Some New Approaches to Smoothing: Convolution Estimators in Regression Models and Backfitting in Panels of Time Series

Dr.Scient. Thesis in Statistics

Bård Støve

Department of Mathematics
University of Bergen
2005

Abstract

A new estimator, called the convolution estimator, is examined for the density estimation of the responses in a nonlinear regression model, and for the marginal density of nonlinear time series. Asymptotic properties are established and it is found that the convolution estimator has a better rate. Simulations indicate that the proposed estimator outperforms the well-known kernel density estimator in many finite sample cases.

Based on the same convolution idea, an estimator for nonparametric regression is proposed. Both asymptotic analysis and simulation results show that this estimator has an overall better asymptotic bias property than standard nonparametric regression estimators.

An iterative algorithm for estimation of the unknown functions in an additive model for a panel of time series observations is proposed. Some asymptotic properties are given, and simulations indicate that the proposed algorithm works well.

Acknowledgments

This work has been funded by the Norwegian Research Council grant 147231/432.

I could never have completed this thesis without the support from my supervisor, Prof. Dag Tjøstheim. I am grateful for his guidance, enthusiasm and encouragement.

My co-supervisor, Associate Prof. Hans A. Karlsen, has been of great support during my work. I would also like to thank the other academic staff in the statistics group at University of Bergen. Many thanks to Prof. Jiti Gao at the University of Western Australia for giving me the opportunity for a research stay "down under" and to Prof. Enno Mammen at the University of Mannheim for making a great effort in our joint project.

I am grateful to the PhD students and Master students in the statistics group for creating a pleasant environment at Kropeliens.

Finally, thanks to my family and friends for always supporting me.

Bård Støve,
Bergen, August 2005.

Contents

INTRODUCTION

The following sections will briefly introduce nonparametric smoothing methods, and a short section will introduce panel data models, which is the other topic of this thesis. Further, this thesis consist of four papers, and short summaries will be given during the introduction, thereafter, the papers are presented.

1 Nonparametric smoothing

Statistics is the science that deals with the collection, summarization, presentation and interpretation of data. The existence of high speed computing has made it easy to look at data in ways that were once impossible. One area is that of nonparametric density and regression function estimation, or also called smoothing methods.

In nonparametric smoothing the aim is to avoid assumptions on the parametric form of a density function or regression function, and let the data speak for themselves and find a function that describes the available data well. This is in contrast to parametric modelling, where a specific model with parameters is assumed to generate the data in question. In this case, it can be easy to do inference and great gains in efficiency are possible, however, only if the model is (almost) true. If the assumed model is incorrect, inferences can be useless, leading to misleading interpretations of the data.

Nonparametric smoothing provides a simple way to find structures in data sets without imposing a parametric model, the only assumption is that the density or regression function in question is a smooth curve.

2 Kernel density estimation

The literature concerning nonparametric smoothing, which include kernel density estimation, is vast. Over the last three decades, nonparametric smoothing has been one of the most active areas of statistical reasearh. Estimating probability density functions can be considered the simplest data smoothing situation and the following section will describe kernel density estimation briefly. Only a few references are mentioned here, but several of the books referred to give excellent overviews of the large amount of literature in this field.

The problem of estimating the probability density function of a sample of univariate and independent observations, $X_1, ..., X_n$, having common density f, is often encountered. A classical method will be to use a histogram, but an improvement of the histogram method is kernel density estimation, which first appeared in a report by Fix & Hodges (1951). Books on density estimation are for example Härdle (1990), Wand & Jones (1995) and Simonoff (1996).

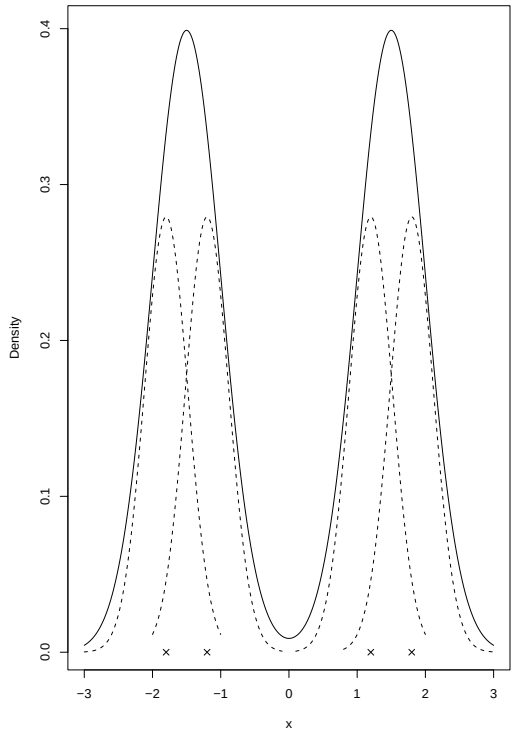

Figure 1: Kernel density estimate

The formula for the univariate kernel density estimator is,

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

where K is a function, called kernel function, satisfying $\int K(x)\mathrm{d}x = 1$, and h is a positive number called the bandwidth or smoothing parameter. Usually K is taken to be a nonnegative symmetric and unimodal probability density function, this ensures that the estimate $\hat{f}(x)$ is also a density.

Figure 1 shows how the kernel density estimator behaves in practice. For each observation X_i the scaled kernel function, here Gaussian, is centered. The value of the kernel estimate is the average of the n kernel ordinates at that point. In regions where there are many observations, this will produce a relatively large value, as expected, and the opposite will occur in regions with few observations. For the sake of clarity, only four observations are used here, but this number is far to small to actually estimate a density reasonably well.

To use the kernel density estimator, one has to choose a kernel and the bandwidth. It is well known that different choices of the kernel function will not have

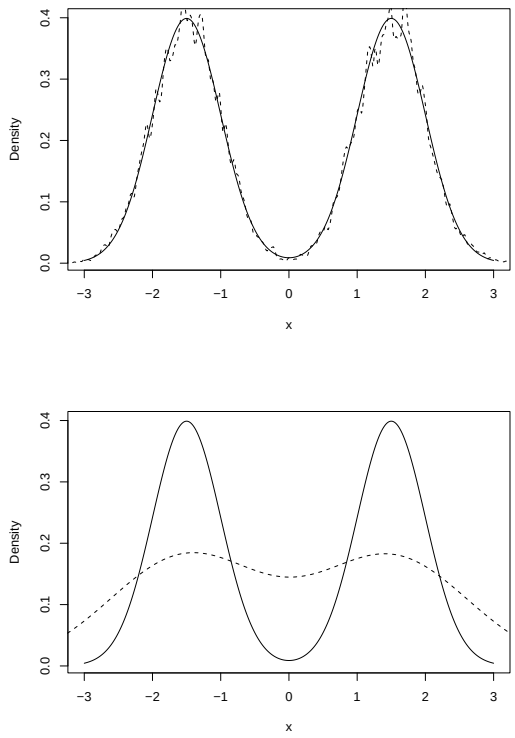

Figure 2: Kernel density estimates with different bandwidhts (solid line - true function, dashed line - the estimated density)

a large influence of the estimator's performance. The bandwidth choice however, is crucial. As figure 2 shows, a large bandwidth will produce an oversmoothed density (lower graph), but a small bandwidth produces a wiggly estimate (upper graph). How to optimally choose the bandwidth will be considered later.

2.1 Properties of the kernel density estimtor

The analysis of the performance of the kernel density estimator is usually based on the size of its mean squared error (MSE), that is,

$$\mathrm{MSE}(\hat{f}(x)) = \mathrm{E}[\hat{f}(x) - f(x)]^2 = \mathrm{Var}[\hat{f}(x)] + [\mathrm{E}(\hat{f}(x) - f(x))]^2. \qquad (2)$$

The variance and bias decompostion makes it easy to interpret the estimators performance. The MSE is a pointwise measure, a global measure is the mean integrated squared error (MISE),

$$\mathrm{MISE}(\hat{f}(x)) = \mathrm{E}[\int (\hat{f}(x) - f(x))^2 \mathrm{d}x]. \qquad (3)$$

Since the MSE and MISE depends on the bandwidth, h, in a complicated way, an approximation of the leading terms in the bias and variance must be calculated. These approximations show in a simple way how the bandwidth actually influence the performance of the estimator, they can be used to obtain rate of convergence of the kernel density estimator, and the approximations can also be used to derive optimal bandwidth choices.

To develop the approximations, the following assumptions is made:

1. The density f has a second derivative f'' which is continous, square integrable and monotone.
2. The bandwidth h satisfies $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.
3. The kernel function K is a bounded density function with $\int xK(x)\mathrm{d}x = 0$ and $\int x^4 K(x)\mathrm{d}x < \infty$.

Simple calculations, see e.g. chapter 2 in Härdle (1990), yield

$$\mathrm{E}(\hat{f}(x)) = f(x) + \frac{1}{2}h^2 f''(x)\int z^2 K(z)\mathrm{d}z + o(h^2), \tag{4}$$

and

$$\mathrm{Var}(\hat{f}(x)) = (nh)^{-1} f(x)\int K^2(z)\mathrm{d}z + o((nh)^{-1}). \tag{5}$$

These two expressions give the MSE of the kernel density estimator,

$$\mathrm{MSE}(\hat{f}(x)) = (nh)^{-1} f(x)\int K^2(z)\mathrm{d}z + \frac{1}{4}h^4 f''(x)^2[\int z^2 K(z)\mathrm{d}z]^2 + o((nh)^{-1}) + o(h^4). \tag{6}$$

From the MSE, to reduce the bias, one should let h get smaller and smaller, but this will increase the variance, since it is proportional to $(nh)^{-1}$. This is known as the bias-variance trade-off. Increasing the number of observations, n, clearly reduces the variance. Observe that the bias is proportional to $f''(x)$, thus the bias is large when the curvature of $f(x)$ is large.

Note that the kernel density estimator is consistent, since $\mathrm{MSE}(\hat{f}(x))$ converges to zero, if $h \to 0$ and $nh \to \infty$, that is

$$\hat{f}(x) \xrightarrow{\mathrm{p}} f(x). \tag{7}$$

Since the MISE can be written,

$$\mathrm{MISE}(\hat{f}(x)) = \int \mathrm{MSE}(\hat{f}(x))\mathrm{d}x, \tag{8}$$

we obtain

$$\text{MISE}(\hat{f}(x)) = \frac{h^4}{4}[\int z^2 K(z)\mathrm{d}z]^2 \int f''(x)^2 \mathrm{d}x + (nh)^{-1} \int K(z)^2 \mathrm{d}z + o((nh)^{-1}) + o(h^4). \quad (9)$$

Minimizing the asymptotic MISE, the so-called A-MISE, obtained by ingoring the higher order terms in (9), with respect to the bandwidth parameter h, results in a bandwidth called the asymptotic optimal bandwidth. This is given by,

$$h_{opt} = [\frac{R(K)}{\mu_2(K)^2 R(f'')n}]^{1/5}, \quad (10)$$

where $\mu_2(K) = \int z^2 K(z)\mathrm{d}z$ and $R(f) = \int f(x)^2 \mathrm{d}x$. However, this bandwidth depends on an uknown parameter $\int f''(x)^2 \mathrm{d}x$. In the next section we will introduce methods that handle this problem. Observe that this bandwidth gives the rate of convergence of $\hat{f}(x)$, substitute h_{opt} in the A-MISE gives,

$$\text{A-MISE}(\hat{f}(x)) = \frac{5}{4}[\mu_2(K)^2 R(K)^4 R(f'')]^{1/5} n^{-4/5}. \quad (11)$$

In other words, the convergence speed of the A-MISE of the kernel density estimator is $(nh)^{-1/2} = n^{-2/5}$, which is slower then the typical rate of convergence of $n^{-1/2}$, for parametric models. The asymptotic MSE and MISE were studied by Rosenblatt (1956), Parzen (1962) and Watson & Leadbetter (1963).

2.2 Bandwidth selection

The choice of the bandwidth h is a main problem in kernel density estimation. Various methods exist, and here just a few methods will be mentioned. See the aforementioned books for details and other methods.

One simple bandwidth rule is based on the asymptotic optimal bandwidth in equation (10). By assuming that f is a Gaussian density with standard deviation σ, then

$$R(f'') \approx 0.212\sigma^{-5}. \quad (12)$$

Thus we can estimate $R(f'')$ by using an estimator $\hat{\sigma}$, the sample standard estimator, for σ. Further, if K is a Gaussian kernel, then we obtain the normal reference bandwidth, see for example Bickel & Doksum (1977) and Silverman (1986).

$$\hat{h}_{opt} \approx 1.06\hat{\sigma}n^{-1/5}. \quad (13)$$

This bandwidth can be a good choice if the data are nearly Gaussian distributed. However, it may oversmooth the estimate if the data are not nearly Gaussian distributed.

Other bandwidth selectors include methods based on cross-validation ideas; least squares cross-validation, introduced by Rudemo (1982) and Bowman (1984); maximum likelihood cross-validation, see Habema et al. (1974) and Duin (1976) and biased cross-validation; introduced by Scott & Terrell (1987).

The normal reference bandwidth was obtained by assuming f to be a Gaussian density, and thus we are able to estimate $R(f'')$. Another approach is to estimate $R(f'')$ directly, and plug-in this estimate in the formula for the asymptotically optimal bandwidth. Thus these bandwidths methods are known as Plug-In mehtods. A good implementation of this method is proposed by Sheather & Jones (1991). This bandwidth selector is also used in several of the simulation experiments conducted in this thesis.

2.3 Modifications of the kernel density estimator

Many authors have proposed adjustments and improvements to the kernel density estimator, and a few will be mentioned here. One difficulty with the estimator is that, while it gives good estimates for many different density shapes, it can be inadequate for other shapes. Particulary since just a single smoothing parameter, the bandwidth, is used over the entire real line. Thus the idea of using variable bandwidths was proposed by Victor (1976) and Breiman et al. (1977).

If the random sample has a density, f, that is difficult to estimate then another possibility is to apply a transformation of the data to obtain a new sample with a density, g, more easily estimated using the standard kernel density estimator. Then backtransform this estimated density, $\hat{g}$, to obtain the density in question, $\hat{f}$. This estimator is called transformation kernel density estimator, and were studied by Wand et al. (1991) and Yang & Marron (1999).

Hjort & Glad (1995), Efron & Tibshirani (1996) and Glad (1998) have studied the possibility of parametrically guided nonparametric density and regression estimation.

In many situations the density in question will have a bounded support. When using the kernel density estimator it will spread point masses smoothly around the observed data points, and some of those near the boundary of the support are distributed outside the support of the density. As a result, the kernel density estimator underestimates the density in boundary regions. Two approaches which deals with this problem is, using boundary kernels, which are weight functions only used within the boundary region; see Gasser & Müller (1979) and Gasser et al. (1985), and the reflection method; see Schuster (1985) and Hall & Wehrly (1991).

Several authors have studied the use of higher order kernels to improve the asymptotic bias; see e.g. Marron & Wand (1992), for a quantification of the practical gain in density estimation. To define a higher order kernel, let

$$\mu_j(K) = \int z^j K(z) \mathrm{d}z$$

be the jth moment of the kernel K. Then we say that K is a $kth-order$ kernel if

$$\mu_0(K) = 1,\ \mu_j(K) = 0 \text{ for } j = 1, ..., k-1, \text{ and } \mu_k(K) \neq 0.$$

However, the estimation by higher order kernels implies two disadvantages. First, we have to assume that the unknown density is more than twice differentiable in x, and second, the kernel must be negative in some intervals, thus the density estimate also might become negative in some intervals.

2.4 Functionals and root n consistent density estimators

As presented above, ususally the kernel density estimator has a convergence rate of $(nh)^{-1/2}$. However, functionals of densities may have a parametric rate of convergence $n^{-1/2}$. This includes smooth functionals of densities, where appropriate estimators for the density are plugged into the functional, see e.g.; Hall & Marron (1987), Birgè & Massart (1995) and Efromovich & Samarov (2000). For continous-time processes root n consistent kernel-type density estimators also exists, see for example; Castellana & Leadbetter (1986), Blanke & Bosq (1997) and Veretennikov (1999).

Frees (1994) introduces density estimation for functions of observations and showed that independent and identical distributed random variables can be estimated at the parametric rate $n^{-1/2}$. This result generalizes to convolution densities $f * l(y) = \int f(y-x)l(x)\mathrm{d}x$. In the paper Saavedra & Cao (1999), the authors introduce a convolution-kernel estimator for the marginal density of a moving average process, $Y_i = X_i - \theta X_{i-1}$ when θ is unknown, and proves that this estimator, for an appropriate choice of bandwidth, is $n^{1/2}$-consistent. Schick & Wefelmeyer (2004a) introduces a slight simplified variant of this estimator, which also has the root n consistent property. However, these papers are quite technical, and only limited simulation experiments are performed. Further, Schick & Wefelmeyer (2004b) shows 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 root n consistent as well.

2.5 Dependent data

In many situations the observed sample $X_1, ..., X_n$ is not independent, e.g. in a time series setting. Assuming that the dependence in the data is weak enough, so-called short-range dependence, the dependence will not affect the leading term of the MISE approximation, see e.g. Györfi et al. (1990), and also the book by Fan & Yao (2003), which examines nonparametric time series methods. Thus, the kernel density estimator is often used to analyse dependent data. However, the results obtained need to be viewed with some caution, since dependencies in the data are certain to have some effect on the performance of the kernel

estimator, since it usually represents loss of information. Note that if the data is long-range dependent, this leads to a worse rate of convergence of the kernel density estimator.

2.6 Summary: Paper A and B

In paper A, "A convolution estimator for the density of nonlinear regression observations", we introduce a density estimator for the responses in a regression model

$$Y_i = m(X_i) + e_i, \tag{14}$$

where $\{X_i\}$ and $\{e_i\}$ consist of indpendent and identical distributed random variables with $\{e_i\}$ independent of $\{X_i\}$ and the function $m(\cdot)$ is unknown. The density of Y_i is of interest. It can of course be estimated by the standard kernel density estimator. But by using the idea of convolution, as explained in section 2.4, we introduce a new density estimator, called the convolution estimator. This estimator has better asymptotic properties than the kernel density estimator, that is, the variance of the estimator has leading terms n^{-1} instead of the standard $(nh)^{-1}$. In this sense it leads to a parametric rate of convergence of the convolution estimator. Several simulation experiments have been performed, and the convolution estimator behaves better in several cases.

Paper B, "A convolution density estimator for nonlinear time series: Simulations and some preliminary analysis", uses the same idea as in paper A, but in this case the model is an autoregressive process,

$$X_t = m(X_{t-1}) + e_t, \tag{15}$$

thus the observed X_t-s are dependent. Here the marginal density of X is of interest. Several simulation experiments are performed, and they indicate that the convolution estimator has a better performance than the kernel density estimator.

3 Nonparametric regression

In regression curve fitting, the aim is to find a relationship between variables X and Y. Assume we have n independent observations of $\{X_i, Y_i\}$, the regression model is

$$Y_i = m(X_i) + e_i, \tag{16}$$

where $\{e_i\}$ consist of i.i.d. random variables with $\{e_i\}$ independent of $\{X_i\}$. Thus $\mathrm{E}(Y|X = x) = m(x)$ and due to independence between X_i and e_i, $\mathrm{Var}(Y|X = x) = \mathrm{Var}(e_i) = \sigma_e^2$.

Nonparametric regression can also be studied in a fixed design. This is the case when the design consists of ordered non-random numbers $x_1, ..., x_n$, however, there is not a big difference between this case and the case we study, so-called

random design. Nonparametric regression estimators have been treated in several books, e.g. Härdle (1990), Wand & Jones (1995) and Fan & Gijbels (1996).

3.1 Common estimators

For estimating the regression function at a given point x, we weight the observations Y_i depending on the distance between X_i to x. Thus, we use the estimator

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} W_h(x; X_1, ..., X_n) Y_i, \tag{17}$$

where $W_h(\cdot)$ is weight function depending on the smoothing parameter or bandwidht h and the sample $X_1, ...X_n$ of explanatory variables. The weights $W_h(\cdot)$ depend on the technique used.

The three most common estimators are the local polynomial estimator; see Stone (1977), Cleveland (1979), Müller (1987) and Fan (1992), the Nadaraya-Watson estimator; see Nadaraya (1964) and Watson (1964) and the Gasser-Müller estimator; see Gasser & Müller (1979).

The local polynomial estimator, $\hat{m}(x; p, h)$, at a point x with p degrees of polynomial fit, is obtained by fitting the polynomial

$$\beta_0 + \beta_1(x - \cdot) + ... + \beta_p(x - \cdot)$$

to the (X_i, Y_i) using weighted least squares with kernel weights $K_h(x - X_i)$. The value of $\hat{m}(x; p, h)$ is the height of the fit $\hat{\beta}_0$ where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$ minimises

$$\sum_{i=1}^{n} [Y_i - \beta_0 - ... - \beta_p(x - X_i)^p]^2 K_h(x - X_i).$$

Assuming invertibility of $\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x$, standard weighted least squares theory gives the solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{Y}, \tag{18}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^T$ is the vector of responses,

$$\mathbf{X}_x = \begin{pmatrix} 1 & x - X_1 & \dots & (x - X_1)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x - X_n & \dots & (x - X_n)^p \end{pmatrix}$$

is an $n \times (p + 1)$ matrix and

$$\mathbf{W}_x = \text{diag}[K_h(x - X_1), ..., K_h(x - X_n)] \tag{19}$$

is an $n \times n$ diagonal matrix of weights. Since the estimator of $m(x)$ is the intercept β_0 we obtain

$$\hat{m}(x; p, h) = \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{Y}, \tag{20}$$

where $\mathbf{e}_1$ is the $(p+1) \times 1$ vector having 1 in the first entry and zeros elsewhere. Choosing $p = 1$ we obtain the so-called local linear estimator. If we choose $p = 0$ we get the Nadaraya-Watson estimator,

$$\hat{m}_{NW}(x; h) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}. \tag{21}$$

The Gasser-Müller estimator is

$$\hat{m}_{GM}(x; h) = \sum_{i=1}^{n} Y_{[i]} \int_{s_i - 1}^{s_i} K_h(x - u) \mathrm{d}u, \tag{22}$$

where $s_i = \frac{1}{2}(X_{[i]} + X_{[i+1]})$, $s_0 = 0$, $s_n = 1$. Here $(X_{[i]}, Y_{[i]})$, $i = 1, ..., n$, denote the (X_i, Y_i) ordered with respect to the X_i values. This estimator is intended for the fixed design case.

3.2 Asymptotic properties

The performance of the different estimators are again studied by examining the MSE. We need the following conditions,

1. The regression function $m(x)$ has a bounded and continous second derivative.

2. The kernel function K is a bounded density function with $\int xK(x)\mathrm{d}x = 0$ and $\int x^4 K(x)\mathrm{d}x < \infty$.

3. The marginal density f_X of the explanatory variable X is continous.

If $h \to 0$, and $nh \to \infty$, calculations give the following bias and variance for the different estimators, see e.g. Wand & Jones (1995) and Härdle (1990). For the local linear estimator,

$$\mathrm{Bias}(\hat{m}(x; 1, h)) = \frac{1}{2} h^2 m''(x) \int z^2 K(z)\mathrm{d}z + o(h^2), \tag{23}$$

$$\mathrm{Var}(\hat{m}(x; 1, h)) = \frac{\sigma_e^2}{nh f_X(x)} \int K^2(z)\mathrm{d}z + o((nh)^{-1}). \tag{24}$$

For the Nadaraya-Watson estimator

$$\mathrm{Bias}(\hat{m}_{NW}(x; h)) = (\frac{1}{2} m''(x) + \frac{m'(x) f_X'(x)}{f_X(x)}) h^2 \int z^2 K(z)\mathrm{d}z + o(h^2), \tag{25}$$

$$\mathrm{Var}(\hat{m}_{NW}(x; h)) = \frac{\sigma_e^2}{nh f_X(x)} \int K^2(z)\mathrm{d}z + o((nh)^{-1}). \tag{26}$$

And last, for the Gasser-Müller estimator,

$$\text{Bias}(\hat{m}_{GM}(x;h)) = \frac{1}{2}h^2 m''(x)\int z^2 K(z)\mathrm{d}z + o(h^2), \tag{27}$$

$$\text{Var}(\hat{m}_{GM}(x;h)) = \frac{3\sigma_e^2}{2nhf_X(x)}\int K^2(z)\mathrm{d}z + o((nh)^{-1}). \tag{28}$$

Observe that the estimators have different leading terms in the variance and bias. However, the variance for the Gasser-Müller estimator is 1.5 times larger than the others. The bias for the Nadaraya-Watson estimator is a more complicated expression than the others, but note that the bias in all cases depends on $m''(x)$, thus when the curvature of the function $m(x)$ is large, the estimators will have a large bias.

Note that the local linear estimator has more appealing properties than the other two estimators. These two estimators have boundary effects, as for the kernel density estimator, that is, bias of order $O(h)$ instead of $O(h^2)$ at the boundary. This is not the case for the local linear estimator, which automatically correct for this boundary bias. Again, the bias converge with the order h^2, the variance has convergence rate $(nh)^{-1}$ and the A-MISE optimal bandwidth is of order $n^{-1/5}$, which implies that the MSE, from equation (2), and MISE, equation (3), has convergence rate $n^{-4/5}$.

As for the kernel density estimtor, similar modifcations of the nonparametric regression estimators can be used to improve the performance of the estimates asymptotically.

3.3 Bandwidth selection

The basic idea of choosing the smoothing parameter is the same as for density estimation; the bandwidth should be choosen to minimise an error measure, usually the MISE. Again, the methods can basically be classified into two categories: cross-validation methods and plug-in methods.

For the Nadaraya-Watson estimator, Härdle & Marron (1985) applied the ideas of least squares cross-validation, and Gasser et al. (1991) and Härdle et al. (1992) proposed bandwidth selectors with better asymptotic properties and practical performance. A simple direct plug-in idea for local linear regression with attractice theoretical and practical properties is proposed by Ruppert et al. (1995). This bandwidth is used in some parts of this thesis.

There exists several other bandwidth selectors; see e.g. the books by Wand & Jones (1995), Härdle (1990) and Fan & Gijbels (1996) for further references.

3.4 Dependent data

As under kernel density estimation, if the data satisfies certain mixing conditions or short-range dependence, the nonparametric regression estimators behaves al-

most as for independent data. If the data do not fulfill these conditions, the variance of the estimator can be affected. The general strategy is then to increase the bandwidth, to reduce variance. Chapter 5 and 6 in Fan & Yao (2003) gives results for nonparametric kernel density and regression estimators used with dependent data under mixing conditions.

3.5 Multivariate nonparametric regression

In a multivariate regression problem we want to study the relationship between the response variable Y and the vector of covariates $\mathbf{X} = (X_1, ..., X_d)^T$ via

$$m(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x}), \tag{29}$$

where $\mathbf{x} = (x_1, ..., x_d)^T$ and $m(\mathbf{x}) = m(x_1, ..., x_d)$. It is straigthforward to generalize the univariate smoothing techniques, see e.g. Fan & Gijbels (1996). However, the main problem is that the sparseness of data in higher-dimensional space makes regression estimation difficult unless the sample size is very large. This is known as the curse of dimensionality. Thus these smoothing techniques are not usually applied with d larger than three. Several approaches have been proposed to handle the curse of dimensionality problem, and one approach, additive modelling, will be introduced in the next section.

3.6 Additive models and backfitting

Additive models are widely used in applied statistics and econometrics, see e.g. the monograph by Hastie & Tibshirani (1990). The setup is as in the section above, but now the regression function $m(\mathbf{x})$ is modelled additively, that is,

$$m(\mathbf{x}) = \sum_{j=1}^{d} m_j(x_j). \tag{30}$$

To ensure identifiability, $m_1, ..., m_d$ are required to satisfy

$$\mathrm{E}[m_j(X_j)] = 0, \quad j = 1, ..., d. \tag{31}$$

Usually an intercept term is added, $\mathrm{E}(Y) = \alpha$, thus the model becomes

$$Y = \alpha + \sum_{j=1}^{d} m_j(x_j) + e, \tag{32}$$

where $\mathrm{E}(e) = 0$, $\mathrm{Var}(e) = \sigma^2$ and e is independent of the vector of covariates $\mathbf{X}$.

Estimation of the unknown functions $m_1, ..., m_d$ is done by the backfitting algorithm, introduced by Breiman & Friedman (1985) and Buja et al. (1989). Note first, that if the additive model, (32), is correct then

$$\mathrm{E}[Y - \alpha - \sum_{j \neq k} m_j(X_j)|X_k] = m_k(X_k), \quad k = 1, ..., d. \tag{33}$$

This relationship suggest an iterative procedure for the estimation of the unknown functions. Thus for a known constant α and given functions m_j, $j \neq k$, the function m_k can be estimated by a univariate regression fit based on the observations (X_k^i, Y_i), $i = 1, ..., n$, where X_k^i is the ith observation of the kth additive variable. Denote the univariate smoother of m_k by S_k. The algorithm works as follows:

Step 1. Initialization: $\hat{\alpha} = n^{-1} \sum_{i=1}^{n} Y_i$, $\hat{m}_k = m_k^0$ for $k = 1, ..., d$.
Step 2. Find new transformations: For $k = 1, ..., d$:

$$\hat{m}_k = S_k[Y - \hat{\alpha} - \sum_{j \neq k} \hat{m}_j(X_j)|X_k];$$

$$\text{centre the estimator to obtain } \hat{m}_k^* = \hat{m}_k - n^{-1} \sum_{i=1}^{n} \hat{m}_k(X_k^i),$$

$$\text{and } \hat{\alpha}^* = \hat{\alpha} + n^{-1} \sum_{i=1}^{n} \hat{m}_k(X_k^i).$$

Step 3. Repeat step 2 until convergence.

The idea behind this algorithm is to carry out a fit, calculate partial residuals from that fit, and refit again. That is why the iteration scheme is called backfitting. The starting functions $m_1^0, ..., m_d^0$ can be obtained in various ways, for example, from a linear regression fit of Y on the covariates X_k.

However, this backfitting algorithm does not reach the oracle efficiency bound, i.e. the additive components are not estimated with the same asymptotic bias and variance as if the other components were known. Mammen et al. (1999) defined a new backfitting-type estimator, called smooth backfitting, that overcome this shortcoming, and also has a number of other tractable properties. Here the estimator is simply the projection of the data on the additive space of interest.

3.7 Other methods

Besides the already mentioned estimators for nonparametric regression, there exist several other methods to estimate a relationship between a response Y and a covariate X, they include wavelet and spline smoothing.

In local polynomial modelling, the estimator models the unknown regression function m locally by a few parameters and uses the bandwidth to control the complexity of the modelling. But it is also possible to model the function globally with a large amount of unknown parameters. Wavelet and spline expansion can be used for such modelling. Wavelet transforms have received great deal of attention in e.g. applied mathematics and signal analysis. For references to these methods; see chapter 2 in Fan & Gijbels (1996).

3.8 Summary: Paper C

In this paper, "A new convolution estimator for nonparametric regression", we present a new type of estimator for standard nonparametric regression. The estimator is introduced by recognizing that the standard estimators presented here, do not take into account the extra information contained in the regression relationship

$$Y_i = m(X_i) + e_i. \tag{34}$$

Again we use the idea of convolution, and by analysing the asymptotic properties of this new estimator, that we have called "the convolution estimator", we find that its asymptotic bias is smaller than for standard methods. We also prove asymptotic normality. Some simulation experiments have been performed, and in several cases, the convolution estimator outperforms the standard methods.

An adjusted kernel function is also introduced, and by using this kernel in our convolution estimator, it seems that the estimator gives even better results. However, more theoretical work has to be done here.

4 Panel data models

Data sets that combine time series and cross sections are common in several fields, especially in economics. An example of such a data set can be time series observations of income, expenditure etc. for several firms simultanously. Usually the number of individuals, here firms, are very large, but the number of observations for each individual are moderate or small. Panel data sets contain more information than just time series or cross sectional data alone, since it is for example possible to model cross-sectional heterogeneity. Thus the analysis of panel data is the subject of one of the most active areas of research in econometrics. Several books have been written on the analysis of panel data; see e.g. Hsiao (1986), Baltagi (2001) og Arrellano (2003). The following presentation is mainly taken from Greene (2003) chapter 13, and I refer to this book for further details.

4.1 General model

The basic framework for panel data analysis is a regression model of the form

$$y_{it} = \alpha_i + \gamma_t + \boldsymbol{\beta}'\mathbf{x}_{it} + \epsilon_{it}, \quad i = 1, ..., n \text{ and } t = 1, ..., T, \tag{35}$$

where $\mathbf{x}_{it}$ contains K regressors, α_i is constant over time, t, and it is the individual effect for the cross-sectional unit i and γ_t is the time effect. ϵ_{it} is the error term corresponding to individ i at time t.

The time effect is sometimes not included, and if we take the α_i's to be the same across all units, then ordinary least squares provides consistent and efficient estimates of α and $\boldsymbol{\beta}$. Otherwise, there are two frameworks used to generalize

the model: the fixed effects approach takes α_i to be a group specific constant that needs to be estimated or the random effects, that specifies that α_i is a group specific disturbance.

4.2 Fixed effects

A common formulation of the model assumes that differences across units can be captured in differences in the constant term. Thus each α_i is an unknown parameter to be estimated. Let $\mathbf{y}_i$ and $\mathbf{X}_i$ be the T observations for the ith unit, $\boldsymbol{\epsilon}_i$ is the $T \times 1$ vector of disturbances, and $\mathbf{i}$ is a $T \times 1$ column of ones, then

$$\mathbf{y}_i = \mathbf{i}\alpha_i + \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i. \tag{36}$$

Collecting all terms we can write,

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{37}$$

where $\mathbf{y}$ is a $nT \times 1$ vector, $\mathbf{D}$ is a $nT \times n$ matrix of dummy variables indicating units, $\boldsymbol{\alpha}$ is a $n \times 1$ vector, $\mathbf{X}$ is a $nT \times K$ matrix, $\boldsymbol{\beta}$ is a $K \times 1$ vector and $\boldsymbol{\epsilon}$ is a $nT \times 1$ vector.

This model is known as the least squares dummy variable model (LSDV), and ordinary least squares can be used to estimate $\boldsymbol{\beta}$, and the $\boldsymbol{\alpha}$. The time specific effect can easily be included.

4.3 Random effects

In some settings it might be appropriate to view the individual specific terms as randomly distributed across cross-sectional units, for example if we believe that the sampled cross-sectional units are drawn from a large population. Consider a reformulation of the model,

$$y_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + u_i + \epsilon_{it}, \tag{38}$$

where there are K regressors in addition to the constant term. The u_i is the random error term characterizing the ith observation and is constant through time. By making assumptions on the error terms u_i and ϵ_{it} it is possible to estimate $\boldsymbol{\beta}$ by using generalized least squares (GLS) or feasible generalized least squares (FGLS).

4.4 Other models

Missing data are common in panel data sets, and panels where the unit sizes differ across units are called unbalanced panels. In this case we have $\sum_{i=1}^{n} T_i$ observations instead of nT. However, analysing an unbalanced panel in the fixed effects

model is fairly straightforward. Panel data models can handle heteroscedasticity, which appears when the error terms in the model have variances which are not constant across observations. And models that handle autocorrelated error terms have also been proposed.

Panel data models are well suited for examining dynamic effects,

$$y_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_i t + \delta y_{i,t-1} + \epsilon_{it}. \tag{39}$$

But substantial complications arise in estimation of this model. The problem is that the lagged dependent variable is correlated with the disturbances, even if it is assumed that ϵ_{it} is not itself autocorrelated. However, General Method of Moments (GMM) estimators have been constructed to handle this problem.

4.5 Nonparametric panel data models

In the panel data setting, most models proposed and examined are linear parametric models, as introduced above. Very little literature exist on nonlinear or nonparametric panel data models, see Arellano & Honorè (2001) for a review of recent work on nonlinear panel data models and Ullah & Roy (1998) for a review of nonparametric models. See also Hjellvik et al. (2004) for a recently proposed nonparametric model.

One example of a nonparametric panel data model is a pooled Nadaraya-Watson kernel estimation. Here the model is

$$y_{it} = m(x_{it}) + u_{it}, \quad i = 1, ...n, \;\; t = 1, ..., T \tag{40}$$

where x_{it} is a vector of K regressors, $m(x_{it}) = \mathrm{E}(y_{it}|x_{it})$, $\mathrm{E}(u_{it}|x_{it}) = 0$, $\mathrm{Var}(u_{it}|x_{it}) = \sigma^2(x_{it})$ and (y_{it}, x_{it}) are assumed to be i.i.d. The nonparametric estimate of $m(x)$, the conditional mean at a point x, is the smoothed average of y values which correspond to the x_{it} values in a small interval of x, i.e. the Nadaraya-Watson estimator, which were introduced earlier.

4.6 Summary: Paper D

In this paper, "Nonparametric additive models for panels of time series", we introduce a general additive model for the panel of observations $\{y_{it}\}$, $i = 1, ..., n$, $t = 1, ..., T$, where i represents individuals and t time. The model is

$$y_{it} = \sum_{j=1}^{p} m_j(x_{it}^j) + \eta_t + \lambda_i + \epsilon_{it}, \tag{41}$$

where $\{x_{it}^j\}$, $j = 1, ..., p$, is a set of explanatory variables, η_t and λ_i are temporary and individual effects, respectively, and ϵ_{it} are error terms. The functions m_j,

$j = 1, ..., p$ are unknown, and the task is to estimate them. We have chosen to use backfitting, as introduced in Mammen et al. (1999).

