A comparative study of monotone nonparametric kernel estimates

Holger Dette, Kay F. Pilz
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: holger.dette@ruhr-uni-bochum.de

March 24, 2004

Abstract

In this paper we present a detailed numerical comparison of three monotone nonparametric kernel regression estimates, which isotonize a nonparametric curve estimator. The first estimate is the classical smoothed isotone estimate of Brunk (1958). The second method has recently been proposed by Hall and Huang (2001) and modifies the weights of a commonly used kernel estimate such that the resulting estimate is monotone. The third estimate was recently proposed by Dette, Neumeyer and Pilz (2003) and combines density and regression estimation techniques to obtain a monotone curve estimate of the inverse of the isotone regression function. The three concepts are briefly reviewed and their finite sample properties are studied by means of a simulation study. Although all estimates are first order asymptotically equivalent (provided that the unknown regression function is isotone) some differences for moderate samples are observed.

AMS Subject Classification: 62G05, 62G20
Keywords and Phrases: isotonic regression, order restricted inference, Nadaraya-Watson estimator, local linear regression, monte carlo simulation

1 Introduction

An important problem in applied statistics is to study the influence of an explanatory variable on a response and numerous linear and nonlinear regression models have been proposed for this purpose in the literature [see e.g. Seber and Wild (1989)]. Because it is often difficult to specify a parametric form of a regression function, smoothing as a means of modeling nonlinear structure in data has become increasingly popular in numerous applications. In many cases monotone estimates of the regression function are required, because physical considerations suggest that the response is a monotone function of the explanatory variable. These include the analysis of dose-response curves in pharmakinetics, growth curves in biology and many specific practical problems discussed in the literature cited below. Since the early work of Brunk (1955) a vast amount of literature has been published on the problem of estimating a regression function m which is believed to be monotone [see Mukerjee (1988), Mammen (1991), Cheng and Lin (1981), Wright (1981), Friedman and Tibshirani (1984), Ramsay (1988), Kelly and Rice (1990), Mammen and Thomas-Agnam (1999), Mammen, Marron, Turlach and Wand (2001)].

As pointed out by Hall and Huang (2001) many of the constrained estimates proposed in the literature reduce the smoothness of the estimator with which they started by projecting an unconstrained curve estiamte onto a constrained subspace of regression functions or by using isotonic regression techniques. As a consequence the monotone estimates appear less smooth as the unconstrained estimates and often have jump discontinuities. For this reason Hall and Huang (2001) proposed a new method for monotonizing a general kernel type estimator, which modifies the weights of the estimator by the least possible amount such that the modified function is monotone. This approach uses only standard computing routines and software, it is applicable to most of the commonly used kernel estimates and is for these reasons particularly appealing to users of conventional kernel methods. Because the method involves tilting of the empirical distribution function by the least possible amount subject to the constraints being enforced it will be called tilting method throughout this paper. An alternative smooth estimate could be obtained by smoothing an isotonized regression estimate obtained from the PAVA (Pool-Adjacent-Violators-Algorithm), which provides a simple and computational efficient technique for the calculation of the isotone least squares estimate. This approach is first order asymptotically equivalent to isotonizing a standard kernel estimate [see Mammen (1991)] and yields for this reason smooth monotone increasing estimates for reasonable sample sizes. Recently Dette, Neumeyer and Pilz (2003) proposed to combine a density with a regression estimate to obtain a monotone estimate of the inverse regression function. The isotone regression estimate is finally obtained by reflecting this function at the line $y = x$ and the amount of smoothness can be controlled by the smoothing parameter in the density estimation step. This method does not require constrained optimization techniques and is therefore attractive to practitioners because of its computational simplicity. Moreover the approach yields strictly isotone estimates which are first order asymptotically equivalent to the unconstrained estimates and to the estimates obtained by the PAVA method. Throughout this paper the estimate of Dette, Neumeyer and Pilz (2003) will also be called monotone density regression estimate.

It is a purpose of the present paper to present a detailed comparative study of the finite sample properties of these three asymptotically first order equivalent kernel based estimates by means of a simulation study. In Section 2 we review some of the properties of the PAVA-method, the tilting approach of Hall and Huang (2001) and the density regression estimate of Dette, Neumeyer and Pilz (2003). Section 3 is devoted to a detailed numerical comparison of the finite sample properties of these estimates. It is demonstrated that all estimation techniques yield similar results which confirms their first order asymptotic equivalence. Moreover, the tilting method of Hall and Huang (2001) and the density regression approach usually yield estimates with a smaller mean squared error than the PAVA estimate. Finally, some conclusions and recommendations for monotone kernel regression estimation are presented in Section 4.

2 Monotone nonparametric regression estimates

Consider the nonparametric regression model

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\{(X_i, Y_i)\}_{i=1}^n$ is a bivariate sample of i.i.d. observations such that X_i has a positive two times continuously differentiable density f with compact support, say $[0,1]$. The variance function $\sigma : [0,1] \to \mathbb{R}^+$ and the regression function $m : [0,1] \to \mathbb{R}$ are also assumed to be two times continuously differentiable, respectively. Throughout this paper $\hat{m}$ denotes a nonparametric regression estimate, which will be specified in concrete situations if necessary. We will briefly review the different concepts to obtain monotone regression estimates, which will be considered in the simulation study in Section 3. For a more complete review we refer to Gijbels (2004).

2.1 Smoothing isotonized estimates

Suppose that $Y_{[1]}, \ldots, Y_{[n]}$ are the observations corresponding to the order statistic $X_{(1)} < \ldots < X_{(n)}$ of $\{X_i\}_{i=1}^n$, then Brunk (1958) proposed

$$Y_i^* = \max_{s \le i} \min_{t \ge i} \frac{1}{t-s+1} \sum_{j=s}^{t} Y_{[j]} \tag{2.2}$$

as an estimate for the regression function at the point $X_{(i)}$. The monotone estimate of the curve m is then obtained by linear interpolation. In order to obtain a smooth monotone curve several authors proposed to apply a smoothing procedure to this linear interpolation, which consists in the application of a nonparametric estimate to the "data" $(X_{(i)}, Y_i^*)$. Mukerjee (1988) used the Nadaraya-Watson estimate

$$\hat{m}_{IS}(x) = \frac{\sum_{i=1}^n K_r\left(\frac{X_{(i)}-x}{h_r}\right) Y_i^*}{\sum_{i=1}^n K_r\left(\frac{X_{(i)}-x}{h_r}\right)} \tag{2.3}$$

