A Convolution Estimator for the Density of Nonlinear Regression Observations

Bård Støve
Department of Finance and Management Science,
Norwegian School of Economics and Business Administration
Helleveien 30, 5045 Bergen, Norway
E mail: bard.stove@nhh.no

Dag Tjøstheim
Department of Mathematics, University of Bergen
Johannes Brunsgate 12, 5008 Bergen, Norway
E mail: dag.tjostheim@mi.uib.no

November 8, 2007

Author's Footnote: Bård Støve is a post doctor, Department of Finance and Management Science, Norwegian School of Economics and Business Administration located in Bergen, and Dag Tjøstheim is Professor, Department of Mathematics, University of Bergen, Norway. This work has been partly supported by grant no. 147231/432 of the Norwegian Research Council.

Abstract

The problem of estimating an unknown density function has been widely studied. In this paper we present a convolution estimator for the density of the responses in a nonlinear regression model. The rate of convergence for the variance of the convolution estimator is of order n^{-1}. This is faster than the rate for the kernel density method. The intuition behind this result is that the convolution estimator uses model information, and thus an improvement can be expected. We also derive the bias of the new estimator and conduct simulation experiments to check the finite sample properties. The proposed estimator performs substantially better than the kernel density estimator for well-behaved noise densities.

KEY WORDS: Convergence rate, Convolution estimator, Kernel function, Mean squared error, Nonparametric density estimation.

1. Introduction

There exists a vast literature on the problem of estimating an unknown density function $f(x)$ from a given sample $X_1, X_2, ..., X_n$ of independent and identically distributed random variables, see e.g.; the books by Härdle (1990), Wand & Jones (1995) and Simonoff (1996). The most used method is kernel density estimation where $f(x)$ is estimated by

$$f^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big),$$

with K being a kernel function and h the bandwidth. It is well known that the asymptotic bias and variance of this estimator are of the order h^2 and $(nh)^{-1}$, respectively.

In this paper we consider the standard nonlinear regression model,

$$Y_i = g(X_i) + e_i, \tag{1}$$

where g is unknown and where $\{X_i\}$ and $\{e_i\}$ consist of independent and identically distributed random variables with $\{e_i\}$ independent of $\{X_i\}$. Denote the density of Y_i by $f_Y(\cdot)$. This is the density of interest. The densities of X_i and e_i are denoted by $f_X(\cdot)$ and $f_e(\cdot)$, respectively. For given observations of (X_i, Y_i) one method of estimating the density of Y_i is using the already mentioned kernel density estimator on $\{Y_i\}$. This estimator does not require the relationship (1) to hold, and if one is able to construct an estimator of f_Y by convolution taking this relationship into account, one would think that it should be possible to make an improvement. This idea was used in Støve & Tjøstheim (2007*b*) for nonparametric estimation of g. For that case, the asymptotic bias and variance were of the same order as the standard nonparametric regression estimators, but an asymptotic bias improvement was obtained. However, in the case of density estimation in equation (1), we are able to obtain a better convergence rate for the variance. Moreover, often the bias properties are better, although asymptotically the order of the bias is the same as for the kernel density estimator.

Other authors have also studied this convolution idea; Frees (1994) introduced density estimation for a symmetric function $Y = g(X_1, ..., X_m)$, with g known, of $m > 1$ independent and identically distributed variables. The density can be estimated at the rate $n^{-1/2}$ for certain functions g. This result generalizes to non-identically distributed random variables, and in particular to convoluted densities $f * l(y) = \int f(y-x)l(x)\mathrm{d}x$. Saavedra & Cao (2000) introduced the convolution-kernel estimator for the marginal density of a moving average process $Y_i = X_i - \theta X_{i-1}$ when θ is known, and proved that this estimator is $n^{1/2}$-consistent. The case when θ is unknown is examined in Saavedra & Cao (1999*b*), and an analogous result is obtained; in this case both θ and the innovations X_i have to be estimated. Further, Schick & Wefelmeyer (2004*a*) introduced a slightly simplified variant of this estimator and proved a stronger result of asymptotic normality. In Schick & Wefelmeyer (2004*b*) it is shown that the density of a sum of independent random variables can be estimated by the convolution of kernel estimators for the marginal densities, and that this estimator is $n^{1/2}$-consistent as well. In Schick & Wefelmeyer (2007) they establish such a result for a general linear process.

A rather different way of taking extra information into account is presented in Gelfand & Smith (1990), who use Markov Chain Monte Carlo methods when there is information available on conditional densities.

Note that we allow a nonlinear model where both the function $g(\cdot)$ and the error terms e_i are unknown, and thus have to be estimated. This is in contrast to the models examined by Frees, Saavedra & Cao and Schick & Wefelmeyer, where the authors assume that the function describing the nonlinearity is known or that the model is linear.

Our proposed estimator is presented in section 2, its asymptotic behaviour is examined in section 3, and some simulation results and a real data example are given in section 4. Conclusions are in section 5. Proofs are deferred to the appendix.

2. The estimator

From equation (1), because $g(X_i)$ and e_i are independent, we have

$$f_Y(y) = \int f_e(y - g(u)) f_X(u) \mathrm{d}u = \mathrm{E}\big[f_e(y - g(X))\big], \tag{2}$$

where f_e is the density of the residuals. Assume we have observations $(X_1, Y_1,), \ldots, (X_n, Y_n)$ of (X, Y). We introduce an estimator based on (2) as

$$\hat{f}_Y(y) = \hat{\mathrm{E}}\big[f^*_{\tilde{e}}(y - \tilde{g}(X))\big]. \tag{3}$$

Here, $\tilde{g}$ is the Nadaraya-Watson estimator, see e.g. Härdle (1990), with bandwidth h_R, and kernel function $K^{NW}_{x,h_R}(X_i) = (1/h_R)K^{NW}\big((x - X_i)/h_R\big)$, i.e.

$$\tilde{g}(x) = \frac{\sum_{i=1}^{n} K^{NW}_{x,h_R}(X_i)Y_i}{\sum_{i=1}^{n} K^{NW}_{x,h_R}(X_i)}. \tag{4}$$

The estimator for e_i is

$$\tilde{e}_i = Y_i - \tilde{g}(X_i),$$

whereas the estimator $f^*_{\tilde{e}}$ of the density of e_i is the kernel estimator

$$f^*_{\tilde{e}}(y) = \frac{1}{nh_D}\sum_{i=1}^{n} K\Big(\frac{y - \tilde{e}_i}{h_D}\Big),$$

with bandwidth h_D and kernel function $K(\cdot)$, not necessarily equal to the kernel $K^{NW}(\cdot)$. Thus, using (3),

$$\hat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} f^*_{\tilde{e}}\big(y - \tilde{g}(X_i)\big) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh_D}\sum_{j=1}^{n} K\Big(\frac{y - \tilde{g}(X_i) - \tilde{e}_j}{h_D}\Big)\Big]. \tag{5}$$

For better understanding of the estimator in (5), we give a simple algorithm for its numerical calculation. Assume we want to estimate the density of Y in the gridpoints t_k, $k = 1, ..., M$.

Step 1: Estimate $\tilde{g}(X_i)$, for all $i = 1, ..., n$, with bandwidth h_R and kernel K^{NW}.

Step 2: Calculate the error terms $\tilde{e}_i = Y_i - \tilde{g}(X_i)$ for all i.

Step 3: For each estimate $\tilde{g}(X_i)$ ($i = 1, ..., n$) calculate the density estimate $f^*_{\tilde{e}}\big(t_k - \tilde{g}(X_i)\big)$ in all gridpoints t_k, with bandwidth h_D and kernel K.

Step 4: For each gridpoint t_k, $k = 1, ..., M$, average across the n estimates from step 3. This produces the final estimates $\hat{f}_Y(t_k)$.

Note that the bandwidths h_R and h_D need not to be the same. Further, observe that other nonparametric estimates for $\tilde{g}$ are possible, e.g. the local polynomial estimator, see Fan (1992). We also believe standard modifications, see e.g. Wand & Jones (1995), of the kernel density estimator in step 3, could lead to improved estimation of $\hat{f}_Y$.