An iterative algorithm has been proposed, and both simulation experiments and a real data study indicates that this algorithm estimates the additive functions well.

References

Arellano, M. & Honorè, B. (2001), Panel data models: Some recent developments, *in* J. J. Heckman & E. Leamer, eds, 'Handbook of Econometrics, vol.5', North-Holland.

Arrellano, M. (2003), *Panel Data Econometrics*, Oxford University Press.

Baltagi, B. H. (2001), *Econometric Analysis of Panel Data*, John Wiley & Sons, Ltd.

Bickel, P. J. & Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.

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

Blanke, D. & Bosq, D. (1997), 'Accurate rates of density estimators for continous-time processes', *Statistics and Probability Letters* **33**, 185–191.

Bowman, A. W. (1984), 'An alternative method of cross-validation for the smoothing of density estimate', *Biometrika* **71**, 353–360.

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

Breiman, L., Meisel, W. & Purcell, E. (1977), 'Variable kernel estimates of probability density estimates', *Technometrics* **19**, 135–144.

Buja, A., Hastie, T. J. & Tibshirani, R. (1989), 'Linear smoothers and additive models (with discussion)', *Annals of Statistics* **17**, 453–555.

Castellana, J. V. & Leadbetter, M. R. (1986), 'On smoothed probability density estimation for stationary processes', *Stochastic Processes and Applications* **21**, 179–193.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Duin, R. P. W. (1976), 'On the choice of smoothing parameters for parzen estimators of probability density functions', *IEEE Trans. Comput.* **C-25**, 1175–1179.

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

Efron, B. & Tibshirani, R. (1996), 'Using specially designed exponential families for density estimation', *The Annals of Statistics* **24**, 2431–2461.

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

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall.

Fan, J. & Yao, Q. (2003), *Nonlinear Time Series*, Springer-Verlag.

Fix, E. & Hodges, J. L. (1951), Discriminatory analysis - nonparametric discrimination: Consistency properties. Report No. 4, Project no. 21-29-004. USAF School of Aviation Medicine, Randolph Field, TX.

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

Gasser, T., Kneip, A. & Kohler, W. (1991), 'A fast and flexible method for automatic smoothing', *Journal of the American Statistical Association* **86**, 643–652.

Gasser, T. & Müller, H. G. (1979), Kernel estimation of regression functions, *in* T. Gasser & M. Rosenblatt, eds, 'Smoothing Techniques for Curve Estimation', Springer-Verlag, Heidelberg, pp. 23–68.

Gasser, T., Müller, H. G. & Mammitzsch, V. (1985), 'Kernels for nonparametric curve estimation', *Journal of the Royal Statistical Society. Series B* **47**, 238–252.

Glad, I. K. (1998), 'Parametrically guided non-parametric regression', *Scandinavian Journal of Statistics* **25**, 649–668.

Greene, W. H. (2003), *Econometric Analysis*, Prentice hall.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1990), *Nonparametric Curve Estimation of Time Series*, Springer-Verlag.

Habema, J. D. F., Hermans, J. & van der Broeck, K. (1974), A stepwise discrimination program using density estimation, *in* G. Bruckman, ed., 'Compstat', Physica Verlag, Vienna, pp. 100–110.

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

Hall, P. & Wehrly, T. E. (1991), 'A geometrical method for removing edge effects from kernel-type nonparametric regression estimators', *Journal of the American Statistical Association* **86**, 665–672.

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

Hjellvik, V., Chen, R. & Tjøstheim, D. (2004), 'Nonparametric estimation and testing in panels of intercorrelated time series', *Journal of Time Series Analysis* **25**, 831–872.

Hjort, N. L. & Glad, I. K. (1995), 'Nonparametric density estimation with a parametric start', *The Annals of Statistics* **23**, 882–904.

Hsiao, C. (1986), *Analysis of Panel Data*, Cambridge University Press.

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

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

Härdle, W. & Marron, J. S. (1985), 'Optimal bandwidth selection in nonparametric regression function estimation', *Annals of Statistics* **13**, 1465–1481.

Mammen, E., Linton, O. & Nielsen, J. (1999), 'The existence and asymptotic properties of a backfitting projection algorithm under weak conditions', *Annals of Statistics* **27**, 1443–1490.

Marron, J. S. & Wand, M. P. (1992), 'Exact mean integrated squared error', *Annals of Statistics* **20**, 712–736.

Müller, H. G. (1987), 'Weigthed local regression and kernel methods for nonparametric curve fitting', *Journal of the American Statistical Association* **82**, 231–238.

Nadaraya, E. A. (1964), 'On estimating regression', *Theory of Probability and its Applications* **9**, 141–142.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *Annals of Mathematical Statistics* **33**, 1065–1076.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *Annals of Mathematical Statistics* **27**, 832–837.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators', *Scandinavian Journal of Statistics* **9**, 65–78.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), 'An effective bandwidth selector for local least squares regression', *Journal of the American Statistical Association* **90**, 1257–1270.

Saavedra, A. & Cao, R. (1999), '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.

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', *Nonparametric Statistics* **16**, 925–935.

Schuster, E. F. (1985), 'Incorporating support constraints into nonparametric estimates of densities', *Communications in Statistics - Theory and Methods* **14**, 1123–1126.

Scott, D. W. & Terrell, G. R. (1987), 'Biased and unbiased cross-validation in density estimation', *Journal of American Statistical Association* **82**, 1131–1146.

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.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman-Hall.

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

Stone, C. J. (1977), 'Consistent nonparametric regression', *Annals of Statistics* **5**, 596–620.

Ullah, A. & Roy, N. (1998), Nonparametric and semiparametric econometrics of panel data, *in* A. Ullah & D. E. A. Giles, eds, 'Handbook on Applied Economic Statistics', Marcel Dekker, pp. 579–604.

Veretennikov, A. Y. (1999), 'On castellana-leadbetter's condition for diffusion density estimators', *Statistical Inference and Stochastical Processes* **2**, 1–9.

Victor, N. (1976), Nonparametric allocation rules, *in* F. T. Dombal & F. Grémy, eds, 'Decision Making and Medical Care: Can Information Science help?', North-Holland, Amsterdam, pp. 515–529.

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

Wand, M. P., Marron, J. S. & Ruppert, D. (1991), 'Transformations in density estimation (with discussion)', *Journal of the American Statistical Association* **86**, 343–361.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhyā Ser. A* **26**, 101–116.

Watson, G. S. & Leadbetter, M. R. (1963), 'On the estimation of a probability density, 1', *Annals of Mathematical Statistics* **34**, 480–491.

Yang, L. & Marron, J. S. (1999), 'Iterated transformations-kernel density estimation', *Journal of the American Statistical Association* **94**, 580–589.

PAPERS

Paper A: "A convolution estimator for the density of nonlinear regression observations".
Will be submitted to *Journal of the American Statistical Association.*

Paper B: "A convolution density estimator for nonlinear time series: Simulations and some preliminary analysis".
In preperation for submission.

Paper C: "A new convolution estimator for nonparametric regression".
Revised for IMS Lecture Notes Series. *A Festschrift for Kjell Doksum.*

Paper D: "Nonparametric additive models for panels of time series".
In preperation for submission.

PAPER A

Will be submitted to *Journal of the American Statistical Association*

A convolution estimator for the density of nonlinear regression observations

Bård Støve Dag Tjøstheim

Department of Mathematics
University of Bergen
Johannes Brunsgate 12
5008 Bergen
Norway

Abstract

We present a convolution estimator for the density of the responses in a standard regression model. The rate of convergence for the variance of this estimator is examined and found to be of order n^{-1}. We also derive the bias of the new estimator and conduct simulation experiments to check the finite sample properties. The proposed estimator performs better than the kernel density estimator for well behaved noise densities.

Some key words: Convolution, Kernel function, Mean squared error, Nonparametric density estimation, Regression.

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

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

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,

$$Y_i = g(X_i) + e_i, \tag{2}$$

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 being the density of interest, and the densities of X_i and e_i, $f_X(\cdot)$ and $f_e(\cdot)$, respectively. For given observations of (X_i, Y_i) a standard 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 (2) to hold, and if one is able to construct an estimator which takes this relationship into account, one should think that it would be possible to make an improvement. This idea was introduced in Støve & Tjøstheim (2005*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 (2), it will be seen that we manage to obtain a better convergence rate for the variance and often better bias properties, 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 $Y = g(X_1, ..., X_m)$ of $m > 1$ independent and identically distributed variables. The density can be estimated at the rate $n^{-1/2}$, in contrast to the standard nonparametric rate of $(nh)^{-1/2}$, if this density and the conditional density of Y given X_1 are sufficiently smooth. 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 this type of 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, for an appropriate choice of bandwidth, is $n^{1/2}$-consistent. The case when θ is unknown is examined in Saavedra & Cao (1999*b*), and an analogous result is obtained, but 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.

Note that we assume that the function $g(\cdot)$ and the error terms e_i are not known, and thus has to be estimated. This is in contradistinction to the models examined in Frees (1994) and Schick & Wefelmeyer (2004*b*), where the authors assume that the function g is known.

Our proposed estimator is presented in section 2, its asymptotic behaviour is examined in section 3, and some simulation results and a real data set are located in section 4. Some conclusions are given in section 5.

2 The estimator

From equation (2) using that $g(X_i)$ and e_i are independent, we get

$$f_Y(y) = \int f_e\big(y - g(u)\big) f_X(u)\mathrm{d}u, \tag{3}$$

where f_e is the density of the residuals. This motivates an estimator of the density f_Y given by the functional in (3), which can be written as

$$f_Y(y) = \mathrm{E}\big[f_e\big(y - g(X)\big)\big]. \tag{4}$$

Assume we have observations $(X_1, Y_1,), ..., (X_n, Y_n)$ of (X, Y). We introduce estimators of g, e, f_e and an estimate of the expectation in (4), thus

$$\hat{f}_Y(y) = \hat{\mathrm{E}}\big[f^*_{\tilde{e}}\big(y - \tilde{g}(X)\big)\big]. \tag{5}$$

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

$$\tilde{g}(x) = \frac{\sum_{i=1}^n K^{NW}_{x,h_R}(X_i)Y_i}{\sum_{i=1}^n K^{NW}_{x,h_R}(X_i)}. \tag{6}$$

Other nonparametric regression estimators, such as the local linear estimator, see Fan & Gijbels (1996) page 20, for g can also be choosen. The estimate for e_i is

$$\tilde{e}_i = Y_i - \tilde{g}(X_i). \tag{7}$$

The estimate $f^*_{\tilde{e}}$ of the density of e_i is by the kernel estimator with bandwidth h_D and kernel function $K(\cdot)$,

$$f^*_{\tilde{e}}(y) = \frac{1}{(n-1)h_D}\sum_{i=2}^{n} K\Big(\frac{y - \tilde{e}_i}{h_D}\Big). \tag{8}$$

Thus the final density estimator for f_Y is, using (5),

$$\hat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} f^*_{\tilde{e}}\big(y - \tilde{g}(X_i)\big) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h_D}\sum_{j=2}^{n} K\Big(\frac{y - \tilde{g}(X_i) - \tilde{e}_j}{h_D}\Big)\Big]. \tag{9}$$

3 Asymptotic properties

Asymptotic analysis of nonparametric estimators is usually based on the asymptotic bias and variance of the estimator. In this section we discuss these asymptotic properties of the estimator (9). The following assumptions are made, and a few others are introduced later when needed,

1. The kernel function K is a bounded non-negative, two times differentiable, symmetric function that integrates to 1, and hence the derivative K' satisfies

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

2. The function g is differentiable and its inverse exists.

3. The density f_X has compact support $S(X)$, is continous and two times differentiable on the support.

4. $\lim_{n\to\infty} h_D = 0$ and $\lim_{n\to\infty} nh_D = \infty$.

Condition 1 is standard in nonparametric estimation, and observe that if the kernel function is the standard normal distribution, this condition is automatically fulfilled. Condition 2 is introduced to obtain simple expressions. We think it can be relaxed. Condition 3 is also standard, and the compact support is introduced for the sake of simplicity. It can probably be removed at the cost of lengthier arguments. Condition 4 is standard.

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,

$$\hat{f}_Y(x) - f_Y(x) = \hat{f}_Y(x) - \tilde{f}_Y(x) + \tilde{f}_Y(x) - f_Y(x), \tag{10}$$

where

$$\tilde{f}_Y(x) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h_D}\sum_{j=2}^{n}K\Big(\frac{x-g(X_i)-e_j}{h_D}\Big)\Big], \tag{11}$$

that is, the proposed estimator with $g(\cdot)$ and e_j for $j = 1, ..., n$, known. We first consider this "estimator". Recall that the MSE can be decomposed as follows,

$$\mathrm{MSE}(\tilde{f}_Y(x)) = \mathrm{E}[(\tilde{f}_Y(x) - f_Y(x))^2] = \mathrm{var}(\tilde{f}_Y(x)) + [\mathrm{E}(\tilde{f}_Y(x)) - f_Y(x)]^2. \tag{12}$$

To ease notation, we now set $h_D = h$. Consider the bias term first. Since X_i and e_j are independent for all i and j,

$$\begin{aligned}\mathrm{E}\big(\tilde{f}_Y(x)\big) &= \frac{1}{n(n-1)h}\mathrm{E}\Big[\sum_{i=1}^{n}\sum_{j=2}^{n}K\big(\frac{x-g(X_i)-e_j}{h}\big)\Big] \\ &= \frac{1}{h}\mathrm{E}\Big[K\big(\frac{x-g(X)-e}{h}\big)\Big]. \end{aligned}\tag{13}$$

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

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

Thus the bias can be written,

$$\mathrm{E}\big(\tilde{f}_Y(x)\big) - f_Y(x) = \frac{h^2}{2}f_Y''(x)\int w^2K(w)\mathrm{d}w + O(h^4). \tag{14}$$

This is in fact equal to the bias of the standard kernel density estimator, see e.g. Wand & Jones (1995) page 20.

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

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

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

$$= \frac{1}{n^2(n-1)^2h^2}\Big[(n-1)\mathrm{var}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big)\Big) \quad (15)$$

$$+(n-1)(n-2)\mathrm{var}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big) \quad (16)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big), K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big) \quad (17)$$

$$+2(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big), K\big(\frac{x-g(X_3)-e_1}{h}\big)\Big) \quad (18)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big), K\big(\frac{x-g(X_3)-e_2}{h}\big)\Big) \quad (19)$$

$$+(n-1)(n-2)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big), K\big(\frac{x-g(X_2)-e_1}{h}\big)\Big) \quad (20)$$

$$+2(n-1)(n-2)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big), K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big) \quad (21)$$

$$+2(n-1)(n-2)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big), K\big(\frac{x-g(X_2)-e_1}{h}\big)\Big) \quad (22)$$

$$+(n-1)(n-2)(n-3)(n-4)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big), K\big(\frac{x-g(X_3)-e_4}{h}\big)\Big) \quad (23)$$

$$+(n-1)(n-2)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big), K\big(\frac{x-g(X_2)-e_2}{h}\big)\Big) \quad (24)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big), K\big(\frac{x-g(X_2)-e_3}{h}\big)\Big) \quad (25)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_2)-e_1}{h}\big), K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big)\Big]. \quad (26)$$

Each variance and covariance term can be studied using

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2, \quad (27)$$

$$\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y). \quad (28)$$

By independence the terms (18), (20), (23), (24), (25) and (26) are equal to zero, and we just have to examine the remaining terms.

Equation (15) is

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

Consider the first term. By change of variables, convolution and Taylor expansion,

$$\mathrm{E}\Big(K^2\big(\frac{x-g(X_1)-e_1}{h}\big)\Big) = \iint K^2\big(\frac{x-g(v)-u}{h}\big) f_X(v) f_e(u) \mathrm{d}v\mathrm{d}u$$

$$= h\iint K^2(z) f_X(v) f_e(x-g(v)-hz)\mathrm{d}v\mathrm{d}z = h\int K^2(z) f_Y(x-zh)\mathrm{d}z$$

$$= h\int K^2(z)\big[f_Y(x) - hzf_Y'(x) + \frac{h^2z^2}{2}f_Y''(x)\big]\mathrm{d}z + O(h^4)$$

$$= hf_Y(x)\int K^2(z)\mathrm{d}z + \frac{h^3}{2}f_Y''(x)\int z^2K^2(z)\mathrm{d}z + O(h^4).$$

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

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

In total,

$$(n-1)\mathrm{var}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big)\Big)$$

$$= (n-1)\Big[hf_Y(x)\int K^2(z)\mathrm{d}z - h^2f_Y^2(x) + O(h^2)\Big]. \tag{29}$$

The second term, (16), is,

$$\mathrm{var}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big) = \mathrm{E}\Big(K^2\big(\frac{x-g(X_1)-e_2}{h}\big)\Big) - \Big[\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big)\Big]^2.$$

This is equal to term (15), thus,

$$(n-1)(n-2)\mathrm{var}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big)$$

$$= (n-1)(n-2)\Big[hf_Y(x)\int K^2(z)\mathrm{d}z - h^2f_Y^2(x) + O(h^2)\Big]. \tag{30}$$

The third term, equation (17), is

$$\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big), K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big)$$

$$= \mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big)$$

$$-\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big)\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big).$$

Further, by change of variables and Taylor expansion,

$$\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big)$$

$$= \iiint K\big(\frac{x-g(v)-u_1}{h}\big)K\big(\frac{x-g(v)-u_2}{h}\big) f_X(v) f_e(u_1) f_e(u_2)\mathrm{d}v\mathrm{d}u_1\mathrm{d}u_2$$

$$= h^2\iiint K(z_1)K(z_2) f_X(v) f_e(x-g(v)-z_1h) f_e(x-g(v)-z_2h)\mathrm{d}v\mathrm{d}z_1\mathrm{d}z_2$$

$$= h^2\Big[\int f_X(v) f_e^2\big(x-g(v)\big)\mathrm{d}v + O(h^2)\Big].$$

As before,

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

In total this gives,

$$(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big),K\big(\frac{x-g(X_1)-e_3}{h}\big)\Big)$$
$$= (n-1)(n-2)(n-3)\Big[h^2\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v-h^2f_Y^2(x)+O(h^4)\Big]. \qquad (31)$$

The fifth term, equation (19), is

$$\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big),K\big(\frac{x-g(X_3)-e_2}{h}\big)\Big). \qquad (32)$$

Using the assumption that the inverse of $g(\cdot)$ exists we obtain

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

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,

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

In total,

$$(n-1)(n-2)(n-3)\mathrm{cov}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big),K\big(\frac{x-g(X_3)-e_2}{h}\big)\Big)$$
$$= (n-1)(n-2)(n-3)\Big[h^2\int\frac{f_X^2(v)f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v-h^2f_Y^2(x)+O(h^4)\Big]. \qquad (33)$$

The seventh term, equation (21), is

$$\operatorname{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big),K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big). \tag{34}$$

We have

$$\begin{aligned}
&\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big)K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big)\\
&=\iiint K\big(\frac{x-g(v)-u}{h}\big)K\big(\frac{x-g(v)-w}{h}\big)f_X(v)f_e(u)f_e(w)\mathrm{d}v\mathrm{d}u\mathrm{d}w\\
&=h^2\iiint K(z_1)K(z_2)f_X(v)f_e(x-g(v)-hz_1)f_e(x-g(v)-hz_2)\mathrm{d}v\mathrm{d}z_1\mathrm{d}z_2\\
&=h^2\int f_X(v)f_e\big(x-g(v)\big)f_e\big(x-g(v)\big)\mathrm{d}v+O(h^4).
\end{aligned}$$

As before,

$$\begin{aligned}
&\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big)\Big)\mathrm{E}\Big(K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big)\\
&=h^2f_Y^2(x)+O(h^4).
\end{aligned}$$

In total,

$$\begin{aligned}
&2(n-1)(n-2)\operatorname{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big),K\big(\frac{x-g(X_1)-e_2}{h}\big)\Big)\\
&=2(n-1)(n-2)\Big[h^2\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v\\
&\qquad -h^2f_Y^2(x)+O(h^4)\Big].
\end{aligned}\tag{35}$$

The eigth term, equation (22), is

$$\operatorname{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big),K\big(\frac{x-g(X_2)-e_1}{h}\big)\Big). \tag{36}$$

This is equal to the fifth term, (19), so in total,

$$\begin{aligned}
&2(n-1)(n-2)\operatorname{cov}\Big(K\big(\frac{x-g(X_1)-e_1}{h}\big),K\big(\frac{x-g(X_2)-e_1}{h}\big)\Big)\\
&=2(n-1)(n-2)\Big[h^2\int\frac{f_X^2(v)f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v\\
&\qquad -h^2f_Y^2(x)+O(h^4)\Big].
\end{aligned}\tag{37}$$

Adding all the expressions stemming from (15)-(26), we get the total variance,

$$\begin{aligned}
\operatorname{var}\big(\tilde{f}_Y(x)\big)&=\frac{1}{n-1}\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v\\
&+\frac{1}{n-1}\int\frac{f_X^2(v)f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v-\frac{2}{n-1}f_Y^2(x)\\
&+\frac{1}{n(n-1)h}f_Y(x)\int K^2(z)\mathrm{d}z+O\big((n-1)^{-1}h^2\big).
\end{aligned}\tag{38}$$

Observe that from this expression the variance of $\tilde{f}_Y(x)$ has leading terms of order n^{-1}, this in contrast to the kernel density estimator which has leading terms $(nh)^{-1}$, see e.g. Wand & Jones (1995) page 21.

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

$$\begin{aligned}\mathrm{MSE}\big(\tilde{f}_Y(x)\big) = \mathrm{E}\Big[\big(\tilde{f}_Y(x) - f_Y(x)\big)^2\Big] &= \frac{1}{4}h^4 f_Y''(x)^2\Big[\int w^2 K(w)\mathrm{d}w\Big]^2 \\ &+ \frac{1}{n-1}\int f_X(v) f_e^2\big(x - g(v)\big)\mathrm{d}v \\ &+ \frac{1}{n-1}\int \frac{f_X^2(v) f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v - \frac{2}{n-1}f_Y^2(x) \\ &+ \frac{1}{n(n-1)h} f_Y(x)\int K^2(z)\mathrm{d}z + O\big((n-1)^{-1}h^2\big) + O(h^6). \qquad (39)\end{aligned}$$

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 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 (10), that is, $\hat{f}_Y(x) - \tilde{f}_Y(x)$. Here some additional assumptions are introduced. Before we list them, let $f_{X,Y}(x,y)$ denote the joint distribution of (X,Y) and define $m(x) = \int y f(x,y)\mathrm{d}y$:

5. $\mathrm{E}|Y|^s < \infty$ and $\sup_x \int |y|^s f(x,y)\mathrm{d}y < \infty$, $s \geq 2$.

6. m is continous on $S(X)$.

Consider the estimator $\hat{f}_Y(x)$ in (9). 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,

$$\begin{aligned}\hat{f}_Y(x) &= \frac{1}{n}\sum_{i=1}^n \Big[\frac{1}{(n-1)h}\sum_{j=2}^n K\big(\frac{x - \tilde{g}(X_i) - \tilde{e_j}}{h}\big)\Big] \\ &= \frac{1}{n}\sum_{i=1}^n \Big[\frac{1}{(n-1)h}\sum_{j=2}^n K\big(\frac{x - g(X_i) - e_j + \tilde{g}(X_j) - g(X_j) - (\tilde{g}(X_i) - g(X_i))}{h}\big)\Big] \\ &= \frac{1}{n}\sum_{i=1}^n \Big[\frac{1}{(n-1)h}\sum_{j=2}^n \big[K\big(\frac{x - g(X_i) - e_j}{h}\big) + K'\big(\frac{x - g(X_i) - e_j}{h}\big) \\ &\times\big(\frac{\tilde{g}(X_j) - g(X_j) - (\tilde{g}(X_i) - g(X_i))}{h}\big) + O_P\big(\big(\frac{\tilde{g}(X_j) - g(X_j) - (\tilde{g}(X_i) - g(X_i))}{h}\big)^2\big)\big]\Big]. \qquad (40)\end{aligned}$$

Thus,

$$\hat{f}_Y(x) - \tilde{f}_Y(x) =$$
$$\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}\sum_{j=2}^{n}\big[K'\big(\frac{x-g(X_i)-e_j}{h}\big)\cdot\frac{\tilde{g}(X_j)-g(X_j)}{h}$$
$$+K'\big(\frac{x-g(X_i)-e_j}{h}\big)\cdot\frac{g(X_i)-\tilde{g}(X_i)}{h}\big]\Big]$$
$$+\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}\sum_{j=2}^{n}O_P\big(\big(\frac{\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))}{h}\big)^2\big)\Big]. \quad (41)$$

Let us first examine the second term in the Taylor expansion above. We observe that for $i = j$ it is zero. For $i \neq j$, if we denote by F_n the common empirical distibution function of X_i and X_j,

$$\frac{1}{h^3}\Big|\frac{1}{n}\sum_{i=1}^{n}\big[\frac{1}{(n-1)}\sum_{j=2}^{n}\big(\tilde{g}(X_j)-g(X_j)-(\tilde{g}(X_i)-g(X_i))\big)\big]^2\Big|$$
$$\leq \frac{1}{h^3}\iint \sup_{x_1\in S(X),x_2\in S(X)}\big(\tilde{g}(x_2)-g(x_2)-(\tilde{g}(x_1)-g(x_1))\big)^2 \mathrm{d}F_n(x_1)\mathrm{d}F_n(x_2)$$
$$\leq \frac{1}{h^3}\sup_{x_1\in S(X),x_2\in S(X)}\big(\tilde{g}(x_2)-g(x_2)-(\tilde{g}(x_1)-g(x_1))\big)^2 \iint \mathrm{d}F_n(x_1)\mathrm{d}F_n(x_2)$$
$$\leq \frac{1}{h^3}\sup_{x_1\in S(X)}\big(g(x_1)-\tilde{g}(x_1)\big)^2 + \sup_{x_1\in S(X),x_2\in S(X)}\Big|2\big(g(x_1)-\tilde{g}(x_1)\big)$$
$$\times\big(\tilde{g}(x_2)-g(x_2)\big)\Big| + \sup_{x_2\in S(X)}\big(\tilde{g}(x_2)-g(x_2)\big)^2. \quad (42)$$

Using a uniform convergence result for the Nadarya-Watson estimator and making use of assumptions 5 and 6, see Mack & Silverman (1982),

$$\sup_{x\in S(X)}|\tilde{g}(x)-g(x)| = O_P\Big(\big[\frac{1}{nh_R}\log\big(\frac{1}{h_R}\big)\big]^{1/2}\Big), \quad (43)$$

and a well-known result from order in probability, we obtain that the order of the term in question is

$$O_P\Big(\frac{1}{nh_R\cdot h^3}\log\big(\frac{1}{h_R}\big)\Big), \quad (44)$$

where h_R is defined in (6). Using the same argument as when evaluating (42) 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 (44). (See below for the existence of these quantities under the assumption $\inf_{x\in S(X)} f_X(x) > 0$.)

There is a potential to improve on (44), since the evalution is quite crude. An alternative would be to try to evaluate the order of the second order term directly, as will be done presently for the first order term using the convolution property, and then include a third order term which can be evaluated (crudely) as above, resulting in a term $O_p\Big(\big[\frac{1}{nh_Rh^3}\log\frac{1}{h_R}\big]^{3/2}\Big)$. For the crude estimate (44) to be of order $O(\frac{1}{\sqrt{n}})$, we must have

$$h_Rh^3 = O(n^{-1/2-\epsilon}) \text{ for some } \epsilon > 0. \quad (45)$$

Further, let ut examine the first term of the Taylor expansion of $\hat{f}_Y(x) - \tilde{f}_Y(x)$. If this is of larger order than (44), this will be the contributing term. We start by examining the expectation of this term. For the expectation of this term 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 exists if the density function f_X is such that $\inf_{x \in S(X)} f_X(x) > 0$. This is assumed in the following. 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

$$\mathrm{E}\Big(\frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=2}^{n} K'\big(\frac{x - g(X_i) - e_j}{h}\big) \cdot \big(\frac{\tilde{g}(X_j) - g(X_j)}{h}\big)\Big)$$

$$\sim \frac{1}{h^2} \iiint K'\big(\frac{x - g(x_1) - u}{h}\big)\big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_e(u) f_X(x_1) f_X(x_2) \mathrm{d}u \mathrm{d}x_1 \mathrm{d}x_2. \quad (46)$$

Now Taylor expanding and using a convolution argument, we obtain

$$\begin{aligned}
&\frac{1}{h} \iiint K'(z_1)\big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_e(x - g(x_1) - z_1 h) f_X(x_1) f_X(x_2) \mathrm{d}z_1 \mathrm{d}x_1 \mathrm{d}x_2 \\
&\qquad = \frac{1}{h} \iint K'(z_1)\big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_Y(x - z_1 h) f_X(x_2) \mathrm{d}z_1 \mathrm{d}x_2 = \\
&\qquad\qquad -f_Y'(x) \int z_1 K'(z_1) \mathrm{d}z_1 \int \big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_X(x_2) \mathrm{d}x_2 \\
&\qquad + \frac{h}{2} f_Y''(x) \int z_1^2 K'(z_1) \mathrm{d}z_1 \int \big(\mathrm{E}(\tilde{g}(x_2)) - g(x_2)\big) f_X(x_2) \mathrm{d}x_2 + O(h^4). \quad (47)
\end{aligned}$$

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 (41), by similar arguments

$$-\mathrm{E}\Big(\frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=2}^{n} K'\big(\frac{x - g(X_i) - e_j}{h}\big) \cdot \big(\frac{\tilde{g}(X_i) - g(X_i)}{h}\big)\Big)$$

$$\sim \int z_2 K'(z_2) \mathrm{d}z_2 \int \big(\mathrm{E}(\tilde{g}(x_1)) - g(x_1)\big) f_e'\big(x - g(x_1)\big) f_X(x_1) \mathrm{d}x_1 + O(h^4). \quad (48)$$

In total, the terms (47) and (48) are of order h^2, but can be reduced by higher order kernels.

Further, we examine the variance of the first order term in (41). The condition $\inf_{x \in S(X)} f_X(x) > 0$ 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,

$$\mathrm{var}\Big(\frac{1}{n(n-1)h^2} \sum_{i=1}^{n} \sum_{j=2}^{n} \Big[K'\big(\frac{x - g(X_i) - e_j}{h}\big)\big(g(X_i) - \tilde{g}(X_i) + \tilde{g}(X_j) - g(X_j)\big)\Big]\Big)$$

$$= \frac{1}{n^2(n-1)^2 h^4} \sum_{i=1}^{n} \sum_{j=2}^{n} \sum_{k=1}^{n} \sum_{l=2}^{n} \mathrm{cov}\Big[K'\big(\frac{x - g(X_i) - e_j}{h}\big)\big(g(X_i) - \tilde{g}(X_i) + \tilde{g}(X_j) - g(X_j)\big),$$

$$K'\big(\frac{x - g(X_k) - e_l}{h}\big)\big(g(X_k) - \tilde{g}(X_k) + \tilde{g}(X_l) - g(X_l)\big)\Big]. \quad (49)$$

The evaluation of these terms can be found in the appendix. 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 $\hat{f}_Y(x)$.

In total the bias of $\hat{f}_Y(x)$ will consists of the terms (14), (47) and (48). This is of order h^2, 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 (45) is fulfilled, the total variance has a leading term given by equation (38), 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 as a smooth function 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 estimatated functionals are obtained.

4 Evaluating the convolution estimator

To evaluate the finite sample properties of the proposed estimator, (9), we use simulation experiments to compare the convolution estimates with the estimates from the classical 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 the convolution estimator is

$$\mathrm{MISE}(\hat{f}) = \mathrm{E}\Big[\int_{-\infty}^{\infty} (\hat{f} - f)^2(x)\mathrm{d}x\Big], \tag{50}$$

and likewise for the kernel density estimator. We have used 100 simulated realizations with sample sizes from 100 to 5000 of the model (2), with different choices of the function $g(\cdot)$ and distributions of X and e. The value of MISE is approximated as an average of the ISE (integrated squared error) of the 100 realizations, and ISE is estimated by numerical integration. If the true density f_Y is not known analytically, we have based our comparisons on an estimated kernel density computed from 1 000 000 generated observations of (X_i, Y_i), using this as the "true" density.

The choice of bandwidth in non-parametric estimation has a considerable impact on the accuracy of the estimator. The bandwidth, h_D, used in the kernel density estimation in our simulation study, is the Solve-the-Equation Plug-in Apparoach proposed in Sheather & Jones (1991), and is the same for all the $(n-1)$ density estimations in equation (9). 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, $\hat{\sigma}$ is the variance of all observations $X_1, ..., X_n$. Thus we may actually obtain better results using a more optimal bandwidth for the non-parametric regression.

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. $g(x) = x^2,\ X \sim \mathrm{U}[0,2],\ e \sim (\frac{1}{2}\mathrm{N}(-3/2, 1/2) + \frac{1}{2}\mathrm{N}(3/2, 1/2))$.

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 would like to see how the estimator performs in some extreme cases. In most cases of the examples the compactness assumption 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. The second parameter given for the normal distributions is the standard deviation.

The simulation results are given in table 1. In some simulations only 30 realizations are performed, and those are marked. The table shows the percentage change by using the convolution estimator compared with the kernel density estimator. For the MISE, this change is calculated by

$$\frac{\mathrm{MISE}(f_Y^*) - \mathrm{MISE}(\hat{f}_Y)}{\mathrm{MISE}(f_Y^*)} \cdot 100, \tag{51}$$

for the squared bias

$$\frac{[\mathrm{Ave}(f_Y^* - f_Y)]^2 - [\mathrm{Ave}(\hat{f}_Y - f_Y)]^2}{[\mathrm{Ave}(f_Y^* - f_Y)]^2} \cdot 100, \tag{52}$$

where

$$[\mathrm{Ave}(f_Y^* - f_Y)]^2 = \frac{1}{k}\sum_{j=1}^{k}\big[\big(\frac{1}{100}\sum_{i=1}^{100} f_Y^{*i}(x_j)\big) - f_Y(x_j)\big]^2, \tag{53}$$

and similarly for the convolution estimator. In (53) 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

$$\frac{\hat{\text{var}}(f_Y^*) - \hat{\text{var}}(\hat{f}_Y)}{\hat{\text{var}}(f_Y^*)} \cdot 100, \tag{54}$$

where

$$\hat{\text{var}}(f_Y^*) = \frac{1}{k}\sum_{j=1}^{k}\Big[\frac{1}{99}\Big(\sum_{i=1}^{100}\big(f_Y^{*i}(x_j) - \text{Ave}\{f_Y^*(x_j)\}\big)^2\Big)\Big] \tag{55}$$

and similarly for the convolution estimator. Here $\text{Ave}\{f_Y^*(x_j)\}$ denotes the average of all 100 (or 30) 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.

In many cases the MISE is smallest for the convolution estimator, however, in cases where the variance of the error terms is relatively small, especially for model 1, 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, thus we obtain the extreme results in table 1.

In the non-linear models 5 and 6 with normally distributed error terms, the convolution estimator is clearly better. But introducing asymmetric and multimodal distributions on 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 1 shows one simulation of sample size 500 from this model. In the upper plot, the simulated X_i and Y_i is given as points, and the estimated $\tilde{g}$ is given as the solid line. 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 line, the kernel density estimate is given as the solid line and the convolution estimate as the dashed line. 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 many simulations, and when the sample size increases, the improvements are also increasing. The squared bias is smallest almost for all cases for the convolution estimator. This is somewhat suprising since from the asymptotic analysis it is of the same order as the kernel estimator. Figure 2 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,

$$\text{E}\big(f_Y^*(x)\big) - f_Y(x) = \frac{h^2}{2}f_Y''(x)\int w^2K(w)\text{d}w + O(h^2), \tag{56}$$

and it behaves as we would expect, since it 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 in fact 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 exact expressions for the bias of the convolution estimator and compare it to the observed bias in figure 2. The bias of the convolution estimator consists of three terms, (14), (47) and (48). Equation (14) is now

$$\mathrm{E}\big(\tilde{f}_Y(x)\big) - f_Y(x) = \frac{h^2}{2} f_Y''(x) \int w^2 K(w)\mathrm{d}w$$
$$= \frac{h^2}{2} \int w^2 K(w)\mathrm{d}w \Big[- \frac{1}{4\sqrt{\pi}} \exp\{-1/4(x-1)^2\}(1 - \frac{(x-1)^2}{16\sqrt{\pi}}) \Big]. \tag{57}$$

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

In equation (47) and (48), the bias of the Nadaraya-Watson estimator of $g(x)$ is a part of the expression. This bias is well-known and is

$$\mathrm{E}\big(\tilde{g}(x)\big) - g(x) = h^2 \Big(\frac{1}{2} g''(x) + \frac{g'(x) f_X'(x)}{f_X(x)} \Big) \int u^2 K(u)\mathrm{d}u. \tag{58}$$

Since $\int u^2 K(u)\mathrm{d}u = 1$, $g(x) = x$ and f_X is a normal distribution, this expression will be equal to $-2(x-1)h^2$. Inserting this in equation (47) and again using the fact that f_X is normal with mean and variance equal to one gives, for the leading term in (47),

$$-f_Y'(x) \int z_1 K'(z_1)\mathrm{d}z_1 \int \big(-2(x_2-1)h^2 \frac{1}{\sqrt{2\pi}} \exp(-(x_2-1)^2/2) \big)\mathrm{d}x_2. \tag{59}$$

Observe that the last integral above is equal to zero. Thus we have no contribution to the bias from this expression.