to obtain a smooth monotone estimate, where K_r denotes a symmetric kernel with existing second moment and h_r is a bandwidth converging to 0 with increasing sample size n. Mammen (1991) showed that this estimate is first order asymptotically equivalent to the estimate, which is obtained by projecting the Nadaraya-Watson estimate from the unconstrained data onto the space of all monotone functions. As a consequence the estimate $\hat{m}_{IS}$ is expected to be monotone for reasonable sample sizes. Moreover, it is demonstrated by means of a simulation study that the mean squared error of $\hat{m}_{IS}$ is usually smaller than the mse of the estimate obtained by isotonizing a smooth curve estimate. Mammen (1991) also showed that at points where the derivative of the regression function is positive the estimate $\hat{m}_{IS}$ is first order asymptotically equivalent to the unconstrained estimate. In general any smoothing procedure as local polynomials [see Wand and Jones (1995), Fan and Gijbels (1996)], other types of kernel estimators [see Müller and Gasser (1979)], smoothing splines or series estimators [see Eubank (1988)] can be applied in the second step of this procedure. Simple computational algorithms for the calculation of the estimate (2.2) can be found in Barlow, Bartholomew, Bremner and Brunk (1972). Throughout this paper the estimate $\hat{m}_{IS}$ will also be denoted as PAVA estimate (Pool-Adjacent-Violators-Algorithm).

2.2 A tilting method

Recently, Hall and Huang (2001) proposed to modify the weights of the commonly used kernel estimators such that the modification becomes (strictly) monotone. Roughly speaking the method proposed by these authors involes tilting of the empirical distribution function subject to the constraint being enforced. To be more precise assume that the unconstrained nonparametric estimate of the regression curve can be represented in the form

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} A_i(x)Y_i, \tag{2.4}$$

where the weights $A_i(x)$ depend only on the explanatory variables X_i but not on the responses Y_i. Note that the Priestley-Chao, the Nadaraya-Watson, the Gasser-Müller estimate and the local linear estimates as well as many of the modified forms of these types can be written in the form (2.4). Hall and Huang (2001) suggest to modify the estimate in (2.4) by replacing the weights $\frac{1}{n}$ with arbitrary weights p_i, i.e.

$$\hat{m}(x \mid p) = \sum_{i=1}^{n} p_i A_i(x)Y_i, \tag{2.5}$$

and to determine the vector of probabilities $p = (p_1, \ldots, p_n)$ such that it is close to the uniform distribution $p_{\text{uniform}} = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ and such that the resulting estimate is monotone. In order to achieve this monotonicity they introduce a distance measure $d(.,.)$ on the set of probability measures on $\{1, 2, \ldots, n\}$. The algorithm proposed by these authors determines a vector $p = (p_1, \ldots, p_n)$ such that the weights satisfy $p_i \geq 0\,(i = 1, \ldots, n), \sum_{i=1}^{n} p_i = 1$ and such that $d(p, p_{\text{uniform}})$ becomes minimal subject to the constraints $\hat{m}'(\cdot \mid p) \geq \varepsilon$ on a grid of N points, where $\varepsilon \geq 0$ is a prespecified constant. Some possible distance measures are given in Cressie and Read (1984) and we mention the Kullback-Leibler divergence

$$d_0(p, q) = -\sum_{i=1}^{n} q_i \log\Big(\frac{p_i}{q_i}\Big), \tag{2.6}$$

which will be used in our numerical study in Section 3. The implementation of this procedure is straightforward using an off-the-shelf quadratic programming routine. Note that the resulting constrained estimate will have the same smoothness properties as the unconstrained estimate, because the smoothness is mainly determined by the properties of the weight functions A_i. Moreover, it was shown by Hall and Huang (2001) that on intervals where the regression function is strictly increasing the constrained estimate coincides with the unconstrained estimate with probability 1 if the sample size is sufficiently large.

2.3 Combining density and regression estimation

Recently Dette, Neumeyer and Pilz (2003) proposed a simple estimate, which combines classical density and regression estimation techniques to obtain an estimate of the inverse of a strictly increasing regression function. The method can easily be motivated by considering an i.i.d. sample of uniformly distributed random variables, say $U_1, \ldots, U_N \sim \mathcal{U}([0, 1])$. If m is a strictly increasing function on the interval $[0, 1]$, K_d is a kernel function with compact support, say $[-1, 1]$, and h_d a bandwidth, then

$$\frac{1}{Nh_d}\sum_{i=1}^{N} K_d\Big(\frac{m(U_i) - u}{h_d}\Big) \tag{2.7}$$

is the classical kernel estimate of the density of the random variable $m(U_1)$, that is

$$(m^{-1})'(u) I_{[m(0),m(1)]}(u).$$

Consequently

$$\frac{1}{Nh_d}\int_{-\infty}^{t}\sum_{i=1}^{N}K_d\Big(\frac{m(U_i)-u}{h_d}\Big)du \tag{2.8}$$

is a consistent estimate of the function m^{-1} at the point t. One can easily utilize this idea to obtain an isotone estimate of the regression function in the model (2.1). For this Dette, Neumeyer and Pilz (2003) define for $N \in \mathbb{N}$

$$\hat{m}_I^{-1}(t) := \frac{1}{Nh_d}\sum_{i=1}^{N}\int_{-\infty}^{t}K_d\Big(\frac{\hat{m}(\frac{i}{N})-u}{h_d}\Big)du \tag{2.9}$$

as an estimate of the inverse function $m^{-1}(t)$, where $\hat{m}(x)$ denotes an unconstrained nonparametric estimate of the regression function at the point x. Note that the uniformly distributed random variables in (2.8) have been replaced by an equidistant design. The number N used in the density step does not necessarily coincide with the sample size n and the summation over the index i can be considered as simple quadrature formula for the integral

$$\frac{1}{h_d}\int_0^1\int_{-\infty}^{t}K_d\Big(\frac{\hat{m}(x)-u}{h_d}\Big)dudx\ .$$