3. Asymptotic properties

The following assumptions are made,

> **A1**: The kernel function K is a non-negative symmetric function that integrates to 1, moreover it is two times differentiable with a bounded second order derivative.
>
> **A2**: The function g is differentiable and its inverse exists.
>
> **A3**: The density f_X has compact support $S(X)$, is continous and two times differentiable on its support.
>
> **A4**: $\lim_{n\to\infty} h_D = 0$ and $\lim_{n\to\infty} nh_D = \infty$.

Condition A1 is standard in nonparametric estimation. If the kernel function is the standard normal distribution, this condition is automatically fulfilled. It implies that

$$\int K'(z)\mathrm{d}z = 0 \quad \text{and} \quad \int z^2 K'(z)\mathrm{d}z = 0.$$

Condition A2 is introduced to obtain simple expressions. It can be relaxed. The compact support in condition A3 is also introduced for the sake of simplicity. It can be removed at the cost of lengthier arguments. An alternative would be to just look at the X-observations falling within a compact set and do the analysis on that compact set. Condition A4 is standard. Other assumptions will be imposed when needed.

To study the mean squared error (MSE) of the estimator, it is useful to decompose the difference between the estimator and the true density in the following manner,

$$\hat{f}_Y(x) - f_Y(x) = \hat{f}_Y(x) - \tilde{f}_Y(x) + \tilde{f}_Y(x) - f_Y(x), \tag{6}$$

where

$$\tilde{f}_Y(x) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh_D}\sum_{j=1}^{n} K\Big(\frac{x - g(X_i) - e_j}{h_D}\Big)\Big],$$

that is, the proposed estimator (5) with $g(\cdot)$ and e_j for $j = 1, ..., n$, known. We first consider this "estimator". To ease notation, we set $h_D = h$.

Theorem 1. *If conditions A1-A4 are fulfilled, the bias of $\tilde{f}_Y$ is*

$$E\big(\tilde{f}_Y(x)\big) - f_Y(x) = \frac{h^2}{2} f_Y''(x) \int w^2 K(w) dw + O(h^4), \tag{7}$$

and the variance is

$$\begin{aligned} var\big(\tilde{f}_Y(x)\big) &= \frac{1}{n} \int f_X(v) f_e^2\big(x - g(v)\big) dv \\ &+ \frac{1}{n} \int \frac{f_X^2(v) f_e\big(x - g(v)\big)}{g'(v)} dv - \frac{2}{n} f_Y^2(x) \\ &+ \frac{1}{n^2 h} f_Y(x) \int K^2(z) dz + O\big(n^{-1} h^2\big). \end{aligned} \tag{8}$$

We note that the bias is equal to the bias of the standard kernel density estimator, see e.g. Wand & Jones (1995) page 20, but that $\mathrm{var}\big(\tilde{f}_Y(x)\big) = O(n^{-1})$.

Thus the MSE for $\tilde{f}_Y(x)$ becomes,

$$\begin{aligned} \mathrm{MSE}\big(\tilde{f}_Y(x)\big) &= \frac{1}{4} h^4 f_Y''(x)^2 \Big[\int w^2 K(w) \mathrm{d}w\Big]^2 \\ &+ \frac{1}{n} \int f_X(v) f_e^2\big(x - g(v)\big) \mathrm{d}v \\ &+ \frac{1}{n} \int \frac{f_X^2(v) f_e\big(x - g(v)\big)}{g'(v)} \mathrm{d}v - \frac{2}{n} f_Y^2(x) \\ &+ \frac{1}{n^2 h} f_Y(x) \int K^2(z) \mathrm{d}z + O\big(n^{-1} h^2\big) + O(h^6). \end{aligned} \tag{9}$$

If h is of order $n^{-1/4}$ it follows trivially that the MSE is of order $O(n^{-1})$.

Note that there are bias reducing techniques, using e.g. a higher order kernel, see Wand & Jones (1995) page 32, so that the squared bias can be reduced to $O(h^6)$ or $O(h^8)$, say, while still keeping the variance at $O(n^{-1})$. This means that there would be a wider choice of bandwidths for which the MSE is of order n^{-1}.

We next study the properties of the other term in equation (6), that is, $\hat{f}_Y(x) - \tilde{f}_Y(x)$. At this point some additional assumptions are introduced. Let $f(x, y) = f_{X,Y}(x, y)$ denote the joint distribution of (X, Y) and define $m(x) = \int y f(x, y) \mathrm{d}y$. We assume

A5: $\mathrm{E}|Y|^s < \infty$ and $\sup_x \int |y|^s f(x,y)\mathrm{d}y < \infty$, for some $s > 2$.

A6: m is continous on the support $S(X)$ of X.

A7: $\left(nh_R h^3\right)^{-1} = O(h^{2-\epsilon})$ for some $\epsilon > 0$, where $h = h_D$.

A8: $\mathrm{Inf}_{x\in S(X)} f_X(x) > 0$.

A9: The kernel K_{NW} is uniformly continous and of bounded variation. K_{NW} is absolutely integrable w.r.t. Lebesgue measure on the line. Further, $K_{NW}(x) \to 0$ as $|x| \to \infty$, and $\int |x \log |x||^{\frac{1}{2}} dK_{NW}(x) < \infty$.

A10: $n^{2\eta-1}h_R \to \infty$ for some $\eta < 1 - s^{-1}$, and $h_R^2 = o\Big(\big[\frac{1}{nh_R}\log\frac{1}{h_R}\big]^{1/2}\Big)$.

These conditions are essentially introduced to secure the uniform convergence of $\tilde{g}(x)$ to $g(x)$. They are taken from Mack & Silverman (1982), and are discussed there. Note that the condition A8 secures the existence of $\mathrm{E}\big(\tilde{g}(x) - g(x)\big)$ for $x \in S(X)$.

Theorem 2. *If conditions A1-A10 are fulfilled, then*

$$\begin{aligned} E\big(\hat{f}_Y(x) - \tilde{f}_Y(x)\big) \sim{}& -f_Y'(x)\int z_1 K'(z_1)dz_1 \int \big(E(\tilde{g}(x_2)) - g(x_2)\big) f_X(x_2)dx_2 \\ &+ \int z_2 K'(z_2)dz_2 \int \big(E(\tilde{g}(x_1)) - g(x_1)\big) f_e'(x - g(x_1)) f_X(x_1)dx_1 + O(h^4) \end{aligned} \tag{10}$$

where the leading term is of order h^2, and

$$var\big(\hat{f}_Y(x) - \tilde{f}_Y(x)\big) \sim O(h^4/n).$$

Using theorem 1 and 2; the total bias of $\hat{f}_Y(x)$ will consists of the terms (7) and (10). This is of order h^2, through the dependence on $\mathrm{E}\big(\tilde{g}(x_2) - g(x_2)\big)$, as for the kernel density estimator, but as we will see in the next section, a bias improvement may actually occur in some cases. If the bandwidth condition (38) is fulfilled, the total variance of $\hat{f}_Y(x)$ has a leading term given by equation (8), i.e. $O(n^{-1})$; it is in fact the rate of the variance for a parametric estimation problem. This result may seem striking.

However, observe that the density of Y is expressed in (2) as a smooth functional of the densities of X and e. This suggests that the density of Y can be estimated by plugging in estimators of the unknown densities and the unknown function g, in the functional. By the plug-in principle we can expect that this estimator converges at the parametric rate, even though the estimators being plugged in have a slower rate of convergence. Some references for smooth functionals of densities are e.g; Hall & Marron (1987), Birgè & Massart (1995) and Efromovich & Samarov (2000). In these cases the parametric convergence rate $n^{-1/2}$ for the estimated functionals are obtained.

4. Evaluating the convolution estimator

To evaluate the finite sample properties of the proposed estimator, (5), we carry out simulation experiments to compare the convolution estimator with the classic kernel estimator in (1).

To avoid looking at separate sets of points, the comparisons are based on the mean integrated squared error (MISE) of the two estimators. The MISE for a density estimator is

$$\mathrm{MISE}(\hat{f}) = \mathrm{E}\Big[\int_{-\infty}^{\infty}(\hat{f}-f)^2(x)\mathrm{d}x\Big].$$

We have used 500 simulated realizations with sample sizes from 100 to 5000 for the model (1), with different choices of the function $g(\cdot)$ and distributions of X and e. The value of the MISE is approximated as an average of the ISE (integrated squared error) of the 500 realizations, and the ISE is estimated by numerical integration. If the true density f_Y is not known analytically, we have based our comparisons on a numerically calculated true density from the convolution integral (2). For models 4, 8 and 9, given below, only 100 realizations have been used, and here the "true" density is taken as the estimated kernel density computed from 1 000 000 generated observations of (X_i, Y_i).

