) vs eigen/singular values for two normal (cov= 0.8) distributed samples (left, right).

the eigenvalue analysis.

3.6 Degrees of Freedom

Another parameter we looked at is the degree of freedom of the smoothers. Hastie and Tibshirani (1990) give various interpretations for degrees of freedom in the context of nonparametric estimation as well as for testing nonparametrically. One of them is that they give us the amount of fitting. Further they can be used to approximate the distribution of test statistics. They also state that we can draw out of them some information about the smoothness of the estimator. So they propose for a fair comparison of different estimators to choose those smoothing parameters that give equal degrees of freedom for the different estimators. Our experience was that this leads to unreasonable bandwidths. So we have to doubt these interpretations at least for the integration estimator.

For all smoothing matrices we calculated the values for three different definitions of degrees of freedom, $\mathrm{tr}(W)$, $\mathrm{tr}(WW^T)$ and $n-\mathrm{tr}(2W-WW^T)$, but restrict ourselves in presenting only $\mathrm{tr}(W)$. The other results can be requested.

As already mentioned at the beginning of this paragraph the chosen asymptotically 'optimal' bandwidth led us to totally different degrees of freedom as defined above.

Looking at table 7, where the degrees are defined as the trace of W, we see that the degrees of freedom for the backfitting are almost always bigger than the degrees for the integration estimator. For both estimators the degrees are bigger in the case of normal

TABLE 7: DEGREES OF FREEDOM MEASURED BY $trace(W)$, USING LOC. LIN.

Distribution:		U^2	$N(.0)$	$N(.4)$	$N(.8)$	U^2	$N(.0)$	$N(.4)$	$N(.8)$
$m = m_\alpha + m_\beta$		$m = m_1 + m_2$				$m = m_1 + m_3$			
$\hat{m}_\alpha$	back.	3.63	4.19	4.01	2.43	3.60	4.19	4.00	2.43
	int.	3.52	3.96	3.31	2.11	3.64	3.61	3.29	2.11
$\hat{m}_\beta$	back.	8.42	7.54	8.30	8.86	12.70	8.15	8.87	8.19
	int.	8.53	8.00	6.66	3.84	12.99	8.65	7.39	3.52
$\hat{m}$	back.	13.05	12.73	13.30	12.29	17.30	13.33	13.87	11.62
	int.	11.05	10.96	8.97	4.95	15.64	11.26	9.68	4.63
$m = m_\alpha + m_\beta$		$m = m_1 + m_4$				$m = m_2 + m_3$			
$\hat{m}_\alpha$	back.	3.65	4.22	4.04	2.45	9.16	9.32	9.52	10.37
	int.	3.71	3.85	3.86	2.24	8.37	8.78	7.21	6.09
$\hat{m}_\beta$	back.	5.88	5.38	6.17	6.75	12.49	8.04	8.05	6.88
	int.	6.17	5.81	4.58	2.48	13.00	8.69	7.36	6.15
$\hat{m}$	back.	10.53	10.59	11.20	10.19	22.64	18.37	18.56	18.25
	int.	8.88	8.66	7.44	3.72	20.37	16.46	13.58	11.24
$m = m_\alpha + m_\beta$		$m = m_2 + m_4$				$m = m_3 + m_4$			
$\hat{m}_\alpha$	back.	9.33	9.37	9.61	10.41	13.73	10.02	10.26	9.51
	int.	9.34	8.78	8.34	6.09	13.80	9.23	9.24	5.69
$\hat{m}_\beta$	back.	5.78	5.31	5.49	5.49	5.70	5.31	5.39	5.67
	int.	6.33	5.73	6.33	5.22	6.25	5.79	6.33	5.22
$\hat{m}$	back.	16.10	15.69	16.10	16.89	20.43	16.32	16.65	16.18
	int.	14.67	13.52	13.66	10.31	19.05	14.02	14.56	9.916

distributed designs but it is hardly possible to detect a systematic difference in the degrees for the increasing correlation of the explanatory variables. What can be seen clearly is that the degrees of freedom are varying strongly with the choice of the model. This holds true for both estimators.

Note that the degrees of the function m in the integration method is the result of summing the degrees of its additive components minus one, as a result of eliminating in each estimation the sample mean. In the backfitting you take the sum of the degrees of the additive components and add one, see Opsomer and Ruppert (1997).

When we considered $\text{tr}(WW^T)$, this is certainly different. Further, in the local linear case, considering $\text{tr}(WW^T)$ led to different results at all. Here now the degrees were often much bigger for the integration method. However, since interpretation is hardly possible in that case, we skipped the presentation of these results.