Further, equation (48) yields,

$$h^2 \int z_2 K'(z_2)\mathrm{d}z_2 \int \big((-2(x_1-1)(\frac{-(x-x_1)}{\sqrt{2\pi}} \exp(-(x-x_1)^2/2)) \frac{1}{\sqrt{2\pi}} \exp(-(x_1-1)^2/2) \big)\mathrm{d}x_1. \tag{60}$$

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 total bias of the convolution estimator will in this case be

$$\mathrm{E}\big(\hat{f}_Y(x)\big) - f_Y(x) = \frac{h^2}{2} \Big[- \frac{1}{4\sqrt{\pi}} \exp\{-1/4(x-1)^2\}(1 - \frac{(x-1)^2}{16\sqrt{\pi}}) \Big]$$
$$-h^2 \int \big((-2(x_1-1)(\frac{-(x-x_1)}{\sqrt{2\pi}} \exp(-(x-x_1)^2/2)) \frac{1}{\sqrt{2\pi}} \exp(-(x_1-1)^2/2) \big)\mathrm{d}x_1. \tag{61}$$

This expression is plotted in figure 3, with a reasonable choice for the bandwidth, $h = 0.3$. And taking the different scaling into account, this graph is comparable to the empirical bias from figure 2. Similar explanations are possible for the other models, although some problems arise in the computation in the cases where the true density is not known.

Model	Sample size	Squared bias	Variance	MISE
1	100	61.5 %	-665.0 %	-504.7 %
1	500	79.9 %	-464.7 %	-346.4 %
1	5000 (30)	-638.8 %	-265.9 %	-314.4 %
2	100	79.2 %	23.5 %	34 %
2	500	97.7 %	48.2 %	59.2 %
2	5000 (30)	97.8 %	77.8 %	81.9 %
3	100	93.4 %	-116.0 %	-80.2 %
3	500	93.1 %	-40.6 %	-13.9 %
3	5000 (30)	43.7 %	15.9 %	19.6 %
4	100	62.1 %	19.1 %	27.5 %
4	500	68.2 %	1.6 %	17.7 %
4	5000	80.3 %	21.9 %	32.5 %
5	100	94.6 %	28.0 %	34.9 %
5	500	83.9 %	42.4 %	50.9 %
5	5000	79.4 %	60.2 %	64.4 %
6	100	87.6 %	-4.2 %	15.5 %
6	500	83.4 %	33.3 %	44.9 %
6	5000 (30)	79.1 %	62.2 %	64.3 %
7	100	97.5 %	25.0 %	37.2 %
7	500	96.4 %	31.5 %	43.4 %
7	5000	86.8 %	47.5 %	55.9 %
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	-49.3 %	-15.3 %	-25.5 %
10	500	-23.1 %	-14.9 %	-16.9 %
10	5000	70.1 %	-5.8 %	2.0 %

Table 1: Percentage improvements in estimations using the convolution density estimator compared with the kernel density estimator. 100 realizations (in some cases 30, as indicated). The MISE, squared bias and variance are explained in formula (51), (52), (54), and below. A minus sign indicates that the kernel density estimator performs best.

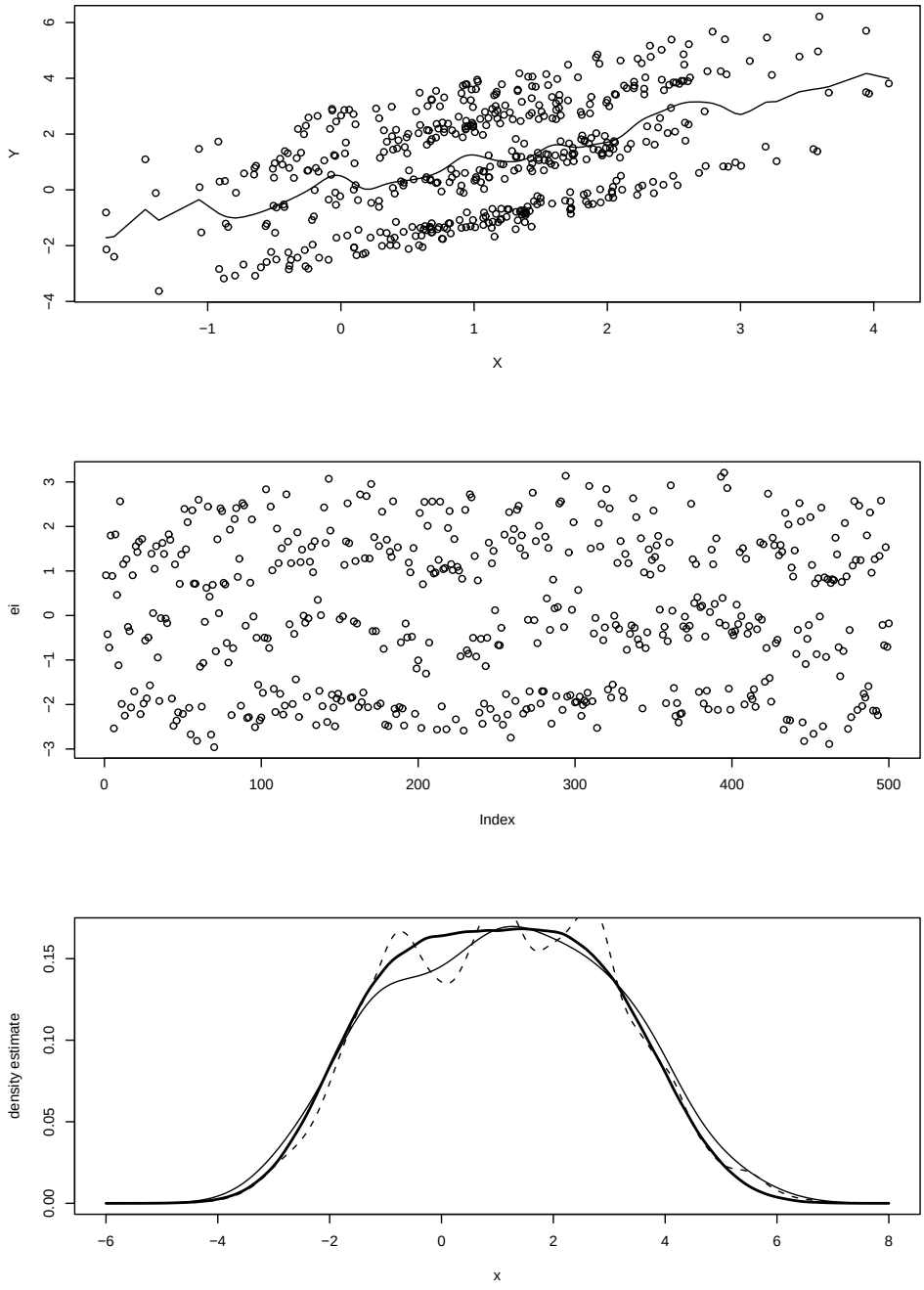

Figure 1: Upper plot: The estimated g function (solid line) and the simulated points (X_i, Y_i) from one simulation of sample size 500 from model 8. Middle plot: The corresponding estimated e_i. Lower plot: The "true" density f_Y (thick solid line) and the estimated densities (dashed line - convolution estimator, solid line - kernel density estimator).

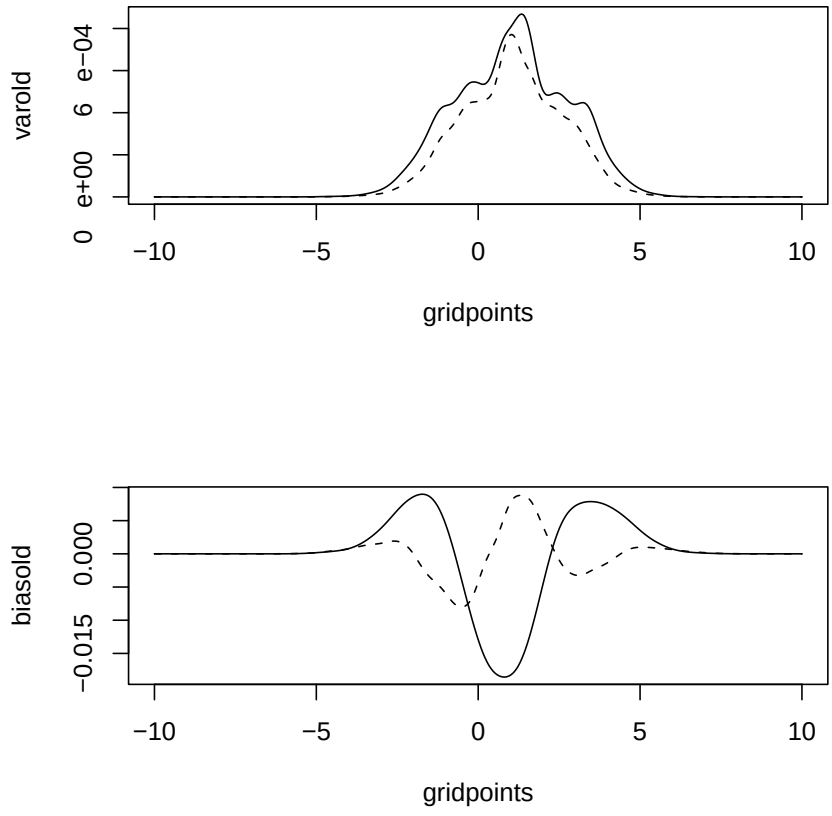

Figure 2: The estimated variance (top) and bias for model 2 (dashed line - convolution estimator, solid line - kernel density estimator).

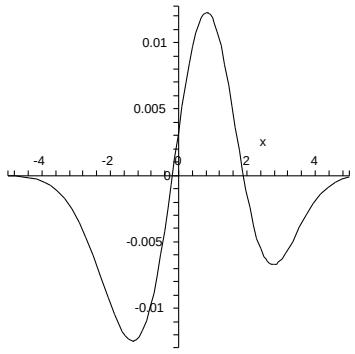

Figure 3: The theoretical bias of the convolution estimator in model 2.

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.

As suggested in section 2, 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. These results are in most cases better than the corresponding results using the Nadaraya-Watson estimator, given in table 1.

A real data set has been studied for completeness. The data is the motorcycle data set, 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 4. It seems that 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 seems to outperform the usual kernel estimator in many situations, especially if the error term density function is smooth and has a relatively large variance. We believe that the cases where it does not perform so

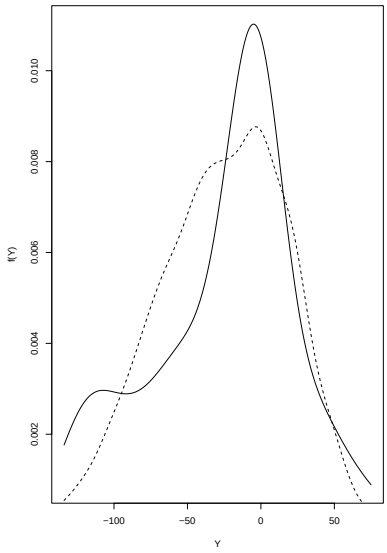

Figure 4: The estimated densities of a real data set (solid line - kernel and dashed line - convolution)

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. Also, by selecting the bandwidth parameters in the convolution estimator more correctly, by e.g. a cross-validation technique, we could possibly improve the estimates even more. We also believe that this estimator can be used in a more general time-series setting, $X_t = g(X_{t-1}) + e_t$, where 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 that the order of the variance of the convolution estimator will also be n^{-1} in this situation, see Støve & Tjøstheim (2005*a*).

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. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall.

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

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', *Nonparametric Statistics* **16**, 925–935.

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. (2005*a*), 'A convolution density estimator for nonlinear time series: Simulations and some preliminary analysis', Work in progress.

Støve, B. & Tjøstheim, D. (2005*b*), 'A new convolution estimator for nonparametric regression', Submitted.

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

Appendix

The variance in (49) can be classified to the following (see also Saavedra & Cao (2000)):

$$\frac{1}{n^2(n-1)^2h^4}\Big[(n-1)\mathrm{var}\Big(K'\big(\frac{x-g(X_1)-e_1}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_1\ )-g(X_1)\big)\Big) \quad (62)$$

$$+(n-1)(n-2)\mathrm{var}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big)\Big) \quad (63)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big),$$
$$K'\big(\frac{x-g(X_1)-e_3}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_3)-g(X_3)\big)\Big) \quad (64)$$

$$+2(n-1)(n-2)(n-3)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big),$$
$$K'\big(\frac{x-g(X_3)-e_1}{h}\big)\big(g(X_3)-\tilde{g}(X_3)+\tilde{g}(X_1)-g(X_1)\big)\Big) \quad (65)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big),$$
$$K'\big(\frac{x-g(X_3)-e_2}{h}\big)\big(g(X_3)-\tilde{g}(X_3)+\tilde{g}(X_2)-g(X_2)\big)\Big) \quad (66)$$

$$+(n-1)(n-2)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big),$$
$$K'\big(\frac{x-g(X_2)-e_1}{h}\big)\big(g(X_2)-\tilde{g}(X_2)+\tilde{g}(X_1)-g(X_1)\big)\Big) \quad (67)$$

$$+2(n-1)(n-2)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_1}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_1)-g(X_1)\big),$$
$$K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big)\Big) \quad (68)$$

$$+2(n-1)(n-2)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_1}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_1)-g(X_1)\big),$$
$$K'\big(\frac{x-g(X_2)-e_1}{h}\big)\big(g(X_2)-\tilde{g}(X_2)+\tilde{g}(X_1)-g(X_1)\big)\Big) \quad (69)$$

$$+(n-1)(n-2)(n-3)(n-4)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big),$$
$$K'\big(\frac{x-g(X_3)-e_4}{h}\big)\big(g(X_3)-\tilde{g}(X_3)+\tilde{g}(X_4)-g(X_4)\big)\Big) \quad (70)$$

$$+(n-1)(n-2)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_1}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_1)-g(X_1)\big),$$
$$K'\big(\frac{x-g(X_2)-e_2}{h}\big)\big(g(X_2)-\tilde{g}(X_2)+\tilde{g}(X_2)-g(X_2)\big)\Big) \quad (71)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K'\big(\frac{x-g(X_1)-e_1}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_1)-g(X_1)\big),$$
$$K'\big(\frac{x-g(X_2)-e_3}{h}\big)\big(g(X_2)-\tilde{g}(X_2)+\tilde{g}(X_3)-g(X_3)\big)\Big) \quad (72)$$

$$+(n-1)(n-2)(n-3)\mathrm{cov}\Big(K'\big(\frac{x-g(X_2)-e_1}{h}\big)\big(g(X_2)-\tilde{g}(X_2)+\tilde{g}(X_1)-g(X_1)\big),$$
$$K'\big(\frac{x-g(X_1)-e_3}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_3)-g(X_3)\big)\Big)\Big]. \quad (73)$$

The terms (70), (71) and (72) are neglible by independence, and the terms (62), (68) and (69) are automatically zero. Hence we study the remaining six terms, using $\mathrm{var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$ and $\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.

Looking at the term (63),

$$(n-1)(n-2)\mathrm{var}\Big(K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big)\Big) \doteq (n-1)(n-2)\mathrm{var}(U). \quad (74)$$

Starting by examining $\mathrm{E}(U^2)$, using the same techniques as when deriving (47) and (48), and since $\int z\big(K'(z)\big)^2\mathrm{d}z = 0$, we obtain,

$$\begin{aligned}\mathrm{E}(U^2) \sim \iiint K'\big(\frac{x-g(v)-u}{h}\big)^2\mathrm{E}\big[\big(g(v)-\tilde{g}(v)+\tilde{g}(w)-g(w)\big)^2\big] \\ \times f_e(u)f_X(v)f_X(w)\mathrm{d}u\mathrm{d}v\mathrm{d}w = h\iiint K'(z)^2\mathrm{E}\big[\big(g(v)-\tilde{g}(v)+\tilde{g}(w)-g(w)\big)^2\big] \\ \times f_X(v)f_X(w)f_e(x-g(v)-zh)\mathrm{d}v\mathrm{d}w\mathrm{d}z = h\iiint K'(z)^2 \\ \times\mathrm{E}\big[\big(g(v)-\tilde{g}(v)+\tilde{g}(w)-g(w)\big)^2\big]f_X(v)f_X(w)f_e(x-g(v))\mathrm{d}v\mathrm{d}w\mathrm{d}z + O(h^5). \quad (75)\end{aligned}$$

Next, using the same technique,

$$\begin{aligned}\big(\mathrm{E}(U)\big)^2 \sim h^2\Big[\iiint K'(z)\mathrm{E}\big[g(v)-\tilde{g}(v)+\tilde{g}(w)-g(w)\big] \\ \times f_X(v)f_X(w)f_e(x-g(v))\mathrm{d}v\mathrm{d}w\mathrm{d}z\Big]^2 + O(h^5). \quad (76)\end{aligned}$$

Thus the leading term of (63) will be $O((n-1)(n-2)h^4)$, since the expectation in the term (76) is the bias of the Nadarya-Watson estimator, which is $O(h^2)$, see e.g. Härdle (1990) page 135.

Examining the term (64)

$$\begin{aligned}(n-1)(n-2)(n-3)\mathrm{cov}\Big[K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big), \\ K'\big(\frac{x-g(X_1)-e_3}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_3)-g(X_3)\big)\Big] \\ \doteq (n-1)(n-2)(n-3)\mathrm{cov}(U,V). \quad (77)\end{aligned}$$

Calculating $\mathrm{E}(UV)$, by conditioning, substituting, Taylor expanding, and since $\int K'(z)\mathrm{d}z =$

0, we obtain,

$$\begin{aligned}
\mathrm{E}(UV) \sim \iiiint \int K'\big(\frac{x-g(v)-u}{h}\big)\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(w)\big)-g(w)\Big) \\
\times K'\big(\frac{x-g(v)-y}{h}\big)\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(a)\big)-g(a)\Big) \\
\times f_e(u)f_e(y)f_X(v)f_X(w)f_X(a)\mathrm{d}u\mathrm{d}y\mathrm{d}v\mathrm{d}w\mathrm{d}a \\
= h^2 \iiiint \int K'(z_1)K'(z_2)\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(w)\big)-g(w)\Big) \\
\times\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(a)\big)-g(a)\Big)f_X(v)f_X(w)f_X(a) \\
\times f_e(x-g(v)-z_1h)f_e(x-g(v)-z_2h)\mathrm{d}v\mathrm{d}w\mathrm{d}a\mathrm{d}z_1\mathrm{d}z_2 \\
= h^2 \iiiint \int K'(z_1)K'(z_2)\Big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}\big(\tilde{g}(w)\big)-g(w)\Big) \\
\times\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(a)\big)-g(a)\Big)f_X(v)f_X(w)f_X(a) \\
\times\big(f_e(x-g(v))-z_1hf_e'(x-g(v))+\frac{z_1^2h^2}{2}f_e''(x-g(v))+O(h^2)\big) \\
\times\big(f_e(x-g(v))-z_2hf_e'(x-g(v))+\frac{z_2^2h^2}{2}f_e''(x-g(v))+O(h^2)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}a\mathrm{d}z_1\mathrm{d}z_2 \\
= h^4 \iiiint \int z_1K'(z_1)z_2K'(z_2)\Big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\Big) \\
\times\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(a)\big)-g(a)\Big)f_X(v)f_X(w)f_X(a) \\
\times f_e'\big(x-g(v)\big)f_e'\big(x-g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}a\mathrm{d}z_1\mathrm{d}z_2+O(h^9). \qquad (78)
\end{aligned}$$

Further,

$$\begin{aligned}
\mathrm{E}(U) \sim -h^2 \iiint z_1K'(z_1)\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(w)\big)-g(w)\Big) \\
\times f_X(v)f_X(w)f_e'\big(x-g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1+O(h^5), \qquad (79)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(V) \sim -h^2 \iiint z_2K'(z_2)\Big(g(v)-\mathrm{E}\big(\tilde{g}(v)\big)+\mathrm{E}\big(\tilde{g}(a)\big)-g(a)\Big) \\
\times f_X(v)f_X(a)f_e'\big(x-g(v)\big)\mathrm{d}v\mathrm{d}a\mathrm{d}z_2+O(h^5). \qquad (80)
\end{aligned}$$

In total, the leading term of (64) is $O((n-1)(n-2)(n-3)h^8)$, again using the order of the bias of the Nadaraya-Watson estimator.

Examining the term (65),

$$\begin{aligned}
2(n-1)(n-2)(n-3)\mathrm{cov}\Big[K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big), \\
K'\big(\frac{x-g(X_3)-e_1}{h}\big)\big(g(X_3)-\tilde{g}(X_3)+\tilde{g}(X_1)-g(X_1)\big)\Big] \\
\doteq 2(n-1)(n-2)(n-3)\mathrm{cov}(U,V). \qquad (81)
\end{aligned}$$

We have

$$\begin{aligned}\mathrm{E}(UV) \sim \iiiint \int & K'\big(\frac{x-g(v)-u}{h}\big)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)\\ & \times K'\big(\frac{x-g(a)-y}{h}\big)\big(g(a)-\mathrm{E}(\tilde{g}(a))+\mathrm{E}(\tilde{g}(v))-g(v)\big)\\ & \times f_e(u)f_e(y)f_X(v)f_X(w)f_X(a)\mathrm{d}u\mathrm{d}y\mathrm{d}v\mathrm{d}w\mathrm{d}a\\ = h^4 \iiiint \int & z_1K'(z_1)z_2K'(z_2)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)\\ & \times\big(g(a)-\mathrm{E}(\tilde{g}(a))+\mathrm{E}(\tilde{g}(v))-g(v)\big)f_X(v)f_X(w)f_X(a)\\ & \times f'_e\big(x-g(v)\big)f'_e\big(x-g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}a\mathrm{d}z_1\mathrm{d}z_2+O(h^9), \qquad (82)\end{aligned}$$

$$\begin{aligned}\mathrm{E}(U) \sim -h^2 \iiint & z_1K'(z_1)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)\\ & \times f_X(v)f_X(w)f'_e\big(x-g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1+O(h^5), \qquad (83)\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(V) \sim -h^2 \iiint & z_2K'(z_2)\big(g(a)-\mathrm{E}(\tilde{g}(a))+\mathrm{E}(\tilde{g}(b))-g(b)\big)\\ & \times f_X(a)f_X(b)f'_e\big(x-g(a)\big)\mathrm{d}a\mathrm{d}b\mathrm{d}z_2+O(h^5). \qquad (84)\end{aligned}$$

The above calculations give that the order of the term (65) is $O((n-1)(n-2)(n-3)h^8)$.

Examining the term (73),

$$\begin{aligned}2(n-1)(n-2)(n-3)\mathrm{cov}\Big[K'\big(\frac{x-g(X_2)-e_1}{h}\big)\big(g(X_2)-\tilde{g}(X_2)+\tilde{g}(X_1)-g(X_1)\big),\\ K'\big(\frac{x-g(X_1)-e_3}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_3)-g(X_3)\big)\Big]\\ \doteq 2(n-1)(n-2)(n-3)\mathrm{cov}(U,V). \qquad (85)\end{aligned}$$

As before,

$$\begin{aligned}\mathrm{E}(UV) \sim h^4 \iiiint \int & z_1K'(z_1)z_2K'(z_2)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)\\ & \times\big(g(w)-\mathrm{E}(\tilde{g}(w))+\mathrm{E}(\tilde{g}(b))-g(b)\big)f_X(v)f_X(w)f_X(b)\\ & \times f'_e\big(x-g(w)\big)f'_e\big(x-g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1\mathrm{d}z_2\mathrm{d}b+O(h^9). \qquad (86)\end{aligned}$$

Further,

$$\begin{aligned}\mathrm{E}(U) \sim -h^2 \iiint & z_1K'(z_1)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)\\ & \times f_X(v)f_X(w)f'_e\big(x-g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1+O(h^5). \qquad (87)\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(V) \sim -h^2 \iiint & z_2K'(z_2)\big(g(a)-\mathrm{E}(\tilde{g}(a))+\mathrm{E}(\tilde{g}(b))-g(b)\big)\\ & \times f_X(a)f_X(b)f'_e\big(x-g(a)\big)\mathrm{d}a\mathrm{d}b\mathrm{d}z_2+O(h^5). \qquad (88)\end{aligned}$$

Thus the term (73) is of order $O((n-1)(n-2)(n-3)h^8)$.

Examining the term (66),

$$2(n-1)(n-2)(n-3)\mathrm{cov}\Big[K'\big(\frac{x-g(X_1)-e_2}{h}\big)\big(g(X_1)-\tilde{g}(X_1)+\tilde{g}(X_2)-g(X_2)\big),$$
$$K'\big(\frac{x-g(X_3)-e_2}{h}\big)\big(g(X_3)-\tilde{g}(X_3)+\tilde{g}(X_2)-g(X_2)\big)\Big]$$
$$\doteq 2(n-1)(n-2)(n-3)\mathrm{cov}(U,V). \quad (89)$$

We have

$$\mathrm{E}(UV)\sim \iiiint K'\big(\frac{x-g(v)-u}{h}\big)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)$$
$$\times K'\big(\frac{x-g(a)-u}{h}\big)\big(g(a)-\mathrm{E}(\tilde{g}(a))+\mathrm{E}(\tilde{g}(w))-g(w)\big)$$
$$\times f_e(u)f_X(v)f_X(w)f_X(a)\mathrm{d}u\mathrm{d}v\mathrm{d}w\mathrm{d}a. \quad (90)$$

Introducing $(g^{-1})'=r$ and $f_X(g^{-1})=l_X$ as in the derivation of (33), and noting that

$$r'\big(g(v)\big)=-\frac{g''(v)}{g'(v)^2} \quad (91)$$

and

$$l_X'\big(g(v)\big)=f_X'(v), \quad (92)$$

we obtain by using $\int K'(z)\mathrm{d}z=0$,

$$\mathrm{E}(UV)\sim h^2\iiiint K'(z_1)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)K'(z_2)$$
$$\times\Big[g\Big(g^{-1}\big(g(v)+h(z_1-z_2)\big)\Big)-\mathrm{E}\Big(\tilde{g}\big(g^{-1}(g(v)+h(z_1-z_2))\big)\Big)$$
$$+\mathrm{E}(\tilde{g}(w))-g(w)\Big]f_e(x-g(v)-z_1h)f_X(v)f_X(w)l_X\big(g(v)+h(z_1-z_2)\big)$$
$$\times r\big(g(v)+h(z_1-z_2)\big)\mathrm{d}z_1\mathrm{d}v\mathrm{d}w\mathrm{d}z_2=h^4\iiiint K'(z_1)$$
$$\times\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)K'(z_2)$$
$$\times\Big[g\big(g^{-1}(g(v)+h(z_1-z_2))\big)-\mathrm{E}\Big(\tilde{g}\big(g^{-1}(g(v)+h(z_1-z_2))\big)\Big)$$
$$+\mathrm{E}(\tilde{g}(w))-g(w)\Big]f_X(v)f_X(w)\Big\{z_1z_2r'\big(g(v)\big)l_X\big(g(v)\big)f_e'\big(x-g(v)\big)$$
$$-2z_1z_2r'\big(g(v)\big)f_e\big(x-g(v)\big)l_X'\big(g(v)\big)$$
$$+z_1z_2f_e'\big(x-g(v)\big)r\big(g(v)\big)l_X'\big(g(v)\big)\Big\}\mathrm{d}v\mathrm{d}w\mathrm{d}z_1\mathrm{d}z_2+O(h^9)=$$
$$h^4\iiiint K'(z_1)\big(g(v)-\mathrm{E}(\tilde{g}(v))+\mathrm{E}(\tilde{g}(w))-g(w)\big)K'(z_2)$$
$$\times\Big[g\big(g^{-1}(g(v)+h(z_1-z_2))\big)-\mathrm{E}\Big(\tilde{g}\big(g^{-1}(g(v)+h(z_1-z_2))\big)\Big)+\mathrm{E}(\tilde{g}(w))-g(w)\Big]$$
$$\times f_X(v)f_X(w)\Big\{-z_1z_2\frac{g''(v)}{g'(v)^2}f_X(v)f_e'\big(x-g(v)\big)+2z_1z_2\frac{g''(v)}{g'(v)^2}f_X'(v)f_e\big(x-g(v)\big)$$
$$+z_1z_2f_e'\big(x-g(v)\big)\frac{f_X'(v)}{g'(v)}\Big\}\mathrm{d}v\mathrm{d}w\mathrm{d}z_1\mathrm{d}z_2+O(h^9). \quad (93)$$

Moreover,

$$\mathrm{E}(U) \sim -h^2 \iiint z_1 K'(z_1)\big(g(v) - \mathrm{E}(\tilde{g}(v)) + \mathrm{E}(\tilde{g}(w)) - g(w)\big) \times f_X(v) f_X(w) f_e'\big(x - g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1 + O(h^5). \tag{94}$$

and

$$\mathrm{E}(V) \sim -h^2 \iiint z_2 K'(z_2)\big(g(a) - \mathrm{E}(\tilde{g}(a)) + \mathrm{E}(\tilde{g}(b)) - g(b)\big) \times f_X(a) f_X(b) f_e'\big(x - g(a)\big)\mathrm{d}a\mathrm{d}b\mathrm{d}z_2 + O(h^5). \tag{95}$$

Thus the term (66) is of order $O((n-1)(n-2)(n-3)h^8)$.

Finally, examining the term (67),

$$(n-1)(n-2)\mathrm{cov}\Big(K'\big(\frac{x - g(X_1) - e_2}{h}\big)\big(g(X_1) - \tilde{g}(X_1) + \tilde{g}(X_2) - g(X_2)\big), K'\big(\frac{x - g(X_2) - e_1}{h}\big)\big(g(X_2) - \tilde{g}(X_2) + \tilde{g}(X_1) - g(X_1)\big)\Big) \doteq (n-1)(n-2)\mathrm{cov}(U, V). \tag{96}$$

Starting with $\mathrm{E}(UV)$,

$$\mathrm{E}(UV) \sim h^4 \iiiint z_1 K'(z_1) z_2 K'(z_2)\big(g(v) - \mathrm{E}(\tilde{g}(v)) + \mathrm{E}(\tilde{g}(w)) - g(w)\big) \times \big(g(w) - \mathrm{E}(\tilde{g}(w)) + \mathrm{E}(\tilde{g}(v)) - g(v)\big) f_X(v) f_X(w) \times f_e'\big(x - g(v)\big) f_e'\big(x - g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1\mathrm{d}z_2 + O(h^9). \tag{97}$$

As before,

$$\mathrm{E}(U) \sim -h^2 \iiint z_1 K'(z_1)\big(g(v) - \mathrm{E}(\tilde{g}(v)) + \mathrm{E}(\tilde{g}(w)) - g(w)\big) \times f_X(v) f_X(w) f_e'\big(x - g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_1 + O(h^5). \tag{98}$$

and

$$\mathrm{E}(V) \sim -h^2 \iiint z_2 K'(z_2)\big(g(w) - \mathrm{E}(\tilde{g}(w)) + \mathrm{E}(\tilde{g}(v)) - g(v)\big) \times f_X(w) f_X(v) f_e'\big(x - g(v)\big)\mathrm{d}v\mathrm{d}w\mathrm{d}z_2 + O(h^5). \tag{99}$$

Thus, the leading term of (67) is $O((n-1)(n-2)h^8)$.

All of these calculations give us that the leading terms of the variance from equation (49) will be of order $O(h^4/n)$.

PAPER B

In preperation for submission

A convolution density estimator for nonlinear time series: Simulations and some preliminary analysis

Bård Støve Dag Tjøstheim

Department of Mathematics
University of Bergen
Johannes Brunsgate 12
5008 Bergen
Norway

Abstract

We present a convolution estimator of the marginal density for a nonlinear time series. Some simulations have been performed, and the results seem to be better than corresponding results for the standard kernel density estimator. Some results concerning the asymptotic bias and variance of the proposed estimator are presented. They imply that the order of the asymptotic variance is n^{-1}.

Some key words: Convolution, Kernel function, Mean squared error, Nonparametric density estimation, Time series.

1 Introduction

There exists a vast literature which considers 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 commonly used method is the kernel density estimator

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

where K is a kernel function and h is the bandwidth.

Often the independence assumption is violated, e.g. in a time-series setting, and early references on kernel density estimation for dependent data are Roussas (1969) and Rosenblatt (1970). For further references, see Györfi et al. (1990) chapter 4 and Fan & Yao (2003) chapter 5.

In Saavedra & Cao (1999*b*) the authors introduced a convolution-kernel estimator for the marginal density of a moving average process, $Y_t = X_t - \theta X_{t-1}$, when θ is unknown. This estimator is proved to have a parametric rate of convergence, $n^{-1/2}$. Schick & Wefelmeyer (2004) introduced a slightly simplified variant of this estimator, and proved an even stronger result of asymptotic normality. Further, in Müller et al. (2005) a similar density estimator is introduced for the innovation density in nonlinear parametric autoregressive models. This article is related to these papers, but we focus on nonparametric density estimation in a more general nonlinear time series situation,

$$X_t = g(X_{t-1}) + e_t, \tag{2}$$

where g is an unknown function and $\{e_t\}$ is a sequence of zero-mean, independent and identically distributed random variables. The process $\{X_t\}$ is stationary; thus all the marginal densities are the same. Denote this common density by f_X, and this is density we want to estimate. If one is able to construct an estimator of the marginal density making explicit use of the extra information contained in (2), then possibly one could improve on the kernel density estimator. In Støve & Tjøstheim (2005) we used this idea in a nonlinear regression model, $Y_t = g(X_t) + e_t$, where the density of the responses Y_t is of interest. There the asymptotic variance of the proposed density estimator was shown to be of order n^{-1}. As we will argue in this paper, the density estimator for a nonlinear time series will also have this property. Note that since the function $g(\cdot)$ and the error terms $\{e_t\}$ are unknown, they both have to be estimated.

The proposed estimator, "the convolution density estimator", is presented in section 2, a preliminary (and incomplete) analysis of its asymptotic behaviour is given in section 3 and some simulation results are located in section 4. A short conclusion is presented in section 5.

2 The estimator

From equation (2) and since $g(X_{t-1})$ and e_t are independent, we get

$$f_X(x) = \int f_e\big(x - g(u)\big) f_X(u) \mathrm{d}u = \mathrm{E}\big[f_e\big(x - g(X)\big)\big], \tag{3}$$

where f_e is the density of the residuals. This motivates an improved estimation of the marginal density f_X by the above functional. Assume we have $X_1, ..., X_n$ observations of the process $\{X_t\}$. We can estimate g in (2), by e.g the Nadaraya-Watson estimator, see Härdle (1990) page 127, with bandwidth h_R, and kernel function $K^{NW}_{x,h_R}(X_i) = (1/h_R)K^{NW}\big((x - X_i)/h_R\big)$,

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

The estimate for e_i is

$$\tilde{e}_i = X_i - \tilde{g}(X_{i-1}). \tag{5}$$

Next we estimate f_e by the kernel estimator with bandwidth h_D and kernel function $K(\cdot)$,

$$f^*_{\tilde{e}}(x) = \frac{1}{(n-1)h_D}\sum_{i=2}^{n} K\big(\frac{x-\tilde{e}_i}{h_D}\big). \tag{6}$$

The final density estimator for f_X is then

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} f^*_{\tilde{e}}\big(x-\tilde{g}(X_i)\big) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h_D}\sum_{j=2}^{n} K\big(\frac{x-\tilde{g}(X_i)-\tilde{e_j}}{h_D}\big)\Big]. \tag{7}$$

3 Asymptotic properties

The asymptotic analysis of the estimator in equation (7) will be based on examining the asymptotic bias and variance. The analysis will be closely related to that in Støve & Tjøstheim (2005). Thus the basic assumptions are the same:

1. The kernel function K is a bounded non-negative, two times differentiable, symmetric function that integrates to 1, and satisfies

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

2. The function g is differentiable and its inverse exists.

3. The density f_X has compact support $S(X)$, is continous and two times differentiable on the support.

4. $\lim_{n\to\infty} h_D = 0$ and $\lim_{n\to\infty} nh_D = \infty$.

Since the main difference from the model in Støve & Tjøstheim (2005) is that we now have dependencies in the observations, we introduce:

5. $\{e_t, X_t\}$ is a strictly stationary process and mixing in a sense to be specified in the sequel.

The idea is to analyse the properties of the estimator as in Støve & Tjøstheim (2005), and by using the mixing, we will argue that the asymptotic properties in this case are essentially as in the independent case. We only outline the derivations. More work is needed to make this complete.