The choice of bandwidth has a considerable impact on the accuracy of an estimator. The bandwidth, h_D, used in the kernel density estimation in our simulation study, is the Solve-the-Equation Plug-in estimator proposed in Sheather & Jones (1991). This is the same for all of the $(n-1)$ density estimations in equation (5), and this estimator is also used for the classic kernel estimator. For ease of computation the bandwidth for the kernel smoothing of g is the rule-of-thumb, see e.g. Härdle (1990) page 91, $1.06 \min(\hat{\sigma}, R/1.34) n^{-1/5}$, where R is the interquartile range, $\hat{\sigma}^2$ is the empirical variance of all of the observations $X_1, ..., X_n$. We might have obtained better results using a more optimal bandwidth for the non-parametric regression. Some of the simulations have also been performed with other bandwidths, but without large changes in the results. It would be interesting to find the MISE for both estimators as a function of a general bandwidth, as this would isolate the effect of the estimator used from the effect of the quality of the bandwidth selector. We leave this for future research.

For both the kernel K of the density estimation, and the kernel K_{NW} of the non-parameteric smoothing estimation, we have used the Gaussian kernel. Further, the Gaussian kernel has also been used in the classic kernel density estimator.

The following models are considered:

1. $g(x) = x,\ X \sim \mathrm{N}(1,1),\ e \sim \mathrm{N}(0, 0.1)$.

2. $g(x) = x,\ X \sim \mathrm{N}(1,1),\ e \sim \mathrm{N}(0,1)$.

3. $g(x) = 3x,\ X \sim \mathrm{N}(1,1),\ e \sim \mathrm{N}(0,1)$.

4. $g(x) = x,\ X \sim \chi^2(3),\ e \sim (\chi^2(3) - 3)$.

5. $g(x) = x^2,\ X \sim \mathrm{U}[0,2],\ e \sim \mathrm{N}(0,1)$.

6. $g(x) = (0.5 + 4e^{-x^2})x,\ X \sim \mathrm{U}[-2,2],\ e \sim \mathrm{N}(0,1)$.

7. $g(x) = x,\ X \sim \mathrm{N}(1,1),\ e \sim \text{Double exponential}(0,1)$.

8. $g(x) = x$, $X \sim \mathrm{N}(1,1)$, $e \sim \sum_{l=0}^{2} \frac{2}{7}\mathrm{N}(\frac{12l-15}{7}, \frac{2}{7}) + \sum_{l=8}^{10} \frac{1}{21}\mathrm{N}(\frac{2l}{7}, \frac{1}{21})$.
9. $g(x) = x^2$, $X \sim \mathrm{U}[0,2]$, $e \sim (\exp(1) - 1)$.
10. $g(x) = x^2$, $X \sim \mathrm{U}[0,2]$, $e \sim (\frac{1}{2}\mathrm{N}(-3/2, 1/2) + \frac{1}{2}\mathrm{N}(3/2, 1/2))$.
11. $g(x) = x^3$, $X \sim U[-2,2]$, $e \sim N(0,1)$.
12. $g(x) = x$, $X \sim N(1,1)$, $e \sim$ t-distributed with 4 degrees of freedom

The second parameter given for the normal distributions is the standard deviation.

Models 1-4 are linear models, with error terms that can be encountered in practice, and models 5 and 6 are non-linear with normally distributed error terms. Models 7-10 are rather unusual and difficult, and seldom met in practice, but we wanted to see how the estimator performs in some extreme cases. Models 11 and 12 give a rather heavy-tailed distribution f_Y. In most cases of the examples the compactness assumption A3 on f_X is not fulfilled. Actually, we do not believe that this assumption is necessary, and we wanted to check performances in cases where it is violated. In figures 1 and 2 the densities f_Y are given for all of the models used, except for model 1 and 3, which are similar to model 2.

The simulation results are given in table 1. The table shows the percentage change by using the convolution estimator $\hat{f}$ compared with the kernel density estimator f^*. For the MISE, this change is calculated by

$$\frac{\mathrm{MISE}(f_Y^*) - \mathrm{MISE}(\hat{f}_Y)}{\mathrm{MISE}(f_Y^*)} \cdot 100. \tag{11}$$

It is composed from the squared bias change and the variance change. The former is given by

$$\frac{[\mathrm{Ave}(f_Y^* - f_Y)]^2 - [\mathrm{Ave}(\hat{f}_Y - f_Y)]^2}{[\mathrm{Ave}(f_Y^* - f_Y)]^2} \cdot 100, \tag{12}$$

where

$$[\mathrm{Ave}(f_Y^* - f_Y)]^2 = \frac{1}{k} \sum_{j=1}^{k} \left[\left(\frac{1}{500} \sum_{i=1}^{500} f_Y^{*i}(x_j) \right) - f_Y(x_j) \right]^2, \tag{13}$$

and similarly for the convolution estimator. In (13) k denotes the number of gridpoints for which the estimators are calculated, usually $k = 500$. Thus $f_Y^{*i}(x_j)$ is the calculated kernel estimate for the ith realization in gridpoint x_j. Further, $f_Y(x_j)$ denotes the true density in gridpoint x_j.

The variance change is calculated as

$$\frac{\hat{\text{var}}(f_Y^*) - \hat{\text{var}}(\hat{f}_Y)}{\hat{\text{var}}(f_Y^*)} \cdot 100, \tag{14}$$

where

$$\hat{\text{var}}(f_Y^*) = \frac{1}{k}\sum_{j=1}^{k}\Big[\frac{1}{499}\Big(\sum_{i=1}^{500}\big(f_Y^{*i}(x_j) - \text{Ave}\{f_Y^*(x_j)\}\big)^2\Big)\Big]$$

and similarly for the convolution estimator. Here $\text{Ave}\{f_Y^*(x_j)\}$ denotes the average of all of the 500 (or 100) kernel estimates in gridpoint x_j.

A minus sign in the table, thus indicates that the kernel density estimator performs better than the convolution estimator.

With the exception of model 1 and the unusual models 8-10 the MISE is smallest for the convolution estimator. For model 1 the variance of the error terms is very small, and the kernel density estimator is best. This is not unexpected, since the convolution effect will not be large here. In fact, the estimates obtained by the convolution estimator in this model are extremely wiggly and almost useless.

In the non-linear models 5 and 6 with normally distributed error terms, the convolution estimator is much better. Also, for model 11 and 12 the results are good. But introducing asymmetric and multimodal distributions for the error terms, as in model 8, 9 and 10, the convolution estimator deteriorates. In model 8, the error distribution is difficult to estimate, but the distributions f_Y and f_X is of much smoother form. Hence the kernel density estimator could be expected to be better. Figure 3 shows one simulation of sample size 500 from this model. In the upper plot, the simulated values X_i and Y_i are given as points, and the estimated $\tilde{g}$ is depicted as the solid curve. Three bands can be discerned in the scatter diagram and the regression estimator is poor. The plot in the middle shows the estimated error terms, $\tilde{e}_i$. There are clear indications of multimodality. In the lower plot, the "true" density is given as the thick solid curve, the kernel density estimate is given as the solid curve and the convolution estimate as the dashed curve. The convolution estimator have several modes and thus behaves worse than the kernel density estimator. Similar problems occurs for model 9. These results corresponds to analogous results found in Saavedra & Cao (1999*a*) for the estimation of the marginal density in a moving average process.

The variance for the convolution estimator is smaller for the majorities of the simulated examples, and when the sample size increases, the improvements are also increasing. This is consistent with the asymptotic analysis of Section 3, but note that there are several terms of similar order in the asymptotic expansion, and n has to be quite large for the leading term to dominate.

The squared bias is smallest in almost all cases for the convolution estimator. This comes as a somewhat unexpected bonus of our method, since from the asymptotic analysis the bias is of the same order as for the kernel estimator. Figure 4 shows the estimated variance and bias for the two estimators from the simulations for model 2 with sample size 100. The upper plot shows that the variance for the convolution estimator is smallest, as expected. The bias for the kernel density estimator is,

$$\mathrm{E}\big(f_Y^*(x)\big) - f_Y(x) = \frac{h^2}{2} f_Y''(x) \int w^2 K(w) \mathrm{d}w + o(h^2), \tag{15}$$

and the plot in figure 4 is as expected, since the bias is proportional to the second derivative of the density in question, here a normal distribution with mean equal to one. The bias for the convolution estimator behaves quite differently, and overall it is considerably smaller. This difference can be explained by the following reasoning.