3.7 The Equivalent Kernel Weights of the Estimators

What price do we pay to overcome the curse of dimensionality by choosing an additive model structure? To examine this we compared the two additive model estimators, backfitting and integration procedure, with the bivariate Nadaraya Watson kernel smoother. Equivalent kernels are defined as the linear weights w of the estimates to fit the regression function at a particular point, in our case at $(0, 0)$. For the integration estimator we used only a diagonal bandwidth matrix as in the beginning, even for the strongly correlated designs. We have considered $n = 100$, bivariate normal distributed designs with mean zero, variance 1 and increasing correlation $\rho = 0.0,\ 0.2,\ 0.4,\ 0.6$ and 0.8, but give only figures for $0.0,\ 0.4$ and 0.8. Please note that equivalent kernel weights depend only on the kernel function, the bandwidths and X but not on Y. So the results in figure 16 to 24 presented hold for any underlying two dimensional model.

Since the local linear smoother is also taking into account the first derivative of the functions, we would get, depending on the data generating functions, positive and negative weights varying from point to point. Thus for the local linear smoother the pictures shown beneath would look like wild mountain scenery and so we skipped their presentation.

As we would have expected, both additive model estimators get their strength from the local panels orthogonal to the axes of X_1 and X_2 instead of uniformly in all directions like the bivariate Nadaraya Watson smoother. Since they are composed by components that behave like univariate smoothers, they can overcome the curse of dimensionality. For the backfitting this was already stated by Hastie and Tibshirani (1990). We can see clearly now that the integration estimator behaves very similar. The pictures for the additive smoothers look almost the same, except that the backfitting can also get some negative weights whereas the integration estimator cannot by its construction.

Both estimators run into deep problems to estimate properly in designs with increasing correlation. In contrast to the bivariate Nadaraya Watson smoother this can be seen in the figures for the backfitting as well as for the integration method. But we are not able to discover visually the reason why the integration estimator is doing worse for highly correlated explanatory variables than e.g. the backfitting.

3.8 Do the Asymptotics hold empirically ?

For restricting our presentation on $n = 100$ observations we had mainly two reasons. First, in our simulations we had the same findings also for n different from 100, what is indicated also in this section, see below. Second, for $n > 100$ the difference between integration and backfitting method decreases in such an amount that it even would be

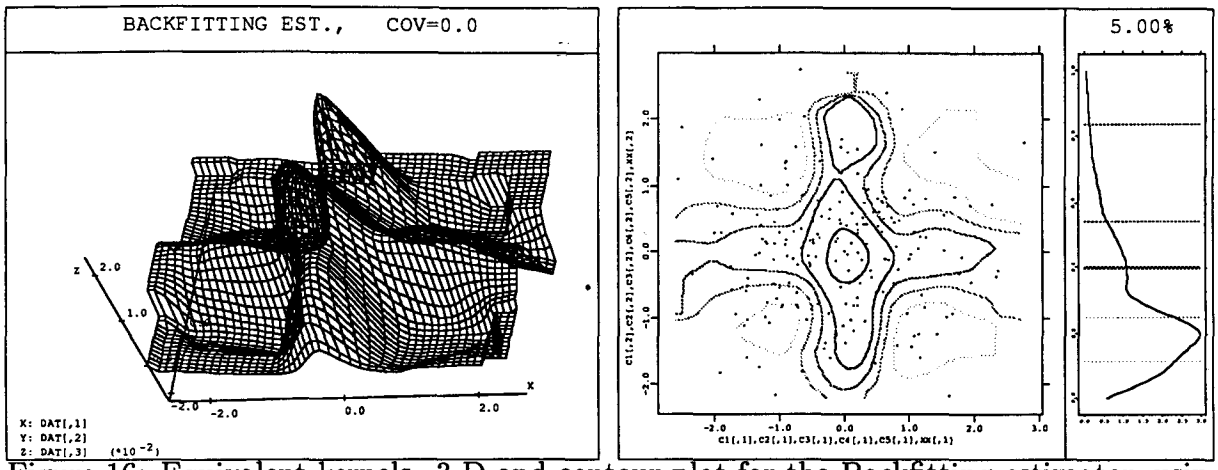

Figure 16: Equivalent kernels. 3-D and contour plot for the Backfitting estimator, using Nadaraya Watson. Regressors are standart normal with $cov = 0.0$.