To study the moments of the estimator, it is useful to decompose the difference between the estimator and the true density in the following manner,

$$\hat{f}_X(x) - f_X(x) = \hat{f}_X(x) - \tilde{f}_X(x) + \tilde{f}_X(x) - f_X(x), \tag{8}$$

where

$$\tilde{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h_D}\sum_{j=2}^{n} K\Big(\frac{x-g(X_i)-e_j}{h_D}\Big)\Big], \tag{9}$$

is the "estimator" with $g(\cdot)$ and e_j for $j = 2, ..., n$, known. We start by examining $\tilde{f}_X(x)$. Note that to ease notation, we set $h_D = h$ in the following.

Consider the bias term first, i.e. we consider

$$\mathrm{E}\big(\tilde{f}_X(x)\big) = \frac{1}{n(n-1)h}\mathrm{E}\Big[\sum_{i=1}^{n}\sum_{j=2}^{n} K\big(\frac{x - g(X_i) - e_j}{h}\big)\Big]. \tag{10}$$

Using that the time series $\{X_t\}$ is strict stationary and mixing we can use standard techniques to prove that the terms in the bias when i and j are far apart and n tends to infinity, will be neglible. Actually we can write (10) as,

$$\begin{aligned}
&\frac{1}{n(n-1)h}\mathrm{E}\Big[\sum_{i=1}^{n}\sum_{j=2}^{n} K\big(\frac{x - g(X_i) - e_j}{h}\big)\Big] \\
&= \frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=2}^{n}\iint K\big(\frac{x - g(v) - u}{h}\big) \\
&\times[f_{X,e}^{i,j}(v,u) - f_X(v)f_e(u) + f_X(v)f_e(u)]\mathrm{d}v\mathrm{d}u.
\end{aligned} \tag{11}$$

We must examine equation (11) in two steps. Using stationarity and the mixing-type condition $|f_{X,e}^{i,j}(v,u) - f_X(v)f_e(u)| \leq C|\alpha|^{|i-j|}$ where $|\alpha| < 1$, we obtain in the first step,

$$\begin{aligned}
&\Big|\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=2}^{n}\iint K\big(\frac{x - g(v) - u}{h}\big)[f_{X,e}^{i,j}(v,u) - f_X(v)f_e(u)]\mathrm{d}v\mathrm{d}u\Big| \\
&\leq C\frac{1}{n(n-1)h}\sum_{k=1}^{n}\iint K\big(\frac{x - g(v) - u}{h}\big)2(n-k)|\alpha|^{k-1}\mathrm{d}v\mathrm{d}u.
\end{aligned} \tag{12}$$

Moreover,

$$\begin{aligned}
&\frac{1}{n(n-1)h}\sum_{k=1}^{n}(n-k)|\alpha|^{k-1} = \frac{1}{(n-1)h}\sum_{k=1}^{n}(1 - \frac{k}{n})|\alpha|^{k-1} \\
&= \frac{1}{(n-1)h}\big[\sum_{k=1}^{n}|\alpha|^{k-1} + \frac{1}{n}\sum_{k=1}^{n}k|\alpha|^{k-1}\big] = O\Big(\frac{1}{(n-1)h}\Big),
\end{aligned} \tag{13}$$

since the two series in (13) both converge. (For $i < j$ the term corresponding to (12) is in fact zero due to independence.)

The second part in (11) gives,

$$\begin{aligned}
&\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=2}^{n}\iint K\big(\frac{x - g(v) - u}{h}\big)f_X(v)f_e(u)\mathrm{d}v\mathrm{d}u \\
&\sim \frac{1}{h}\iint K\big(\frac{x - g(v) - u}{h}\big)f_X(v)f_e(u)\mathrm{d}v\mathrm{d}u \\
&= \iint K(w)f_X(v)f_e(x - g(v) - hw)\mathrm{d}v\mathrm{d}w \\
&= \int K(w)f_X(x - hw)\mathrm{d}w \\
&= f_X(x) + \frac{h^2}{2}f_X''(x)\int w^2K(w)\mathrm{d}w + O(h^4).
\end{aligned} \tag{14}$$

The total bias can be written, putting $(n-1) = n$,

$$\mathrm{E}\big(\tilde{f}_X(x)\big) - f_X(x) = \frac{h^2}{2} f_X''(x) \int w^2 K(w) \mathrm{d}w + O(h^4) + O\big(\frac{1}{nh}\big), \tag{15}$$

the same as the bias in Støve & Tjøstheim (2005), except for the last term.

The variance term can be decomposed into several covariance terms,

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

$$= \frac{1}{n^2(n-1)^2h^2} \sum_{i=1}^{n}\sum_{j=2}^{n}\sum_{k=1}^{n}\sum_{l=2}^{n} \mathrm{cov}\Big(K\big(\frac{x - g(X_i) - e_j}{h}\big), K\big(\frac{x - g(X_k) - e_l}{h}\big)\Big). \tag{16}$$

We must consider all combinations of i, j, k and l. By using mixing properties as above, it will be seen that only independent terms will contribute to the leading term of this variance expression. Here only some of all combinations will be examined. The remaining terms can be treated similarly.

First, examine $i = j = k = l$. As we will see, this term will not contribute to the overall variance, but is included to demonstrate the techniques. The corresponding term in (16) becomes,

$$\frac{1}{n^2(n-1)^2h^2} \sum_{i=1}^{n-1} \mathrm{var}\Big(K\big((\frac{x - g(X_i) - e_i}{h}\big)\Big)$$

$$= \frac{n-1}{n^2(n-1)^2h^2}\Big[\mathrm{E}\big(K^2(\frac{x - g(X_1) - e_1}{h})\big) - \big[\mathrm{E}\big(K(\frac{x - g(X_1) - e_1}{h})\big)\big]^2\Big]. \tag{17}$$

Consider the first term. By change of variables and Taylor expansion,

$$\mathrm{E}\big(K^2(\frac{x - g(X_1) - e_1}{h})\big) = \iint K^2(\frac{x - g(v) - u}{h}) f_{X,e}(v,u) \mathrm{d}v\mathrm{d}u$$

$$= h \iint K^2(z) f_{X,e}(v, x - g(v) - hz) \mathrm{d}v\mathrm{d}z =$$

$$h\Big[\iint K^2(z) f_{X,e}\big(v, x - g(v)\big)\mathrm{d}v\mathrm{d}z - h \iint z K^2(z) \frac{\partial(f_{X,e}(v, x - g(v)))}{\partial(x - g(v))} \mathrm{d}v\mathrm{d}z$$

$$+ \frac{h^2}{2} \iint z^2 K^2(z) \frac{\partial^2(f_{X,e}(v, x - g(v)))}{\partial(x - g(v))^2} \mathrm{d}v\mathrm{d}z + o(h^2)\Big].$$

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

$$\big[\mathrm{E}\big(K(\frac{x - g(X_1) - e_1}{h})\big)\big]^2 = \Big[h\Big[\int f_{X,e}\big(v, x - g(v)\big)\mathrm{d}v$$

$$+ \frac{h^2}{2} \iint z^2 K(z) \frac{\partial^2(f_{X,e}(v, x - g(v)))}{\partial(x - g(v))^2} \mathrm{d}v\mathrm{d}z + o(h^2)\Big]\Big]^2$$

$$= h^2\big[\big(\int f_{X,e}(v, x - g(v))\mathrm{d}v\big)^2 + O(h^2)\big].$$

In total,

$$\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n-1}\mathrm{var}\Big(K\big((\frac{x-g(X_i)-e_i}{h}\big)\Big)$$

$$=\frac{1}{n^2(n-1)h^2}\big[h\iint K^2(z)f_{X,e}\big(v,x-g(v)\big)\mathrm{d}v\mathrm{d}z+O(h^2)\big]. \quad (18)$$

The next combination we study is $k=i\neq j=l$ and the covariance term in (16) will be

$$\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ j\neq i}}^{n}\mathrm{var}\Big(K\big((\frac{x-g(X_i)-e_j}{h}\big)\Big)$$

$$=\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ j\neq i}}^{n}\Big[\mathrm{E}\big(K^2(\frac{x-g(X_i)-e_j}{h})\big)-\big[\mathrm{E}\big(K(\frac{x-g(X_i)-e_j}{h})\big)\big]^2\Big]. \quad (19)$$

Here X_i and e_j are dependent when $i\geq j$ and independent when $i<j$. As for the bias analysis, we use mixing and in total we get

$$\frac{(n-1)(n-2)}{n^2(n-1)^2h^2}\big[hf_X(x)\int K^2(z)\mathrm{d}z-h^2f_X^2(x)+O(h^2)\big], \quad (20)$$

as for the indpendent case in Støve & Tjøstheim (2005).

In the following combination, $i=l\neq j\neq k$, there will be three summation indexes. We can write

$$\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j\neq k}}^{n}\sum_{k=2}^{n}\mathrm{cov}\Big(K\big((\frac{x-g(X_i)-e_j}{h}\big),K\big((\frac{x-g(X_k)-e_i}{h}\big)\Big)$$

$$=\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j\neq k}}^{n}\sum_{k=2}^{n}\Big[\mathrm{E}\big(K(\frac{x-g(X_i)-e_j}{h})K(\frac{x-g(X_k)-e_i}{h})\big)$$

$$-\mathrm{E}\big(K(\frac{x-g(X_i)-e_j}{h})\big)\mathrm{E}\big(K(\frac{x-g(X_k)-e_i}{h})\big)\Big]. \quad (21)$$

Define

$$Q(X_i,e_j)=K\big(\frac{x-g(X_i)-e_j}{h}\big). \quad (22)$$

By writing out the summation over k above we can write the covariance as, note still that $i\neq j$,

$$\frac{1}{n^2(n-1)^2h^2}\Big[\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\mathrm{cov}\big(Q(X_i,e_j),Q(X_2,e_i)\big)$$

$$+\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\mathrm{cov}\big(Q(X_i,e_j),Q(X_3,e_i)\big)+...+\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\mathrm{cov}\big(Q(X_i,e_j),Q(X_n,e_i)\big)\Big]. \quad (23)$$

We can study each of the above terms. Asymptotically these terms will have the same order, thus we study one term for a fixed k,

$$\frac{1}{n^2(n-1)^2h^2}\Big[\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\mathrm{cov}\big(Q(X_i,e_j),Q(X_k,e_i)\big)\Big]$$

$$=\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\Big[\mathrm{E}\big(Q(X_i,e_j)Q(X_k,e_i)\big)-\mathrm{E}\big(Q(X_i,e_j)\big)\mathrm{E}\big(Q(X_k,e_i)\big)\Big]. \quad (24)$$

Again, we can use mixing techniques. Note that dependence will always occur between X_i and e_i, thus we must use a joint distribution here. Since k is fixed, the dependence between X_k and the other variables will vanish due to mixing when i and j becomes large. Thus the first part of (24) can be written, (suppressing the k-dependence)

$$\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\mathrm{E}\big(Q(X_i,e_j)Q(X_k,e_i)\big)$$

$$=\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\iiiint K\big(\frac{x-g(v_1)-u_1}{h}\big)K\big(\frac{x-g(w)-u_2}{h}\big)$$

$$\times\Big(f^{i,j}_{X,e,X,e}(v_1,u_1,w,u_2)-f_{X,e}(v_1,u_2)f_X(w)f_e(u_1)$$

$$+f_{X,e}(v_1,u_2)f_X(w)f_e(u_1)\Big)\mathrm{d}v\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}w. \quad (25)$$

We next introduce a mixing-type condition on $f^{i,j}_{X,e,X,e}(v_1,u_1,w,u_2)-f_{X,e}(v_1,u_2)f_X(w)f_e(u_1)$,

$$\Big|\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\iiiint K\big(\frac{x-g(v_1)-u_1}{h}\big)K\big(\frac{x-g(w)-u_2}{h}\big)$$

$$\times\Big(f^{i,j}_{X,e,X,e}(v_1,u_1,w,u_2)-f_{X,e}(v_1,u_2)f_X(w)f_e(u_1)\Big)\mathrm{d}v\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}w\Big|$$

$$\sim\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j}}^{n}\iiiint K\big(\frac{x-g(v_1)-u_1}{h}\big)K\big(\frac{x-g(w)-u_2}{h}\big)$$

$$\times|\alpha_2|^{|i-j|}\mathrm{d}v\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}w\sim\frac{1}{n(n-1)^2}\sum_{l=1}^{n}2\big(1-\frac{l}{n}\big)|\alpha_2|^l=O\Big(\frac{1}{n(n-1)^2}\Big), \quad (26)$$

where $|\alpha_2| < 1$ implies that the last sum converges. Consider the remaining part of (25),

$$\begin{aligned}
&\frac{(n-1)(n-2)}{n^2(n-1)^2h^2}\iiiint K\big(\frac{x-g(v_1)-u_1}{h}\big)K\big(\frac{x-g(w)-u_2}{h}\big)\\
&\qquad\qquad\times f_e(u_1)f_X(w)f_{X,e}(v_1,u_2)\mathrm{d}v_1\mathrm{d}w\mathrm{d}u_1\mathrm{d}u_2\\
&=\frac{(n-1)(n-2)h^2}{n^2(n-1)^2h^2}\iiiint K(z_1)K(z_2)f_e(x-g(v_1)-z_1h)\\
&\qquad\qquad\times f_X(w)f_{X,e}(v_1,x-g(w)-z_2h)\mathrm{d}v_1\mathrm{d}w\mathrm{d}z_1\mathrm{d}z_2\\
&=\frac{1}{n^2}\Big[\iint f_X(w)f_e\big(x-g(v_1)\big)f_{X,e}\big(v_1,x-g(w)\big)\mathrm{d}v_1\mathrm{d}w+O(h^2)\Big]. \qquad (27)
\end{aligned}$$

As before, the term (27) is of lower order than the term (26). Thus the term (26) will not contribute to the covariance.

The last part of (24) can also be studied using mixing techniques, and again, just the indpendent part will contribute. Using result from Støve & Tjøstheim (2005), we can write

$$\mathrm{E}\big(Q(X_1,e_2)\big)\mathrm{E}\big(Q(X_3,e_1)\big)=h^2f_X^2(x)+O(h^4). \qquad (28)$$

Using the fact that $k=2,...,n$, and since the order of the terms (27) and (28) will be the same for any k, the total covariance from (21) will be,

$$\begin{aligned}
&\frac{1}{n-1}\Big[\iint f_X(w)f_e\big(x-g(v_1)\big)f_{X,e}\big(v_1,x-g(w)\big)\mathrm{d}v_1\mathrm{d}w\\
&\qquad\qquad -f_X^2(x)+O(h^2)\Big]. \qquad (29)
\end{aligned}$$

The term, $j=k\neq i\neq l$, can be analysed as above. Thus, it can be shown that

$$\begin{aligned}
&\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j\neq l}}^{n}\sum_{l=2}^{n}\mathrm{cov}\Big(K\big((\frac{x-g(X_i)-e_j}{h}\big),K\big((\frac{x-g(X_j)-e_l}{h}\big)\Big)\\
&=\frac{1}{(n-1)h^2}\Big[h^2\iint f_X(v_2)f_e\big(x-g(v_1)\big)f_{X,e}\big(v_1,x-g(v_2)\big)\mathrm{d}v_1\mathrm{d}v_2\\
&\qquad\qquad -h^2f_X^2\ (x)+O(h^4)\Big]. \qquad (30)
\end{aligned}$$

Similarly, for the term, $i\neq k\neq j=l$,

$$\begin{aligned}
&\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j\neq k}}^{n}\sum_{k=2}^{n}\mathrm{cov}\Big(K\big((\frac{x-g(X_i)-e_j}{h}\big),K\big((\frac{x-g(X_k)-e_j}{h}\big)\Big)\\
&=\frac{1}{(n-1)h^2}\Big[h^2\int\frac{f_X^2(v)f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v-h^2f_X^2(x)+O(h^4)\Big], \qquad (31)
\end{aligned}$$

and for the term $i=k\neq j\neq l$,

$$\frac{1}{n^2(n-1)^2h^2}\sum_{i=1}^{n}\sum_{\substack{j=2\\ i\neq j\neq l}}^{n}\sum_{l=2}^{n}\operatorname{cov}\Big(K\big(\frac{x-g(X_i)-e_j}{h}\big),K\big(\frac{x-g(X_i)-e_l}{h}\big)\Big)$$
$$=\frac{1}{(n-1)h^2}\big[h^2\int f_X(v)f_e^2(x-g(v))\mathrm{d}v-h^2f_X^2(x)+O(h^4)\big]. \tag{32}$$

The terms from the remaining combinations of i, j, k and l, in the covariance (16) will not contribute to the overall variance of $\tilde{f}_X(x)$, since they are of higher order. Thus the variance from (16) can be written, adding all combinations that contribute,

$$\begin{aligned}\operatorname{var}\big(\tilde{f}_X(x)\big)&=\frac{1}{n-1}\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v\\ &+\frac{2}{n-1}\iint f_X(v_2)f_e(x-g(v_1))f_{X,e}(v_1,x-g(v_2))\mathrm{d}v_1\mathrm{d}v_2\\ &+\frac{1}{n-1}\int\frac{f_X^2(v)f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v-\frac{4}{n-1}f_X^2(x)\\ &+\frac{1}{n(n-1)h}f_X(x)\int K^2(z)\mathrm{d}z+O\big((n-1)^{-1}h^2\big),\end{aligned} \tag{33}$$

which corresponds to the variance found in Støve & Tjøstheim (2005), except for the term with a joint distribution for e and X. However, the variance will reduce correctly to the independent case since X and e are independent and by using a convolution argument;

$$\frac{2}{n-1}\iint f_X(v_2)f_e(x-g(v_1))f_X(v_1)f_e(x-g(v_2))\mathrm{d}v_1\mathrm{d}v_2=\frac{2}{n-1}f_X^2. \tag{34}$$

Thus the variance in the independent case is,

$$\begin{aligned}&\frac{1}{n-1}\int f_X(v)f_e^2\big(x-g(v)\big)\mathrm{d}v\\ &+\frac{1}{n-1}\int\frac{f_X^2(v)f_e\big(x-g(v)\big)}{g'(v)}\mathrm{d}v-\frac{2}{n-1}f_X^2(x)\\ &+\frac{1}{n(n-1)h}f_X(x)\int K^2(z)\mathrm{d}z+O\big((n-1)^{-1}h^2\big),\end{aligned} \tag{35}$$

Observe that from expression (33) the variance of $\tilde{f}_X(x)$ has leading terms of order n^{-1}, this in contrast to the kernel density estimator which has leading terms $(nh)^{-1}$, see e.g. Wand & Jones (1995) page 21.

As in Støve & Tjøstheim (2005), we next study the properties of the other term in equation (8), that is, $\hat{f}_X(x)-\tilde{f}_X(x)$. Consider the estimator $\hat{f}_X(x)$ in (7). By substituting for $\tilde{e}_j$, Taylor expanding of $K(\cdot)$ around $(x-g(X_i)-e_j)/h$ and using the mean value theorem, we obtain,

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h_D}\sum_{j=2}^{n}K\big(\frac{x-\tilde{g}(X_i)-\tilde{e_j}}{h_D}\big)\Big]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}\sum_{j=2}^{n}K\big(\frac{x-\tilde{g}(X_i)-X_j+\tilde{g}(X_{j-1})}{h}\big)\Big] = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}$$

$$\times\sum_{j=2}^{n}K\big(\frac{x-g(X_i)-e_j+g(X_i)-\tilde{g}(X_i)-g(X_{j-1})+\tilde{g}(X_{j-1})}{h}\big)\Big]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}\sum_{j=2}^{n}\big[K\big(\frac{x-g(X_i)-e_j}{h}\big)+K'\big(\frac{x-g(X_i)-e_j}{h}\big)$$

$$\times\big(\frac{g(X_i)-g(X_{j-1})+\tilde{g}(X_{j-1})-\tilde{g}(X_i)}{h}\big)$$

$$+O_P\big(\big(\frac{g(X_i)-g(X_{j-1})+\tilde{g}(X_{j-1})-\tilde{g}(X_i)}{h}\big)^2\big)\big]\Big]. \qquad (36)$$

Thus,

$$\hat{f}(x)-\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}\sum_{j=2}^{n}\big[K'\big(\frac{x-g(X_i)-e_j}{h}\big)$$

$$\times\frac{\tilde{g}(X_{j-1})-g(X_{j-1})}{h}+K'\big(\frac{x-g(X_i)-e_j}{h}\big)\cdot\frac{g(X_i)-\tilde{g}(X_i)}{h}\big]\Big]$$

$$+\sum_{i=1}^{n}\Big[\frac{1}{(n-1)h}\sum_{j=2}^{n}O_P\big(\big(\frac{g(X_i)-g(X_{j-1})+\tilde{g}(X_{j-1})-\tilde{g}(X_i)}{h}\big)^2\big)\Big]. \qquad (37)$$

As in Støve & Tjøstheim (2005) under appropriate regularity conditions, we can show that the second order term in the Taylor expansion above is of higher order under the conditions stated there. Thus we focus on the expectation and variance of the first order term. The following condition guarantees the existence of these moments;

6. $\inf_{x\in S(X)}f_X(x) > 0$, where $S(X)$ is the compact support of the density f_X.

The expectation can be written

$$\mathrm{E}\big(\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=2}^{n}K'\big(\frac{x-g(X_i)-e_j}{h}\big)$$

$$\times\big(\frac{\tilde{g}(X_{j-1})-g(X_{j-1})}{h}-\frac{\tilde{g}(X_i)-g(X_i)}{h}\big)\big). \qquad (38)$$

Note that this disappears for $i = j-1$. Using a conditioning and a convolution argument, as in Støve & Tjøstheim (2005) and mixing arguments as for the expectation of $\tilde{f}_X$, the

first part of (38) is

$$\mathrm{E}\big(\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=2}^{n}K'(\frac{x-g(X_i)-e_j}{h})\cdot\frac{\tilde{g}(X_{j-1})-g(X_{j-1})}{h}\big)\sim$$

$$-f'_X(x)\int z_1K'(z_1)\mathrm{d}z_1\int\big(\mathrm{E}\big(\tilde{g}(x_2)\big)-g(x_2)\big)f_X(x_2)\mathrm{d}x_2 \tag{39}$$

$$+\frac{h}{2}f''_X(x)\int z_1^2K'(z_1)\mathrm{d}z_1\int\big(\mathrm{E}\big(\tilde{g}(x_2)\big)-g(x_2)\big)f_X(x_2)\mathrm{d}x_2+O(h^2). \tag{40}$$

Examining the second part by similar arguments,

$$-\mathrm{E}\big(\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=2}^{n}K'(\frac{x-g(X_i)-e_j}{h})\cdot\frac{\tilde{g}(X_i)-g(X_i)}{h}\big)\sim$$

$$\int z_2K'(z_2)\mathrm{d}z_2\int\big(\mathrm{E}(\tilde{g}(x_1))-g(x_1)\big)f'_e\big(x-g(x_1)\big)f_X(x_1)\mathrm{d}x_1 \tag{41}$$

$$-\frac{h}{2}\int z_2^2K'(z_2)\mathrm{d}z_2\int\big(\mathrm{E}(\tilde{g}(x_1))-g(x_1)\big)f''_e\big(x-g(x_1)\big)f_X(x_1)\mathrm{d}x_1+O(h^2). \tag{42}$$

Note that since the kernel is symmetric, $\int z^2K'(z)\mathrm{d}z=0$. In total, the bias of $\hat{f}_X(x)$ will be the sum of (15), (39) and (41) and is given by

$$\frac{h^2}{2}f''_X(x)\int w^2K(w)\mathrm{d}w \tag{43}$$

$$-f'_X(x)\int z_1K'(z_1)\mathrm{d}z_1\int\big(\mathrm{E}\big(\tilde{g}(x_2)\big)-g(x_2)\big)f_X(x_2)\mathrm{d}x_2 \tag{44}$$

$$+\int z_2K'(z_2)\mathrm{d}z_2\int\big(\mathrm{E}(\tilde{g}(x_1))-g(x_1)\big)f'_e\big(x-g(x_1)\big)f_X(x_1)\mathrm{d}x_1. \tag{45}$$

The whole term will be of order $O(h^2)$ through the dependence on $\mathrm{E}\big(\tilde{g}(x_i)\big)-g(x_i)$, $i=1,2$, which is of the same order as the Nadaraya-Watson estimator. The bias expression (43)-(45) is identical to the bias obtained for the convolution estimator in Støve & Tjøstheim (2005), expect that there the density of interest is f_Y, as explained in the introduction.

Further, examining the variance of the first order term in the Taylor expansion, these calculations are similar to the calculations where we found the variance of $\tilde{f}(x)$. The variance in question is,

$$\mathrm{var}\Big(\frac{1}{n(n-1)h^2}\sum_{i=1}^{n}\sum_{j=2}^{n}\big[K'(\frac{x-g(X_i)-e_j}{h})$$

$$\times\big(g(X_i)-\hat{g}(X_i)+\hat{g}(X_{j-1})-g(X_{j-1})\big)\big]\Big)=\frac{1}{n^2(n-1)^2h^4}\sum_{i=1}^{n}\sum_{j=2}^{n}\sum_{k=1}^{n}\sum_{l=2}^{n}$$

$$\times\mathrm{cov}\Big[K'(\frac{x-g(X_i)-e_j}{h})\big(g(X_i)-\hat{g}(X_i)+\hat{g}(X_{j-1})-g(X_{j-1})\big),$$

$$K'(\frac{x-g(X_k)-e_l}{h})\big(g(X_k)-\hat{g}(X_k)+\hat{g}(X_{l-1})-g(X_{l-1})\big)\Big]. \tag{46}$$

As before, this covariance can be decomposed into several variance and covariance terms. Based on the results in Støve & Tjøstheim (2005) and using mixing assumptions, the terms arising from this covariance will be $O(h^4/n)$. Thus, it will only contribute higher order effects to the variance of $\hat{f}_X(x)$. In total the variance will be $O(n^{-1})$, and given by expression (33).

4 Simulation study

To evaluate the proposed estimator, (7), in finite samples, we compare the estimates obtained with the estimates from the classical kernel estimator in (1).

The comparisons are based on the mean integrated squared error (MISE) of the two estimators. That is,

$$\mathrm{MISE}(\hat{f}) = \mathrm{E}\big[\int_{-\infty}^{\infty} (\hat{f} - f)^2(x)\mathrm{d}x\big], \tag{47}$$

and likewise for the kernel density estimator. We have generated 500 simulated realizations with sample size 500 of the process $\{X_t\}$. The estimated MISE, $\widehat{\mathrm{MISE}}$, is approximated as an average of the integrated squared error (ISE) of the 500 realizations, and the ISE is estimated by numerical integration. If the true marginal density is not known, we have based our comparisons on an estimated density computed from 1 000 000 generated observations. The results are given as the percentage change by using the convolution estimator compared with the kernel density estimator. For the MISE, this change is calculated by

$$\frac{\widehat{\mathrm{MISE}}(f_X^*) - \widehat{\mathrm{MISE}}(\hat{f}_X)}{\widehat{\mathrm{MISE}}(f_X^*)} \cdot 100. \tag{48}$$

Note that instead of using the MISE error measure we could have choosen a measure which is dependent on where the observations are located, for example by introducing a nonnegative weight function. This could have been used for weighting down observations at the boundaries.

The choice of bandwidth in nonparametric estimation has a considerable impact of the accuracy of the estimator. The bandwidth used in the kernel density estimation in our simulation study, is the Solve-the-Equation Plug-in bandwidth proposed in Sheather & Jones (1991). This bandwidth has also been used in the estimation of the density, f_e, if nothing else is stated. The bandwidth for the kernel smoothing of g is the rule-of-thumb, see e.g. Härdle (1990), $1.06\min(\hat{\sigma}, R/1.34)n^{-1/5}$, where R is the interquartile range, $\hat{\sigma}$ is the empirical variance of all observations $X_1, ..., X_n$. This bandwidth selection is usually used as a rule-of-thumb in density estimation, but we have chosen it here for computational purposes. Thus we may actually obtain better results using a more refined bandwidth for the nonparametric time series regression.

We first examine the autoregressive model,

$$X_t = \phi X_{t-1} + e_t, \tag{49}$$

where $e_t \sim N(0, \sigma^2)$. Here the true marginal density of X is $N(0, \frac{\sigma^2}{1-\phi^2})$. Table 1 shows the estimated MISE for the ordinary kernel density estimator and the new estimator for

various choices of ϕ and σ. The convolution estimator has the smallest MISE in all cases, thus it seems that we improve the density estimate. Observe that the results are best for negative autocorrelations. This was also observed in Saavedra & Cao (1999*a*), for the estimation of the marginal density of a MA(1) process, and Hart (1984) noted that a negative correlation in an autoregressive process makes a "balanced" sample, that is, the observations are more likely to be symmetric about the mean than an independent sample.

Parameters	Convolution	Kernel	Improvement in %
$\phi = 0.8$, $\sigma = 2$	$1.88 \cdot 10^{-3}$	$2.76 \cdot 10^{-3}$	31.9
$\phi = 0.8$, $\sigma = 1$	$3.48 \cdot 10^{-3}$	$5.14 \cdot 10^{-3}$	32.3
$\phi = 0.5$, $\sigma = 2$	$8.36 \cdot 10^{-4}$	$1.59 \cdot 10^{-3}$	47.4
$\phi = 0.5$, $\sigma = 1$	$1.79 \cdot 10^{-3}$	$3.28 \cdot 10^{-3}$	45.4
$\phi = 0.2$, $\sigma = 2$	$9.82 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$	33.2
$\phi = 0.2$, $\sigma = 1$	$2.04 \cdot 10^{-3}$	$3.01 \cdot 10^{-3}$	32.2
$\phi = -0.5$, $\sigma = 2$	$4.25 \cdot 10^{-4}$	$1.18 \cdot 10^{-3}$	64.0
$\phi = -0.5$, $\sigma = 1$	$7.98 \cdot 10^{-4}$	$2.75 \cdot 10^{-3}$	71.0

Table 1: The estimated MISE for an AR(1) process

For a further examination of the estimator, several other densities for the error term e_t has been used, see table 2. Five of the densities proposed by Marron & Wand (1992) have been selected. These densities typify many different challenges to curve estimation. The first three densities represent challenges that can arise for unimodal densities. The separated bimodal density is mildly multimodal, while the discrete comb is strongly multimodal.

Density	
(1) Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2}, \frac{2}{3}) + \frac{3}{5}N(\frac{12}{15}, \frac{5}{9})$
(2) Strongly skewed	$\sum_{l=0}^{7} \frac{1}{8}N(3[(\frac{2}{3})^l - 1], (\frac{2}{3})^l)$
(3) Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0, \frac{1}{10})$
(4) Separated bimodal	$\frac{1}{2}N(-\frac{3}{2}, \frac{1}{2}) + \frac{1}{2}N(\frac{3}{2}, \frac{1}{2})$
(5) Discrete comb	$\sum_{l=0}^{2} \frac{2}{7}N(\frac{12l-15}{7}, \frac{2}{7}) + \sum_{l=8}^{10} \frac{1}{21}N(\frac{2l}{7}, \frac{1}{21})$

Table 2: Normal mixture densities used in the simulation study

Since the true marginal density is not known analytically in these cases, we estimate it both by kernel density estimation, with bandwidth choosen to be the rule-of-thumb, and the proposed convolution estimator. The rule-of-thumb bandwidth has here been used for estimating the error distributions as well. The convolution estimator is more computer intensive, thus just 100 000 generated observations are used to estimate the true distribution here. Only the results using the kernel density estimator is given, see table 3. However, we get the same conclusions in both cases.

Density and parameter	Convolution	Kernel	Improvement in %
(1), $\phi = 0.8$	$3.12 \cdot 10^{-3}$	$3.50 \cdot 10^{-3}$	10.9
(1), $\phi = 0.2$	$2.12 \cdot 10^{-3}$	$2.80 \cdot 10^{-3}$	32.1
(2), $\phi = 0.8$	$4.19 \cdot 10^{-3}$	$3.22 \cdot 10^{-3}$	-30.1
(2), $\phi = 0.2$	$1.03 \cdot 10^{-2}$	$6.19 \cdot 10^{-3}$	-66.4
(3), $\phi = 0.8$	$5.32 \cdot 10^{-3}$	$4.57 \cdot 10^{-3}$	-16.4
(3), $\phi = 0.2$	$1.46 \cdot 10^{-2}$	$8.06 \cdot 10^{-3}$	-81.1
(4), $\phi = 0.8$	$2.02 \cdot 10^{-3}$	$1.61 \cdot 10^{-3}$	-25.5
(4), $\phi = 0.2$	$4.21 \cdot 10^{-3}$	$2.58 \cdot 10^{-3}$	-63.2
(5), $\phi = 0.8$	$1.52 \cdot 10^{-3}$	$1.66 \cdot 10^{-3}$	8.4
(5), $\phi = 0.2$	$5.23 \cdot 10^{-3}$	$2.93 \cdot 10^{-3}$	-78.5

Table 3: The estimated MISE for an AR(1) process with different error term densities, with true density estimated by the kernel method

The table 3 shows that for the skewed unimodal density, the convolution-type estimator is better. But for the strongly skewed and kurtotic unimodal the kernel density seems to be more accurate for a sample size of 500. For the separated bimodal density, the kernel estimator is better. Since we have a bimodal error density, the scatter diagram will consist of two clusters of points, thus the kernel regression function g is poorly estimated and creates large error terms, and a bad density estimate. Note that such time series are not often met in practice. We are introducing it here to demonstrate the limitations of our method. Observe further the perhaps surprising outcome under discrete comb with $\phi = 0.8$, that the convolution estimator is better. Observe also that for low values of ϕ the convolution estimator tends to be less accurate then for high values of ϕ. This seems to be because the error terms, e_t, increase and give a poorer convolution effect.

We next study the exponential AR model, which is

$$X_t = (a + b\, e^{-\gamma X_{t-1}^2}) X_{t-1} + e_t, \tag{50}$$

where a, b and γ are constants, with $|a| < 1$, $\gamma > 0$ and $e_t \sim N(0, \sigma^2)$. Table 4 upper part shows the estimated MISE for the ordinary kernel density estimator and the convolution estimator for several choices of the parameters. Since the true density f_X is not known, it has been estimated by the kernel density estimator, using 1 000 000 samples from the specific model, and the rule-of-thumb as the bandwidth. As above, the convolution estimator has also been used to estimate the true distribution, using 100 000 generated observations. Both scenarios gave equal conclusions, see table 4 upper and lower part.

The convolution estimator has the lowest estimated MISE, except when the exponential AR process has a large b parameter, that is, it is almost an exploding process. Then it is difficult to estimate g close to the origin.

The next simulation is a logistic AR process, given as

$$X_t = \theta_1 X_{t-1} + \theta_2 X_{t-1} \left[\left(1 + e^{-\theta_3 (X_{t-1} - \theta_4)}\right)^{-1} - 0.5 \right] + e_t, \quad \theta_3 > 0. \tag{51}$$

Parameters	Convolution	Kernel	Improvement in %
$a = 0.8$, $b = 0.2$, $\gamma = 5$, $\sigma = 1$	$2.33 \cdot 10^{-3}$	$2.68 \cdot 10^{-3}$	13.1
$a = 0.4$, $b = 0.6$, $\gamma = 0.5$, $\sigma = 2$	$5.52 \cdot 10^{-4}$	$1.02 \cdot 10^{-3}$	45.9
$a = 0.5$, $b = 4$, $\gamma = 1$, $\sigma = 1$	$3.21 \cdot 10^{-2}$	$2.95 \cdot 10^{-2}$	-8.8
$a = 0.8$, $b = 0.2$, $\gamma = 5$, $\sigma = 1$	$2.56 \cdot 10^{-3}$	$3.15 \cdot 10^{-3}$	18.7
$a = 0.4$, $b = 0.6$, $\gamma = 0.5$, $\sigma = 2$	$5.57 \cdot 10^{-4}$	$1.03 \cdot 10^{-3}$	45.9
$a = 0.5$, $b = 4$, $\gamma = 1$, $\sigma = 1$	$3.09 \cdot 10^{-2}$	$2.85 \cdot 10^{-2}$	-8.4

Table 4: The estimated MISE for an exponential AR model, (upper part: true density estimated by the kernel method, lower part: true density estimated by the convolution estimator)

The simulations are performed as above, 500 simulated realizations with a sample size of 500. The error term in all simulations is normally distributed with mean equal to zero and variance equal to one. The results are given in table 5, and in all cases the convolution estimator is better than the kernel density estimator.