In cases, where this integral can be evaluated directly, a summation is in fact not necessary. Obviously, the estimate $\hat{m}_I^{-1}$ is isotonic if the kernel K_d is positive, which will be assumed whenever the estimate $\hat{m}_I^{-1}$ (or its inverse) is used in this paper. In this case an isotonic estimate of the regression function m is simply obtained by reflection of the function $\hat{m}_I^{-1}$ at the line $y = x$. This estimate will be denoted by $\hat{m}_I$ and is called density regression estimate throughout this paper. It can be shown [see Dette, Neumeyer and Pilz (2003)] that on any interval $[a, b]$, where the unconstrained estimate $\hat{m}$ is strictly increasing such that $\hat{m}^{-1}(\{a\})$ and $\hat{m}^{-1}(\{b\})$ are singletons the estimate $\hat{m}_I$ coincides with $\hat{m}$ on the interval $[a + h_d, b - h_d]$ for sufficiently large $N \to \infty$ and small $h_d \to 0$. If the Nadaraya and Watson estimate (2.3) or a local linear estimate is used for the preliminary regression estimator $\hat{m}$ these authors also show that the monotone estimate $\hat{m}_I$ (appropriately standardized) is asymptotically normal distributed, where the standardization depends on the relative behaviour of the ratio of bandwidths h_d/h_r. In particular, if $\lim_{n\to\infty} h_d/h_r = 0$, then the constrained estimator of the regression function is asymptotically first order equivalent to the unconstrained estimator; see Dette, Neumeyer and Pilz (2003) for more details. As pointed also out by these authors the choice of the bandwidth h_d is less critical compared to the choice of the bandwidth h_r for the regression estimate. Usually, a substantially smaller bandwidth $h_d = h_r^{\gamma}$ with $\gamma = 3, 4$ is recommended for applications.

3 A comparison of smooth monotone regression estimates

It follows from Mammen (1991), Hall and Huang (2001) and Dette, Neumeyer and Pilz (2003) that for a strictly monotone regression function and for the commonly used kernel estimators in the preliminary step the three smooth monotone nonparametric kernel estimates are asymptotically first order equivalent in the sense that for any $t \in [0, 1]$ with $m'(t) > 0$

$$\sqrt{nh_r}\Big(\hat{m}_{\text{monotone}}(t) - m(t) - h_r^2 b(t)\Big) \overset{\mathcal{D}}{\Rightarrow} \mathcal{N}\Big(0, \frac{\sigma^2(t)}{f(t)}\int K_r^2(u)du\Big), \tag{3.1}$$

where f is the density of the explanatory variables X_i, $b(t)$ depends on the particular kernel estimator $\hat{m}$, K_r is the kernel used in this estimator and $\hat{m}_{\text{monotone}}$ is any of the estimators discussed in Section 2 (note that this statement requires $h_d = o(h_r)$ for the density regression estimator $\hat{m}_I$). In this section we investigate the finite sample properties of the three estimators $\hat{m}_{IS}$ (the PAVA estimator), $\hat{m}(\cdot|p)$ (obtained from the tilting method) and $\hat{m}_I$ (obtained by the inversion of the combination of a density and regression estimate). Note that all techniques require a preliminary nonparametric (unconstrained) estimate of the regression function with corresponding smoothing parameter. For the the sake of comparison we use for all three methods the same regression estimate $\hat{m}$ in the first step, namely a local linear estimator [see Wand and Jones (1995), Fan and Gijbels (1996)] with Epanechnikov kernel $K_r(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$. The bandwidth h_r of this estimate is chosen as

$$\hat{h}_r = \Big(\frac{\hat{\sigma}^2}{n}\Big)^{1/5}, \tag{3.2}$$

where $\hat{\sigma}^2$ denotes the variance estimator of Rice (1984), that is

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \Big(Y_{[i+1]} - Y_{[i]}\Big)^2 .$$

Note that this statistic converges almost surely to the integrated variance

$$\sigma^2 = \int_0^1 \sigma^2(x)dx$$

[see Rice (1984)], which appears in the asymptotically optimal global bandwidth

$$h_r = \left(\frac{\int_{-1}^1 K_r^2(u)\,du \,\int_0^1 \sigma^2(x)\,dx}{n(\int_{-1}^1 u^2 K_r(u)du)^2 \int_0^1 (m''(x))^2\,dx} \right)^{1/5} \tag{3.3}$$

for the estimation of the regression curve under the uniform design [see Fan and Gijbels (1996)]. We used the PAVA [Pool-Adjacent-Violators Algorithm, see Barlow, Bartholomew, Bremner and Brunk (1972)] to obtain an isotone estimate of the regression function. This estimate was then smoothed by a local linear estimator with bandwidth $\hat{h}_r$ in order to obtain a strictly isotonic and smooth estimate $\hat{m}_{IS}$ [see Mammen (1991)]. Note that the resulting estimate is not necessarily monotone increasing, because the smoothing step is performed after the isotonization. However, in most of the examples of our empirical study the resulting estimate was monotone, which is in accordance to the theoretical result that the estimate $\hat{m}_{IS}$ is asymptotically first order equivalent to the estimate obtained by projecting a smooth curve on the space of monotone functions [see Mammen (1991)]. Moreover, this author compared the finite sample performance of the two isotone regression estimates obtained by interchanging the order of smoothing and isotonizing and concluded that the estimate $\hat{m}_{IS}$ usually yields a smaller mean squared error than the estimate obtained by projecting a non-monotone curve on the space of monotone functions [see Mammen (1991); Table 1 for more details]. Thus we included the more efficient method of these two montone regression estimates in our numerical study.
For the calculation of the estimator of Hall and Huang (2001) we minimized the function $d_0(p, p_{\text{uniform}})$ in (2.6) subject to the constraints $p_i \geq 0$ $(i = 1, \ldots, n)$, $\sum_{i=1}^n p_i = 1$ and

$$\hat{m}'(x,p) \geq 0 \quad x \in [0,1]. \tag{3.4}$$

It should be pointed out here that the number of constraints introduced by the tilting method yields some technical restrictions in the application of standard numerical software for constrained optimization. We used the NAG routine E04UCC, which had problems to handle optimization problems with more than 100 constraints. As pointed out by Hall and Huang (2001) the modification of the non-monotone estimate has only to be applied on those intervals (and a little beyond) where the original estimate is not monotone. However, this fact does not facilitate the implementation of the tilting method in a simulation study because in each iteration the non-monotone estimate $\hat{m}(\cdot, p_{\text{uniform}})$ would have to be inspected visually or numerically in order to determine the intervals, where it is not isotone. Therefore, in our simulation, the algorithm was implemented on the entire interval. The estimate $\hat{m}_I$ obtained from the inversion of a combined density and regression estimate requires the additional specification of a bandwidth h_d and the number of design points N for the step of density estimation. For the number of design points we use $N = n$, which provides a reasonable approximation of the integral by the quadrature formula. Some recommendations for choosing h_d can be found in Dette, Neumeyer and Pilz (2003), who pointed out that the choice of the bandwidth h_d for the density estimate is less critical compared to the choice of the bandwidth h_r for the regression estimate. In particular h_d should be chosen substantially smaller than h_r and Dette, Neumeyer and Pilz (2003) did not observe substantial differences between the isotone estimates as long as this condition is met. Following these authors we use $h_d = h_r^3$ in the present simulation study.