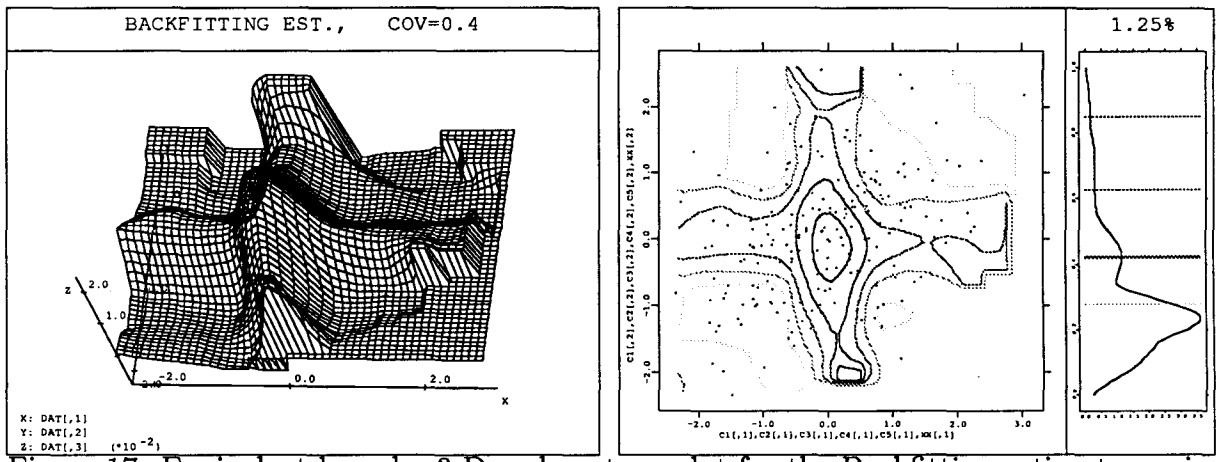

Figure 17: Equivalent kernels. 3-D and contour plot for the Backfitting estimator, using Nadaraya Watson. Regressors are standart normal with $cov = 0.4$.

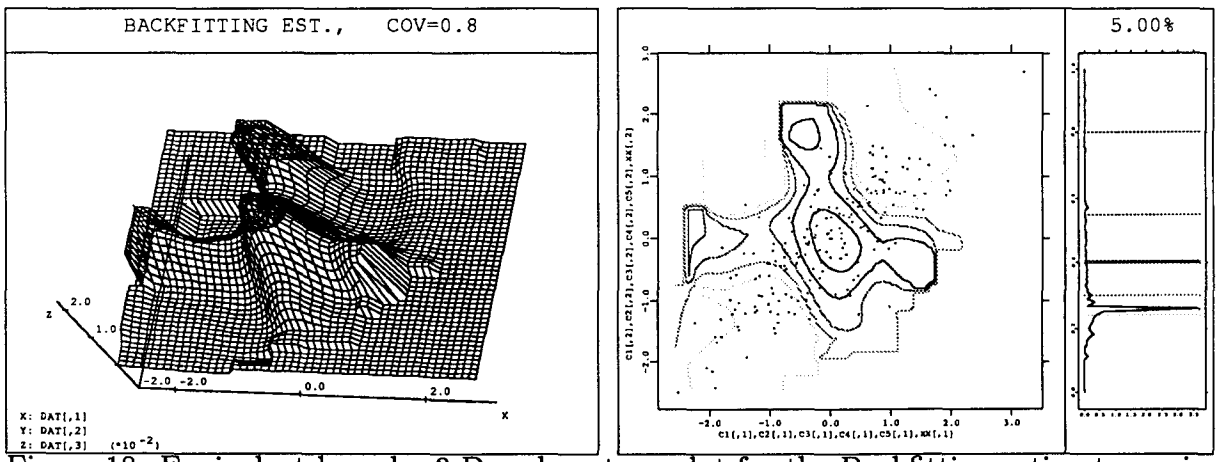

Figure 18: Equivalent kernels. 3-D and contour plot for the Backfitting estimator, using Nadaraya Watson. Regressors are standart normal with $cov = 0.8$.

hard to illustrate them at all. To answer the question about the asymptotics, we did a simulation study, using the local linear smoother, as follows.