Since f_X and f_e are normal distributions, it also means that the true f_Y will be normal with mean equal to one and variance equal to two. From this information it is possible to calculate the expressions for the bias of the convolution estimator and compare it to the observed bias in figure 4. The bias of the convolution estimator consists of three terms, as seen from (7) and (10). Ignoring higher order terms, equation (7) is now

$$\mathrm{E}\big(\tilde{f}_Y(x)\big) - f_Y(x) = \frac{h^2}{2} f_Y''(x) \int w^2 K(w)\mathrm{d}w$$

$$= \frac{h^2}{2} \int w^2 K(w)\mathrm{d}w \Big[-\frac{1}{4\sqrt{\pi}} \exp\{-1/4(x-1)^2\}(1 - \frac{(x-1)^2}{16\sqrt{\pi}})\Big]. \tag{16}$$

This expression is identical to the bias for the kernel density estimator.

In equation (10) the bias of the Nadaraya-Watson estimator of $g(x)$ is a part of the expression. This bias is well-known and its leading term is

$$\mathrm{E}\big(\tilde{g}(x)\big) - g(x) = h^2 \Big(\frac{1}{2} g''(x) + \frac{g'(x) f_X'(x)}{f_X(x)}\Big) \int u^2 K(u)\mathrm{d}u.$$

Since $\int u^2 K(u)\mathrm{d}u = 1$, $g(x) = x$ and f_X is a normal distribution, this expression equals $-2(x-1)h^2$. Inserting this in the first term of the right hand side of (10) and again using the fact that f_X is normal with mean and variance equal to one, gives for the leading term,

$$-f_Y'(x) \int z_1 K'(z_1)\mathrm{d}z_1 \int \big(-2(x_2-1)h^2 \frac{1}{\sqrt{2\pi}} \exp(-(x_2-1)^2/2)\big)\mathrm{d}x_2.$$

Observe that the last integral in this expression equals zero.

Further, the second term on the right hand side of (10) yields,

$$h^2 \int z_2 K'(z_2)\mathrm{d}z_2 \int \big((-2(x_1-1)(\frac{-(x-x_1)}{\sqrt{2\pi}} \exp(-(x-x_1)^2/2)) \\ \times \frac{1}{\sqrt{2\pi}} \exp(-(x_1-1)^2/2)\big)\mathrm{d}x_1.$$

If we choose to use a Gaussian kernel function with mean zero and variance one, then $\int z_2 K'(z_2)\mathrm{d}z_2$ is equal to minus one. Thus, the leading term of the bias of the convolution estimator in this case is

$$\mathrm{E}\big(\hat{f}_Y(x)\big) - f_Y(x) = \frac{h^2}{2}\Big[-\frac{1}{4\sqrt{\pi}} \exp\{-1/4(x-1)^2\}(1 - \frac{(x-1)^2}{16\sqrt{\pi}})\Big]$$

$$-h^2 \int \big((-2(x_1-1)(\frac{-(x-x_1)}{\sqrt{2\pi}} \exp(-(x-x_1)^2/2)) \frac{1}{\sqrt{2\pi}} \exp(-(x_1-1)^2/2)\big)\mathrm{d}x_1.$$

This expression is plotted in figure 5, with a reasonable choice for the bandwidth, $h = 0.3$. And taking the different scaling into account, this graph compares well to the empirical bias from figure 4. Similar explanations are possible for the other models, although problems arise in the computation in the cases where the true density is not known.

Other nonparametric regression estimators may be used to estimate $g(x)$. In table 2, results from simulations from model 2, using the local linear estimator for estimating $g(x)$ are given. For smaller sample sizes these results are better than the corresponding results using the Nadaraya-Watson estimator, given in table 1.

A real data set has been considered as well. It is the motorcycle data, from Härdle (1990) page 70. The X-values represent time after a simulated impact with motorcycles and the response variable Y is the head acceleration of a post human test object. The density of the response Y has been estimated by the kernel density estimator, where the bandwidth is the rule-of-thumb given in Härdle (1990), and the convolution estimator. The estimated densities are given in figure 6. The convolution estimator smooths more than the kernel density estimator, but both estimators seem to give reasonable results.

5. Conclusions

The proposed convolution density estimator substantially outperforms the usual kernel estimator in the majority of cases examined by us, especially if the error term density function is smooth and has a relatively large variance. We believe that the situations where it does not perform so well are of less practical importance.

One should expect that if the g-function is more correctly estimated, then a better density estimate will be obtained. Thus using e.g. local polynomial regression may improve the density estimation, as is indicated by table 2. Also, by selecting the bandwidth parameters in the convolution estimator in a more optimal way, by e.g. a cross-validation technique, one could possibly improve the estimates even more. We also believe that this estimator can be used in a more general time-series setting where $X_t = g(X_{t-1}) + e_t$, and the marginal density of the process X_t is of interest. Some simulation experiments indicate that the convolution estimator will outperform the kernel density estimator and preliminary theoretical derivations show that the order of the variance of the convolution estimator will again be n^{-1}, cf. Støve & Tjøstheim (2007*a*).

Appendix A: Proofs

Proof of theorem 1. Consider the bias term first. Since X_i and e_j are independent for all i and j,

$$\begin{aligned} \mathrm{E}\big(\tilde{f}_Y(x)\big) &= \frac{1}{n^2h}\mathrm{E}\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} K\big(\frac{x-g(X_i)-e_j}{h}\big)\Big] \\ &= \frac{1}{h}\mathrm{E}\Big[K\big(\frac{x-g(X)-e}{h}\big)\Big]. \end{aligned}$$

Further, by a change of variable, the convolution property and Taylor expansion, we obtain

$$\begin{aligned} \frac{1}{h}\mathrm{E}\Big[K\Big(\frac{x-g(X)-e}{h}\Big)\Big] &= \frac{1}{h}\iint K\Big(\frac{x-g(v)-u}{h}\Big) f_X(v) f_e(u)\mathrm{d}v\,\mathrm{d}u \\ &= \iint K(w) f_X(v) f_e(x-g(v)-hw)\mathrm{d}v\mathrm{d}w \\ = \int K(w) f_Y(x-hw)\mathrm{d}w &= f_Y(x) + \frac{h^2}{2} f_Y''(x) \int w^2 K(w)\mathrm{d}w + O(h^4), \end{aligned}$$

and (7) is proved.

The variance term can be decomposed into several covariance terms; see a similar argument in Saavedra & Cao (2000),

$$\operatorname{var}\big(\tilde{f}_Y(x)\big) = \frac{1}{n^4h^2}\operatorname{var}\Big[\sum_{i=1}^n\sum_{j=1}^n K\Big(\frac{x-g(X_i)-e_j}{h}\Big)\Big]$$

$$= \frac{1}{n^4h^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n \operatorname{cov}\Big(K(\frac{x-g(X_i)-e_j}{h}),K(\frac{x-g(X_k)-e_l}{h})\Big)$$

$$= \frac{1}{n^4h^2}\Big[n\operatorname{var}\Big(K(\frac{x-g(X_1)-e_1}{h})\Big) \tag{17}$$

$$+n(n-1)\operatorname{var}\Big(K(\frac{x-g(X_1)-e_2}{h})\Big) \tag{18}$$

$$+n(n-1)(n-2)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_1)-e_3}{h})\Big) \tag{19}$$

$$+2n(n-1)(n-2)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_3)-e_1}{h})\Big) \tag{20}$$

$$+n(n-1)(n-2)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_3)-e_2}{h})\Big) \tag{21}$$

$$+n(n-1)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_2)-e_1}{h})\Big) \tag{22}$$

$$+2n(n-1)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_1}{h}),K(\frac{x-g(X_1)-e_2}{h})\Big) \tag{23}$$

$$+2n(n-1)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_1}{h}),K(\frac{x-g(X_2)-e_1}{h})\Big) \tag{24}$$

$$+n(n-1)(n-2)(n-3)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_3)-e_4}{h})\Big) \tag{25}$$

$$+n(n-1)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_1}{h}),K(\frac{x-g(X_2)-e_2}{h})\Big) \tag{26}$$

$$+n(n-1)(n-2)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_1}{h}),K(\frac{x-g(X_2)-e_3}{h})\Big) \tag{27}$$

$$+n(n-1)(n-2)\operatorname{cov}\Big(K(\frac{x-g(X_2)-e_1}{h}),K(\frac{x-g(X_1)-e_3}{h})\Big)\Big]. \tag{28}$$

By independence the terms (20), (22), (25), (26), (27) and (28) are equal to zero, and we just have to examine the remaining terms. In the following the derivations for the contributing terms, (19) and (21), are shown. One example of a non-contributing term, (17), is also included. The derivations of the other terms are similar, see Støve & Tjøstheim (2007*c*).