We investigate the regression model (2.1) with a uniform design, i.e. $X_i \sim \mathcal{U}([0,1])$, normally distributed errors, i.e. $\varepsilon_i \sim \mathcal{N}(0,1)$, standard deviations $\sigma = 0.2$, $\sigma = 1$ and sample sizes $n = 50$ and $n = 80$. The regression functions are chosen as

$$m_1(x) = \frac{1}{10}x + \frac{1}{2} \tag{3.5}$$

$$m_2(x) = \frac{\exp(20(x - 1/2))}{1 + \exp(20(x - 1/2))} \tag{3.6}$$

$$m_3(x) = \frac{1}{2}(2x - 1)^3 + \frac{1}{2} \tag{3.7}$$

$$m_4(x) = \sin(\frac{\pi}{2}x) \tag{3.8}$$

$$m_5(x) = x^2 \tag{3.9}$$

$$m_6(x) = x + \frac{1}{6\pi}\sin(6\pi x) \tag{3.10}$$

and correspond to flat curve (3.5), a continuous "jump" (3.6), a strictly increasing curve with some plateau (3.7) a concave and convex function [see (3.8) and (3.9), respectively] and a function which changes several times from a strongly increasing part to a flat part (3.10). The different functions are displayed in Figure 1.

We use 2000 simulation runs to calculate the bias, variance and mean squared error (mse) of the three estimates $\hat{m}_{IS}(\cdot), \hat{m}(\cdot \mid p), \hat{m}_I(\cdot)$ for the 6 regression functions defined in (3.5) - (3.10). In the following, we first present curves for the mse, squared bias and variance of the three estimates, where only the results for the sample size $n = 50$ and $\sigma = 0.2$ are displayed. The results corresponding to the cases $n = 50, \sigma = 1; n = 80, \sigma = 0.2$ and $n = 80, \sigma = 1$ are quite similar and available from the authors [see also Table 1 for results of the integrated mse, squared bias and variance in these cases]. The dashed curves in Figure 2-4 correspond to the PAVA estimate $\hat{m}_{IS}$, the dotted curves represent Hall and Huang's (2001) estimate, respectively, and the estimate of Dette, Neumeyer and Pilz (2003) corresponds to the solid curves.

The upper and lower panel of Figure 2 show the squared bias-, mse- and variance-curves corresponding to the regression functions (3.5) and (3.6), respectively. In the first case the estimator $\hat{m}_I(\cdot)$ and $\hat{m}(\cdot, p)$ have a very similar variance behaviour, while the bias of the estimator obtained from the tilting method is substantially larger which leads some advantages of $\hat{m}_I(\cdot)$ with respect to the mse-criterion. Here the PAVA estimator $\hat{m}_{IS}$ has a substantially larger variance but a smaller bias than the estimate obtained from the tilting method. Compared to the density regression estimate $\hat{m}_I$ the estimate $\hat{m}(\cdot \mid p)$ yields larger values with respect to all three criteria. Visually, the PAVA method yields the worst results while the differences to the tilting method are only minor. For the function (3.6) the situation is not so clear. The bias is strongly influenced by the magnitude of the second derivative, i.e. $|m_2''(x)|$, which is maximal at $x \approx 0.43, x \approx 0.57$ and minimal for $x = 0.5$ (note that $m_2''(0.5) = 0$). The PAVA estimate $\hat{m}_{IS}$ yields the smallest variance except at the boundary of the interval $[0, 1]$, but usually has a substantially larger bias compared to the estimators $\hat{m}_I$ and $\hat{m}(\cdot, p)$. The latter estimates behave similar with respect to bias criterion, but the density regression estimate $\hat{m}_I$ yields a smaller squared bias at the boundary and at points where the second derivative of the regression function is large. The tilting estimate $\hat{m}(\cdot, p)$ tends to produce slightly smaller variances in the interval $[0, 1]$. A comparison of the mse curves shows that the PAVA estimate yields the largest error over a broad range of the interval $[0, 1]$, while the estimates $\hat{m}_I$ and $\hat{m}(\cdot, p)$ behave very similar with slight advantages of $\hat{m}_I$, especially at points, where the second derivative of the regression function is large. It is interesting to note that except for the boundary regions the simulations reflect the expected properties from the asymptotic theory. The simulated variance is nearly constant in most parts of the interval $[0, 1]$, while the limit theorem (3.1) yields the asymptotic variance $\sigma^2 \int K^2(u)du$. Because all estimates are first order asymptotically equivalent to the unconstrained estimate $\hat{m}$, the asymptotic squared bias at the point x is proportional to $(m''(x))^2$, which is also reflected in our simulation results presented in Figure 2.