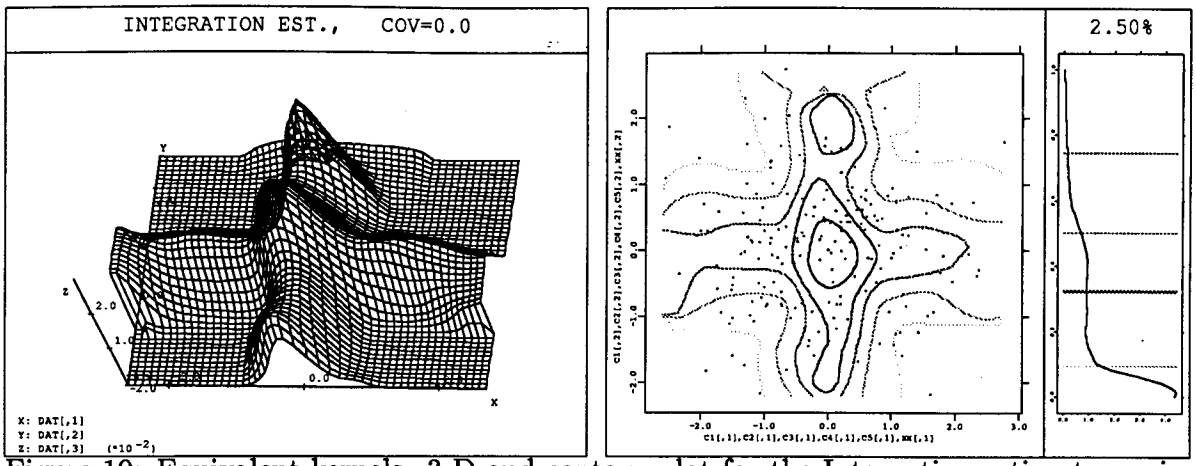

Figure 19: Equivalent kernels. 3-D and contour plot for the Integration estimator, using Nadaraya Watson. Regressors are standart normal with $cov = 0.0$.

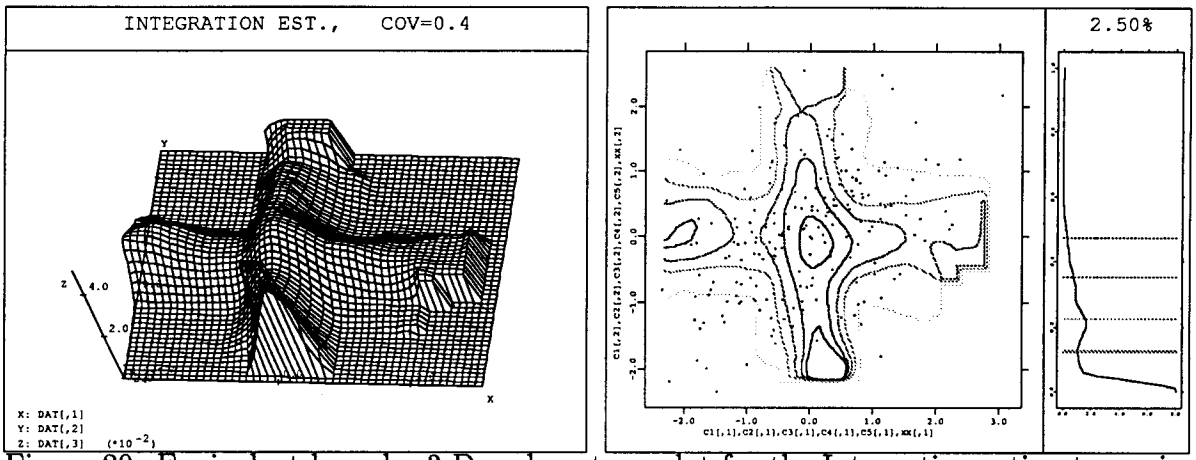

Figure 20: Equivalent kernels. 3-D and contour plot for the Integration estimator, using Nadaraya Watson. Regressors are standart normal with $cov = 0.4$.