We start by examining the expression (19),

$$\begin{aligned}
&\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_1)-e_3}{h})\Big)\\
&=\mathrm{E}\Big(K(\frac{x-g(X_1)-e_2}{h})K(\frac{x-g(X_1)-e_3}{h})\Big)\\
&-\mathrm{E}\Big(K(\frac{x-g(X_1)-e_2}{h})\Big)\mathrm{E}\Big(K(\frac{x-g(X_1)-e_3}{h})\Big).
\end{aligned}$$

By change of variables and Taylor expansion,

$$\begin{aligned}
&\mathrm{E}\Big(K(\frac{x-g(X_1)-e_2}{h})K(\frac{x-g(X_1)-e_3}{h})\Big)\\
&=\iiint K(\frac{x-g(v)-u_1}{h})K(\frac{x-g(v)-u_2}{h})f_X(v)f_e(u_1)f_e(u_2)\mathrm{d}v\mathrm{d}u_1\mathrm{d}u_2\\
&=h^2\iiint K(z_1)K(z_2)f_X(v)f_e(x-g(v)-z_1h)f_e(x-g(v)-z_2h)\mathrm{d}v\mathrm{d}z_1\mathrm{d}z_2\\
&=h^2\Big[\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v+O(h^2)\Big].
\end{aligned}$$

The second term in the covariance expression is, using exactly the same techniques,

$$\begin{aligned}
&\mathrm{E}\Big(K(\frac{x-g(X_1)-e_2}{h})\Big)\mathrm{E}\Big(K(\frac{x-g(X_1)-e_3}{h})\Big)\\
&=\Big[h\big[f_Y(x)+\frac{h^2}{2}f_Y''(x)\int z^2K(z)\mathrm{d}z+o(h^2)\big]\Big]^2=h^2f_Y^2(x)+O(h^4).
\end{aligned}$$

In total this gives,

$$\begin{aligned}
&n(n-1)(n-2)\operatorname{cov}\Big(K(\frac{x-g(X_1)-e_2}{h}),K(\frac{x-g(X_1)-e_3}{h})\Big)\\
&=n(n-1)(n-2)\Big[h^2\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v-h^2f_Y^2(x)+O(h^4)\Big]. \qquad (29)
\end{aligned}$$

For (21), using the assumption that the inverse of $g(\cdot)$ exists we obtain

$$\begin{aligned}
&\mathrm{E}\Big(K(\frac{x-g(X_1)-e_2}{h})K(\frac{x-g(X_3)-e_2}{h})\Big)\\
&=\iiint K(\frac{x-g(v)-u}{h})K(\frac{x-g(w)-u}{h})f_e(u)f_X(v)f_X(w)\mathrm{d}u\mathrm{d}v\mathrm{d}w\\
&=h^2\iiint K(z_1)K(z_2)f_X(v)l_X(g(v)+h(z_1-z_2))f_e(x-g(v)-hz_1)\\
&\qquad\times r\big(g(v)+h(z_1-z_2)\big)\mathrm{d}v\mathrm{d}z_1\mathrm{d}z_2\\
&=h^2\int r\big(g(v)\big)f_X(v)f_e\big(x-g(v)\big)l_X\big(g(v)\big)\mathrm{d}v+O(h^4),
\end{aligned}$$

where $(g^{-1})' = r$ and $f_X(g^{-1}) = l_X$. Note that

$$r(v) = \frac{\mathrm{d}}{\mathrm{d}v}\big(g^{-1}(v)\big) = \frac{1}{g'\big(g^{-1}(v)\big)},$$

and $g^{-1}\big(g(v)\big) = v$. Thus

$$r\big(g(v)\big) = \frac{1}{g'(v)}$$

and

$$l_X\big(g(v)\big) = f_X\big(g^{-1}(g(v))\big) = f_X(v).$$

As before,

$$\mathrm{E}\Big(K(\frac{x - g(X_1) - e_2}{h})\Big)\mathrm{E}\Big(K(\frac{x - g(X_3) - e_1}{h})\Big) = h^2 f_Y^2(x) + O(h^4),$$

and hence

$$\begin{aligned} &n(n-1)(n-2)\mathrm{cov}\Big(K(\frac{x - g(X_1) - e_2}{h}), K(\frac{x - g(X_3) - e_2}{h})\Big) \\ &= n(n-1)(n-2)\Big[h^2 \int \frac{f_X^2(v) f_e\big(x - g(v)\big)}{g'(v)}\mathrm{d}v - h^2 f_Y^2(x) + O(h^4)\Big]. \end{aligned} \tag{30}$$

The expressions (29) and (30) give the three leading terms in the theorem.

To consider one example of a non-contributing term, we turn to the expression (17),

$$\mathrm{var}\Big(K(\frac{x - g(X_1) - e_1}{h})\Big) = \mathrm{E}\Big(K^2(\frac{x - g(X_1) - e_1}{h})\Big) - \Big[\mathrm{E}\Big(K(\frac{x - g(X_1) - e_1}{h})\Big)\Big]^2.$$

By change of variables, convolution and Taylor expansion,

$$\begin{aligned} \mathrm{E}\Big(K^2(\frac{x - g(X_1) - e_1}{h})\Big) &= \iint K^2(\frac{x - g(v) - u}{h}) f_X(v) f_e(u)\mathrm{d}v\mathrm{d}u \\ = h \iint K^2(z) f_X(v) f_e(x - g(v) - hz)\mathrm{d}v\mathrm{d}z &= h \int K^2(z) f_Y(x - zh)\mathrm{d}z \\ &= h \int K^2(z)\,[f_Y(x) - hz f_Y'(x) + \frac{h^2 z^2}{2} f_Y''(x)]\,\mathrm{d}z + O(h^4) \\ &= h f_Y(x) \int K^2(z)\mathrm{d}z + \frac{h^3}{2} f_Y''(x) \int z^2 K^2(z)\mathrm{d}z + O(h^4). \end{aligned}$$

Using exactly the same techniques,

$$\Big[\mathrm{E}\Big(K(\frac{x - g(X_1) - e_1}{h})\Big)\Big]^2 = \Big[h\big[f_Y(x) + \frac{h^2}{2} f_Y''(x) \int z^2 K(z)\mathrm{d}z + o(h^2)\big]\Big]^2,$$

and hence

$$
\begin{aligned}
&n\mathrm{var}\Big(K(\frac{x-g(X_1)-e_1}{h})\Big)\\
&= n\Big[hf_Y(x)\int K^2(z)\mathrm{d}z - h^2 f_Y^2(x) + O(h^2)\Big].
\end{aligned}
$$

The other non-contributing terms is derived similarly, see Støve & Tjøstheim (2007*c*). Adding all the expressions stemming from (17)-(28), we get the variance expression (8) in the theorem.