Parameters	Convolution	Kernel	Improvement in %
$\theta_1 = 0.5$, $\theta_2 = 0.5$, $\theta_3 = 1$, $\theta_4 = 1$	$1.28 \cdot 10^{-3}$	$2.13 \cdot 10^{-3}$	39.8
$\theta_1 = 0.1$, $\theta_2 = 0.5$, $\theta_3 = 1$, $\theta_4 = 1$	$1.39 \cdot 10^{-3}$	$1.87 \cdot 10^{-3}$	25.4
$\theta_1 = -0.2$, $\theta_2 = -0.8$, $\theta_3 = 4$, $\theta_4 = 1$	$1.25 \cdot 10^{-3}$	$1.79 \cdot 10^{-3}$	30.3

Table 5: The estimated MISE for a logistic AR model, with true density estimated by the kernel method

To show how the proposed estimator's bias and variance behave, the last simulation is from an AR(1) model with parameter $\phi = 0.8$, and normally distributed error terms with mean zero and variance one. Here 100 realizations with sample size equal to 500 is simulated. Based on this simulation, the MISE, and also the squared bias and variance of both the kernel density estimator and the convolution estimator is estimated. The squared bias and variance are calculated as in Støve & Tjøstheim (2005). The results are given in table 6, and the convolution estimator is clearly better. In figure 1 the bias and variance of both estimators is plotted. As expected, the variance of the convolution estimator is smallest, the dotted line in the upper plot. The bias of the kernel density estimator behaves as we could expect, the solid line in the lower plot, since from the well-known bias formula, see e.g. Härdle (1990) page 56, the bias is proportional to the second derivative of the density in question. The bias of the convolution estimator is harder to interpret, dotted line in lower plot, but if one looks at the bias formula, (43)-(45), more closely, it is possible to explain its behaviour as in Støve & Tjøstheim (2005), section 4.

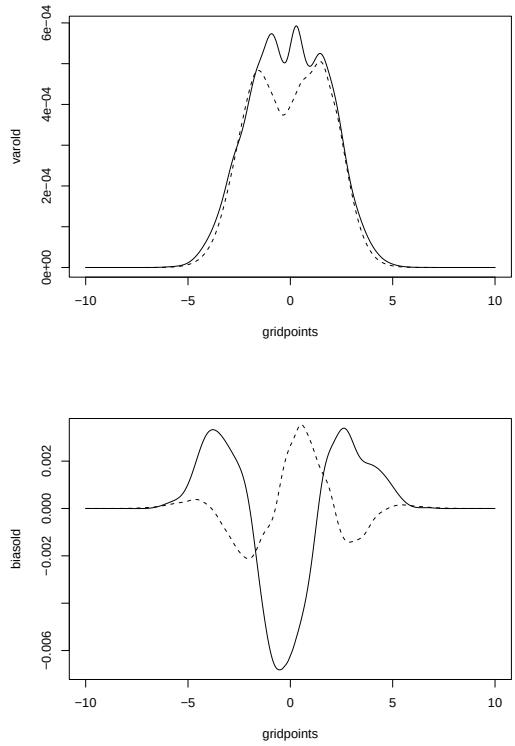

Figure 1: The estimated variance (upper plot) and bias (lower plot) of both the kernel density estimator (solid line) and the convolution estimator (dotted line)

Estimator	Squared bias	Variance	MISE
Convolution	$1.20 \cdot 10^{-6}$	$1.33 \cdot 10^{-4}$	$2.67 \cdot 10^{-3}$
Kernel	$5.86 \cdot 10^{-6}$	$1.53 \cdot 10^{-4}$	$3.15 \cdot 10^{-3}$

Table 6: The estimated squared bias, variance and MISE for an AR(1) model

5 Conclusions

The proposed convolution density estimator seems to outperform the usual kernel estimator in many situations, especially if the error term density function is smooth, and the g function is relatively easy to estimate. The degree of autocorrelation may also account for the behaviour of the estimator.

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 observed in Støve & Tjøstheim (2005). Also, by selecting the bandwidth parameters in the convolution estimator more correctly, by e.g. a cross-validation technique, this could improve the estimate.

The preliminary analysis of the estimator's asymptotic performance with an asymptotic variance of order n^{-1} is supported by the simulation experiments. However, a more detailed mathematical analysis is needed to put this on a firm theoretical basis.

References

Fan, J. & Yao, Q. (2003), *Nonlinear Time Series*, Springer-Verlag.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1990), *Nonparametric Curve Estimation of Time Series*, Springer-Verlag.

Hart, J. D. (1984), 'Efficiency of a kernel density estimator under an autoregressive dependence model', *Journal of the American Statistical Association* **79**, 110–117.

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

Marron, J. S. & Wand, M. P. (1992), 'Exact mean integrated squared error', *Annals of Statistics* **20**, 712–736.

Müller, U. U., Schick, A. & Wefelmeyer, W. (2005), 'Weighted residual-based density estimators for nonlinear autoregressive models', *Statistica Sinica* **15**, 177–195.

Rosenblatt, M. (1970), 'Density estimates and Markov sequences', *Nonparametric Techniques in Statistical Inference* pp. 199–213.

Roussas, G. (1969), 'Nonparametric estimation in Markov processes', *Annals of the Institute of Statistical Mathematics* **21**, 73–87.

Saavedra, A. & Cao, R. (1999*a*), 'A comparative study of two convolution-type estimators of the marginal 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.

Schick, A. & Wefelmeyer, W. (2004), 'Root n consistent and optimal density estimators for moving average processes', *Scandinavian Journal 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. (2005), 'A convolution estimator for the density of nonlinear regression observations', Submitted.

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

PAPER C

Revised for IMS Lecture Notes Series. *A Festschrift for Kjell Doksum*

IMS Lecture Notes–Monograph Series

A New Convolution Estimator for Nonparametric Regression

Bård Støve and Dag Tjøstheim

University of Bergen

Abstract: We present a convolution smoother for nonparametric regression. Its asymptotic behaviour is examined, and its asymptotic bias is found to be smaller than standard kernel estimators, such as Nadaraya-Watson and local linear regression. Some simulation studies have been performed. Asymptotic normality for the proposed estimator is proved.

1. Introduction.

There are many nonparametric estimators of the conditional mean $\mathrm{E}(Y|X = x)$ for independent identically distributed observations (X_i, Y_i), $i = 1, ..., n$, with a joint density $f(\cdot, \cdot)$, and a marginal density $f(\cdot)$ of X_i. The three most common are the local polynomial estimator; see [26], [2], [19] and [4], the Nadaraya-Watson estimator; see [20] and [28] and the Gasser-Müller estimator; see [9]. In the case

$$Y_i = m(X_i) + \epsilon_i, \tag{1.1}$$

where $\{X_i\}$ and $\{\epsilon_i\}$ consist of i.i.d. random variables with $\{\epsilon_i\}$ independent of $\{X_i\}$, $\mathrm{E}(Y|X = x) = m(x)$. Neither of the three above mentioned estimators require a regression relationship like (1.1) to work, and one might think that if one was able to construct an estimator of $m(x)$ making explicit use of the extra information contained in (1.1), then possibly one could improve on the standard nonparametric regression estimators. This is the basic idea of this paper, and it leads to what we have called "the convolution estimator". We will show that this new estimator generally has a smaller asymptotic bias than the standard estimators, and also that in a number of finite sample experiments it gives better results, although in many cases these improvements are not dramatic.

Before we define the new estimator, let us briefly mention that several authors have proposed adjustments and improvements of the kernel estimators. Both the Gasser-Müller and Nadaraya-Watson estimator have a large order bias when estimating a curve in its boundary region. Thus the idea of boundary kernels, which are weight functions that are used only within the boundary region, were introduced and studied by [9] and [10]. Another approach has been the reflection method; see [23] and [12]. In the papers [16], [3] and [11] the possibility of parametrically guided nonparametric density and regression estimation are examined. Several authors have studied the use of higher order kernels to improve the asymptotic bias; see e.g. [18] for a quantification of the practical gain, in density estimation.

A brief summary of the paper is as follows: In section 2 the estimator is introduced, and its asymptotic behaviour is examined and discussed in sections 3, 4 and 5. In section 6 some simulation results are given. Section 7 introduces a variant of the new estimator, and finally, section 8 gives some concluding remarks.

AMS 2000 subject classifications: Primary 62G08; secondary 62G20

Keywords and phrases: convolution, kernel function, mean squared error, nonparametric estimation

2. The estimator.

The regression function of interest is

$$m(x) = \mathrm{E}(Y|X = x) = \frac{\int yf(y,x)\mathrm{d}y}{f(x)} = \int yf(y|x)\mathrm{d}y. \tag{2.1}$$

Under the assumption that the equation (1.1) holds, $f(y|x)$ can be written

$$f(y|x) = f_\epsilon\big(y - m(x)\big), \tag{2.2}$$

where f_ϵ is the density of ϵ. Inserting this into (2.1) gives the convolution integral equation

$$m(x) = \int yf_\epsilon\big(y - m(x)\big)\mathrm{d}y, \tag{2.3}$$

where both $m(x)$ and f_ϵ are unknown. However, $m(x)$ may be replaced by a standard estimator $\tilde{m}(x)$, e.g. the local linear estimator, and f_ϵ can be estimated using a kernel estimate of $f_{\hat{\epsilon}}$ with $\hat{\epsilon}_i = Y_i - \tilde{m}(X_i)$. Based on the relation (2.3), the proposed estimator is

$$\begin{aligned}\hat{m}(x) &= \int y\hat{f}_{\hat{\epsilon}}\big(y - \tilde{m}(x)\big)\mathrm{d}y \\ &= n^{-1}\sum_{i=1}^{n}\int yK_{h_D}\big(y - \tilde{m}(x) - Y_i + \tilde{m}(X_i)\big)\mathrm{d}y,\end{aligned} \tag{2.4}$$

where $\tilde{m}(x)$ is, as mentioned, a nonparametric regression estimator and $K_{h_D}(\cdot) = K(\cdot/h_D)/h_D$, where K is a kernel function. We have choosen the local linear estimator, that is, the local polynomial estimator of degree 1, with bandwidth h_R and kernel function $K^L(\cdot)$,

$$\tilde{m}(x) = n^{-1}\sum_{i=1}^{n}\frac{\{\hat{s}_2(x) - \hat{s}_1(x)(x - X_i)\}K^L_{h_R}(x - X_i)Y_i}{\hat{s}_2(x)\hat{s}_0(x) - \hat{s}_1(x)^2}, \tag{2.5}$$

where

$$\hat{s}_r(x) = n^{-1}\sum_{i=1}^{n}(x - X_i)^r K^L_{h_R}(x - X_i), \quad r = 0, 1, 2, \tag{2.6}$$

and $K^L_{h_R}(\cdot) = K^L(\cdot/h_R)/h_R$, see e.g. [7]. The expression $\hat{f}_{\hat{\epsilon}}\big(y - \tilde{m}(x)\big)$ in (2.4) is the kernel density estimator with bandwidth equal to h_D,

$$\hat{f}_{\hat{\epsilon}}\big(y - \tilde{m}(x)\big) = \frac{1}{nh_D}\sum_{i=1}^{n}K\Big(\frac{y - \tilde{m}(x) - \hat{\epsilon}_i}{h_D}\Big) \tag{2.7}$$

with $\hat{\epsilon}_i = Y_i - \tilde{m}(X_i)$, see e.g. [27] page 11. Observe that the new estimator is computationally more demanding than standard methods.

It is also possible to iterate the estimator (2.4) using a previous estimate of $m(x)$ as input for the next iteration. Set $\tilde{m}_0$ equal to the local linear estimator for the regression curve. Then the convolution estimator is,

$$\hat{m}_1(x) = \int y\hat{f}_{\hat{\epsilon}^0}\big(y - \tilde{m}_0(x)\big)\mathrm{d}y, \tag{2.8}$$

where $\hat{\epsilon}_i^0 = Y_i - \tilde{m}_0(X_i)$. Iterating further gives, for $j = 1, 2, ...$

$$\hat{m}_{j+1}(x) = \int y \hat{f}_{\hat{\epsilon}^j}\big(y - \hat{m}_j(x)\big)\mathrm{d}y, \tag{2.9}$$

where again $\hat{\epsilon}_i^j = Y_i - \hat{m}_j(X_i)$. Note that in this estimator x can only be equal to the observed x_i for $i = 1, ..., n$, because we update $\hat{\epsilon}_i^j$ at each iteration. One would perhaps believe that this "iterated convolution estimator" will give better results than the convolution estimator. However, this is not the case, unless one uses a special type of kernel function in the estimation of $\hat{f}_{\hat{\epsilon}^j}$. This special kernel is introduced, and some simulation results are given in section 7.

3. Intuitive discussion of bias reduction.

Asymptotic analysis of nonparametric estimators is usually based on the asymptotic bias and variance of the estimator. It is well known that the Gasser-Müller estimator has an asymptotic variance 1.5 times that of the Nadaraya-Watson estimator, but its asymptotic bias is superior; see [17] and [1]. The local polynomial estimator of order one and higher has been examined by several authors, and has been found to have better properties than the above mentioned estimators. In particular, it provides automatic boundary bias correction; see [4], [6], [5] and [15]. For a more complete discussion of the different estimators and comparisons, see the books [27], [25], [7], [8] and references therein.

We now discuss the asymptotic properties of the estimator (2.4). In the sequel, the bandwidth h refers to both h_D and h_R, since most of the time we assume that these two bandwidths are equal. Standard conditions on the kernels, the random variables and the regression function are assumed to be fulfilled, see e.g. [27] page 120.

The relation (2.4) can be written as

$$\hat{m}(x) = \sum_{i=1}^{n} \int y \frac{1}{nh_D} K\big(\frac{y - \tilde{m}(x) - \hat{\epsilon}_i}{h_D}\big)\mathrm{d}y. \tag{3.1}$$

By a simple substitution and using assumptions on $K(\cdot)$, (3.1) gives

$$\hat{m}(x) = \tilde{m}(x) + \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i. \tag{3.2}$$

Further,

$$\begin{aligned} \hat{\epsilon}_i = Y_i - \tilde{m}(X_i) &= Y_i - m(X_i) + m(X_i) - \tilde{m}(X_i) \\ &= \epsilon_i + m(X_i) - \tilde{m}(X_i). \end{aligned} \tag{3.3}$$

Substituting (3.3) in (3.2), we obtain

$$\hat{m}(x) - m(x) = \tilde{m}(x) - m(x) - \frac{1}{n}\sum_{i=1}^{n}[\tilde{m}(X_i) - m(X_i)] + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i. \tag{3.4}$$

From this relation it is possible to find the asymptotic bias of $\hat{m}(x)$.

Let us recall the asymptotic bias formula for the local linear estimator (e.g. [27] page 124). With a slight abuse of notation,

$$\text{As.Bias}\big(\tilde{m}(x)\big) = \text{E}\big(\tilde{m}(x) - m(x)\big) \sim \frac{h^2}{2} m''(x) \int z^2 K(z)\text{d}z. \tag{3.5}$$

Since

$$\begin{aligned}\text{E}\big(\tilde{m}(X_i) - m(X_i)\big) &= \text{E}[\text{E}\big(\tilde{m}(X_i)\big) - m(X_i)|X_i] \\ &\sim \text{E}\big(\frac{h^2}{2} m''(X_i) \int z^2 K(z)\text{d}z\big),\end{aligned} \tag{3.6}$$

we obtain

$$\text{E}\big(\frac{1}{n}\sum_{i=1}^{n}[\tilde{m}(X_i) - m(X_i)]\big) \sim \frac{h^2}{2} \int z^2 K(z)\text{d}z \int m''(y) f(y)\text{d}y. \tag{3.7}$$

Further,

$$\text{E}\big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\big) = 0. \tag{3.8}$$

Hence the asymptotic bias of $\hat{m}(x)$ is

$$\begin{aligned}\text{As.Bias}\big(\hat{m}(x)\big) &\sim \text{As.Bias}\big(\tilde{m}(x)\big) - \frac{h^2}{2} \int z^2 K(z)\text{d}z \int m''(y) f(y)\text{d}y \\ &= \frac{h^2}{2} \int z^2 K(z)\text{d}z [\int \big(m''(x) - m''(y) f(y)\big)\text{d}y].\end{aligned} \tag{3.9}$$

Let us consider the following special cases:

1. $m''(x) =$ constant (i.e. $m(x) = a + bx + cx^2$). The bias of $\hat{m}(x)$ is of higher order and improvement of the bias can be expected.
2. $m''(x) = 0$ (linear case). The bias is of higher order, both for $\hat{m}(x)$ and $\tilde{m}(x)$, and it is uncertain whether improvement is obtained.
3. x close to maximum and minimum values (peaks and valleys). If $\hat{m}(x) - m(x)$ has maxima at these points even though the bias correction in (3.9) is x-independent, visually the reduction will be large in these points (cf. figure 2, which is explained in more detail in section 6).

Observe, however, that if one has a curve with several peaks and valleys it may be difficult to gain any bias reduction. This is because the integral $\int m''(y) f(y)\text{d}y$ can be equal to zero in this case. Thus the bias reduction gained in some places (e.g. in a peak) will be balanced by an increased bias other places (e.g. in a valley).

As mentioned before, performing iterations of the estimator in equation (2.9) will not give any improved bias effect. In this case, the equation (3.2) will be

$$\hat{m}_{j+1}(x) = \hat{m}_j(x) + \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^j. \tag{3.10}$$

For $j = 1$, the same argument as above gives

$$\text{As.Bias}\big(\hat{m}_2(x)\big) = \text{As.Bias}\big(\hat{m}_1(x)\big), \tag{3.11}$$

and further

$$\text{As.Bias}\big(\hat{m}_{j+1}(x)\big) = \text{As.Bias}\big(\hat{m}_j(x)\big). \tag{3.12}$$

However, we introduce a special kernel in section 7, such that a bias reduction may occur at each iteration.

Another possible improvement suggested by a referee, is that instead of (3.2) one could think about localized corrections where the average over all residuals is replaced by a locally weighted average of residuals in the neighborhood of x only. This could alleviate some of the disadvantages that are associated with the current global adjustment, and should be a part of further research.

4. Distributional properties.

By (3.4)

$$\hat{m}(x) - \mathrm{E}\big(\hat{m}(x)\big) = \tilde{m}(x) - \mathrm{E}\big(\tilde{m}(x)\big) - \Big[\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big) - \mathrm{E}\big[\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)\big]\Big] + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i. \tag{4.1}$$

We have

$$\tilde{m}(x) - \mathrm{E}\big(\tilde{m}(x)\big) = O_p(\frac{1}{\sqrt{nh}}) \tag{4.2}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i = O_p(\frac{1}{\sqrt{n}}). \tag{4.3}$$

If we can show that the convergence in probability of the remaining terms in (4.1) is of higher order, then the asymptotic distribution of $\hat{m}(x) - \mathrm{E}(\hat{m}(x))$ is the same as $\tilde{m}(x) - \mathrm{E}(\tilde{m}(x))$, but with a different asymptotic bias.

Although this is not necessary, we do our formal calculations as if all expectations exist. Let us consider again

$$\mathrm{E}\big[\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)\big]^2 = \mathrm{E}\big[\frac{1}{n^2}\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)^2\big] + \mathrm{E}\big[\frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)\big(\tilde{m}(X_j) - m(X_j)\big)\big]. \tag{4.4}$$

Further

$$\mathrm{E}\big(\tilde{m}(X_i) - m(X_i)\big)^2 = \mathrm{E}\big[\mathrm{E}\big(\tilde{m}(X_i) - m(X_i)|X_i\big)^2\big] = o(1). \tag{4.5}$$

Thus

$$\mathrm{E}\big[\frac{1}{n^2}\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)^2\big] = o(n^{-1}). \tag{4.6}$$

Let us examine the second term in (4.4). By independence,

$$\begin{aligned}&\mathrm{E}\big[\big(\tilde{m}(X_i) - m(X_i)\big)\big(\tilde{m}(X_j) - m(X_j)\big)\big]\\ &= \mathrm{E}\Big[\mathrm{E}\big[\big(\tilde{m}(X_i) - m(X_i)\big)\big(\tilde{m}(X_j) - m(X_j)\big)|X_i, X_j\big]\Big]\\ &\sim \int \mathrm{E}\big[\big(\tilde{m}(x) - m(x)\big)\big(\tilde{m}(y) - m(y)\big)\big]f(x)f(y)\mathrm{d}x\mathrm{d}y.\end{aligned} \tag{4.7}$$

The expectation in the above integral satisifies

$$\mathrm{E}\big[(\tilde{m}(x)-m(x))(\tilde{m}(y)-m(y))\big]=\mathrm{E}\big[\tilde{m}(x)\tilde{m}(y)\big]$$
$$-m(x)\mathrm{E}\big[\tilde{m}(y)-m(y)\big]-m(y)\mathrm{E}\big[\tilde{m}(x)-m(x)\big]-m(x)m(y). \tag{4.8}$$

Let us examine the term $\mathrm{E}[\tilde{m}(x)\tilde{m}(y)]$. By a conditioning argument and by independence, we obtain

$$\mathrm{E}[\tilde{m}(x)\tilde{m}(y)]=\mathrm{E}\Big[n^{-2}\sum_{i=1}^{n}\frac{\{\hat{s}_2(x)-\hat{s}_1(x)(x-X_i)\}K_{h_R}^L(x-X_i)Y_i}{\hat{s}_2(x)\hat{s}_0(x)-\hat{s}_1(x)^2}$$
$$\cdot\sum_{j=1}^{n}\frac{\{\hat{s}_2(y)-\hat{s}_1(y)(y-X_j)\}K_{h_R}^L(y-X_j)Y_j}{\hat{s}_2(y)\hat{s}_0(y)-\hat{s}_1(y)^2}\Big]\sim\frac{n-1}{n^2}\mathrm{E}\big(\tilde{m}(x)\big)\mathrm{E}\big(\tilde{m}(y)\big)$$
$$+\mathrm{E}\Big[n^{-2}\sum_{i=1}^{n}\frac{Y_i^2\{\hat{s}_2(x)-\hat{s}_1(x)(x-X_i)\}K_{h_R}^L(x-X_i)}{[\hat{s}_2(x)\hat{s}_0(x)-\hat{s}_1(x)^2][\hat{s}_2(y)\hat{s}_0(y)-\hat{s}_1(y)^2]}$$
$$\cdot K_{h_R}^L(y-X_i)\{\hat{s}_2(y)-\hat{s}_1(y)(y-X_i)\}\Big]. \tag{4.9}$$

From [27] p. 123, asymptotically

$$\hat{s}_l(x)\sim\begin{cases}h^l\int z^lK^L(z)\mathrm{d}zf(x)+o_P(h^l) & l \text{ even},\\ h^{l+1}\int z^{l+1}K^L(z)\mathrm{d}zf'(x)+o_P(h^{l+1}) & l \text{ odd}.\end{cases}$$

Therefore the order of the denominator in (4.9) is

$$h^4\Big[\int z^2K^L(z)\mathrm{d}z\Big]^2f^2(x)f^2(y)+o_P(h^4). \tag{4.10}$$

The only term contributing to the last term of the numerator in (4.9) is

$$\frac{1}{n^2h^2}\hat{s}_2(x)\hat{s}_2(y)\mathrm{E}\Big[\sum_{i=1}^{n}K^L(\frac{x-X_i}{h})K^L(\frac{y-X_i}{h})Y_i^2\Big], \tag{4.11}$$

since all the other terms are of higher order.

Further, by a simple substitution, Taylor expansion, using the equation (4.10) and since

$$\hat{s}_2(x)\hat{s}_2(y)=h^4\Big[\int z^2K^L(z)\mathrm{d}z\Big]^2f(x)f(y)+o_P(h^4), \tag{4.12}$$

we obtain that the last term in (4.9) becomes asymptotically,

$$\frac{1}{n^2h^2f(x)f(y)}\mathrm{E}\Big[\sum_{i=1}^{n}K^L(\frac{x-X_i}{h})K^L(\frac{y-X_i}{h})Y_i^2\Big]$$
$$=\frac{1}{nhf(x)f(y)}\int v^2K^L(z)K^L(\frac{zh+y-x}{h})f_{X,Y}(x-zh,v)\mathrm{d}z\mathrm{d}v$$
$$=\frac{1}{nhf(y)}K_2^L(\frac{y-x}{h})\mathrm{E}(Y^2|X=x), \tag{4.13}$$

where $K_2^L(w)=\int K(z)K(z+w)\mathrm{d}z$.

Writing

$$\mathrm{E}\big(\tilde{m}(x)\big)\mathrm{E}\big(\tilde{m}(y)\big)=m(x)m(y)+m(x)\mathrm{E}[\tilde{m}(y)-m(y)]$$
$$+m(y)\mathrm{E}[\tilde{m}(x)-m(x)]+\mathrm{E}[\tilde{m}(y)-m(y)]\mathrm{E}[\tilde{m}(x)-m(x)], \tag{4.14}$$

the equation (4.8) becomes

$$\begin{aligned}\mathrm{E}\big[\big(\tilde{m}(x)-m(x)\big)\big(\tilde{m}(y)-m(y)\big)\big] \sim \\ (1-\frac{1}{n})\mathrm{E}\big[\tilde{m}(y)-m(y)\big]\mathrm{E}\big[\tilde{m}(x)-m(x)\big] \\ +\frac{1}{nhf(y)}K_2^L\big(\frac{y-x}{h}\big)\mathrm{E}(Y^2|X=x).\end{aligned} \tag{4.15}$$

Inserting this in (4.7), gives by substitution

$$\begin{aligned}\mathrm{E}\big[\big(\tilde{m}(X_i)-m(X_i)\big)\big(\tilde{m}(X_j)-m(X_j)\big)\big] \sim \\ (1-\frac{1}{n})\int \mathrm{E}[\tilde{m}(y)-m(y)]\mathrm{E}[\tilde{m}(x)-m(x)]f(x)f(y)\mathrm{d}x\mathrm{d}y \\ +\frac{1}{nh}\int K_2^L\big(\frac{y-x}{h}\big)\mathrm{E}(Y^2|X=x)f(x)\mathrm{d}x\mathrm{d}y \\ =(1-\frac{1}{n})(\frac{h^2}{2})^2[\int z^2K_2^L(z)\mathrm{d}z]^2[\int m''(x)f(x)\mathrm{d}x]^2 \\ +\frac{1}{n}\int K_2^L(z)\mathrm{d}z\int \mathrm{E}(Y^2|X=x)f(x)\mathrm{d}x.\end{aligned} \tag{4.16}$$

Clearly the first term in (4.16) can be identified with the squared expectation in the decomposition

$$\begin{aligned}\mathrm{E}\big[\frac{1}{n}\sum_{i=1}^n\big(\tilde{m}(X_i)-m(X_i)\big)\big]^2=\mathrm{Var}\big[\frac{1}{n}\sum_{i=1}^n\big(\tilde{m}(X_i)-m(X_i)\big)\big] \\ +\big[\mathrm{E}\big(\frac{1}{n}\sum_{i=1}^n[\tilde{m}(X_i)-m(X_i)]\big)\big]^2.\end{aligned} \tag{4.17}$$

The term

$$\begin{aligned}\frac{1}{n}\int K_2^L(z)\mathrm{d}z\int \mathrm{E}(Y^2|X=x)f(x)\mathrm{d}x= \\ \frac{1}{n}\int K_2^L(z)\mathrm{d}z[\sigma_\epsilon^2+\int m^2(x)f(x)\mathrm{d}x]\end{aligned} \tag{4.18}$$

can be identified with the variance. We have

$$\frac{1}{n}\sum_{i=1}^n[\tilde{m}(X_i)-m(X_i)]-\mathrm{E}\big(\frac{1}{n}\sum_{i=1}^n\tilde{m}(X_i)-m(X_i)\big)\sim o_P(\frac{1}{\sqrt{n}}). \tag{4.19}$$

From the equation (4.1), it follows that $\hat{m}(x)-\mathrm{E}(\hat{m}(x))$ has the same asymptotic normal distribution as $\tilde{m}(x)-\mathrm{E}(\tilde{m}(x))$, i.e. the asymptotic variance is the same for the estimators $\hat{m}(x)$ and $\tilde{m}(x)$, but

$$\mathrm{As.Bias}\big(\hat{m}(x)\big)=\mathrm{As.Bias}\big(\tilde{m}(x)\big)-\frac{h^2}{2}\int z^2K(z)\mathrm{d}z\int m''(y)f(y)\mathrm{d}y. \tag{4.20}$$

5. Total squared error.

We would like to compare the asymptotic total squared error, i.e.

$$\mathrm{E}\big[\sum_{i=1}^n\big(\hat{m}(X_i)-m(X_i)\big)^2\big]\ \text{ against }\ \mathrm{E}\big[\sum_{i=1}^n\big(\tilde{m}(X_i)-m(X_i)\big)^2\big]. \tag{5.1}$$

From the equation (3.4)

$$\hat{m}(X_i) - m(X_i) = \tilde{m}(X_i) - m(X_i) - \frac{1}{n}\sum_{j=1}^{n}[\tilde{m}(X_j) - m(X_j)] + \frac{1}{n}\sum_{j=1}^{n}\epsilon_j. \tag{5.2}$$

Further,

$$\begin{aligned}[\hat{m}(X_i) - m(X_i)]^2 &= [\tilde{m}(X_i) - m(X_i)]^2 \\ -\frac{2}{n}[\tilde{m}(X_i) - m(X_i)]&\sum_{j=1}^{n}[\tilde{m}(X_j) - m(X_j)] \\ +\frac{1}{n^2}\big[\sum_{j=1}^{n}\big(\tilde{m}(X_j) - m(X_j)\big)\big]^2 &+ \frac{2}{n}[\tilde{m}(X_i) - m(X_i)]\sum_{j=1}^{n}\epsilon_j \\ -\frac{2}{n^2}\sum_{j=1}^{n}[\tilde{m}(X_j) - m(X_j)]&\sum_{k=1}^{n}\epsilon_k + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\epsilon_j\epsilon_k. \end{aligned} \tag{5.3}$$

This implies

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[\hat{m}(X_i) - m(X_i)]^2 &= \frac{1}{n}\sum_{i=1}^{n}[\tilde{m}(X_i) - m(X_i)]^2 \\ -\frac{1}{n^2}\big[\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)\big]^2 &+ \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\epsilon_j\epsilon_k. \end{aligned} \tag{5.4}$$

Taking expectation in (5.4) gives us the total squared error of $\hat{m}(x)$. Thus the order of the different terms is

$$\begin{aligned}\mathrm{E}\big[\frac{1}{n}\sum_{i=1}^{n}[\tilde{m}(X_i) - m(X_i)]^2\big] &\sim \int \Big[\mathrm{Var}\big(\tilde{m}(x)\big) + \mathrm{Bias}^2\big(\tilde{m}(x)\big)\Big] f(x)\mathrm{d}x \\ &= O(\frac{1}{nh} + h^4), \end{aligned} \tag{5.5}$$

from the decomposition, (4.17), and the calculated bias and variance

$$\mathrm{E}\Big[\frac{1}{n^2}\big[\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)\big]^2\Big] = O(h^4) \tag{5.6}$$

and at last

$$\mathrm{E}\big[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\epsilon_j\epsilon_k\big] = \frac{\sigma_\epsilon^2}{n}. \tag{5.7}$$

This means that if $\mathrm{E}\big[\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)\big] \neq 0$ asymptotically, then

$$\mathrm{E}\big[\sum_{i=1}^{n}\big(\hat{m}(X_i) - m(X_i)\big)^2\big] < \mathrm{E}\big[\sum_{i=1}^{n}\big(\tilde{m}(X_i) - m(X_i)\big)^2\big] \tag{5.8}$$

i.e. the total asymptotic squared error of $\hat{m}(x)$ is smaller than $\tilde{m}(x)$.

6. Simulation study.

We compare the estimator (2.4) with the local linear estimator in several situations. The comparisons are based on the mean squared error (MSE) of the estimators. For $\hat{m}(x)$ the MSE is, if it exists,

$$\mathrm{MSE}\big(\hat{m}(x)\big) = \mathrm{E}\big[\{\hat{m}(x) - m(x)\}^2\big]. \tag{6.1}$$

In the simulations we use the empirical mean squared error

$$\hat{\mathrm{MSE}}(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{m}(X_i) - m(X_i)\}^2, \tag{6.2}$$

and likewise for the local linear estimator.

We have simulated between 100 and 500 realizations of sample size 100 to 10 000 of (X_i, Y_i), and calculated the empirical MSE (6.2). The bandwidth choice in the kernel density estimation for f_ϵ is the Solve-the-Equation Plug-in approach proposed in [24], while the bandwidth used in the local linear estimator is the Direct Plug-In methodology described in [22], if nothing else is stated. Gaussian kernels are always used, both in the regression estimation of $m(x)$ and the density estimation of f_ϵ. The integration in the estimator (2.4) is calculated using the trapezoidal rule between $[2\min(Y_i), 2\max(Y_i)]$, when $\min(Y_i) < 0$ and $\max(Y_i) > 0$. We consider the case where observations of (X_i, Y_i) are independent.

Our first simulation experiment is based on the model $Y_i = X_i^2 + \epsilon_i$, where ϵ_i are i.i.d normal with expectation 0 and variance 0.1 and X_i is uniformly distributed on $[-2, 2]$. A hundred realizations each with sample size 500 have been simulated, and the convolution and local linear estimators have been used to estimate the regression curves. In this case, the estimated MSE for the local linear estimator is: $2.301 \cdot 10^{-3}$ and for the convolution estimator: $1.986 \cdot 10^{-3}$. Thus we obtain an improvement of 13.690%. Figure 1 shows the estimated variance and bias of the two estimators. The upper figure displays the estimated variance; here there is no difference between the two estimators. The lower figure shows the estimated bias. The dashed line is the estimated bias for the convolution estimator. This is clearly smaller than for the local linear estimator; thus our predictions that the improvement occurs in the bias is supported. The results are similar for the other simulation experiments. See table 1 with sample size equal to 100, 1000 and 5000. In these cases the improvement is also considerable.

Table 1: The estimated MSE for a parabola using the convolution estimator

Sample size	Local linear	Convolution type	Improvement in %
100	$9.086 \cdot 10^{-3}$	$7.997 \cdot 10^{-3}$	11.985
500	$2.301 \cdot 10^{-3}$	$1.986 \cdot 10^{-3}$	13.690
1000	$1.321 \cdot 10^{-3}$	$1.120 \cdot 10^{-3}$	15.216
5000	$3.489 \cdot 10^{-4}$	$2.957 \cdot 10^{-4}$	15.248

The next simulation is based on the same model, except that the interval is now $[-0.5, 0.5]$. Only one realization with 500 sample points has been simulated, and the estimated lines are given in figure 2. The solid line is the true function, the non-filled points are the local linear estimates and the black points are the convolution estimates. These results clearly indicate that the asymptotic bias formula in

Table 2: The estimated MSE for a parabola using the convolution estimator and the Nadaraya-Watson estimator with a fourth order kernel

Sample size	Nadaraya-Watson	Convolution type	Improvement in %
100	$2.448 \cdot 10^{-2}$	$1.002 \cdot 10^{-2}$	59.069
500	$5.839 \cdot 10^{-3}$	$2.175 \cdot 10^{-3}$	62.750
1000	$3.396 \cdot 10^{-3}$	$1.243 \cdot 10^{-3}$	63.398
5000	$1.020 \cdot 10^{-3}$	$2.978 \cdot 10^{-4}$	70.804

Table 3: The estimated MSE for a straight line using the convolution estimator

Sample size	Local linear	Convolution type	Improvement in %
100	$5.675 \cdot 10^{-3}$	$5.653 \cdot 10^{-3}$	0.388
500	$1.240 \cdot 10^{-3}$	$1.246 \cdot 10^{-3}$	-0.484
1000	$5.627 \cdot 10^{-4}$	$5.630 \cdot 10^{-4}$	-0.053
5000	$1.009 \cdot 10^{-4}$	$1.098 \cdot 10^{-4}$	-8.821

equation (3.9) is reasonable. For each estimated local linear point, the estimated convolution point is below by a fixed amount, but the visual impression is that the convolution estimator does much better at the bottom points of the parabola.

As mentioned in the introduction, a common bias reduction technique in non-parametric estimation is the use of higher order kernels, see e.g. [27] page 32. Thus we have included a comparison between the Nadaraya-Watson estimator with a fourth order kernel and the proposed convolution estimator. The fourth order kernel used is $K_4(x) = 0.5(3-x^2)\phi(x)$, where $\phi(x)$ is the standard normal distribution. The bandwidth used in the Nadaraya-Watson estimator is the same as for the above local linear estimator. It is thus not optimal in this situation, but other choices of the bandwidth has been examined without a large impact on the results. Again we perform simulations for the parabola model on the interval $[-2, 2]$. The results from the simulations are given in table 2. The convolution estimator clearly outperforms the fourth order kernel method when comparing the MSE. Figure 3 may explain these results. Here the bias and variance of both estimator are plotted for the simulations with sample size 500. The bias of both estimators, the lower plot, seems to be reasonably equal. But the variance, the upper plot, for the fourth order kernel method is much larger, the solid line, than the variance for the convolution estimator, the dashed line. This behaviour of the fourth order kernel method is not unexpected, see e.g. [25] page 60.

The last simulation experiment in this section is based on a straight line regression, $Y_i = a + bX_i + \epsilon_i$, with $a = 1$, $b = 1$, ϵ as before, and X_i uniformly distributed on $[0, 2]$. From this model, 100 realizations of sample size 100 to 5000 have been simulated. The integration in the estimator is now performed on the interval $[-2\max(Y_i), 2\max(Y_i)]$. The results given in table 3 for the convolution estimator, indicate that the convolution estimator is almost as good as the local linear estimator. We cannot expect the convolution estimator to do better here, since $m''(x)$ is zero if $m(x)$ is a straight line, thus no bias improvement occurs in the formula (3.9).