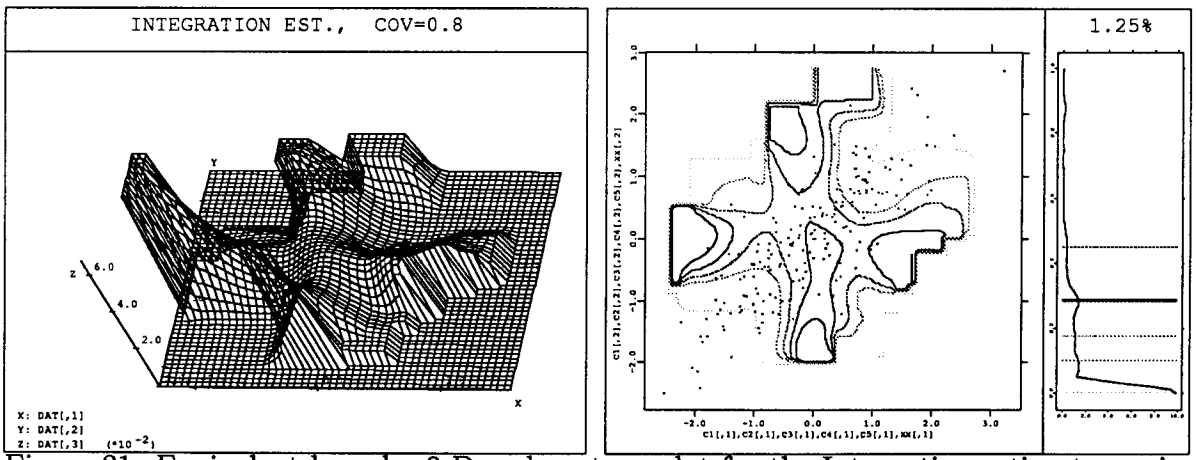

Figure 21: Equivalent kernels. 3-D and contour plot for the Integration estimator, using Nadaraya Watson. Regressors are standart normal with $cov = 0.8$.

We considered the model

$$Y = c + m_1(x_1) + m_2(x_2) + \varepsilon$$

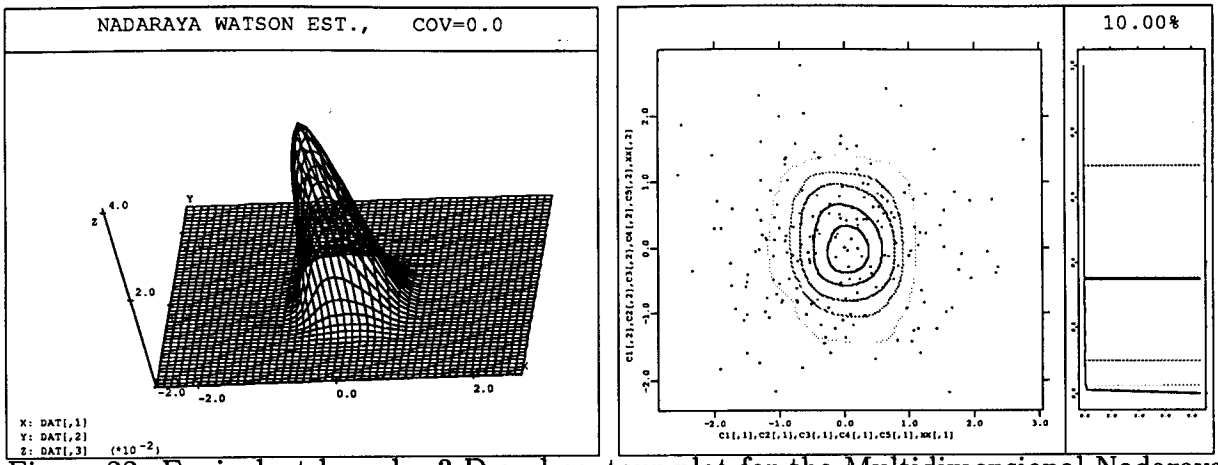

Figure 22: Equivalent kernels. 3-D and contour plot for the Multidimensional Nadaraya Watson estimator. Regressors are standart normal with $cov = 0.0$.