Proof of Theorem 2. Consider the estimator $\hat{f}_Y(x)$ in (5). By substituting for $\tilde{e}_j$, Taylor expanding $K(\cdot)$ around $(x - g(X_i) - e_j)/h$ and using the mean value theorem, we obtain,

$$
\begin{aligned}
\hat{f}_Y(x) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh}\sum_{j=1}^{n}K(\frac{x-\tilde{g}(X_i)-\tilde{e}_j}{h})\Big]\\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh}\sum_{j=1}^{n}K(\frac{x-g(X_i)-e_j+\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))}{h})\Big]\\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh}[\sum_{j=1}^{n}K(\frac{x-g(X_i)-e_j}{h}) + K'(\frac{x-g(X_i)-e_j}{h})\\
&\quad\times(\frac{\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))}{h}) + A_n(\xi)]\Big],
\end{aligned}
$$

where for some ξ determined by the mean value theorem

$$
\begin{aligned}
A_n(\xi) &= K''(\xi)\Big(\frac{\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))}{h}\Big)^2\\
&\le M\cdot\Big(\frac{\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))}{h}\Big)^2. \qquad (31)
\end{aligned}
$$

Here M is a constant determined by condition A1. Thus,

$$
\begin{aligned}
&\hat{f}_Y(x) - \tilde{f}_Y(x) =\\
&\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh}\sum_{j=1}^{n}[K'(\frac{x-g(X_i)-e_j}{h})\cdot\frac{\tilde{g}(X_j)-g(X_j)}{h}\\
&\quad +K'(\frac{x-g(X_i)-e_j}{h})\cdot\frac{g(X_i)-\tilde{g}(X_i)}{h}]\Big]\\
&+\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{nh}(\sum_{j=1}^{n}K''(\xi)(\frac{\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))}{h})^2)\Big]. \qquad (32)
\end{aligned}
$$

Let us first examine the last term of (32). We observe that for $i = j$ it is zero. For $i \neq j$, if we denote by F_n the joint empirical distribution function of X_i and X_j. We have

$$
\begin{aligned}
&\frac{1}{h^3}\Big|\frac{1}{n}\sum_{i=1}^{n}\big[\frac{1}{n}\sum_{j=1}^{n}\big(\tilde{g}(X_j) - g(X_j) - (\tilde{g}(X_i) - g(X_i))\big)\big]^2\Big| \\
&\leq \frac{1}{h^3}\iint \sup_{x_1\in S(X), x_2\in S(X)}\big(\tilde{g}(x_2) - g(x_2) - (\tilde{g}(x_1) - g(x_1))\big)^2 \mathrm{d}F_n(x_1)\mathrm{d}F_n(x_2) \\
&\leq \frac{1}{h^3}\sup_{x_1\in S(X), x_2\in S(X)}\big(\tilde{g}(x_2) - g(x_2) - (\tilde{g}(x_1) - g(x_1))\big)^2 \iint \mathrm{d}F_n(x_1)\mathrm{d}F_n(x_2) \\
&\leq \frac{1}{h^3}\sup_{x_1\in S(X)}\big(g(x_1) - \tilde{g}(x_1)\big)^2 + \sup_{x_1\in S(X), x_2\in S(X)}\Big|2\big(g(x_1) - \tilde{g}(x_1)\big) \\
&\qquad \times\big(\tilde{g}(x_2) - g(x_2)\big)\Big| + \sup_{x_2\in S(X)}\big(\tilde{g}(x_2) - g(x_2)\big)^2. \qquad (33)
\end{aligned}
$$

Using a uniform convergence result for the Nadaraya-Watson estimator and making use of assumptions A5, A6, A9 and A10; see Mack & Silverman (1982),

$$
\sup_{x\in S(X)}|\tilde{g}(x) - g(x)| = O_P\Big(\big[\frac{1}{nh_R}\log(\frac{1}{h_R})\big]^{1/2}\Big),
$$

and hence the order in probability of the expression in (33) is

$$
O_P\Big(\frac{1}{nh_R \cdot h^3}\log(\frac{1}{h_R})\Big), \qquad (34)
$$

where h_R is defined in (4). Using the same argument as when evaluating (33) it will also be seen that the mean of the absolute value and the standard deviation of this term is of the order given in (34). (See below for the existence of these quantities under the assumption $\inf_{x\in S(X)} f_X(x) > 0$.)

Next, we examine the first order term of (32). We start by looking at the expectation of this term. For the expectation to exist we need the existence of $\mathrm{E}(\tilde{g}(X_i) - g(X_i))$, but using the definition of the Nadaraya-Watson estimator, this follows from condition A8. We again note that the expectation disappears for $i = j$, and for $i \neq j$ we have, using independence, for the first part of the first order term

$$
\begin{aligned}
&\mathrm{E}\Big(\frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n}K'\big(\frac{x - g(X_i) - e_j}{h}\big)\cdot\big(\frac{\tilde{g}(X_j) - g(X_j)}{h}\big)\Big) \\
&\sim \frac{1}{h^2}\iiint K'\big(\frac{x - g(x_1) - u}{h}\big)\big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big)f_e(u)f_X(x_1)f_X(x_2)\mathrm{d}u\mathrm{d}x_1\mathrm{d}x_2.
\end{aligned}
$$

Now Taylor expanding and using a convolution argument, we obtain

$$\frac{1}{h}\iiint K'(z_1)\big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_e(x - g(x_1) - z_1 h) f_X(x_1) f_X(x_2) \mathrm{d}z_1 \mathrm{d}x_1 \mathrm{d}x_2$$
$$= \frac{1}{h}\iint K'(z_1)\big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_Y(x - z_1 h) f_X(x_2) \mathrm{d}z_1 \mathrm{d}x_2 =$$
$$-f_Y'(x)\int z_1 K'(z_1)\mathrm{d}z_1 \int \big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_X(x_2)\mathrm{d}x_2$$
$$+\frac{h}{2} f_Y''(x)\int z_1^2 K'(z_1)\mathrm{d}z_1 \int \big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_X(x_2)\mathrm{d}x_2 + O(h^4). \tag{35}$$

Observe that since the kernel is symmetric, $\int z_1^2 K'(z_1)\mathrm{d}z_1 = 0$, so there is no term of order $O(h^3)$. The whole term is of order $O(h^2)$ through the dependence on $\mathrm{E}\big(\tilde{g}(x_2) - g(x_2)\big)$.

Examining the second part of the first order term in (32), by similar arguments

$$-\mathrm{E}\Big(\frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K'\big(\frac{x - g(X_i) - e_j}{h}\big)\cdot\big(\frac{\tilde{g}(X_i) - g(X_i)}{h}\big)\Big)$$
$$\sim \int z_2 K'(z_2)\mathrm{d}z_2 \int \big(\mathrm{E}(\tilde{g}(x_1)) - g(x_1)\big) f_e'\big(x - g(x_1)\big) f_X(x_1)\mathrm{d}x_1 + O(h^4). \tag{36}$$

In total, the terms (35) and (36) are of order h^2, but can be reduced by higher order kernels.

Further, we examine the variance of the first order term in (32). The condition A8 again guarantees the existence of this variance. The calculations are similar to the calculations where we found the variance of $\tilde{f}_Y(x)$. The variance in question is

$$\mathrm{var}\Big(\frac{1}{n^2 h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[K'\big(\frac{x - g(X_i) - e_j}{h}\big)$$
$$\times\big(g(X_i) - \tilde{g}(X_i) + \tilde{g}(X_j) - g(X_j)\big)\Big]\Big) = \frac{1}{n^4 h^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}$$
$$\mathrm{cov}\Big[K'\big(\frac{x - g(X_i) - e_j}{h}\big)\big(g(X_i) - \tilde{g}(X_i) + \tilde{g}(X_j) - g(X_j)\big),$$
$$K'\big(\frac{x - g(X_k) - e_l}{h}\big)\big(g(X_k) - \tilde{g}(X_k) + \tilde{g}(X_l) - g(X_l)\big)\Big]. \tag{37}$$

The evaluation of these terms can be found in Støve & Tjøstheim (2007*c*). It turns out that they are of order $O(h^4/n)$, thus they will only contribute higher order effects to the overall variance of $\hat{f}_Y(x)$. It follows that under condition A7 the first order term of the Taylor expansion in (32) dominates the second order term.

Remark. There is a potential to improve on (34), since the evalution in (33)-(34) is quite crude. An alternative would be to try to evaluate the second order term directly, as was done for the first order term in the Taylor expansion (32), using the convolution property, and then include a third order term which can be evaluated (crudely) as above, resulting in a term $O_p\left(\left[\frac{1}{nh_R h^3}\log\frac{1}{h_R}\right]^{3/2}\right)$. For the crude estimate (34) to be of order $O(\frac{1}{\sqrt{n}})$, we must have

$$h_R h^3 = O(n^{-1/2-\epsilon}) \text{ for some } \epsilon > 0. \tag{38}$$

References

Birgè, L. & Massart, P. (1995), 'Estimation of integral functionals of a density', *Annals of Statistics* **23**, 11–29.