We now discuss the upper and lower panel in Figure 3 corresponding to the regression functions (3.7) and (3.8), respectively. For model (3.7) we observe that the estimate $\hat{m}_I$ leads the smallest squared bias, while the PAVA estimate is the worst with respect to this criterion. The estimate $\hat{m}_{IS}$ obtained from the PAVA method and the tilting estimate $\hat{m}(\cdot, p)$ have the smallest variance in this case but $\hat{m}_I$ yields only slightly larger values for the variance. The mse comparison shows not too substantial differences between the three estimates $\hat{m}_{IS}$, $\hat{m}_I$ and $\hat{m}(\cdot, p)$ in regions where the bias of the two estimates is similar (thus in these regions the different performance with respect to bias and variance is compensated in the mse). However, there are two intervals (approximately $[0.1, 0.25]$ and $[0.7, 0.9]$) where the mse of $\hat{m}_{IS}$ is larger than that of $\hat{m}(\cdot, p)$ and $\hat{m}_I$. This difference is mainly caused by the bias of the PAVA estimator. For model (3.8) we observe exactly the opposite behaviour, the PAVA estimate yields the smallest mse's except at the right boundary of the interval $[0, 1]$, but $\hat{m}_I$ and $\hat{m}(\cdot, p)$ are not much worse. There are also no substantial differences between the estimates $\hat{m}_I$ and $\hat{m}(\cdot, p)$. The differences are mainly caused by the variances of the estimators and no clear pattern can be obtained for the squared bias which is substantially smaller compared to the variance. Finally, the bias, variance and mse curves corresponding to the cases (3.9) and (3.10) are presented in the upper and lower panel of Figure 4. In these cases we also do not observe substantial differences between the three estimators. For the model (3.9) all estimates yield a very small squared bias and the mse is mainly dominated by the variance of the estimators. This confirms asymptotic theory, because $|m_4''(x)| = \frac{\pi^2}{4}|\sin\frac{\pi}{2}x| \leq \frac{\pi^2}{4} \leq 2.5$ for all $x \in [0,1]$. The PAVA estimate appears to be slightly better with respect to the mse compared to the estimates $\hat{m}_I$ and $\hat{m}(\cdot,p)$, provided that estimation is not performed at the left boundary of the interval $[0,1]$. The other estimates $\hat{m}_I$ and $\hat{m}(\cdot,p)$ behave very similar with slight advantages for density regression method at the right boundary of the interval $[0,1]$. The situation is similar for the regression model (3.10), the PAVA estimate yields the smallest variances, but larger squared bias than the density regression estimate $\hat{m}_I$ near the boundaries. The estimate $\hat{m}(\cdot,p)$ obtained by the tilting method has larger bias than $\hat{m}_I$ and slightly smaller variance. The mse-curves of the three estimates are very similar, except at the regions, where the bias of $\hat{m}_{IS}$ or $\hat{m}(\cdot,p)$ is large compared to $\hat{m}_I$.

These results indicate that a superiority of one estimator in general cannot be established. The estimators obtained by the tilting method and by the inversion of the density-regression estimate behave very similar, but the application of the PAVA method can produce a larger mse in some cases. In order to investigate if these cases are representative with respect to the choice of the variance and sample size we have listed in Table 1 the simulated mean integrated squared error, squared bias and variance for the six regression functions corresponding to the remaining cases $n = 50, \sigma = 1; n = 80, \sigma = 0.2$ and $n = 80, \sigma = 1$. The integrated squared error was calculated for the interval $[0.05, 0.95]$ in order to avoid a domination by boundary effects. We observe that in our study the estimate $\hat{m}_I$ has the smallest integrated squared bias in nearly all cases, while the PAVA estimate yields the largest values. On the other hand the integrated variances of the estimate $\hat{m}_I$ are larger than the corresponding values for $\hat{m}_{IS}$ and $\hat{m}(\cdot,p)$, where the estimate $\hat{m}(\cdot,p)$ obtained by the tilting method usually produces the smallest variances. A comparison of the three estimation techniques with respect to the mean integrated squared error criterion yields the largest values for the PAVA estimate in most cases (75% of the considered scenarios), while there is no clear winner between the estimates $\hat{m}_I$ and $\hat{m}(\cdot,p)$. The superiority of one particular monotone estimation procedure depends on the regression model under consideration, but the differences between the estimates $\hat{m}_I$ and $\hat{m}(\cdot,p)$ are usually very small. For example in model (3.9) with $n = 80$ and $\sigma^2 = 1$ the estimator $\hat{m}(\cdot,p)$ obtained by the tilting method is the best, but the estimator $\hat{m}_I$ has only a 0.9% larger mean integrated squared error. The worst difference between the integrated mean squared errors of $\hat{m}_I$ and $\hat{m}(\cdot,p)$ that we have observed in our study was 30.7% and corresponds to the regression model (3.5) with $n = 80, \sigma = 0.2$, where $\hat{m}_I$ yields a substantially smaller integrated mse. The differences between the PAVA estimate and $\hat{m}_I$ and $\hat{m}(\cdot,p)$ can be larger and vary between 1% and 43%. Consider as a typical example the regression function m_3 where the integrated mse of the PAVA estimate is 10% ($n = 50, \sigma = 0.2$), 43% ($n = 50, \sigma = 1$), 10% ($n = 80, \sigma = 0.2$) and 43% ($n = 80, \sigma = 1$) larger compared to the best case. Note also that there is only one case, (3.9) with $n = 50$ and $\sigma = 0.2$, where the PAVA estimate is the best with respect to the mean integrated squared error criterion. However, in this case the estimates $\hat{m}_I$ and $\hat{m}(\cdot,p)$ produce only 4.2% and 7.6% larger values for the integrated mean squared error, respectively.

4 Conclusions

In this note we presented a finite sample comparison between three estimation techniques of a monotone regression function. All methods monotonize a nonparametric curve estimate and are first order asymptotically equivalent to a local linear or kernel estimate, if this type of estimate is used in the step of smoothing. It is demonstrateted that the PAVA estimate (obtained by smoothing of a monotone estimate) usually yields a larger mean squared error than the estimates of Hall and Huang (2001) and Dette, Neumeyer and Pilz (2003), which are based on the concept of tilting and the combination of density and regression estimation techniques. The lastnamed two estimates behave very similar with respect to the mean squared error criterion in our study. On the other hand the estimate of Dette, Neumeyer and Pilz (2003) has substantial computational advantages, because it does not rely on constrained optimization methods. Therefore this estimate is recommended for the kernel estimation of a smooth monotone regression curve. It is particularly attractive to users of conventional kernel methods, because of its simplicity and is at least competitive to the estimation techniques which have been proposed so far for the smooth estimation of an isotone regression function (in many cases of our study it was even better than the estimates obtained by the PAVA and tilting method).

Acknowledgements. The authors are grateful to Isolde Gottschlich who typed parts of this paper with considerable technical expertise and to E. Mammen and W. Polonik for useful discussions and some help with the references. The work of the authors was supported by the Sonderforschungsbereich 475, Komplexitätsreduktion in multivariaten Datenstrukturen. Parts of the paper were written during a visit of the second author at the Institute of Statistics in Louvain-la-Neuve and this author would like to thank the institute for its hospitality.

References

R.E. Barlow, D.J. Bartholomew, J.M. Bremner, H.D. Brunk (1972). Statistical inference under order restrictions: The theory and application of isotonic regression. Wiley, New York.

H.D. Brunk (1955). Maximum likelihood estimates of monotone parameters. Ann. Math. Statist. 26, 607-616.

H.D. Brunk (1958). On the estimation of parameters restricted by inequalities. Ann. Math. Statist. 29, 437-454.

K.F. Cheng, P.E. Lin (1981). Nonparametric estimation of a regression function. Z. Wahrsch. Verw. Gebiete 57, 223-233.