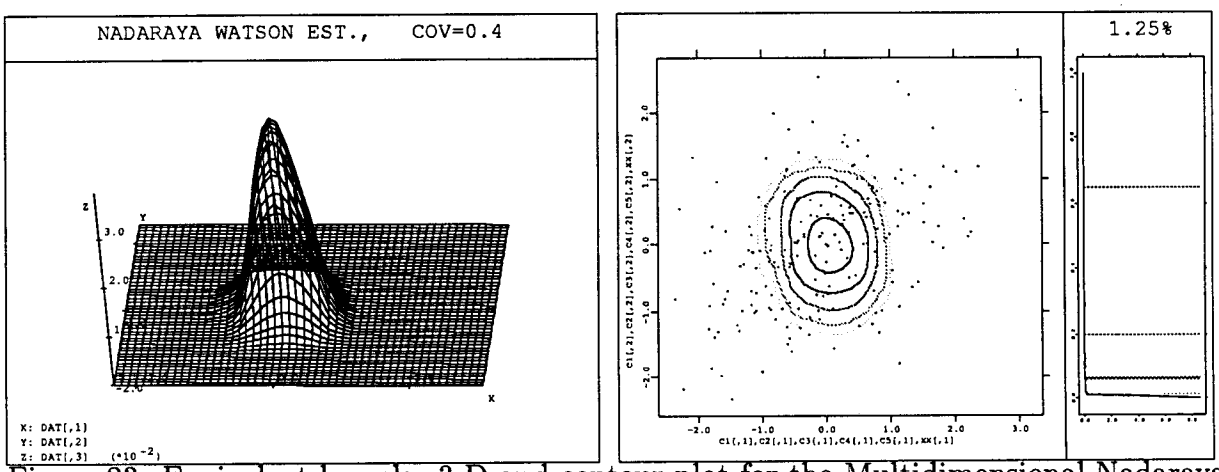

Figure 23: Equivalent kernels. 3-D and contour plot for the Multidimensional Nadaraya Watson estimator. Regressors are standart normal with $cov = 0.4$.

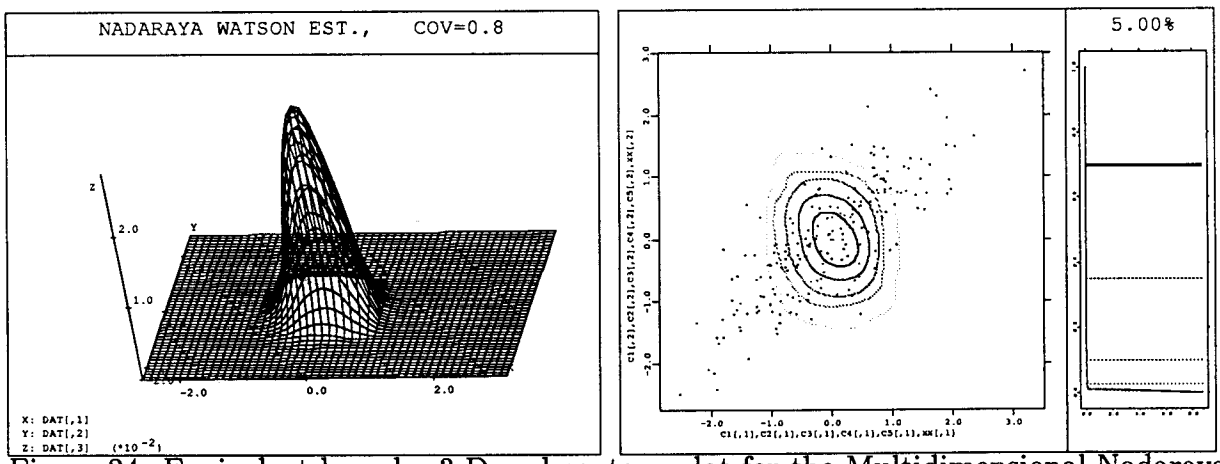

Figure 24: Equivalent kernels. 3-D and contour plot for the Multidimensional Nadaraya Watson estimator. Regressors are standart normal with $cov = 0.8$.

with $m_1(x) = 2x$, $m_2(x) = x^2 - E(x^2)$ and $c = 0$. The error term ε has been normal distributed with mean zero and variance 0.5, the design X was uniform on $[-3, 3]^2$ dis-

tributed. For $n = 250, 500, 1000$ and 2000 observations we calculated the estimates $\widehat{m}_1$, $\widehat{m}_2$ at $x = -1.5, -0.75, 0.0, 0.75$ and 1.5 and determined their biases B (which is always mentioned as $b \cdot h^2$ in theory) and variances V for each n. The bandwidths have been $h_1 := h_n = h_0 n^{-1/5}$ with $h_0 \approx 0.69$ and $h_2 := g_n = 2h_n/3$ for the nuisance direction.

Our first question was whether the rate of convergence mentioned by the theory holds also empirically. Therefore we considered the following regression

$$\begin{aligned} \ln(V) &= \beta_1 \ln(nh_n) \\ \ln(B) &= \beta_2 \ln(nh_n). \end{aligned}$$

For the integration estimator we got for all five points $\beta_1 \approx \beta_2 \approx -1.003$, for the backfitting $\beta_1 \approx -1.02$ and $\beta_2 \approx -1.05$ which ensures the theory concerning the rate of convergence.

The second question we were interested in was the comparison of the empirical biases and variances calculated in our simulation study with the analytical ones. We present results only for the function m_2 in the above mentioned setting, but have to remark that the biases certainly depend on the particular data generating model as well as on the chosen design, at least in practice. To consider the function m_1 that is linear in this model is useless since we know that a local linear estimator is fitting such a function almost always exactly by definition and thus this would not be typical in practice. For the comparison see table 8 for the analytical values and table 9, 10 for the empirical values.