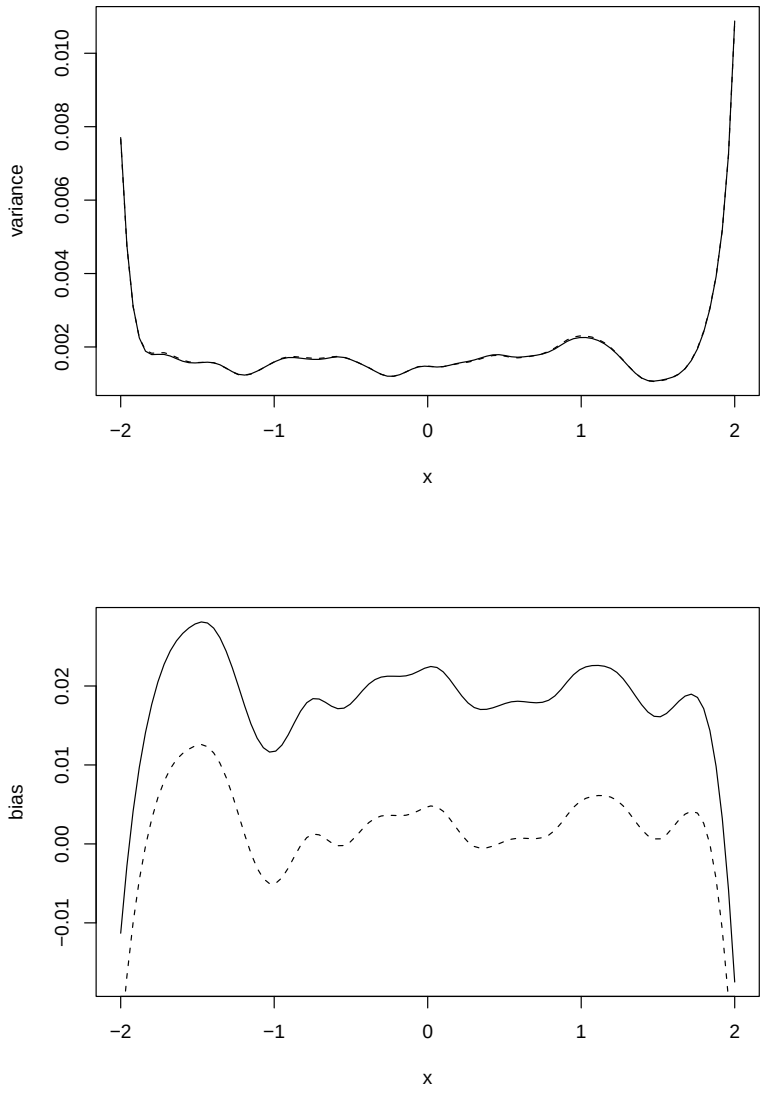

Figure 1: The estimated variance (top) and bias for the parabola experiment (dashed line - convolution estimator, solid line - local linear estimator)

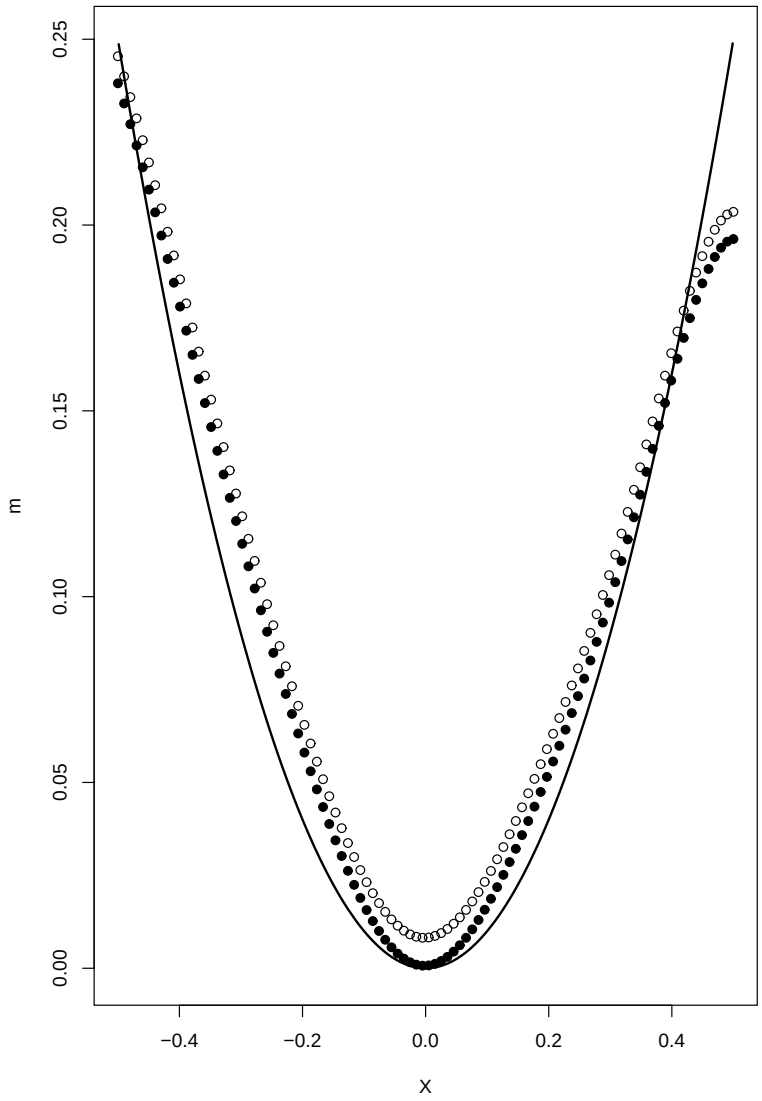

Figure 2: The solid true line, the estimated local linear points (non-filled) and the estimated convolution points (black) from one realization for the parabola model

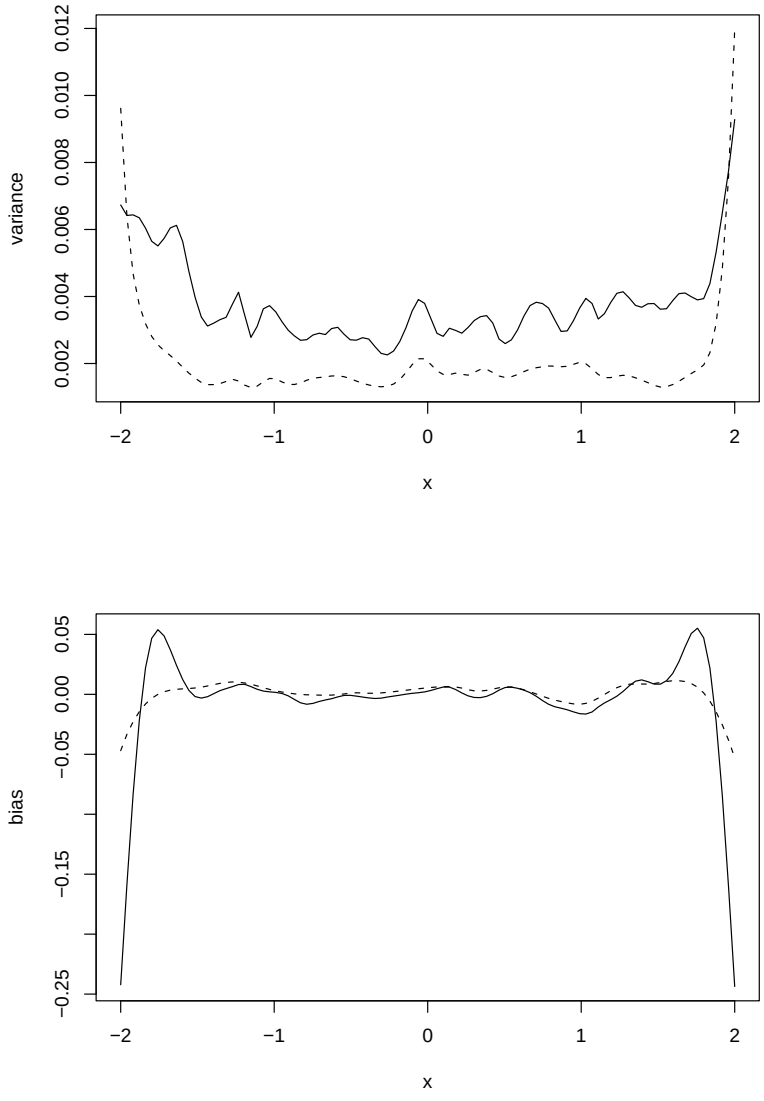

Figure 3: The estimated variance (top) and bias for the parabola experiment (dashed line - convolution estimator, solid line - Nadaraya-Watson estimator with fourth order kernel)

7. A special kernel variant.

Let us consider equation (2.4) again. With a substitution we obtain,

$$\hat{m}(x) = \int y\hat{f}_{\hat{\epsilon}}(y - \tilde{m}(x))\mathrm{d}y = \int z\hat{f}_{\hat{\epsilon}}(z)\mathrm{d}z + \tilde{m}(x)\int \hat{f}_{\hat{\epsilon}}(z)\mathrm{d}z. \tag{7.1}$$

If $\int \hat{f}_{\hat{\epsilon}}(z)\mathrm{d}z > 1$, the estimator $\hat{m}(x)$ will clearly be closer to the true function than $\tilde{m}(x)$ in locations where the function has a large curvature due to the bias formula (3.5). Of course one could adjust with an estimate $\tilde{m}''(x)$ of $m''(x)$, but this increases the variance. Instead we have chosen to introduce a kernel function with the property

$$\int K(z)\mathrm{d}z > 1. \tag{7.2}$$

This could be considered as an alternative to allowing the kernel to be negative, which is a known device for reducing bias, as seen in section 6 for the higher order kernel used there.

In this case, it may pay to perform iterations, as equation 2.9 suggests. However, $\hat{m}(x)$ will also be larger than $\tilde{m}(x)$ in absolute value in locations where $m''(x) \approx 0$, and this is not desirable. The following simulation experiments should therefore just be considered as a part of a preliminary investigation where at least some promising results are obtained, but where more work is needed to find a more optimal procedure. In these experiments we have chosen the kernel K such that $\int \hat{f}_{\hat{\epsilon}}(z)\mathrm{d}z = 1.001$, a very modest overestimation indeed, and clearly other choices can be examined.

Two regression functions have been studied: from [14] chapter 5,

$$m_1(x) = \sin^3(2\pi x^3) \tag{7.3}$$

and from [21]

$$m_2(x) = \sin(2x) + 2\exp(-16x^2). \tag{7.4}$$

See figure 4 for these curves. The observations have been generated by simulating X_i as uniformly distributed on an interval $[0, 1]$ for $m_1(x)$ and $[-2, 2]$ for $m_2(x)$. The response observations have been generated through

$$Y_i = m(X_i) + \epsilon_i, \tag{7.5}$$

where ϵ_i are i.i.d normal with expectation 0 and variance 0.1. Here 100 realizations of sample size 100 to 10 000 of (X_i, Y_i) have been simulated. Since both functions have at least one peak and one valley, we may not expect to get much bias reduction. The estimated MSE, using the convolution estimator with the adusted kernel (7.2), is shown in table 4 for m_1. As expected, the results show only a very modest improvement. The results for m_2 were similar.

In table 5 and 6, the iterated convolution estimator (2.9) with the adjustment (7.2) has been used with $i = 10$ iterations, for the regression of $m_1(x)$ and $m_2(x)$. Again 100 realizations have been simulated with different sample size, and the MSE has been estimated. Here the normal reference bandwidth selector, see e.g. [14] page 91, has been used for the selection of the bandwidth in the density estimation.

Both tables show that the iterated convolution type estimator performes better than the local linear estimator. The results also indicate that performing iterations improves the first order convolution estimator $\hat{m}_1(x)$. Clearly, one cannot improve

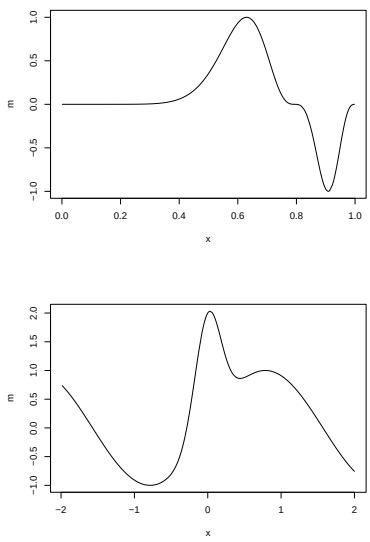

Figure 4: The curve $m_1(x)$ at the top, $m_2(x)$ at the bottom

Table 4: The estimated MSE for $m_1(x)$

Sample size	Local linear	Convolution type	Improvement in %
100	$1.442 \cdot 10^{-2}$	$1.445 \cdot 10^{-2}$	-0.208
500	$4.071 \cdot 10^{-3}$	$4.049 \cdot 10^{-3}$	0.540
1000	$2.382 \cdot 10^{-3}$	$2.374 \cdot 10^{-3}$	0.336
5000	$6.766 \cdot 10^{-4}$	$6.723 \cdot 10^{-4}$	0.636
10 000	$4.028 \cdot 10^{-4}$	$3.988 \cdot 10^{-4}$	0.993

Table 5: The estimated MSE for $m_1(x)$, using the iteration estimator $\hat{m}_i(x)$ with $i = 10$

Sample size	Local linear	Iterated convolution type	Improvement in %
100	$1.438 \cdot 10^{-2}$	$1.434 \cdot 10^{-2}$	0.278
500	$4.292 \cdot 10^{-3}$	$4.235 \cdot 10^{-3}$	1.328
1000	$2.464 \cdot 10^{-3}$	$2.387 \cdot 10^{-3}$	3.125
5000	$6.648 \cdot 10^{-4}$	$6.319 \cdot 10^{-4}$	4.949

Table 6: The estimated MSE for $m_2(x)$, using the iteration estimator $\hat{m}_i(x)$ with $i = 10$

Sample size	Local linear	Iterated convolution type	Improvement in %
100	$1.773 \cdot 10^{-2}$	$1.741 \cdot 10^{-2}$	1.801
500	$4.669 \cdot 10^{-3}$	$4.531 \cdot 10^{-3}$	2.956
1000	$2.696 \cdot 10^{-3}$	$2.576 \cdot 10^{-3}$	4.451
5000	$7.245 \cdot 10^{-4}$	$6.887 \cdot 10^{-4}$	4.941

Table 7: The estimated MSE for $m_1(x)$ on the interval $[0.55, 0.7]$ using the iterated convolution estimator with $i = 10$

Sample size	Local linear	Iterated convolution type	Improvement in %
100	$1.373 \cdot 10^{-2}$	$1.373 \cdot 10^{-2}$	0
500	$3.273 \cdot 10^{-3}$	$3.127 \cdot 10^{-3}$	4.461
1000	$2.241 \cdot 10^{-3}$	$2.058 \cdot 10^{-3}$	8.166
5000	$4.766 \cdot 10^{-4}$	$4.205 \cdot 10^{-4}$	11.771

Table 8: The estimated MSE for $m_1(x)$ on the interval $[0.85, 0.95]$ using the iterated convolution estimator with $i = 10$

Sample size	Local linear	Iterated convolution type	Improvement in %
100	$3.621 \cdot 10^{-2}$	$3.414 \cdot 10^{-2}$	5.717
500	$1.005 \cdot 10^{-2}$	$8.802 \cdot 10^{-3}$	12.418
1000	$5.740 \cdot 10^{-3}$	$4.849 \cdot 10^{-3}$	15.523
5000	$1.653 \cdot 10^{-3}$	$1.229 \cdot 10^{-3}$	25.650

the estimates indefinitely. The improvement will be smaller and smaller, and at the same time there will be more and more higher order terms which may lead to trouble unless n is increased. It remains an important task to carry out more detailed calculations and to find a good stopping criterion. Possibly some cross-validation type criterion can be used, but this is left for future research.

The point of the adjustment (7.2) is to improve results in peaks and valleys. To check this, simulations with 100 realizations with sample sizes from 100 to 5000 have been performed for $m_1(x)$ and $m_2(x)$. However, the MSE has been calculated only for specific intervals, which contain the peaks and valleys of the curves.

In table 7 and 8, the results from simulations of $m_1(x)$ are given, using the iterated convolution estimator, with $i = 10$ and adjusted with (7.2). The intervals considered are $[0.55, 0.7]$, where the curve has a peak, and $[0.85, 0.95]$, which is a valley. The curve has not a very large curvature on the interval $[0.55, 0.7]$, thus the results in the table 7 show a modest improvement. However, the table 8 shows that the new estimator is much better when estimating on the parts of the function with high curvature. Similar results were obtained for $m_2(x)$.

The above results show that the proposed estimator is better in estimating parts of functions with high curvature, and in some cases this improvement is substantial. This is in contrast to the modest improvement we obtained when we compared the MSE of the whole curve.

A real data set with peaks and valleys has been examined for completeness. We use the motorcycle data set from [13], page 70, where the X-values represent time after a simulated impact with motorcycles and the response variable Y is the head acceleration of a post mortem human test object. Figure 5 shows the results, together with the data points. The upper graphs show the result from the local linear estimator (solid line) and the estimator in equation (2.4) (dashed line) with the adjusted kernel (7.2). These two estimators give approximately the same result. However, the lower graphs are different. Here the dashed line is the iteration estimator from (2.9), using 50 iterations and adjusted kernel, the solid line is as above. Again the convolution estimator is lower in valleys and higher in peaks compared to the local linear estimator.

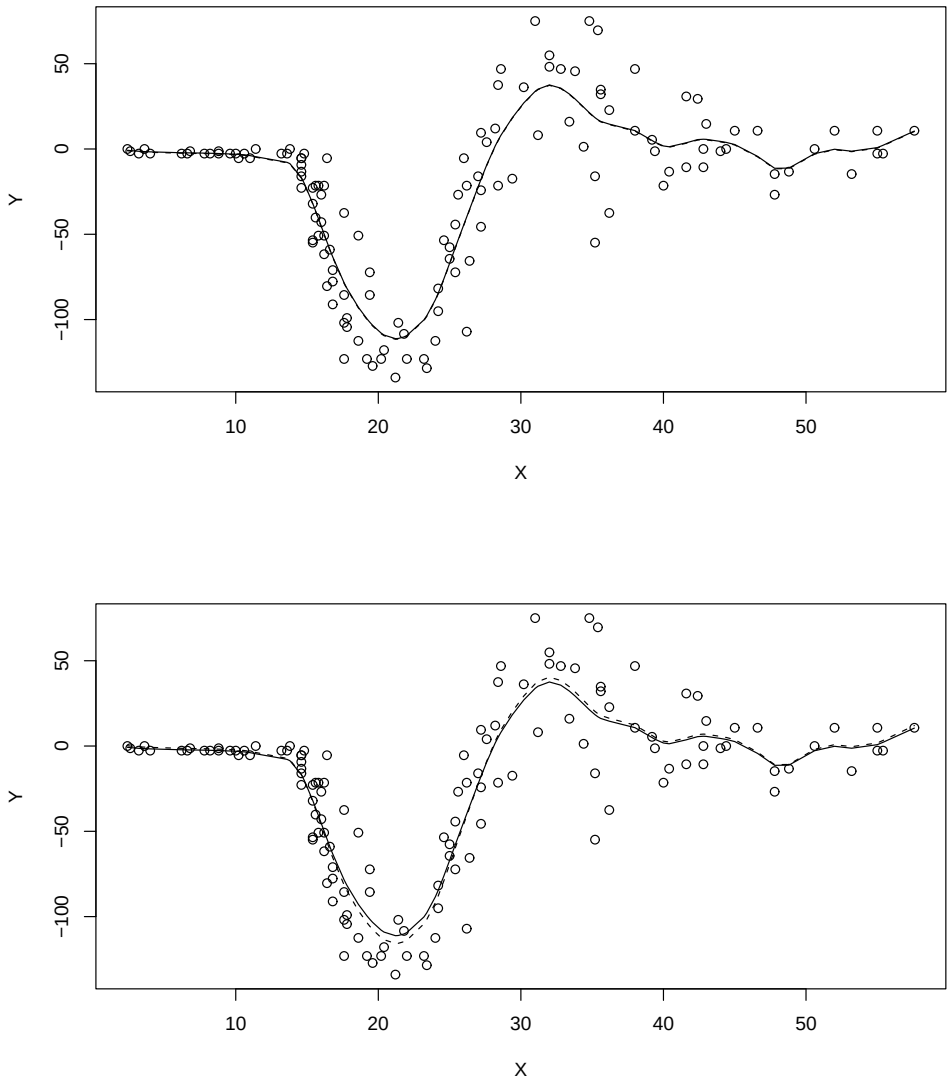

Figure 5: Nonparametric regression of the motorcycle data set: The data is given as points in both plots. Upper plot: solid line - local linear estimate, dashed line - convolution estimator with special kernel. Lower plot: solid line - local linear estimate, dashed line - iterated convolution estimator with special kernel (50 iterations).

8. Concluding remarks.

In this paper we have introduced a convolution estimator for nonparametric regression. Its asymptotic bias is proved to be smaller than standard kernel methods. The bias reduction will be large in cases where the function of interest has only one maximum (i.e. a peak) or one minimum (i.e. a valley).

Since the convolution estimator has two bandwidths, their choice is important. This has not been studied in this paper, and one might believe that the bias reduction can be larger if one is able to choose more optimal bandwidths.

An adjusted kernel has also been introduced and simulation results indicate that by using this kernel, even more bias reduction can be achieved. However, more theoretical analysis is needed here.

Acknowledgments.

This work is supported by grant 147231/432 from the Norwegian Research Council.

References

[1] Chu, C. K. and Marron, J. S (1991). Choosing a kernel regression estimator (with discussion). *Statist. Science* **6** 404-436. MR1146907

[2] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829-836. MR556476

[3] Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24** 2431-2461. MR1425960

[4] Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998-1004. MR1209561

[5] Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21** 196-216. MR1212173

[6] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008-2036. MR1193323

[7] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall. MR1383587

[8] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series*. Springer-Verlag. MR1964455

[9] Gasser, T. and Müller, H. G. (1979). Kernel Estimation of Regression Functions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.) Springer-Verlag, Heidelberg, 23-68. MR564251

[10] Gasser, T., Müller, H.-G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Series B* **47** 238-252. MR816088

[11] Glad, I. K. (1998). Parametrically guided non-parametric regression. *Scand. J. Statist.* **25** 649-668. MR1666776

[12] Hall, P. and Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *J. Amer. Statist. Assoc.* **86** 665-672. MR1147090

[13] Härdle, W. (1989). *Applied Nonparametric Regression*. Cambridge University Press. MR1161622

[14] Härdle, W. (1990). *Smoothing Techniques: With Implementation in S*. Springer-Verlag. MR1140190

[15] Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry (with comments). *Statist. Science* **8** 120-143.
[16] Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23** 882-904. MR1345205
[17] Mack, Y. P. and Müller, H. G. (1989). Convolution type estimators for nonparametric regression. *Statist. Probab. Lett.* **7** 229-239. MR980926
[18] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712-736. MR1165589
[19] Müller, H. G. (1987). Weigthed local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231-238. MR883351
[20] Nadaraya, E. A. (1964). On estimating regression. *Theo. Probab. and Applic.* **9** 141-142.
[21] Prewitt, K. A. (2003). Efficient bandwidth selection in non-parametric regression. *Scand. J. of Statist.* **30** 75-92. MR1963894
[22] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257-1270. MR1379468
[23] Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimates of densities. *Comm. Statist. - Theo. Meth.* **14** 1123-1126. MR797636
[24] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Series B* **53** 683-690. MR1125725
[25] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag. MR1391963
[26] Stone, C. J. (1977). Consistent Nonparametric Regression. *Ann. Statist.* **5** 596-620. MR443204
[27] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall. MR1319818
[28] Watson, G. S. (1964). Smooth Regression Analysis. *Sankhyā Ser. A* **26** 359-372. MR185765

University of Bergen
Department of Mathematics
Johannes Bruns gt.12
5008 Bergen
Norway
E-mail: {baards,dagt}@mi.uib.no

PAPER D

In preperation for submission

NONPARAMETRIC ADDITIVE MODELS FOR PANELS OF TIME SERIES *

Enno Mammen
Department of Economics, University of Mannheim
L7, 3-5, 68131 Mannheim, Germany
E mail: emammen@rumms.uni-mannheim.de

Bård Støve and Dag Tjøstheim
Department of Mathematics, University of Bergen
Johannes Brunsgate 12, 5008 Bergen, Norway
E mail: {baards, dag.tjostheim}@mi.uib.no

August 26, 2005

Keywords and phrases: *additive functions, backfitting, panel data, nonparametric estimation, nonlinear functions*

Short title: *'Nonparametric models for panel data'*

*This research was supported by grant no. 147231/432 of the Norwegian Research Council

1 Introduction.

There is already a substantial literature on nonlinear models and nonparametric methods in a regression and time series setting. But almost without exception these developments have been limited to univariate or multivariate models of moderate dimensions. Very little has been done for panels, where the dimension, often corresponding to a number of individuals, typically is very large, but where the number of observations for each individual may be small or moderate. One would expect that the practical need for nonlinear models will be no less in a panel situation, but a systematic theoretical foundation has been lacking. It is the aim of this paper to start establishing such a theory. The emphasis will be on a nonparametric approach, in particular on additive models. Extending existing methodology to the panel situation is by no means trivial (actually also in the linear case of interrelated time series in a panel there are many unsolved problems; see e.g. Hjellvik and Tjøstheim (1999), Fu et al. (2002), Wooldridge (2005a) and Chang (2004)).

To be more specific, let $\{Y^{it}\}$, $i = 1, \ldots, n;\ \ t = 1, \ldots, T$, be stochastic variables representing observations on a set of n individuals over T time periods. A rather general additive model for the panel $\{Y^{it}\}$ will be given by

$$Y^{it} = \sum_{j=1}^{p} m_j(X_j^{it}) + \eta_t + \lambda_i + \varepsilon^{it}. \tag{1.1}$$

Here $\{X_j^{it}\}$, $j = 1, \ldots, p$, is a set of explanatory variables (time lagged values of Y^{it} may be included among them), $\{\eta_t\}$ and $\{\lambda_i\}$ are temporary and individual effects, respectively, not captured by the explanatory variables, and $\{\varepsilon^{it}\}$ are error terms. The functions $\{m_j,\ j = 1, \ldots, p\}$ are unknown, and it is our task to estimate them. In the linear case $m_j(x) = \alpha_j x$ for some unknown parameters α_j, and our model reduces to the more familiar time series panel regression model treated in many papers and in the books by Hsiao (1986), Mátyás and Sevestre (1992), Baltagi (1995), Arellano (2003) and Arellano and Honoré (2001).

One of the difficulties arising in trying to establish an estimation theory for (1.1) is that the asymptotics can be imagined in several ways. We may let $n \to \infty$, $T \to \infty$ or both. Another problem is the presence of the individual effects λ_i and/or temporal effects η_t.

There are several ways of approaching the problem of estimating m_j. We have chosen to use backfitting (Hastie and Tibshirani 1990) to take advantage of the recent developments in this area in Opsomer and Ruppert (1997) and Mammen et al. (1999). In particular, we will rely heavily on the smoothed backfitting approach, the theory of which is introduced in Mammen et al. (1999), and where a simplified version with several applications is given in Nielsen and Sperlich (2005). An alternative would have

been to use marginal integration (Newey 1994, Tjøstheim and Auestad 1994, Linton and Nielsen 1995). The latter method only works for low dimensional covariates and for well behaved designs. But in case of deviations from the additive model (1.1) it is fairly robust and can be extended to include interaction terms (Sperlich et al. 2002). Other recent contributions to nonlinear and non/semiparametric modeling of panels are given in Hjellvik et al. (2004), Fan and Li (2004), Racine and Li (2004), Wooldridge (2005b), Baltagi and Li (2002) and Honoré and Lewbel (2002).

It will be seen that we do not at all manage to obtain a complete theory, and open problems will be pointed out as we proceed. An overview of the paper is as follows. In Section 2 we discuss Nadaraya-Watson smoothing in model (1.1). This will be done for models with and without individual effects ($\lambda_i \neq 0$ or $\lambda_i = 0$, respectively) and for two asymptotic settings. In the first asymptotic approach we fix T and let $n \to \infty$. In the second setting we let $T \to \infty$ and $n \to \infty$. Nadaraya-Watson smoothing in this model can be easily introduced and leads to an intuitive clear asymptotic theory. The more complicated theory for local linear smoothing will be developed in Section 3. Simulations and a real data example are presented in Section 4. The paper concludes in Section 5. Proofs are deferred to Section 6.

2 Local constant smoothing of panels of time series.

We will start by looking at model (1.1) when there are no individual effects $\{\lambda_i\}$ present. The presence of the temporary effects $\{\eta_t\}$ means that the time series of the panels will be interrelated (but not necessarily intercorrelated), even conditionally on the explanatory variables. A linear version similar to this case was considered in Hjellvik and Tjøstheim (1999) and Fu et al. (2002), and as in those publications we first put the emphasis on asymptotics for the case $n \to \infty$ with T fixed. As mentioned, this is a situation which frequently occurs for panels

2.1 Definition of local constant smoothers and asymptotics for $n \to \infty$ and T fixed.

We consider the following model

$$Y^{it} = \sum_{j=1}^{p} m_j(X_j^{it}) + \eta_t + \varepsilon^{it} \quad (i = 1, \ldots, n; t = 1, \ldots, T), \tag{2.1}$$

where one observes Y^{it} and covariables X_j^{it} with $j = 1, \ldots, p$. In particular, we think of the case where partially the covariables X_j^{it} are lagged observations Y^{it-j} (for $j \leq p'$, say) and where the other covariable X_j^{it} ($p' < j \leq p$) are external covariables. The variables $\eta_1, \ldots, \eta_T$ are constants that are unknown. They model the influence of some additional external variables onto the development of the time series Y^{it}. Typically they could be considered as nuissance parameters. For the error variables ε^{it} we assume that

(A1) We assume that $(\mathbf{Y}^1, \mathbf{X}^1, \varepsilon^1), ..., (\mathbf{Y}^n, \mathbf{X}^n, \varepsilon^n)$ are i.i.d., where $\mathbf{Y}^i = (Y^{it} : 1 \leq t \leq T)$, $\mathbf{X}^i = (X_j^{it} : 1 \leq j \leq p, 1 \leq t \leq T)$ and $\varepsilon^i = (\varepsilon^{it} : 1 \leq t \leq T)$ Given $(X_j^{is} : 1 \leq j \leq p,\ 1 \leq s \leq t; \varepsilon^{is} : 1 \leq s \leq t-1)$, the variable ε^{it} has conditional mean 0, and conditional variance σ^2 and they have a conditional absolute moment of order 5/2 that is uniformly (in i and t) bounded by a constant.

We could treat more complicated data generating processes. In particular, one could allow for conditional heteroscedasticity and one could allow for weak conditional dependence of $\varepsilon^{1t}, \ldots, \varepsilon^{nt}$. We do not treat such general cases mostly for simplicity of notation. Essentially, for the proof of our results we only need asymptotic normality of $\sqrt{nTh} \sum_{i=1}^n \sum_{t=1}^T K_h(x_j - X_j^{it})\varepsilon^{it}$ for a kernel $K_h(u) = h^{-1}K(uh^{-1})$ with bandwidth h and rates of convergence for terms $(Tn)^{-1} \sum_{i=1}^n \sum_{t=1}^T g_n(X_j^{it})\varepsilon^{it}$ for certain sequences of functions g_n. Clearly, this could be studied in a much more general set up. Furthermore, one could treat non i.i.d. $\mathbf{X}^i$. In fact, we only make use of some ergodicity properties of the covariable vectors that are needed to get asymptotic formulas for bias and variance expressions.

We will discuss asymptotics of a backfitting estimator in an asymptotic framework where

(A2) T is fixed and $n \to \infty$.

This covers a large class of applications where one observes a large number of individual time series for a moderate number of time points. We will comment on the case $T \to \infty,\ n \to \infty$ in the next subsection.

For identifiability of m_j we use the following norming conditions

(A3) It holds that

$$\sum_{t=1}^T Em_j(X_j^{it}) \quad = \quad 0.$$

We make the following smoothness assumptions on the density of the covariables and the regression functions $m_1, \ldots, m_p$:

(A4) We assume that X^{it} has a density on $[0,1]^p$. The conditional density of X^{it} given that X^{it} lies in $[0,1]^p$ is denoted by f^t. We assume that f^t has continuous partial derivatives of order one and that it is bounded from below (on $[0,1]^p$). The regression functions m_j have continuous second derivatives on $[0,1]$.

We restrict the smoothness conditions here on the interval $[0,1]^p$ because we will consider estimation of m_1, ...,m_p only on compact intervals $[a_1, b_1], \ldots, [a_p, b_p]$, respectively. For simplicity of notation we put $a_1 = \ldots = a_p = 0$ and $b_1 = \ldots = b_p = 1$.

We use the following additional notation. The one and two dimensional marginals of f^t are denoted by $f^t_j(x_j)$ or $f^t_{j,k}(x_j, x_k)$, respectively. Furthermore, we put $f(x) = \sum_{t=1}^T f^t(x)$, $f_j(x_j) = \sum_{t=1}^T f^t_j(x_j)$ and $f_{j,k}(x_j, x_k) = \sum_{t=1}^T f^t_{j,k}(x_j, x_k)$. We denote by N the number of covariables X^{it} in $[0,1]^p$ for $1 \le i \le n,\ 1 \le t \le T$. The number N_t is the number for fixed t. By the law of large numbers we have that in probability $N/n \to d, N_t/n \to d_t$ where $d_t = P(X^{it} \in [0,1]^p)$ and $d = \sum_{t=1}^T d_t$.

In our smoothing estimates we use bandwidths $h(1), \ldots, h(p)$ that are of order $n^{-1/5}$. This is motivated by our smoothness conditions. Furthermore, we assume the following conditions for the smoothing kernel K.

(A5) The bandwidths $h(1), \ldots, h(p)$ fulfill $n^{1/5} h(j) \to c_j$ for some constants $c_1, \ldots, c_p$. The kernel K is non-negative, Lipschitz continuous, symmetric around 0 and has compact support ($[-1,1]$, say).

We now define our estimates $\widehat{m}_1, \ldots, \widehat{m}_p$ of $m_1, \ldots, m_p$. For the performance of our estimates we need that the kernel K integrates to one over the interval $[0,1]$. This is achieved by the following modification of a convolution kernel

$$K_h(u,v) \;=\; \begin{cases} \frac{K_h(u-v)}{\int_0^1 K_h(w-v)\ dw} & \text{if } u, v \in [0,1], \\ 0 & \text{else} \end{cases} .$$

Our estimators $\widehat{m}_1, \ldots, \widehat{m}_p, \widehat{\eta}_1, \ldots, \widehat{\eta}_T$ are defined as minimizer of a smoothed least squares criterion. They minimize the following criterion function

$$\sum_{i=1}^n \sum_{t=1}^T \int [Y^{it} - \sum_{j=1}^p \widehat{m}_j(u_j) - \widehat{\eta}_t]^2 \\ K_{h(1)}(u_1, X_1^{it}) \cdot \ldots \cdot K_{h(p)}(u_p, X_p^{it}) du_1 \ldots du_p \tag{2.2}$$

under the constraints

$$\sum_{t=1}^T \sum_{i=1}^n \widehat{m}_j(X_j^{it}) = 0$$

for $j = 1, ..., p$. The criterion function (2.2) is a smoothed version of a sum of squared residuals. Smoothing is done with respect to the covariables X_1^{11}, ..., X_p^{nT}, but not with respect to t. Smoothing with respect to t could be easily included by minimizing

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\int\sum_{s=1}^{T}[Y^{it}-\sum_{j=1}^{p}\widehat{m}_j(u_j)-\widehat{\eta}_s]^2$$
$$\pi_{st}^g K_{h(1)}(u_1, X_1^{it})\cdot\ldots\cdot K_{h(p)}(u_p, X_p^{it})du_1\ldots du_p,$$

where π_{st}^g are smoothing weights depending on a "bandwidth" g and fulfilling $\sum_{s=1}^{T}\pi_{st}^g = 1$ for $t = 1, ..., T$. In particular, such an approach could be used in an asymptotic setting with $T \to \infty$. We do not follow this approach and in the following we only will discuss the partially smoothed version (2.2).