N.A.C. Cressie, T.R.C. Read (1984). Multinomial goodness-of-fit tests. J. Roy. Statist. Soc., Ser. B, 46, 4404-464.

H. Dette, N. Neumeyer, K.F. Pilz (2003). A simple non-parametric estimator of a monotone regression function. Technical report, Department of Mathematics.
http://www.ruhr-uni-bochum.de/mathematik3/preprint.htm

R.L. Eubank (1988). Spline Smoothing and Nonparametric Regression. Marcel Dekker, N.Y.

J. Fan, I. Gijbels (1996). Local polynomial modelling and its applications. Chapman and Hall, London.

J. Friedman, R. Tibshirani (1984). The monotone smoothing of scatterplots. Technometrics 26, 243-250.

I. Gijbels (2004). Monotone regression. Discussion paper 0334, Institute de Statistique, Université Catholique de Louvain. http://www.stat.ucl.ac.be.

P. Hall, L.-S. Huang (2001). Nonparametric kernel regression subject to monotonicity constraints. Ann. Statist. 29, 624-647.

C. Kelly, J. Rice (1990). Monotone smoothing with application to dose response curves and the assessment of synergism. Biometrics 46, 1071-1085.

E. Mammen (1991). Estimating a smooth monotone regression function. Ann. Statist. 19, 724-740.

E. Mammen, J.S. Marron, B.A. Turlach, M.P. Wand (2001). A general projection framework for constrained smoothing. Statist. Sci. 16, 232-248.

E. Mammen, C. Thomas-Agnam (1999). Smoothing splines and shape restrictions. Scand. J. Statist. 26, 239-252.

H.G. Müller, T. Gasser (1979). Kernel estimates of regression functions. In: Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics. Springer-Verlag, New York.

H. Mukerjee (1988). Monotone nonparametric regression. Ann. Statist. 16, 741-750.

J.O. Ramsay (1988). Monotone regression splines in action (with comments). Statist. Sci. 3, 425-461.

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

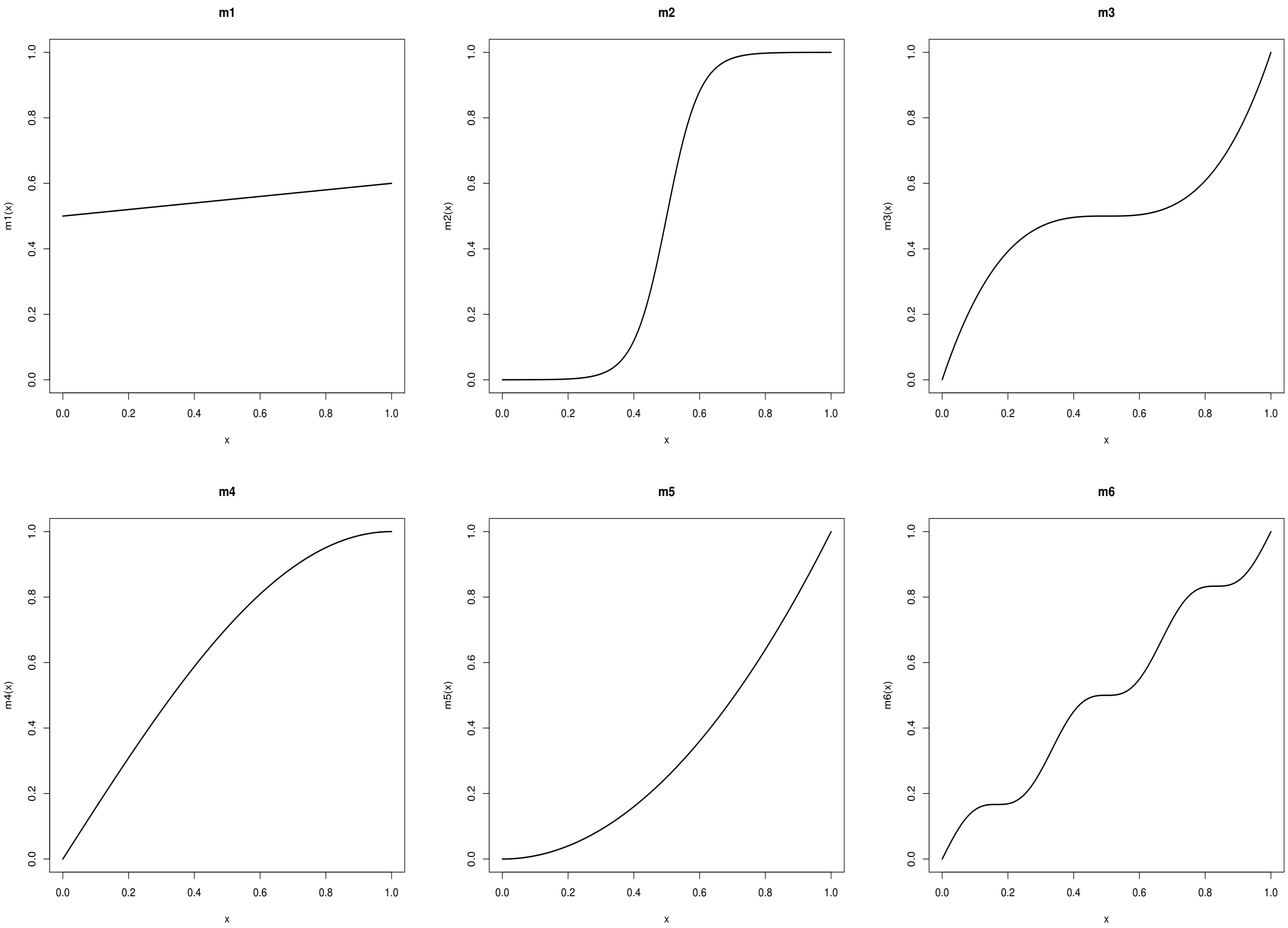

Figure 1: *Regression function* m_1 *-* m_6 *used in the simulation study. The functions are defined in (3.5) - (3.10), respectively.*

G.A.F. Seber, C.J. Wild (1989). Nonlinear regression. Wiley, N.Y.

M.P. Wand, M.G. Jones (1995). Kernel smoothing. Chapman and Hall, London.

F.T. Wright (1981). The asymptotic behavior of monotone regression estimates. Ann. Statist. 9, 443-448.