TABLE 8: ANALYTICAL BIAS AND VARIANCE

n	250	500	1000	2000
variance (equal for all points)	0.0147	0.0085	0.0048	0.0028
bias (equal for all points)	0.0529	0.0400	0.0306	0.0225

As we can see the estimator is doing very well for an increasing number of observations and at least for a low dimensional model the integration estimator obviously reaches his asymptotics pretty fast.

Since we could not calculate (in GAUSS) with weight matrices for the backfitting procedure when n was ≥ 1000, we had to determine the empirical bias and variance by doing 400 replications for huge n and did the regression described above separately for 250 and 500, respectively for 1000 and 2000. We can conclude from β_1 and β_2 that bias and variance also diminish almost in the theoretical one dimensional rate. Obviously the constant h_0 of the bandwidth is chosen too big here, as can be seen in table 19. The variance calculated with the aid of the weight matrices is smaller than expected whereas the bias is much bigger.

TABLE 9: SMALL SAMPLE BIAS AND VARIANCE FOR INTEGRATION ESTIMATOR

n		250	500	1000	2000
	-1.5	0.01897	0.00986	0.00535	0.00305
	-0.75	0.01813	0.00999	0.00536	0.00303
variance at	$+0.0$	0.01825	0.01019	0.00548	0.00306
	$+0.75$	0.01807	0.00996	0.00535	0.00301
	$+1.5$	0.01765	0.00978	0.00546	0.00309
	-1.5	0.06237	0.04866	0.03206	0.02552
	-0.75	0.06681	0.04932	0.03196	0.02509
bias at	$+0.0$	0.05892	0.04705	0.03303	0.02567
	$+0.75$	0.06410	0.04853	0.03139	0.02453
	$+1.5$	0.07067	0.05071	0.03313	0.02471

TABLE 10: SMALL SAMPLE BIAS AND VARIANCE FOR BACKFITTING ESTIMATOR

n		250	500	1000	2000
	-1.5	0.01431	0.00793	0.01503	0.00619
	-0.75	0.01411	0.00798	0.01231	0.00684
variance at	$+0.0$	0.01404	0.00801	0.01509	0.00755
	$+0.75$	0.01409	0.00796	0.01073	0.00528
	$+1.5$	0.01391	0.00791	0.01417	0.00684
	-1.5	0.31041	0.22552	0.16990	0.11953
	-0.75	0.30895	0.22489	0.16621	0.12171
bias at	$+0.0$	0.31176	0.22576	0.17181	0.13314
	$+0.75$	0.31097	0.22529	0.17411	0.13399
	$+1.5$	0.31137	0.22625	0.18929	0.12787

Since in this subsection we were not interested in the direct comparison of the MSE or something similar for backfitting and integration method, we did not look for an optimal bandwidth in each direction neither for each method. So one should only look on the tables respectively the asymptotic behavior of the estimates, but not for a comparison of the absolute values.

3.9 In higher dimensions

Due to the excess of information in this paper we only present results for $d = 4$, $n = 500$. Other simulations we did result in the same statements made for this special case. Here we did 100 replications and calculated bias and variance empirically by doing 400 replications. We took the analytically optimal bandwidth for the estimation of the additive functions, compare our discussion at the very beginning of our simulation study. The additive functions in our model have been

$$\begin{aligned} m_1(x) &= 2x \quad , \quad m_2(x) = x^2 - E(x^2) \ , \\ m_3(x) &= \exp(x) - E\{\exp(x)\} \quad \text{and} \quad m_4(x) = 0.5 \cdot \sin(-1.5x) \ . \end{aligned}$$

Some final results are presented in table 11 together with the bandwidth we used. The bandwidth for the directions not of interest in the integration estimator has been chosen as 0.45. The trends already discovered in the simpler cases were enforced in that study. The regression function itself is estimated well by backfitting whereas the marginal influences of the explanatory variables sometimes are better estimated by the integration estimator. Since the integration estimator suffers much more from boundary effects and data sparseness, what is especially the case in higher dimensions, the average mean squared error looks quite often worse. This concerns mainly the simulation example where the design is normal distributed.