Definition (2.2) is very similar to the definition of a backfitting estimate in an additive model in Mammen et al. (1999). As there, one gets by simple arguments using derivatives of the criterion function (2.2) that the estimates fulfill for $0 \le x_j \le 1$

$$\begin{aligned}(2.3)\qquad \widehat{m}_j(x_j) &= \widetilde{m}_j(x_j)-\sum_{t=1}^{T}\frac{N_t}{N}\widehat{\eta}_t\frac{\widehat{f}_j^t(x_j)}{\widehat{f}_j(x_j)}\\ &\quad -\sum_{l\neq j}\int\widehat{m}_l(x_l)\frac{\widehat{f}_{jl}(x_j,x_l)}{\widehat{f}_j(x_j)}\,dx_l \quad (j=1,\ldots,p),\\ (2.4)\qquad \widehat{\eta}_t &= \widetilde{\eta}_t-\sum_{j=1}^{p}\int\widehat{m}_j(x_j)\widehat{f}_j^t(x_j)\,dx_j \quad (t=1,\ldots,T),\end{aligned}$$

where $\widetilde{m}_j, \widetilde{\eta}_t$ are the following "marginal" estimates

$$\begin{aligned}(2.5)\qquad \widetilde{m}_j(x_j) &= N^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{1}(X^{it}\in[0,1]^p)\\ &\qquad K_h(x_j, X_j^{it})Y^{it}/\widehat{f}_j(x_j),\\ \widetilde{\eta}_t &= N_t^{-1}\sum_{i=1}^{n}\mathbf{1}(X^{it}\in[0,1]^p)Y^{it}.\end{aligned}$$

The functions $\widehat{f}_j, \widehat{f}_{jk}, \widehat{f}_j^t$ are kernel density estimates based on the covariables:

$$\begin{aligned}\widehat{f}_j(x_j) &= N^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{1}(X^{it}\in[0,1]^p)K_{h(j)}(x_j, X_j^{it}),\\ \widehat{f}_{jk}(x_j,x_k) &= N^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{1}(X^{it}\in[0,1]^p)\\ &\qquad K_{h(j)}(x_j, X_j^{it})\;K_{h(k)}(x_k, X_k^{it}),\\ \widehat{f}_j^t(x_j) &= N_t^{-1}\sum_{i=1}^{n}\mathbf{1}(X^{it}\in[0,1]^p)K_{h(j)}(x_j, X_j^{it}).\end{aligned}$$

The following theorem describes the asymptotic behavior of $\widehat{m}_1, \ldots, \widehat{m}_p$. A uniform approximation of the estimates is stated. This immediately implies asymptotic normality of the estimates (see Corollary 1). Furthermore, it can be used to construct uniform confidence bands. We will not state such a result here. It could be done along the lines of the treatment of standard classical nonparametric regression models.

Theorem 1 *Assume (A1) - (A5). Put* $\gamma_{n,j} = -c_j^2 \frac{1}{2} \int m_j''(x_j) f_j(x_j) dx_j \int u^2 K(u) du$ *and define* $\beta^t, \beta_1, \ldots, \beta_p$ *as the minimizers of*

$$\int_{[0,1]^p} \sum_{t=1}^{T} [\beta(x,t) - \beta^t - \beta_1(x_1) - \ldots - \beta_p(x_p)]^2 \; f(x,t) \; dx$$

where the minimization runs over constants $\beta^1, ..., \beta^T$ *and functions* $\beta_1, \ldots, \beta_p : [0,1] \to \mathbf{R}$ *with* $\int \beta_j(x_j) f_j(x_j) \; dx_j = 0$ *and where* $\beta(x,t)$ *is defined as*

$$\begin{aligned} \beta(x,t) \;\; &= \;\; \sum_{j=1}^{p} c_j^2 \, [\, \frac{\partial f^t}{\partial x_j}(x) \; m_j'(x_j) f^t(x)^{-1} \\ &\quad + \frac{1}{2} m_j''(x_j) \,] \int u^2 K(u) \; du \, . \end{aligned}$$

Then the following uniform expansion holds

$$\sup_{x_j \in [h_j, 1-h_j]} \left| \widehat{m}_j(x_j) - \left[\widetilde{m}_j^A(x_j) - \gamma_{n,j} + n^{-2/5} \beta_j(x_j) \right] \right| = o_p(n^{-2/5}) \, ,$$

$$\widehat{\eta}_t = \eta_t - \sum_{s=1}^{T} \frac{d_s}{d} \eta_s + n^{-2/5} (\beta^t - \sum_{s=1}^{T} \frac{d_s}{d} \beta^s) + o_p(n^{-2/5}).$$

Here

$$\widetilde{m}_j^A(x_j) = N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X_j^{it}) \varepsilon^{it} / \widehat{f}_j(x_j) \; . \tag{2.6}$$

Note that $\widetilde{m}_j^A$ is defined as $\widetilde{m}_j$, but with Y^{it} replaced by ε^{it}. It is easy to check under our conditions that $n^{2/5} \widetilde{m}_j^A$ has an asymptotic zero mean normal limit. Furthermore, $\beta_j(x_j)$ is a deterministic function. This shows that $\widehat{m}_j$ has the following asymptotic limit.

Corollary 1 *Under conditions (A1) - (A5) for* x_1, ..., $x_p \in (0,1)$ *the following convergence in distribution holds*

$$n^{2/5} \begin{bmatrix} \widehat{m}_1(x_1) - m_1(x_1) \\ \vdots \\ \widehat{m}_p(x_p) - m_p(x_p) \end{bmatrix} \to N \left(\begin{bmatrix} \beta_1(x_1) - \gamma_{n,1} \\ \vdots \\ \beta_p(x_p) - \gamma_{n,p} \end{bmatrix}, \begin{bmatrix} v_1(x_1) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & v_p(x_p) \end{bmatrix} \right)$$

where $v_j(x_j) = \int K(u)^2 \, du \; \sigma^2 [f_j(x_j) c_j]^{-1}$.

According to Theorem 1 in general the parameter η_t is only estimated with an order of $n^{-2/5}$. This is caused by a bias of the estimate that is of this order. We conjecture that the stochastic part of $\widehat{\eta}_t$ is of parametric order $n^{-1/2}$. As in other semiparametric problems use of smoothed likelihood functions leads to an estimate of the parametric component that does not achieve an $n^{-1/2}$ rate. Estimates of the parametric component with $n^{-1/2}$ rate can be achieved by application of a two step procedure where an unsmoothed likelihood is used for the parametric component and a smoothed likelihood for the nonparametric part.

We now shortly comment on the numerical calculation of our estimate. Equations (2.3) and (2.4) suggest an iterative calculation of the estimates. Application of (2.3) for $j = 1, \ldots, p$ can be used for an update of $\widehat{m}_j$. In each application one plugs the current values of $\widehat{m}_l$ ($l \neq j$) and of $\widehat{\eta}_t$ ($t = 1, \ldots, T$) into the right hand side of (2.3). Afterwards one applies (2.4) for updates of $\widehat{\eta}_t$ for $t = 1, \ldots, T$. Again this is done by using the actual values of $\widehat{m}_j$ ($j = 1, \ldots, p$) on the right hand side of the equation. Let us call these iterative values $\widehat{m}_j^{[a]}$ and $\widehat{\eta}_t^{[a]}$ where a is the number of cycles of the algorithm that have been applied. The starting values are denoted by $\widehat{m}_j^{[0]}$ and $\widehat{\eta}_t^{[0]}$. Convergence of this algorithm is stated in the next theorem.

Theorem 2 *Suppose that (A1) - (A5) hold. Then there exists constants $c > 0$ and $0 < \gamma < 1$ such that with probability tending to one*

$$\int [\widehat{m}_j^{[a]}(x_j) - \widehat{m}_j(x_j)]^2 p_j(x_j)\ dx_j \leq c\gamma^{2a} R \qquad \textit{for} \quad j = 1, \ldots, p.$$

and

$$\left| \widehat{\eta}_t^{[a]} - \widehat{\eta}_t \right|^2 \leq c\gamma^a R \qquad \textit{for} \quad t = 1, \ldots, T,$$

where

$$R = 1 + \sum_{j=1}^{p} \int \widehat{m}_j^{[0]}(x_j)^2 p_j(x_j)\ dx_j + \sum_{t=1}^{T} (\widehat{\eta}_t^{[0]})^2 \ .$$

Our estimates only use response variables in the smoothing if the corresponding covariables lie in $[0,1]^p$. All other responses are thrown away and are not used in the construction of the estimate. In particular if p is large a relatively large portion of the observations may be lost. Clearly, one can use an arbitrarily large but fixed compact set to weaken this effect. Asymptotically, this could be more carefully analyzed by using sets that, depending on the sample size, converge to the whole space. Such an analysis would be technically very involved because one has to take care on the tail behavior of the covariables and of the regression functions. Another way out could be based on semiparametric modifications of the estimate. This could be done by fitting a constant value for $\widehat{m}_j(x_j)$ with x_j outside $[0,1]$ (or more generally to fit a parametric shape outside $[0,1]$). Then no observations are lost in the calculation of the estimate. We do not pursue the discussion of this estimate here.

2.2 Asymptotics with $T \to \infty, n \to \infty$.

The mathematical arguments of the last section make essential use of the assumption that T is fixed. In some applications of panels of time series this assumption may not be justified. We now discuss asymptotics for the estimates $\widehat{m}_j$ of the last section for the case that $T \to \infty$. Again, the estimates are defined as minimizers of (2.2) and they fulfill equations (2.3), (2.4) by the same arguments as in Section 2.1. By plugging (2.4) into (2.3) we get

$$\begin{aligned} \widehat{m}_j(x_j) &= \bar{m}_j(x_j) - \sum_{l \neq j} \int \widehat{m}_l(x_l) \frac{\widehat{f}_{jl}(x_j, x_l)}{\widehat{f}_j(x_j)} \, dx_l \\ &\quad + \sum_{l=1}^{p} \sum_{t=1}^{T} \frac{N_t}{N} \int \widehat{m}_l(u_l) \frac{\widehat{f}_l^t(u_l) \widehat{f}_j^t(x_j)}{\widehat{f}_j(x_j)} \, du_l \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} \bar{m}_j(x_j) &= N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X_j^{it}) [Y^{it} - \bar{Y}^{\cdot t}] / \widehat{f}_j(x_j), \\ \widehat{Y}^{\cdot t} &= N_t^{-1} \sum_{i=1}^{n} \mathbf{1}(X^{it} \in [0,1]^p) Y^{it}. \end{aligned}$$

With $\widehat{m}(x) = (\widehat{m}_1, (x_1), \ldots, \widehat{m}_p(x_p))^T$, $\bar{m}(x) = (\bar{m}_1(x_1), \ldots, \bar{m}_p(x_p))^T$ this can be rewritten as

$$\widehat{m}(x) = \bar{m}(x) + \int \widehat{H}(x, y) \, \widehat{m}(y) \, dy \tag{2.8}$$

where $\widehat{H}(x, y)$ is a $p \times p$ matrix with off diagonal entries

$$\widehat{H}_{jl}(x, y) = \frac{\widehat{f}_{jl}(x_j, y_l)}{\widehat{f}_j(x_j)} + \sum_{t=1}^{T} \frac{N_t \widehat{f}_l^t(y_l) \widehat{f}_j^t(x_j)}{N \widehat{f}_j(x_j)}$$

and diagonal elements

$$\widehat{H}_{jj}(x, y) = \sum_{t=1}^{T} \frac{N_t \widehat{f}_j^t(y_j) \widehat{f}_j^t(x_j)}{N \widehat{f}_j(x_j)} \, .$$

For (2.8) we use the following short hand notation

$$\widehat{m} = \bar{m} + \widehat{H} \widehat{m} \, . \tag{2.9}$$

The operator $\widehat{H}$ is related to the operator $\widehat{S}$ that has appeared in the proof of Theorems 1 and 2. The operator $\widehat{S}$ consists in an iterative application of orthogonal projections (or more precisely of orthogonal projections that depend on n but converge to fixed

orthogonal projections). This was the central argument to show that the operator norm of $\widehat{S}$ is strictly smaller than 1. If the operator norm of $\widehat{H}$ would be strictly smaller than 1 then iterative application of (2.9) would give

$$\widehat{m} = \sum_{r=0}^{\infty} \widehat{H}^r \bar{m}$$

and we could treat asymptotics for $\widehat{m}$ as in the last section. Unfortunately we are not aware of an argument that implies that the norm of $\widehat{H}$ is bounded away from 1 (with probability tending to one for n large enough). However, there exists a simple idea in the theory of integral equations how the integral equation can be modified to achieve an integral operator norm smaller than one, see Luchka (1965). This idea has also been used in Linton and Mammen (2005). It will be our main tool to carry over the results of the last section to the case where $T \to \infty$. For a detailed discussion see the proof of Theorem 3. We will assume that the operator $\widehat{H}$ converges to a deterministic operator H, see (A4′). Let us denote the eigenvalues of H by $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \ldots$. Put $r = \max\{j : |\lambda_j| < 1\}$ and denote by π the projection onto the space L of eigenfunctions of $\lambda_1, \ldots, \lambda_{r-1}$ (in the space $L_2([0,1]^p)$). Then we can write

$$\widehat{m} = \bar{m} + \widehat{H}(I - \pi)\widehat{m} + \widehat{H}\pi\widehat{m}\,.$$

This gives

$$\widehat{m} = (I - \widehat{H}\pi)^{-1}\bar{m} + \widehat{R}\widehat{m}\,, \tag{2.10}$$

where $\widehat{R} = (I - \widehat{H}\pi)^{-1}\widehat{H}(I - \pi)$.

In the proof of the following theorem we will argue that $\widehat{H} \approx H$ implies that

$$\widehat{R} \approx R = (I - H\pi)^{-1}H(I - \pi).$$

We will then use that R has operator norm (largest absolute eigenvalue) strictly less than 1. To see this statement note that $(I - H\pi)^{-1}$ coincides on L with I, that $H(I - \pi)$ maps all functions into L and that $H(I - \pi)$ has operator norm strictly less than 1 by construction of the space L. Iterative application of (2.10) now gives

$$\widehat{m} = \sum_{r=0}^{\infty} \widehat{R}^r (I - \widehat{H}\pi)^{-1}\bar{m} \tag{2.11}$$

with probability tending to one. This expansion is the main tool in the proof of the following theorem.

We now state the assumptions of the following theorem. Instead of (A2) we use

(A2') It holds that $n, T \to \infty$.

Instead of (A4) we will assume:

(A4') There exist $p \times p$ matrices $H(x,y)$ for $x, y \in [0,1]^p$ with

$$\sup_{x,y} |\widehat{H}(x,y) - H(x,y)| = o_p((\log n)^{-1/2}).$$

The matrices $H(x,y)$ are Lipschitz continuous for $x, y \in [0,1]^p$.

We are now in the position to state an expansion for $\widehat{m}$.

Theorem 3 *Assume (A1), (A2'), (A3), (A4') and (A5). Suppose additionally that there exist bounded functions $\mu^r_{n,1}, \ldots, \mu^r_{n,p}$ $(r = 1, 2)$ such that for $1 \leq j \leq p$*

$$\sup_{x_j \in [0,1]} \left| \bar{m}^B_j(x_j) - \mu^1_{n,j}(x_j) - n^{-2/5}\mu^2_{n,j}(x_j) \right| = o_p(n^{-2/5}), \tag{2.12}$$

where

$$\begin{aligned} \bar{m}^B_j(x_j) &= N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X^{it}_j) \sum_{l=1}^{p} \left\{ m_l(X^{it}_l) \right. \\ &\quad \left. - \frac{1}{N_t} \sum_{r=1}^{n} \mathbf{1}(X^{rt} \in [0,1]^p) m_l(X^{rt}_l) \right\} / \widehat{f}_j(x_j). \end{aligned}$$

Then the following uniform expansion holds:

$$\begin{aligned} \sup_{x_j \in [0,1]} | \, \widehat{m}_j(x_j) - \{ \, \bar{m}^A_j(x_j) + \sum_{r=0}^{\infty} \left[\widehat{R}^r (I - \widehat{H}\pi)^{-1} \mu^1_n \right]_j (x_j) \\ + n^{-2/5} \left[R^r (I - H\pi)^{-1} \mu^2_n \right]_j (x_j) \, \} \, | = o_p(n^{-2/5}), \end{aligned}$$

where

$$\bar{m}^A_j(x_j) = N^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X^{it}_j) [\varepsilon^{it} - \bar{\varepsilon}^{\cdot t}] / \widehat{f}_j(x_j)$$

and $\mu^k_n = (\mu^k_{n,1}, \ldots, \mu^k_{n,p})^T, k = 1, 2 \ldots$. Furthermore $[\quad]_j$ denotes the j-element.

Note that $\bar{m} = \bar{m}^A + \bar{m}^B$, by definition with (2.11) we get

$$\widehat{m} = \widehat{m}^A + \widehat{m}^B$$

with

$$\widehat{m}^A = \sum_{r=0}^{\infty} \widehat{R}^r (I - \widehat{H}\pi)^{-1} \bar{m}^A$$

and

$$\widehat{m}^B = \sum_{r=0}^{\infty} \widehat{R}^r (I - \widehat{H}\pi)^{-1} \bar{m}^B.$$

Under the assumptions of the theorem $\widehat{m}^B - m$ is asymptotically equivalent to the bias of $\widehat{m}$, and $\widehat{m}^A$ represents the stochastic part up to terms that are asymptotically negligible. The proof of the theorem is based on two facts. Firstly, $\widehat{m}^A$ is asymptotically equivalent to the first term of its asymptotic expansion, i.e. to $\bar{m}^A$. Secondly, in the expansion of $\widehat{m}^B$ the random operator $\widehat{R}^r(I-\widehat{H}\pi)^{-1}$ can be (partially) replaced by the deterministic operator $R^r(I-H\pi)^{-1}$ and the random function $\bar{m}^B$ can be replaced by the deterministic functions μ_n^1 and μ_n^2 . The bias cannot be so nicely interpreted as in Theorem 1. An interpretation of $\widehat{m}^B$ as least square additive fit to a function that is nonadditive is in general not available. This requires additional assumptions on the distribution of the covariables. In particular, we cannot assume that the X^{it} are stationary. This would exclude the case that X^{it} is a lagged observation (Lagged observations cannot be stationary in our model, unless the parameters η_t are constant). Also a check of the assumption (2.12) would require additional assumptions on the distribution of the covariables. We do not want to specify such assumptions here. Asymptotic normality of $\widehat{m}_j$ can be easily followed from Theorem 3.

2.3 Inclusion of individual effects into the model

In a more general model individual effects are included. The model then becomes

$$Y^{it} = \sum_{j=1}^{p} m_j(X_j^{it}) + \eta_t + \lambda_i + \varepsilon^{it} \qquad (i = 1,\dots,n; \quad t = 1,\dots,T). \tag{2.13}$$

Here $\lambda_1,\dots,\lambda_n$ are constants that are unknown. For identifiability, we assume $\sum_{i=1}^n N^i\lambda_i = 0$, where N^i is the number of covariables X^{it} that lie in $[0,1]^p$ for fixed i and for $t = 1,\dots,T$. . In a first attempt one can again fit a model (2.13) by minimizing a smoothed least squares criterion. The estimates $\widehat{m}_1,\dots,\widehat{m}_p,\widehat{\eta}_1,\dots,\widehat{\eta}_T,\widehat{\lambda}_1,\dots,\widehat{\lambda}_n$ now are defined as minimizers of

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\int [Y^{it} - \sum_{j=1}^{p}\widehat{m}_j(u_j) - \widehat{\eta}_t - \widehat{\lambda}_i]^2 K_{h(1)}(u_1, X_1^{it})\cdot\dots \cdot K_{h(p)}(u_p, X_p^{it})\, du_1\cdot\dots\cdot du_p \tag{2.14}$$

under the constraints

$$\sum_{i=1}^{n}\frac{N^i}{N}\widehat{\lambda}_i = 0,$$

$$\sum_{i=1}^{T}\sum_{i=1}^{n}\widehat{m}_j(X_j^{it}) = 0.$$

By taking derivatives one gets that the estimates fulfill linear equations that are similar to (2.3), (2.4). By plugging the formulas of $\widehat{\eta}_t$ and $\widehat{\lambda}_i$ into the expression for $\widehat{m}_j$ one

can show that $\widehat{m} = (\widehat{m}_1, \ldots, \widehat{m}_p)^T$ fulfills the following integral equation (compare (2.8))

$$\widehat{m}(x) = m^*(x) + \int \widehat{H}(x,y)\widehat{m}(y)\ dy$$

where

$$\begin{aligned} m_j^*(x_j) &= N^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X_j^{it}) \left[Y^{it} - \bar{Y}^{\cdot t} - \bar{Y}^{i\cdot} + \bar{Y}\right], \\ \bar{Y}^{i\cdot} &= (N^i)^{-1}\sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) Y^{it}, \end{aligned}$$

$$\begin{aligned} \widehat{H}_{jl}(x,y) &= \frac{\widehat{f}_{jl}(x_j, y_l)}{\widehat{f}_j(x_j)} + \sum_{t=1}^{T} \frac{N^t \widehat{f}_l^t(y_l)\widehat{f}_j^t(x_j)}{N\widehat{f}_j(x_j)} + \sum_{i=1}^{n} \frac{N^i \widetilde{f}_l^i(y_l)\widetilde{f}_j^i(x_j)}{N\widehat{f}_j(x_j)} \quad \text{for } j \neq l, \\ \widehat{H}_{jj}(x,y) &= \sum_{t=1}^{T} \frac{N^t \widehat{f}_l^t(y_l)\widehat{f}_j^t(x_j)}{N\widehat{f}_j(x_j)} + \sum_{i=1}^{n} \frac{N^i \widetilde{f}_l^i(y_l)\widetilde{f}_j^i(x_j)}{N\widehat{f}_j(x_j)}, \\ \widetilde{f}_j^i(x_j) &= \frac{1}{N^i}\sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p)\ K_{h(j)}(X_j^{it} - x_j). \end{aligned}$$

Under appropriate conditions one can proceed for the discussion of $\widehat{m}_j(x_j)$ as in the last section. For the "stochastic"term of $\widehat{m}_j(x_j)$ one now gets

$$N^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X_j^{it})[\varepsilon^{it} - \bar{\varepsilon}^{\cdot t} - \bar{\varepsilon}^{i\cdot} + \bar{\varepsilon}]$$

with an obvious definition of $\bar{\varepsilon}^i$.

Compared with $\bar{m}_j^A$ (see Theorem 3) this now contains the additional term $\bar{\varepsilon}^{i\cdot}$. Clearly, $\bar{\varepsilon}^{i\cdot}$ is not independent of X_j^{it} in case that the covariables contain lagged observations and for T small it is not asymptotically negligible (as is always the case for $\bar{\varepsilon}$). A more careful analysis shows that $\bar{m}_j^A$ only has asymptotic mean 0 if T grows fast enough and appropriate mixing conditions apply (or if all covariables are external variables).

3 Local linear estimates.

The local constant estimates $\widehat{m}_j$ have a complicated bias, see Theorem 1 and Corollary. The bias of $\widehat{m}_j$ depends on the shape of the other regression functions m_l $(l \neq j)$. This complicates the statistical inference based on $\widehat{m}_j$. This disadvantage is known from backfitting estimates of additive models. There it can be avoided by using local linear instead of local constant smoothing. The same holds true for panels of time

series. We will discuss this in this section. We will only do it for the setting of section 2.1 (i.e. T constant, no individual effects). Local linear estimates

$$\widehat{m}_1, \ldots, \widehat{m}_p, \widehat{m}^1, \ldots, \widehat{m}^p, \widehat{\eta}_1, \ldots, \widehat{\eta}_T$$

are defined as minimizers of

$$\sum_{i=1}^{n}\sum_{t=1}^{n}\int [Y^{it} - \sum_{j=1}^{p}\widehat{m}_j(u_j) - \frac{X_j^{it}-u_j}{h(j)}\widehat{m}^j(u_j) - \widehat{\eta}_t]^2 \tag{3.1}$$
$$K_{h(1)}(u_1, X_1^{it}) \cdot \ldots \cdot K_{h(p)}(u_p, X_p^{it})\ du_1 \cdot \ldots \cdot du_p.$$

Again, $\widehat{m}_1, \ldots, \widehat{m}_p$ are estimates of $m_1, \ldots, m_p$. The estimates $\widehat{m}^1, \ldots, \widehat{m}^p$ fit the local slope and can be used as estimates of the derivatives $m_1', \ldots, m_p'$. By differentiation of the left hand side of (3.1) (w.r.t. $\widehat{m}_j(u_j), \widehat{m}^j(u_j)$ and $\widehat{\eta}_t$) one gets the following linear equations for the estimates

$$\widehat{M}_j(x_j)\begin{pmatrix}\widehat{m}_j(x_j)\\ \widehat{m}^j(x_j)\end{pmatrix} = \widehat{M}_j(x_j)\begin{pmatrix}\widetilde{m}_j(x_j)\\ \widetilde{m}^j(x_j)\end{pmatrix} - \widehat{m}_{0,j}\begin{pmatrix}\widehat{V}_{0,0}^j(x_j)\\ \widehat{V}_{j,0}^j(x_j)\end{pmatrix}$$
$$-\sum_{l\neq j}\int \widehat{S}_{l,j}(x_{l,j}(x_l, x_j)\begin{pmatrix}\widetilde{m}_l(x_l)\\ \widetilde{m}^l(x_l)\end{pmatrix}\ dx_l - \sum_{t=1}^{T}\widehat{\eta}_t\begin{pmatrix}\widehat{V}_{0,0}^{j,t}(x_j)\\ \widehat{V}_{j,0}^{j,t}(x_j)\end{pmatrix},$$
$$\widehat{\eta}_t = \widetilde{\eta}_t - \frac{N}{N_t}\sum_{j=1}^{p}\widehat{V}_{0,0}^{j,t}(u_j)\widehat{m}_j(u_j)\ du_j - \frac{N}{N_t}\sum_{j=1}^{p}\int \widehat{V}_{0,1}^{j,t}(u_j)\widehat{m}^j(u_j)\ du_j,$$

where

$$\widehat{M}_j(x_j) = \begin{pmatrix}\widehat{V}_{0,0}^j(x_j) & \widehat{V}_{j,0}^j(x_j)\\ \widehat{V}_{j,0}^j(x_j) & \widehat{V}_{j,j}^j(x_j)\end{pmatrix}$$
$$= \frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{1}(X^{it}\in[0,1]^p)K_{h(j)}(x_j, X_j^{it})$$
$$\begin{pmatrix}1 & h(j)^{-1}[X_j^{it}-x_j]\\ h(j)^{-1}[X_j^{it}-x_j] & h(j)^{-2}[X_j^{it}-x_j]^2\end{pmatrix},$$
$$\widehat{S}_{l,j}(x_l, x_j) = \frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{1}(X^{i}\in[0,1]^p)K_{h(j)}(x_j, X_j^{it})$$
$$K_{h(l)}(x_l, X_l^{it})\begin{pmatrix}1 & h(l)^{-1}[X_l^{it}-x_l]\\ h(j)^{-1}[X_j^{it}-x_j] & h(j)^{-1}h(l)^{-1}[X_l^{it}-x_l][X_j^{it}-x]\end{pmatrix}$$
$$\widehat{V}_{0,0}^{j,t}(x_j) = \frac{1}{N}\sum_{i=1}^{n}\mathbf{1}(X^{it}\in[0,1]^p)K_{h(j)}(x_j, X_j^{it}),$$
$$\widehat{V}_{j,0}^{j,t}(x_j) = \frac{1}{N}\sum_{i=1}^{n}\mathbf{1}(X^{it}\in[0,1]^p)K_{h(j)}(x_j, X_j^{it})h(j)^{-1}(X_j^{it}-x_j).$$

We now state asymptotic normality of the local linear estimate.

Theorem 4 *Under conditions (A1) - (A5) the following convergence in distribution holds*

$$n^{2/5}\begin{bmatrix} \widehat{m}_1(x_1) - m_1(x_1) + \nu_{n,1} \\ \vdots \\ \widehat{m}_p(x_p) - m_p(x_p) + \nu_{n,p} \end{bmatrix} \to N\left(\begin{bmatrix} \delta_1(x_1) \\ \vdots \\ \delta_p(x) \end{bmatrix}, \begin{bmatrix} v_1(x_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & v_p(x_p) \end{bmatrix}\right),$$

where $v_j(x_j)$ is defined as in Corollary 1 and where

$$\delta_j(x_j) \;=\; c_j^2 \frac{1}{2}\int u^2 K(u)\, du[m_j''(x_j) - \int m_j''(x_j) p_j(x_j)\, dx_j],$$

and $\nu_{n,j} = \int m_j(x_j) K_h(x_j, u_j) f_j(u_j)\, du_j dx_j$. Furthermore it holds that

$$\widehat{\eta}_t = \eta_t - \sum_{s=1}^{T} \frac{d_s}{d}\eta_s + o_p(n^{-2/5}).$$

Note that the nonparametric bias term in the estimation of η_t does not depend on t. This can be explained by the following heuristics. For a full dimensional local linear fit of the regression function $m(x,t) = \eta_t + m_1(x_1) + ... + m_p(x_p)$ one gets a bias term for the estimation of an additive component m_j that only depends on the second derivative of m_j but that does not depend on the density f_t. In particular, it does not depend on t. Therefore $\widehat{\eta}_t$ has only a nonparametric bias that does not depend on t and that could be removed by norming.

4 Simulations and a real data application.

In this section we carry out a simulation study and use a real data set to see how one of the proposed estimation procedures behaves. The estimation algorithm in equations (2.3)-(2.5) has been implemented. In the implementation however, we do not use the interval $[0,1]$, but take $N = nT$ and $N_t = T$. The convergence criterion we use in the simulations is: if for all $j = 1, ..., p$

$$\frac{\sum[\widehat{m}_j^{(k+1)}(x_j) - \widehat{m}_j^{(k)}(x_j)]^2}{\sum[\widehat{m}_j^{(k)}(x_j)]^2 + 0.0001} < 0.0001 \tag{4.1}$$

then stop. Here k indicates the number of iterations performed. This convergence criterion is taken from Nielsen and Sperlich (2005). The Gaussian kernel has been used in the nonparametric estimate of $\widetilde{m}(x_j)$ and in the density estimates. Using a compact kernel (cf.A5) produces similar results. The bandwidths used in these estimations are the "rule-of-thumb" bandwidth selector, see e.g. Härdle (1990) page 91. This bandwidth choice may not be optimal, especially for dependent data, thus another choice of bandwidth may improve the final estimates, as will be seen in our last simulation example.

4.1 Simulation study.

The first simulations are similar to the set-up in Nielsen and Sperlich (2005), although there the authors only used cross-section data. We consider the model

$$Y^{it} = \sum_{j=1}^{p} m_j(X_j^{it}) + \eta_t + \epsilon^{it}, \tag{4.2}$$

with $m_j(x_j^{it}) = \sin(\pi x_j^{it})$ and $\epsilon \sim N(0,1)$. In the simulation, we first draw variables z_j^{it}, $j = 1, ..., p$, $t = 1, ..., T$, for i fixed, from $N(0,1)$ with given correlation ρ_{jk}, $j \neq k$ for each combination of covariables. That is, for each i, the variables z_j^{it} are correlated with z_k^{it}. This drawing of variables is repeated for all $i = 1, ..., n$. Then a transformation is made, $x_j^{it} = 2.5 \arctan(z_j^{it})/\pi$. Thus the variables are projected into the interval $[-1.25, 1.25]$.

In the first simulation, 100 realizations with $n = 100$, $T = 10$, $p = 4$, $\eta_t = 0$ for all t and $\rho_{jk} = 0.1$ for all $j \neq k$ is carried out. One example of the estimated functions $\widehat{m}_j(x_j)$ and $\widetilde{m}_j(x_j)$ are given as dashed and thin solid lines, respectively, in figure 1. The thick solid lines are the true functions. The estimated functions are quite close to the true lines, observe also that the iterated functions, $\widehat{m}_j(x_j)$ is more accurate than the first estimate $\widetilde{m}_j(x_j)$, as expected. The standard deviations of $\widehat{m}_j(x_j)$ are estimated based on the 100 realizations in eleven different points, and $\pm$ one standard deviation are plotted in the figures as vertical dashed lines at each point. As expected, the standard deviations are quite large at both boundaries for all components, but are reasonably small at interior points. The average and the standard deviation of the estimated $\widehat{\eta}_t$ are given in table 1 upper part.

Using Corollary 1, we can construct confidence intervals of $\widehat{m}(x_j)$. Assuming that the bias of $\widehat{m}(x_j)$ is of negligible size compared with the variance, we set $\beta_1(x_1) - \gamma_{n,1}, ..., \beta_4(x_4) - \gamma_{n,4}$ in Corallary 1 equal to zero. Further we assume $c_j^h = 1$ for all j, this correspons to the coefficent in the "rule-of-thumb" bandwidth used (cf. Härdle (1990) page 91). Then the confidence intervals for all j components are equal to

$$[\widehat{m}_j(x_j) - \frac{z_{\alpha/2} v_j(x_j)^{1/2}}{n^{2/5}}, \widehat{m}_j(x_j) + \frac{z_{\alpha/2} v_j(x_j)^{1/2}}{n^{2/5}}], \tag{4.3}$$

where $z_{\alpha/2}$ is a quantile in the standard normal distribution and

$$v_j(x_j) = \int K(u)^2 \mathrm{d}u \; \sigma^2 f_j(x_j)^{-1}. \tag{4.4}$$

Observe that we also have to estimate σ. This is done by calculating the empirical standard deviation of the estimated ϵ^{it}, given by

$$\hat{\epsilon}^{it} = Y^{it} - \sum_{j=1}^{p} \widehat{m}_j(X_j^{it}) - \widehat{\eta}_t. \tag{4.5}$$

t	1	2	3	4	5	6	7	8	9	10
η_t	0	0	0	0	0	0	0	0	0	0
$\hat{\mathrm{E}}(\widehat{\eta}_t)$	-0.005	0.010	-0.023	-0.001	0.013	-0.006	0.028	-0.002	0.003	-0.017
$\hat{\mathrm{SD}}(\widehat{\eta}_t)$	0.099	0.102	0.100	0.102	0.094	0.105	0.111	0.119	0.097	0.095
η_t	-5	-4	-3	-2	-1	0	2	3	4	6
$\hat{\mathrm{E}}(\widehat{\eta}_t)$	-5.001	-4.013	-3.029	-1.989	-1.006	-0.017	2.004	3.003	3.985	5.983
$\hat{\mathrm{SD}}(\widehat{\eta}_t)$	0.122	0.130	0.109	0.133	0.116	0.095	0.107	0.117	0.124	0.119

Table 1: The estimated $\widehat{\eta}_t$-s for the first model

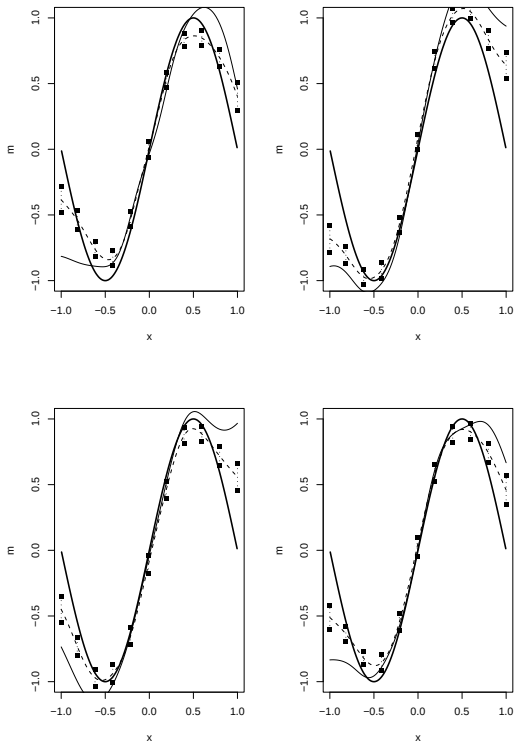

Figure 1: The estimated functions in the first simulation (m_1 and m_2 - upper plots, m_3 and m_4 - lower plots)

One realization of the 100 simulated realizations above, has been choosen, and the confidence intervals for $\alpha = 0.05$ have been plotted for all four components in figure 2. The confidence intervals in this figure are a bit wider than $\pm$ two standard deviations in figure 1.

The second simulation is identical to the first, except that the correlation, $\rho_{jk} = 0.7$ for all $j \neq k$. One example of the estimated functions $\widehat{m}_j(x_j)$ and $\widetilde{m}_j(x_j)$ are again given as dashed and thin solid lines, respectively, in figure 3. In this case the initial estimates $\widetilde{m}_j(x_j)$ do not work at all, whereas the iterated estimates work well. This is

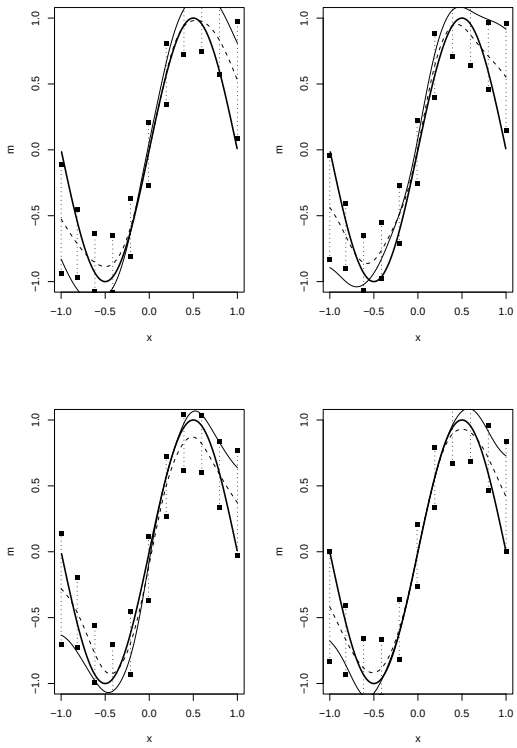

Figure 2: The estimated functions and confidence intervals (m_1 and m_2 - upper plots, m_3 and m_4 - lower plots)

expected because of the increased correlation ρ_{jk}. Again, $\pm$ one standard deviation of $\widehat{m}_j(x_j)$ is given in the figure as the dashed vertical lines in eleven points. The average and standard deviations of the estimated η_t were similar to those in table 1.