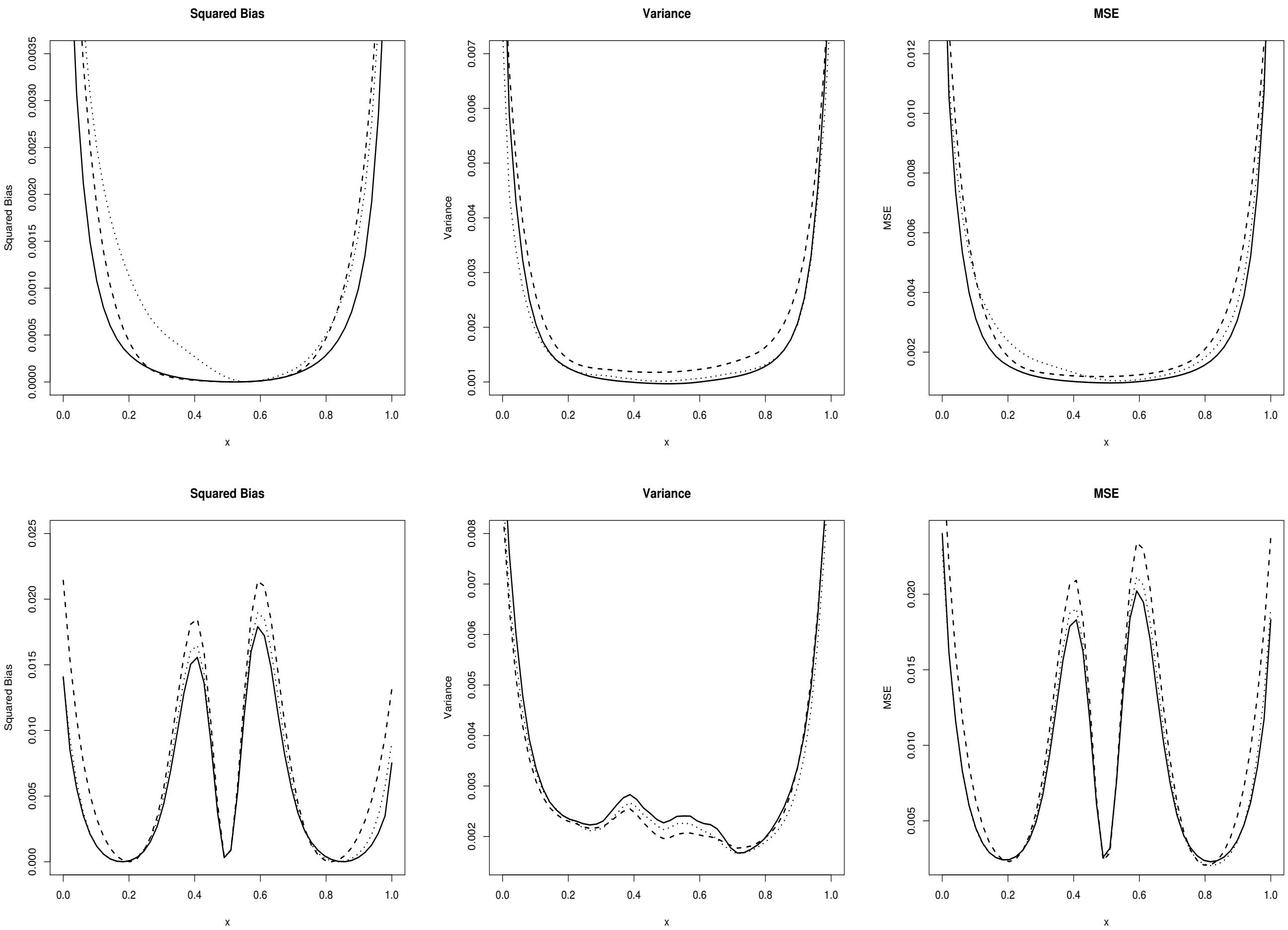

Figure 2: *Simulated mean squared error, bias and variance of the PAVA estimator* $\hat{m}_{IS}$ *(dashed line), the estimator* $\hat{m}(\cdot, p)$ *obtained by the tilting method (dotted line) and the estimator* $\hat{m}_I$ *obtained by the combination of a density and regression estimate (solid line). The sample size is* $n = 50$*, the standard deviation is* $\sigma = 0.2$ *while the regression functions are given by (3.5) (upper panel) and (3.6) (lower panel), respectively.*