Efromovich, S. & Samarov, A. (2000), 'Adaptive estimation of the integral of squared regression derivatives', *Scandinavian Journal of Statistics* **27**, 335–351.

Fan, J. (1992), 'Design-adaptive nonparametric regression', *Journal of the American Statistical Association* **87**, 998–1004.

Frees, E. W. (1994), 'Estimating densities of functions of observations', *Journal of the American Statistical Association* **89**, 517–525.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-based approach to calculating marginal densities', *Journal of The American Statistical Association* **85**, 398–409.

Hall, P. & Marron, J. S. (1987), 'Estimation of integrated squared density derivatives', *Statistics and Probability Letters* **6**, 109–115.

Härdle, W. (1990), *Smoothing Techniques: With Implementation in S*, Springer-Verlag.

Mack, Y. P. & Silverman, B. W. (1982), 'Weak and strong uniform consistency of kernel regression estimates', *Zeitschrift für Wahrscheinlichkeitstheorie verw. Gebiete* **61**, 405–415.

Saavedra, A. & Cao, R. (1999*a*), 'A comperative study of two convolution-type estimators of the marignal density of moving average processes', *Computational Statistics* **14**, 355–373.

Saavedra, A. & Cao, R. (1999*b*), 'Rate of convergence of a convolution-type estimator of the marginal density of a MA(1) process', *Stochastic Processes and their Applications* **80**, 129–155.

Saavedra, A. & Cao, R. (2000), 'On the estimation of the marginal density of a moving average process', *The Canadian Journal of Statistics* **28**, 799–815.

Schick, A. & Wefelmeyer, W. (2004*a*), 'Root n consistent and optimal density estimators for moving average processes', *Scandinavian Journal of Statistics* **31**, 63–78.

Schick, A. & Wefelmeyer, W. (2004*b*), 'Root n consistent density estimators for sums of independent random variables', *Jounal of Nonparametric Statistics* **16**, 925–935.

Schick, A. & Wefelmeyer, W. (2007), 'Uniformly root-n consistent density estimators for weakly dependent invertible linear processes', *Annals of Statistics* **31**, 63–78.

Sheather, S. J. & Jones, M. C. (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society. Series B* **53**, 683–690.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer-Verlag.

Støve, B. & Tjøstheim, D. (2007*a*), 'A convolution density estimator for nonlinear time series: Simulations and some preliminary analysis', Work in progress.

Støve, B. & Tjøstheim, D. (2007*b*), 'A new convolution estimator for nonparametric regression', *in* V. Nair, ed., Advances in Statistical Modeling and Inference. World Scientific, 363-384.

Støve, B. & Tjøstheim, D. (2007*c*), 'Some proofs for the convolution estimator for the density of nonlinear regression observations', Technical report. Available on *www.nhh.no/for/cv/stove-bard.htm*.

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

Model	Sample size	Squared bias	Variance	MISE
1	100	85.4 %	-706.8 %	-582.1 %
1	500	89.9 %	-498.1 %	-385.8 %
1	5000	-225.9 %	-278.0 %	-267.7 %
2	100	85.7 %	30.8 %	38.0 %
2	500	99.2 %	50.6 %	59.8 %
2	5000	95.6 %	71.0 %	75.9 %
3	100	90.2 %	-96.3 %	-68.6 %
3	500	94.0 %	-41.7 %	-14.7 %
3	5000	94.2 %	16.7 %	32.8 %
4	100	62.1 %	19.7%	27.5 %
4	500	68.2 %	1.6 %	17.7 %
4	5000	80.3 %	21.9 %	32.5 %
5	100	89.7 %	24.9 %	34.7 %
5	500	85.7 %	44.3 %	52.5 %
5	5000	80.7 %	62.6 %	66.4 %
6	100	84.6 %	0.4 %	19.6 %
6	500	82.6 %	33.9 %	44.4 %
6	5000	77.4 %	57.2 %	62.1 %
7	100	99.4 %	19.2 %	29.4 %
7	500	98.5 %	25.9 %	42.5 %
7	5000	86.8 %	47.5 %	56.4 %

Model	Sample size	Squared bias	Variance	MISE
8	100	-18.5 %	1.7 %	-3.1 %
8	500	13.6 %	-130.3 %	-96.9 %
8	5000	54.5 %	-319.4 %	-259.6 %
9	100	49.1 %	-37.9 %	-9.1 %
9	500	58.1 %	-33.2 %	-0.8 %
9	5000	57.5 %	-38.5 %	-20.1 %
10	100	-57.4 %	-0.5 %	-18.9 %
10	500	9.3 %	-21.7 %	-11.7 %
10	5000	57.7 %	-3.5 %	11.8 %
11	100	19.6 %	-3.5%	1.5 %
11	500	9.0 %	23.7 %	20.3 %
11	5000	-25.1 %	53.8 %	37.2 %
12	100	96.2 %	22.4%	34.9 %
12	500	98.3 %	38.0 %	49.5 %
12	5000	97.7 %	63.4 %	69.8%

Table 1: Percentage improvements in estimations using the convolution density estimator compared with the kernel density estimator. The MISE, squared bias and variance are explained in formula (11), (12), (14). A minus sign indicates that the kernel density estimator performs best.

Sample size	Squared bias	Variance	MISE
100	92.6 %	36.7 %	49.8 %
500	89.2 %	58.9 %	66.2 %
5000	85.8 %	70.9 %	74.3 %

Table 2: Percentage improvements in estimations using convolution density estimator with local linear estimator, compared with kernel density estimator. Model 2.

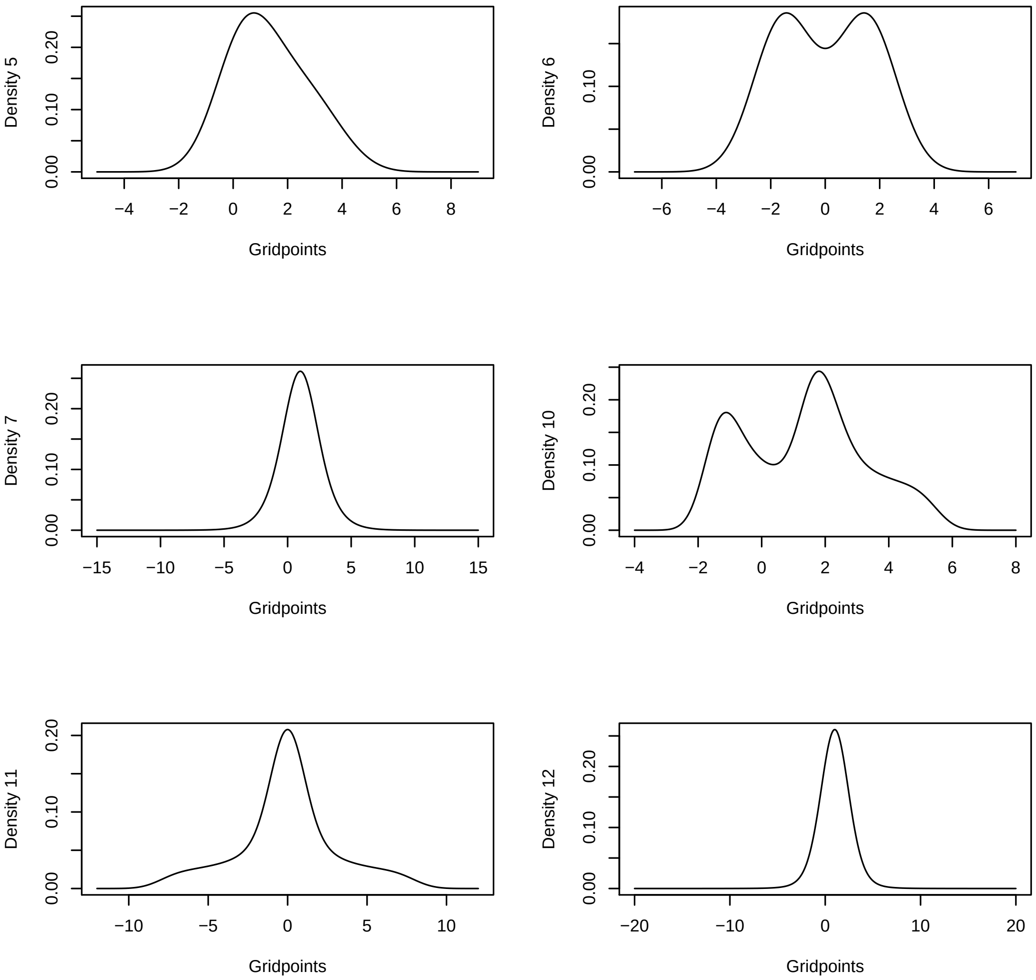

Figure 1: The densities f_Y for models (from top left to bottom right) 5, 6, 7, 10, 11 and 12.