TABLE 11: MASE IN HIGHER DIMENSIONS ($d = 4$)
FOR ADDITIVE COMPONENTS AND REGRESSION FUNCTION

Estimated m_j :		$\hat{m}_1$		$\hat{m}_2$		$\hat{m}_3$		$\hat{m}_4$		$\hat{m}$	
Distribution		U^2	N(0.0)	U^2	N(0.0)	U^2	N(0.0)	U^2	N(0.0)	U^2	N(0.0)
h_1		20	20	0.212	0.211	0.138	0.194	0.309	0.307		
MASE	back.	0.051	0.018	0.180	0.159	0.078	0.036	0.074	0.075	0.028	0.024
	int.	0.041	0.056	0.100	0.037	0.156	0.057	0.135	0.106	0.250	0.540

4 Conclusion

A common misunderstanding of the integration method is that it must inherit the poor properties of the high dimensional regression estimator. Of course, this is absurd. It amounts to saying that the sample mean must behave poorly because the individual observations from which it is constructed are inconsistent estimates of the mean themselves. In any event, we have not found this to be the case. In fact, we have found many similarities between the integration and backfitting methodologies in terms of what they do

to the data (for example the eigenanalysis) and indeed their statistical performance. In particular, both integration and backfitting suffer some small sample cost. The backfitting method seems to work better at boundary points and when there is high correlation among the covariates, while the integration method works better in most of the other cases and especially in estimating the components as opposed to the function itself.

ACKNOWLEDGEMENTS

We would like to thank R.J. Carroll, J. Horowitz, J.P. Nielsen, W. Härdle and two anonymous referees for helpful comments. We thank the National Science Foundation, NATO, and the Deutsche Forschungsgemeinschaft, SFB 373.

References:

Breiman, L. and J.H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580-619.

Buja, A., T. Hastie and R. Tibshirani (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics*, **17** , 453-555.

Deaton, A. and J. Muellbauer (1980). *Economics and Consumer Behavior.* Cambridge University Press: Cambridge.

Härdle, W., and P. Hall (1993). On the backfitting algorithm for additive regression models. *Statistica Neederlandica*, **47**, 43-57.

Härdle, W., and O.B. Linton (1994). Applied nonparametric methods, *The Handbook of Econometrics, vol. IV,* (R.F. Engle and D.F. McFadden, eds.). Elsevier: Amsterdam, Ch.38.

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

Ibragimov, I.A. and R.Z. Hasminskii (1980). On nonparametric estimation of regression, *Soviet Math. Dokl.* **21**, 810-4.

Linton, O.B. (1997). Efficient Estimator. Forthcoming

Linton, O.B. , R. Chen, N. Wang, and W. Härdle (1995). An analysis of transformation for additive nonparametric regression. *Journal of the American Statistical Association*, **92**, 1512-1521.

Linton, O.B. and W. Härdle (1996). Estimation of additive regression models with known links. *Biometrika*, **83**, 529-540.

Linton, O.B., E. Mammen and J. Nielsen (1998). The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under weak conditions. Manuscript, Yale University.

Linton, O.B. and J.P. Nielsen (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93-100.

Masry, E. and D. Tjøstheim (1995). Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258-289.

Masry, E. and D. Tjøstheim (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214-252.

Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*. **5**, 99-135.

Newey, W.K. (1994). Kernel estimation of partial means. *Econometric Theory*, **10**, 233-253.

Nielsen, J.P. (1996). Multiplicative and additive marker dependent hazard estimation based on marginal integration. Manuscript, PFA Pension.

Nielsen, J.P. and O.B. Linton (1997). An optimization interpretation of integration and backfitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, Series B,* Forthcoming.

Opsomer, J.D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**, 212-243.

Porter, J. (1996). Essays in Semiparametric Econometrics. *PhD Thesis, MIT.*

Powell, J.L. (1994). Estimation in semiparametric models. *The Handbook of Econometrics, vol. IV*, (R.F. Engle and D.F. McFadden, eds.). Elsevier: Amsterdam, Ch.41.

Ruppert, D. and M.P. Wand (1995). Multivariate Locally Weighted Least Squares. *The Annals of Statistics*, **22**, 1346-1370.

Severance-Lossin, E. and S. Sperlich (1997). Estimation of Derivatives for Additive Separable Models. *Discussion Paper, SFB 373, Humboldt-University Berlin, Germany*

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

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **8**, 1040-1053.

Stone, C.J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 685-705.

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**, 592-606.

Tjøstheim, D. and B. Auestad (1994). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association*, **89**, 1398-1409.

Venables, W.N. and B. Ripley (1994). *Modern applied statistics with S-Plus.* Springer Verlag: New York.

Wand, M.P. and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall: London. Vol. 60 of Monographs on Statistics and Applied Probability.