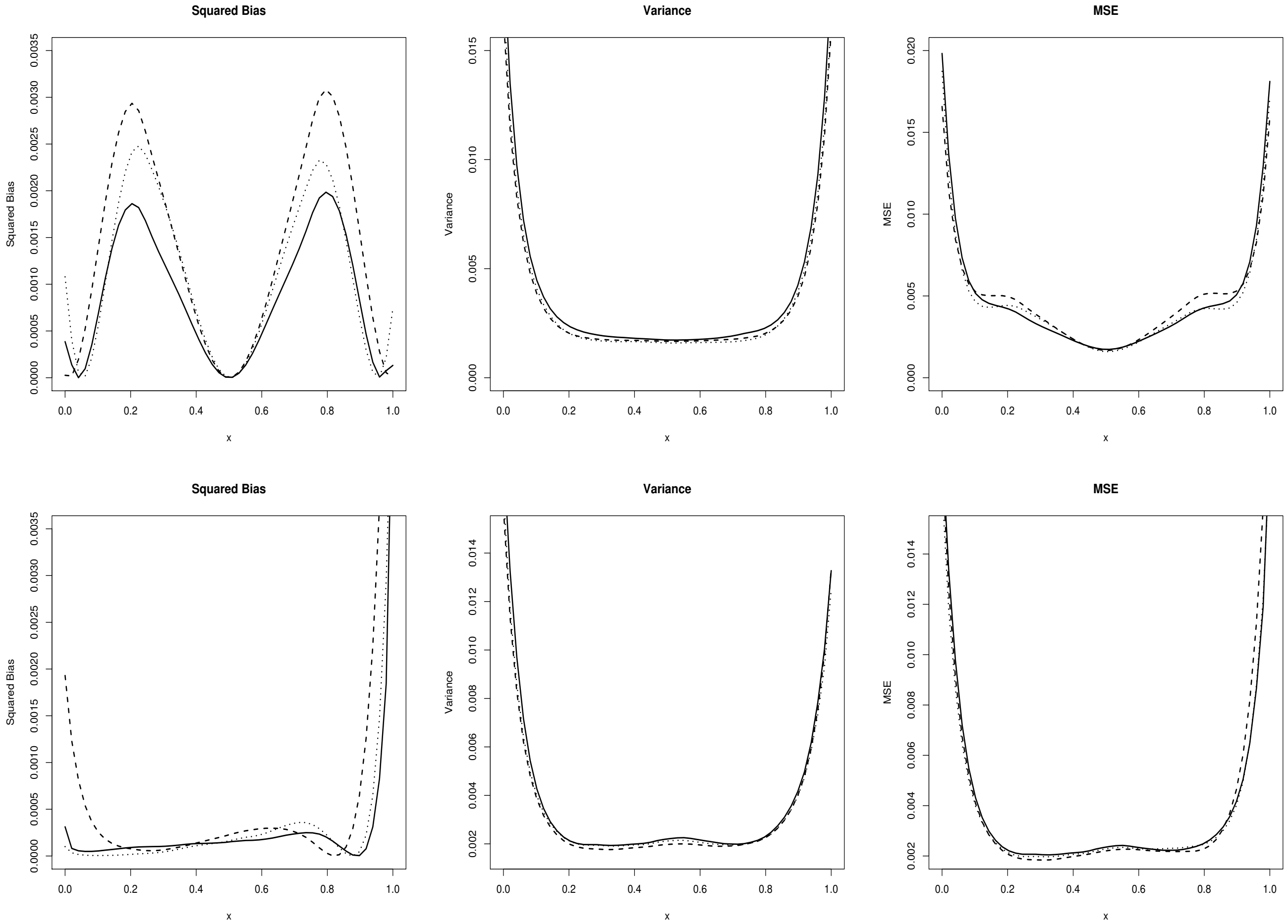

Figure 3: *Simulated mean squared error, bias and variance of the PAVA estimator* $\hat{m}_{IS}$ *(dashed line), the estimator* $\hat{m}(\cdot, p)$ *obtained by the tilting method (dotted line) and the estimator* $\hat{m}_I$ *obtained by the combination of a density and regression estimate (solid line). The sample size is* $n = 50$*, the standard deviation is* $\sigma = 0.2$ *, while the regression functions are given by (3.7) (upper panel) and (3.8) (lower panel), respectively.*

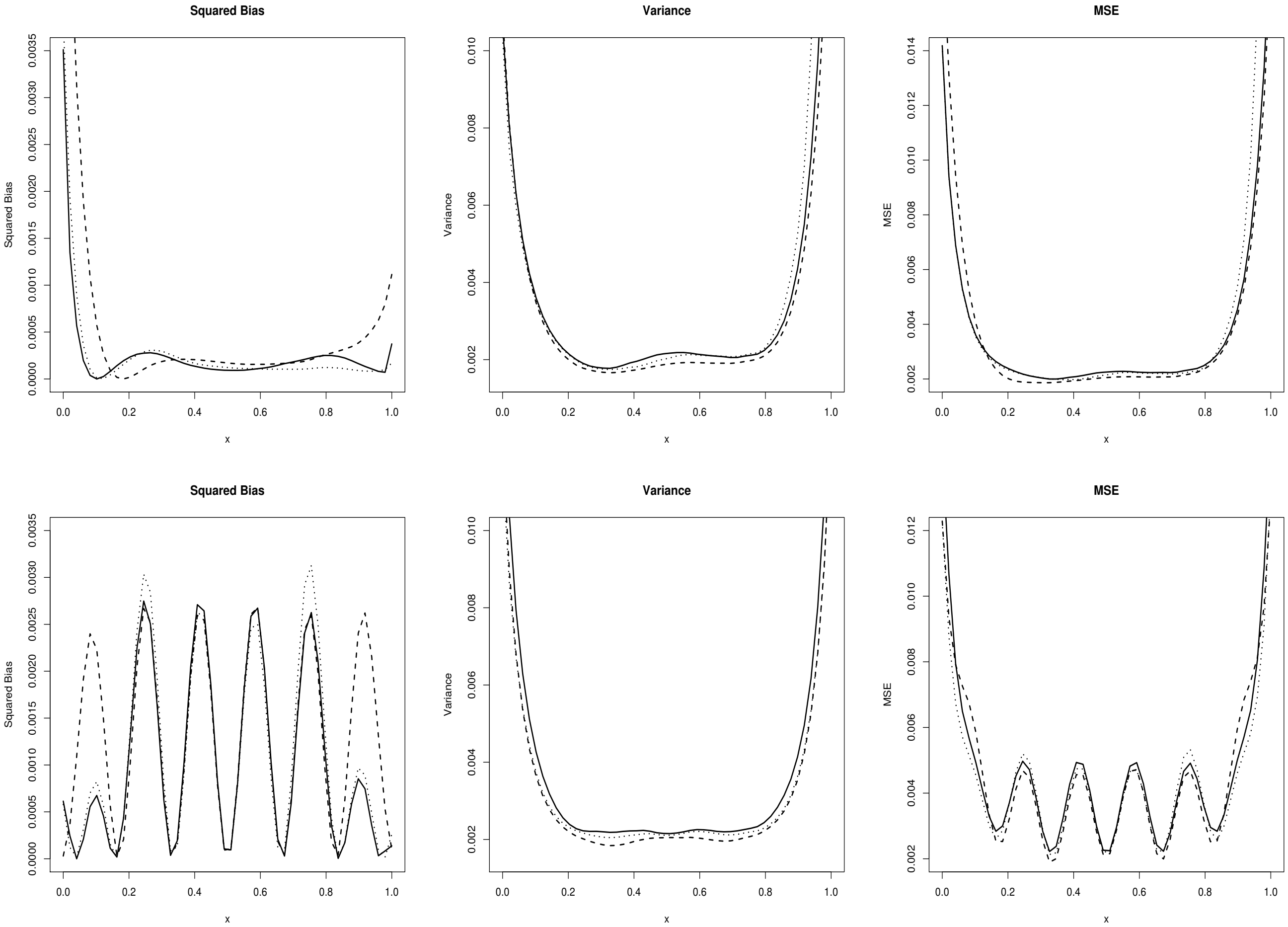

Figure 4: *Simulated mean squared error, bias and variance of the PAVA estimator* $\hat{m}_{IS}$ *(dashed line), the estimator* $\hat{m}(\cdot, p)$ *obtained by the tilting method (dotted line) and the estimator* $\hat{m}_I$ *obtained by the combination of a density and regression estimate (solid line). The sample size is* $n = 50$*, the standard deviation is* $\sigma = 0.2$ *, while the regression functions are given by (3.9) (upper panel) and (3.10) (lower panel), respectively.*