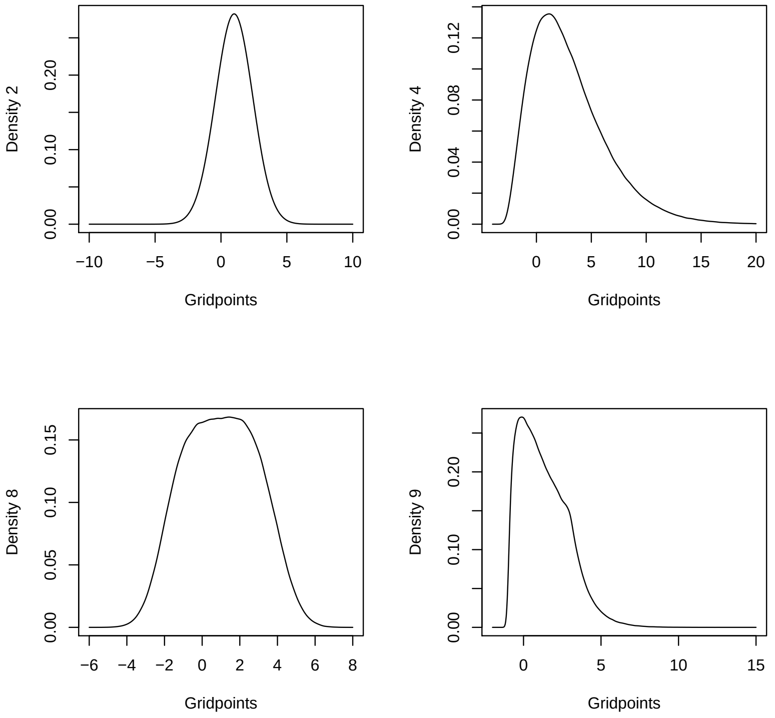

Figure 2: The densities f_Y for models 2, 4, 8 and 9.

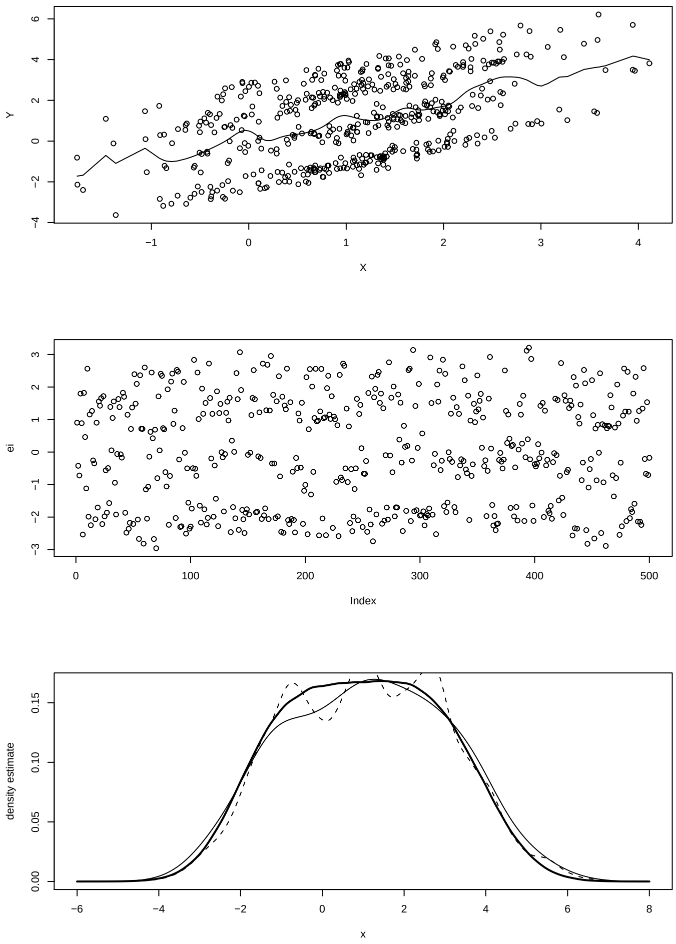

Figure 3: Upper plot: The estimated g function (solid curve) and the simulated points (X_i, Y_i) from one simulation of sample size 500 from model 8. Middle plot: The corresponding estimated e_i. Lower plot: The "true" density f_Y (thick solid curve) and the estimated densities (dashed curve - convolution estimator, solid curve - kernel density estimator).

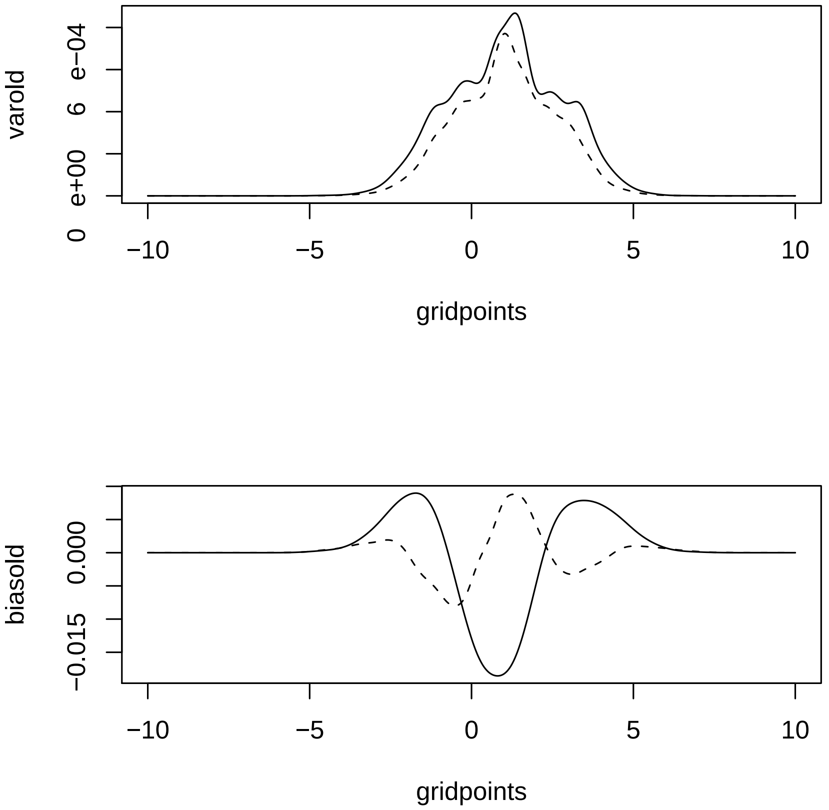

Figure 4: The estimated variance (top) and bias for model 2 (dashed curve - convolution estimator, solid curve - kernel density estimator).

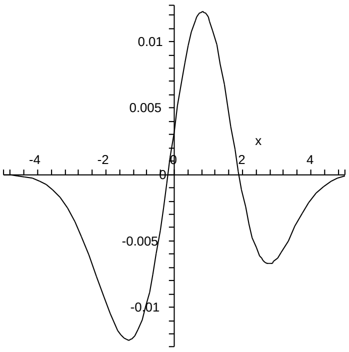

Figure 5: The theoretical bias of the convolution estimator in model 2.

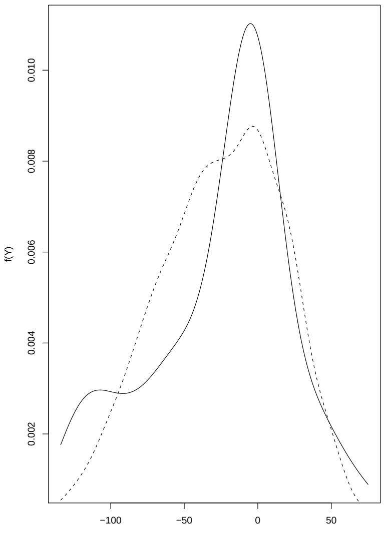

Figure 6: The estimated densities of a real data set (solid line - kernel and dashed line - convolution)