In the last simulation for this model non-zero values of η_t were allowed, as given in the lower part of table 1. Now ρ_{jk} is 0.1. One example of the estimated functions with standard errors are given in figure 4, and the average and standard deviations for the estimated η_t-s in the lower part of table 1.

Next, we look at a panel of time series. The time series lag dependence is taken from Fan and Yao (2003) page 356. Now,

$$X^{it} = \sum_{j=1}^{2} m_j(X_j^{it-j}) + \eta_t + \epsilon^{it}, \tag{4.6}$$

with

$$m_1(X_1^{it-1}) = 4X_{t-1}/(1 + 0.8X_{t-1}^2), \tag{4.7}$$

$$m_2(X_2^{it-2}) = \exp\{3(X_{t-2} - 2)\}/[1 + \exp\{3(X_{t-2} - 2)\}]. \tag{4.8}$$

In the first simulation, we ran 100 realizations with $n = 200$, $T = 10$, $\eta_t = 0$ for all t and ϵ uniform on $[-1, 1]$. An example of the estimated $\widehat{m}_j(x_j)$ functions are

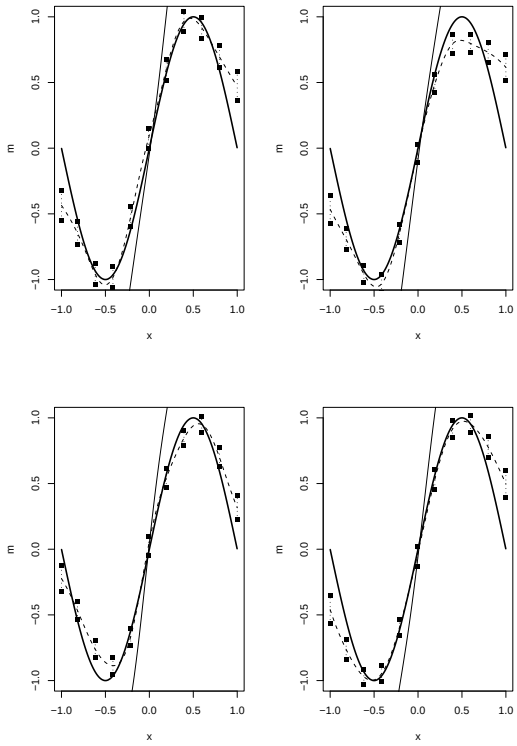

Figure 3: The estimated functions in the second simulation (m_1 and m_2 - upper plots, m_3 and m_4 - lower plots)

t	1	2	3	4	5	6	7	8	9	10
η_t	0	0	0	0	0	0	0	0	0	0
$\hat{\mathrm{E}}(\widehat{\eta}_t)$	0.017	0.018	0.024	0.024	0.018	0.026	0.022	0.024	0.020	0.019
$\hat{\mathrm{SD}}(\widehat{\eta}_t)$	0.047	0.040	0.045	0.046	0.041	0.043	0.047	0.044	0.038	0.040
η_t	-5	-4	-3	-2	-1	0	2	3	4	6
$\hat{\mathrm{E}}(\widehat{\eta}_t)$	-4.989	-3.994	-2.994	-1.996	-0.998	0.007	2.005	2.994	4.002	6.007
$\hat{\mathrm{SD}}(\widehat{\eta}_t)$	0.323	0.337	0.329	0.325	0.333	0.332	0.307	0.349	0.317	0.322

Table 2: The estimated $\widehat{\eta}_t$-s for the second model

given in figure 5 as the dashed lines, and they give a good approximation to the true functions, thick solid lines. Again, $\widetilde{m}_j(x_j)$ are given as thin solid lines. Note that the additive components can only be identified up to a constant, thus the true lines in this figure and figure 6 are the re-centered versions of the functions m_j, i.e. $m_j - \bar{m}_j$, plotted against X_j^{it-j}, where $\bar{m}_j$ indicates averaging. The standard deviations of the estimated components are plotted in the figure. The average of the estimated η_t and their standard deviations are given in the upper part of table 2.

The last simulation is identical to the above, except that η_t is now non-zero, as

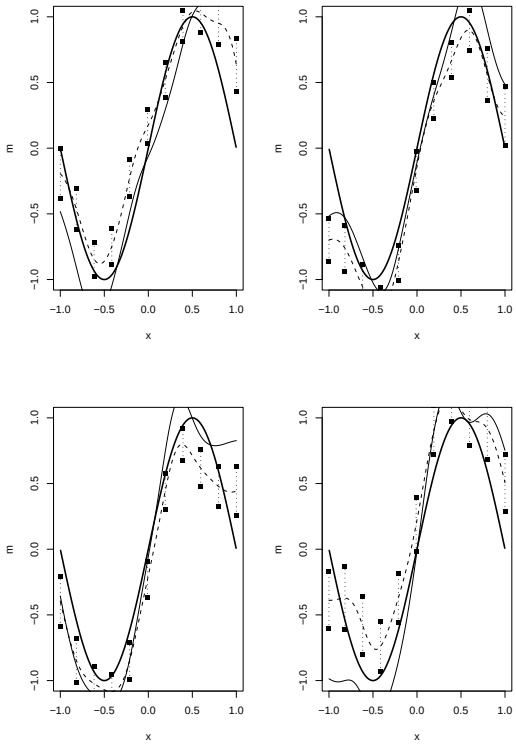

Figure 4: The estimated functions in the third simulation(m_1 and m_2 - upper plots, m_3 and m_4 - lower plots)

given in the lower part of table 2. This introduces a quite strong relationship between the time series, especially at the ends. In figure 6 an example of the estimates of the components is given. Now the estimated functions, the dashed lines, deteriorates in quality. Standard deviations of the estimated components are much larger and not included. The estimated η_t however, are quite good, with relatively small standard deviations, see the lower part of table 2. Since the estimates in figure 6 are quite wiggly, one might believe that a larger bandwidth in the estimates of $\widetilde{m_j}(x_j)$ may improve the estimates. And indeed that is the case. In figure 7 the estimated components are given, based on one realization as above, but now using a bandwidth 1.5 times larger.

4.2 Application to a real data set.

The estimation procedure is also used on a real data set. When using the algorithm on real data, one has to consider the assumptions on which it is based. Thus the following questions must be adressed:

1. Is Y explained by an additive model?

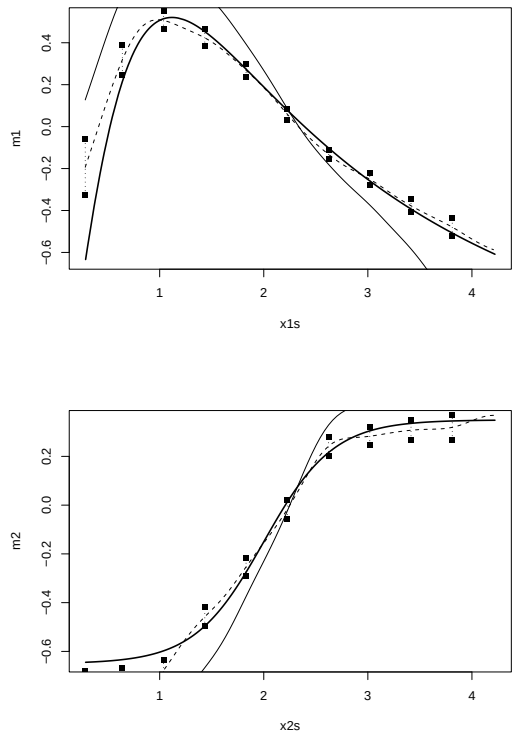

Figure 5: The estimated functions in the second model, first simulation

2. If the norming condition on the expectation of $m_j(x_j)$ is not fulfilled, what happens to the estimates?

3. What about the ergodicity type conditions for our covariates? Are the covariates stationary?

Since convergence of the algorithm rests on these assumptions, one may expect some problems when using the algorithm in practice. Also the number of observations may be an issue.

The real data set is from Baltagi et al. (2000). They estimate a dynamic demand model for cigarettes based on panel data from 46 American states over the period 1963-1992. The estimated equation is

$$\ln Y^{it} = \alpha + \beta_1 \ln Y_1^{it-1} + \beta_2 \ln X_2^{it} + \beta_3 \ln X_3^{it} + \beta_4 \ln X_4^{it} + u^{it}, \tag{4.9}$$

where $i = 1, ..., 46$ denotes the ith state and $t = 1, ..., 29$ denotes the tth year. Y^{it} is real per capita sales of cigarettes, X_2^{it} is the average retail price of a pack of cigarettes measured in real terms, X_3^{it} is real per capita disposable income and X_4^{it} denotes the minimum real price of cigarettes in any neighboring state. The disturbance term is specified as

$$u^{it} = \lambda_i + \eta_t + \epsilon^{it}, \tag{4.10}$$

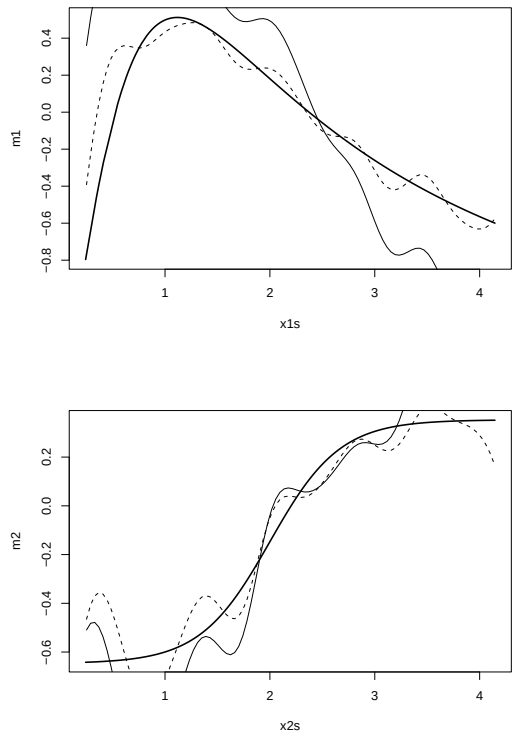

Figure 6: The estimated functions in the second model, last simulation

as before. Baltagi et al. (2000) use different techniques to estimate the unknown β-s and also in some cases λ_i and η_t. We assume that the model is

$$\ln Y^{it} = m_1(\ln Y_1^{it-1}) + m_2(\ln X_2^{it}) + m_3(\ln X_3^{it}) + m_4(\ln X_4^{it}) + \eta_t + \epsilon^{it}, \tag{4.11}$$

and we would like to estimate the unknown m_j and η_t for all j and t. The presence of a λ_i-term would require a large T as indicated in section 2.3.

Since both the response $\ln Y^{it}$ and all the log-transformed covariates exhibit a trend in t, they are not stationary. We therefore de-trend our observations as follows,

$$Y_d^{it} = \ln Y^{it} - g_Y(t) \tag{4.12}$$

where $g_Y(t)$ is a nonparamtric estimator based on all observations of $\ln Y^{it}$. Likewise the different log-transformed covariates were de-trended using

$$X_{jd}^{it} = \ln X_j^{it} - g_{X_j}(t). \tag{4.13}$$

The detrended Y_d^{it} observations are plotted against the different covariates, X_{jd}^{it} in figure 8. Clearly the first covariate (the lagged variable), in the upper plot, is linear. The other covariates do not display strong nonlinearities. But, nevertheless, these detrended observations are then used as input to the implemented algorithm.

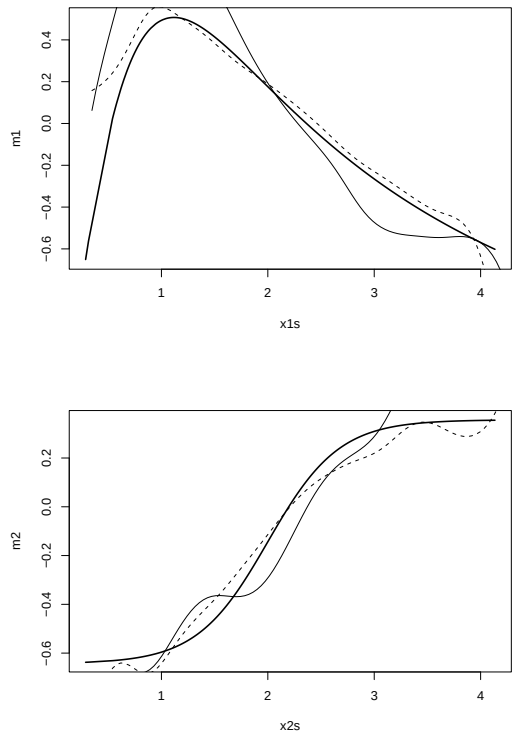

Figure 7: The estimated functions in the second model, last simulation, with a larger bandwidth

Since we have rather few observations and some nonstationarity present, we relax our convergence criteria.

The estimated functions $\widehat{m}_j$ are given in figure 9. Here the thick solid lines corresponds to the estimated linear model, ordinary least squares with time-effects, in Baltagi et al. (2000). The estimated $\widehat{m}_j(x_j)$ and $\widetilde{m}_j(x_j)$ are again given as dashed and thin solid lines, respectively. As could be expected from figure 8, the estimated components using our algorithm only displays weak nonlinearities and the $\widehat{m}_j(x_j)$ estimates are also quite close to the estimated linear model. The initial $\widetilde{m}_j(x_j)$ estimates seem rather unstable in this situation and it is encouraging that the $\widehat{m}_j(x_j)$ estimates do better in this respect.

The confidence intervals from equation (4.3), for $\alpha = 0.05$, have been plotted in eleven points in figure 9 for all components. As before, the intervals are quite small at interior points, but fairly wide at the boundaries.

The detrending of the responses, $\ln Y^{it}$, implies that the estimated η_t should be close to zero, since $g_Y(t)$ can be thought of as an adjusted η_t for all t. And the estimation gives that the interval $[\min(\widehat{\eta}_t), \max(\widehat{\eta}_t)]$ is equal to $[-0.033, 0.025]$.

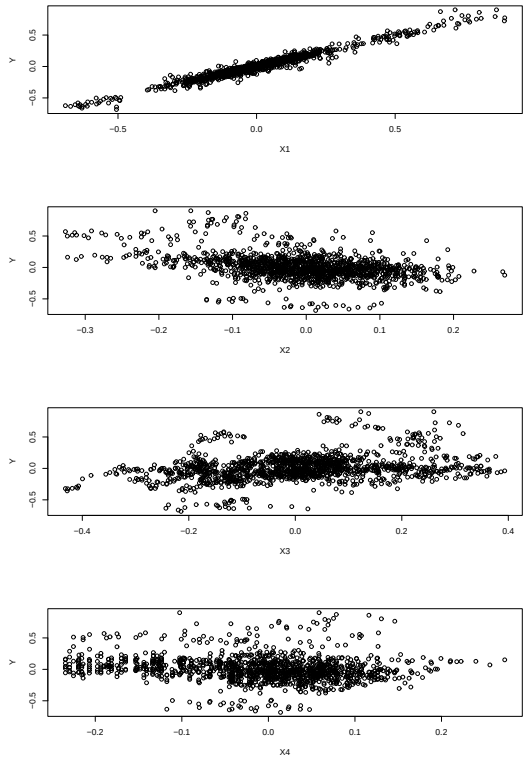

Figure 8: The detrended observations

5 Summary.

The greatest part of the literature examining panels of time series considers parametric models. In this paper we have introduced iterative nonparametric estimators for additive nonlinear panel data models. The estimators presented are based on the smoothed backfitting approach of Mammen et al. (1999) for estimating in additive regression models.

The asymptotic behaviour of the proposed estimators has been studied, mainly in the case where the number of individuals goes to infinity and the time periods are fixed. The presence of individual effects and/or temporal effects complicates the analysis.

Several simulation experiments are done, and they indicate that the proposed estimators perform well. A real data set is also studied, with a reasonable result.

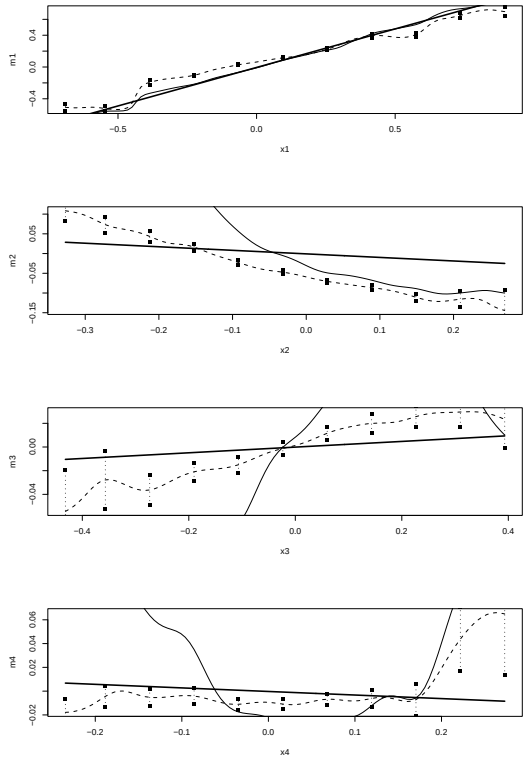

Figure 9: The estimated components in the real data set

6 Proofs.

Proof of Theorems 1 and 2. We proceed similarly as in the proofs of Theorems 1 - 4 in Mammen et al. (1999). We now define the space of additive functions

$$\begin{aligned}\mathcal{H} &= \{m : m(x,t) = m_1(x_1) + \ldots + m_p(x_p) + m_0(t) \\ &\quad \text{with } m_j \in L_2(f_j) \quad (1 \le j \le p)\}.\end{aligned}$$

For $j = 1, \ldots, p$ we consider the following operator $\widehat{\psi}_j : \mathcal{H} \to \mathcal{H}$. For a function $m(x,t) = m_1(x_1) + \ldots + m_p(x_p) + m_0(t)$ the operator does not change m_r for $r \neq j$ and it replaces m_j by

$$-\sum_{t=1}^{T} \frac{N_t}{N} m_0(t) \frac{\widehat{f}_j^t(x_j)}{\widehat{f}_j(x_j)} - \sum_{l \neq j} \int m_l(x_l) \frac{\widehat{f}_{jl}(x_j, x_l)}{\widehat{f}_j(x_j)} \, dx_l.$$

Note that for the index j equation (2.3) can be rewritten as

$$\widehat{m}(x,t) = \widetilde{m}(x,t) + \widehat{\psi}_j \widehat{m}(x,t), \tag{6.1}$$

where $\widehat{m}(x,t) = \widehat{m}_1(x_1) + \ldots + \widehat{m}_p(x_p) + \widehat{\eta}_t$ and $\widetilde{m}(x,t) = \widetilde{m}_1(x_1) + \ldots + \widetilde{m}_p(x_p) + \widetilde{\eta}_t$.

Furthermore (2.4) can be rewritten as

$$\widehat{m}(x,t) = \widetilde{m}(x,t) + \widehat{\psi}_0 \widetilde{m}(x,t) \tag{6.2}$$

where for a function $m(x,t) = m_1(x_1) + \ldots + m_p(x_p) + m_0(t)$ the function $r(x,t) = \widehat{\psi}_0 m(x,t)$ is defined as $r_1(x_1) + \ldots + r_p(x_p) + r_0(t)$ with $r_j(x_j) = m_j(x_j)$ and

$$r_0(t) = -\sum_{j=1}^{p} \int m_j(x_j) \widehat{f}_j^t(x_j).$$

The algorithm studied in Theorem 2 is based on iterative application of the operator $\widehat{S} = \widehat{\psi}_0 \widehat{\psi}_p, \ldots, \widehat{\psi}_1$. We compare this operator with $S = \psi_0 \psi_p \ldots \psi_1$ where ψ_j is defined as $\widehat{\psi}_j$ but with $\widehat{f}_{jk}, \widehat{f}_j, \widehat{f}_j^t$ and N/N_t replaced by f_{jk}, f_j, f_j^t or d/d_t, respectively. This operator does not depend on n and it can be easily checked that ψ_j is the orthogonal projection in $\mathcal{H} \subset L_2(f)$ onto the orthogonal complement of $\mathcal{H}_j = \{m \in \mathcal{H} : m(x,t) = m_j(x_j)$ for a function m_j with $\int m_j(x_j) f_j(x_j)\, dx_j = 0$ (if $j \neq 0$) or $m(x,t) = m_0(t)$ for a function m_t with $\sum_{t=1}^{T} d_t m_0(t) = 0 \quad$ (if $j = 0)\}$. The operator S consists in an iterative application of orthogonal projections. As in Mammen et al. (1999) this can be used to show that the operator norm

$$\rho := \sup\{\|S(f_0 + \ldots + f_p)\|_2 : f_j \in \mathcal{H}_j, \quad \|f_0 + \ldots + f_p\|_2 \leq 1\}$$

is strictly less than 1. Here $\| \quad \|_2$ is the norm of the space $L_2(f)$. Furthermore using uniform consistency or $\widehat{f}_{jk}$ and $\widehat{f}_j$ we get that

$$\begin{aligned} \widehat{\rho} \;:=\; & \sup\{\|\widehat{S}(f_0 + \ldots + f_p)\|_2 : f_j \in \mathcal{H}_j^n, \\ & \|f_0 + \ldots + f_p\|_2 \leq 1\} = \rho + o_p(1), \end{aligned}$$

where $\mathcal{H}_j^n$ is defined as $\mathcal{H}_j$ but with the restriction $\int m_j(x_j)\ f_j(x_j)\ dx_j = 0$ replaced by $\int m_j(x_j) \widehat{f}_j(x_j)\ dx_j = 0$ for $1 \leq j \leq p$ and with $\sum_{j=1}^{T} d_t m_0(t) = 0$ replaced by $\sum_{j=1}^{T} N_t m_0(t) = 0$ for $j = 0$. (Compare Lemma 2 in Mammen et al. (1999)). In particular, for a ρ' with $\rho' < 1$ we get that

$$\widehat{\rho} < \rho' \tag{6.3}$$

with probability tending to one. By iterative application of (6.1) and (6.2) (or equivalently (2.3) and (2.4) we get that (with probability tending to one)

$$\widehat{m} = \sum_{a=0}^{\infty} \widehat{S}^a \widehat{\tau}$$

with $\widehat{\tau} = \widehat{\psi}_0 \widehat{\psi}_p \ldots \widehat{\psi}_2 \widetilde{m}_1 + \ldots + \widehat{\psi}_0 \widetilde{m}_d + \widetilde{m}_0$, where (in abuse of notation) $\widetilde{m}_j(x,t) = \widetilde{m}_j(x_j)$ and $\widetilde{m}_0(x,t) = \widetilde{\eta}_t$, see also Lemma 3 in Mammen et al. (1999).

For the proof of Theorem 2 one now uses that $\widehat{m}^{[a+1]} = \widehat{S}\widehat{m}^{[a]} + \widehat{\tau}$ and therefore $\widehat{m}^{[a+1]} = \sum_{s=0}^{a} \widehat{S}^s \widehat{\tau} + \widehat{S}^{a+1} \widetilde{m}^{[0]}$. Because of (6.3) this shows Theorem 2. For the proof

of Theorem 1 one decomposes $\widetilde{m}$ into $\widetilde{m} = \widetilde{m}^A + \widetilde{m}^B$, where $\widetilde{m}^A$ was defined in (2.6). Put now $\widehat{\tau}^j = \widehat{\psi}_0 \widehat{\psi}_d \ldots \widehat{\psi}_2 \widetilde{m}_1^j + \ldots + \widetilde{m}_0^j$ (for $j = A, B$). Then we have $\widehat{m} = \widehat{m}^A + \widehat{m}^B$ with $\widehat{m}^j = \sum_{a=0}^{\infty} \widehat{S}^a \widehat{\tau}^j$ (for $j = A, B$). The term $\widehat{m}^A$ is now analyzed as in Lemma 4 in Mammen et al. (1999). The basic fact that was used there was that an integral operator that is applied to a *local* average of ε^i (as is $\widetilde{m}^A$) leads to a *global* average of ε^i. A global average is of order $n^{-1/2}$ and therefore of lower order than $n^{-2/5}$. We now give a more detailed discussion of this argument. The basic step is to show that for $1 \le j, k \le d$

$$\sup_{0 \le x_k \le 1} \left| \int_0^1 \frac{\widehat{f}_{j,k}(x_j, x_k)}{\widehat{f}_k(x_k)} \widetilde{m}_j^A(x_j) dx_j \right| = o_P(n^{-2/5}). \tag{6.4}$$

Similarly as in Lemma 4 in Mammen et al. (1999) equation (6.4) implies that also iterative applications of the integral operator are of order $o_P(n^{-2/5})$. This gives that

$$\widehat{m}_j^A(x_j) - \widetilde{m}_j^A(x_j) = o_P(n^{-2/5}). \tag{6.5}$$

For (6.5) it remains to check (6.4). For the proof we use that

$$\widetilde{m}_j^A(x_j) = \sum_{t=1}^{T} \widetilde{r}_{j,t}^A(x_j) / \widehat{f}_j(x_j),$$

where

$$\widetilde{r}_{j,t}^A(x_j) = N^{-1} \sum_{i=1}^{n} \mathbf{1}(X^{it} \in [0,1]^p) K_h(x_j, X_j^{it}) \varepsilon^{it}. \tag{6.6}$$

Using assumption (A1) we get that

$$\sup_{0 \le x_j \le 1} \left| \widetilde{r}_{j,t}^A(x_j) \right| = O_P(\log(n)^{1/2} n^{-2/5}).$$

By application of (A1) and standard smoothing theory we get that

$$\sup_{0 \le x_k \le 1} \left| \int_0^1 \frac{\widehat{f}_{j,k}(x_j, x_k)}{\widehat{f}_k(x_k)} \widetilde{m}_j^A(x_j) dx_j - \sum_{t=1}^{T} \int_0^1 \frac{E[\widehat{f}_{j,k}(x_j, x_k)]}{E[\widehat{f}_j(x_j)] E[\widehat{f}_k(x_k)]} \widetilde{r}_{j,t}^A(x_j) dx_j \right| = o_P(n^{-2/5}).$$

Claim (6.4) now follows from

$$\int_0^1 \frac{E[\widehat{f}_{j,k}(x_j, x_k)]}{E[\widehat{f}_j(x_j)] E[\widehat{f}_k(x_k)]} \widetilde{r}_{j,t}^A(x_j) dx_j = N^{-1} \sum_{i=1}^{n} \mathbf{1}(X^{it} \in [0,1]^p) g_h(X_j^{it}) \varepsilon^{it} = o_P(n^{-2/5}),$$

where

$$g_h(u_j) = \int_0^1 \frac{E[\widehat{f}_{j,k}(x_j, x_k)]}{E[\widehat{f}_j(x_j)] E[\widehat{f}_k(x_k)]} K_h(x_j, u_j) dx_j.$$

For the statement of Theorem 2 it remains to prove that

$$\sup_{x_j \in [h_j, 1-h_j]} \left| \widehat{m}_j^B(x_j) - \left[-\gamma_{n,j} + n^{-2/5} \beta_j(x_j) \right] \right| = o_p(n^{-2/5}). \tag{6.7}$$

For a proof of (6.7) one can apply Theorem 3 in Mammen et al. (1999) with $\alpha_{n,j}(x_j) = m_j(x_j) + m_j'(x_j) \int K_{h(j)}(x_j, u)(u - x_j)\ du\ [\int K_{h(j)}(x_j, v)\ dv]^{-1}$. The conditions of Theorem 3 can be verified as in the proof of Theorem 4 in Mammen et al. (1999).

Proof of Corollary 1. It can be easily checked that the conditional distribution of $\widetilde{r}_{j,t}^A(x_j)/\widehat{f}_j(x_j)$ (see (6.6)), given $(X_j^{is} : 1 \le j \le p, 1 \le s \le t, 1 \le i \le n, \varepsilon^{is} : 1 \le s \le t-1, 1 \le i \le n)$, has a limiting distribution that is nonrandom and that therefore does not depend on the conditioning set. This implies that $\widetilde{m}_j^A(x_j) = \sum_{t=1}^T \widetilde{r}_{j,t}^A(x_j)/\widehat{f}_j(x_j)$ converges in distribution to the convolution of these limiting distributions. The corollary then follows by application of (6.7).

Proof of Theorem 3. Theorem 3 can be proved by similar arguments as Theorem 1. As noted in the discussion of Section 2.2 we have that

$$\rho := \|R\| = \sup\{\|Rf\|_2 : \int \|f(x)\|^2\ dx \le 1\} < 1.$$

We will show that

$$\Delta_\rho := \|R - \widetilde{R}\| = o_p(1). \tag{6.8}$$

This immediately implies that $\widehat{\rho} = \|\widehat{R}\| < \rho'$ with probability tending to one for $\rho < \rho' < 1$.

Furthermore we will prove that

$$\widehat{s} := \max\{\|\widehat{R}f\|_\infty : \int \|f(u)\|^2\ du \le 1, \} = O_p(1), \tag{6.9}$$

$$\Delta_s := \max\{\|(\widehat{R} - R)f(x)\| : \int \|f(u)\|^2\ du \le 1, \ x \in [0,1]\} = o_p(1). \tag{6.10}$$

where $\|g\|_\infty = \max_{x\in[0,1]^p} \|g(x)\|$ and where $\|g(x)\|$ is the Euclidean norm of $\mathbf{R}^p$.

We will use these properties of $\widehat{R}$ to show that uniformly in x

$$\sum_{r=0}^{\infty} \widehat{R}^r \bar{m}^B(x) = \sum_{r=0}^{\infty} \widehat{R}^r \mu_n(x) + o_p(n^{-2/5}), \tag{6.11}$$

where $\mu_n(x) = \mu_n^1 + n^{-2/5}\mu_n^2(x)$

$$\sum_{r=0}^{\infty} \widehat{R}^r \mu_n^2(x) = \sum_{r=0}^{\infty} R^r \mu_n^2(x) + o_p(1), \tag{6.12}$$

$$\widehat{m}^A(x) = \bar{m}^A(x) + o_p(n^{-2/5}). \tag{6.13}$$

These claims imply the theorem, see also the discussion after the statement of the theorem.

We start with a proof of (6.11). Note first that

$$\begin{aligned}
&\|\sum_{r=0}^{\infty} \widehat{R}^r(\bar{m}^B - \mu_n)\|_2 \\
&\quad \leq \sum_{r=0}^{\infty} \widehat{\rho}^r \|\bar{m}^B - \mu_n\|_\infty = o_p(n^{-2/5})
\end{aligned}$$

and, by assumption,

$$\|(\bar{m}^\beta - \mu_n)\|_\infty = o_p(n^{-2/5}).$$

By application of (6.9) we get

$$\begin{aligned}
&\|\sum_{r=0}^{\infty} \widehat{R}^r(\bar{m}^B - \mu_n)\|_\infty \\
&\leq \|\bar{m}^B - \mu_n\|_\infty + \|\widehat{R}(\sum_{r=0}^{\infty} \widehat{R}^r(\bar{m}^B - \mu_n))\|_\infty \\
&\leq o_p(n^{-2/5}) + \widehat{s}\|\sum_{r=0}^{\infty} \widehat{R}^r(\bar{m}^B - \mu_n)\|_2 \\
&= o_p(n^{-2/5}).
\end{aligned}$$

This shows (6.11). For the proof of (6.12) we apply a telescope argument. We get that with probability tending to one

$$\begin{aligned}
&\|\sum_{r=0}^{\infty} (\widehat{R}^r - R^r)\mu_n^2\|_\infty \\
&\leq \sum_{r=0}^{\infty} \sum_{k=0}^{r-1} \|\widehat{R}^k(\widehat{R} - R)R^{r-1-k}\mu_n^2\|_\infty \\
&\leq \sum_{r=0}^{\infty} r(\rho')^{r-1}\Delta_s\|\mu_n\|_2 = o_p(1).
\end{aligned}$$

We now come to the proof of (6.13). Note that

$$\widehat{m}^A = \sum_{r=0}^{\infty} \widehat{R}^r(I - \widehat{H}\pi)^{-1}\bar{m}^A.$$

We will show

$$\|(I - \widehat{H}\pi)^{-1}\bar{m}^A - \bar{m}^A\|_\infty = o_p(n^{-2/5}), \tag{6.14}$$

$$\|\widehat{R}(I - \widehat{H}\pi)^{-1}\bar{m}^A\|_\infty = o_p(n^{-2/5}), \tag{6.15}$$

$$\|\sum_{r=1}^{\infty} \widehat{R}^r(I - \widehat{H}\pi)^{-1}\bar{m}^A\|_\infty = o_p(n^{-2/5}). \tag{6.16}$$

Clearly, these claims imply (6.13). Claim (6.16) immediately follows from (6.15) with help of (6.8) and (6.9). Because of (6.14) for claim (6.15) it suffices to show

$$\|\widehat{R}\bar{m}^A\|_\infty = o_p(n^{-2/5}). \tag{6.17}$$

We start by showing

$$\|\widehat{H}(I-\pi)\bar{m}^A\|_\infty = o_p(n^{-2/5}). \tag{6.18}$$

Because of $\|\bar{m}^A\|_\infty = O_p(\sqrt{\log n}\, n^{-2/5}$ and (A4') this would follow from

$$\|H(I-\pi)\bar{m}^A\|_\infty = o_p(n^{-2/5})\,.$$

It can be easily checked that the left hand side is of order $O_p(n^{-1/2}\sqrt{\log n})$. Note that global average of ε^i are taken in $H(I-\pi)\bar{m}^A$. This shows (6.18). So for (6.15) it remains to check (6.17). For this consider the following expansion

$$\begin{aligned}(I-\widehat{H}\pi)^{-1} &= [(I-H\pi)(I-\Delta)]^{-1} \\ &= \sum_{r=0}^{\infty}\Delta^r(I-H\pi)^{-1}, \\ \Delta &= (I-H\pi)^{-1}(\widehat{H}-H)\pi,\end{aligned}$$

where claim (6.17) follows from (A4') and

$$\max\{\|(I-H\pi)^{-1}f\|_\infty : \|f\|_\infty \le 1\} = O(1). \tag{6.19}$$

We now give a proof of (6.19). Denote the eigenfunction of H with eigenvalue λ_j by e_j. For a function $f = \sum_{j=1}^{\infty}\gamma_j e_j$ we have

$$\begin{aligned}(I-H\pi)^{-1}f &= \sum_{j=r}^{\infty}\gamma_j e_j + \sum_{j=1}^{r-1}\gamma_j(1-\lambda_j)^{-1}e_j \\ &= f + \sum_{j=1}^{r-1}\gamma_j[1-(1-\lambda_j)^{-1}]e_j.\end{aligned} \tag{6.20}$$

Suppose that $\|f\|_\infty \le 1$. Then $\|f\|_2 \le 1$ and we get

$$\begin{aligned}\|(I-H\pi)^{-1}f\|_\infty &\le \|f\|_\infty + \|\sum_{j=1}^{r-1}\gamma_j[1-(1-\lambda_j)^{-1}]e_j\|_\infty \\ &\le 1 + (\sum_{j=1}^{r-1}\gamma_j^2)^{1/2}(\sum_{j=1}^{r-1}[1-(1-\lambda j)^{-1}]^2)^{1/2} \\ &\qquad \max_{1\le j\le r-1, x\in[0,1]^p}\|e_j(x)\|.\end{aligned}$$

Claim (6.17) now follows from

$$\sum_{j=1}^{r-1} \gamma_j^2 \leq \|f_0\|_2^2 \leq 1$$

and

$$\max_{1\leq j\leq r-1, x\in[0,1]^p} \|e_j(x)\| = O(1). \tag{6.21}$$

For the proof of (6.21) note that $e_j(x)$ is a continuous function. This follows from continuity of H and the inequality

$$\begin{aligned} \|e_j(x) - e_j(y)\| &\leq |\lambda_j| \| \int [H(x,z) - H(y,z)] e_j(z)\ dz \| \\ &\leq |\lambda_j| \int \|H(x,z) - H(y,z)\|\ dz. \end{aligned}$$

We now come to the proof of (6.14). With similar arguments as above it can easily be checked that

$$\|(I - \widehat{H}\pi)^{-1}\bar{m}_A - (I - H\pi)^{-1}\bar{m}_A\|_\infty = o_p(n^{-2/5}).$$

So it suffices to show

$$\|(I - H\pi)^{-1}\bar{m}_A - \bar{m}_A\|_\infty = o_p(n^{-2/5}). \tag{6.22}$$

It holds that (see (6.20))

$$\begin{aligned} [(I - H\pi)^{-1} - I]\ f(x) = \int \sum_{j=1}^{r-1} [(1-\lambda_j)^{-1} - 1] \\ e_j(y) e_j(x) f(y)\ dy. \end{aligned}$$

So (6.22) follows by standard smoothing arguments because $[(I - H\pi)^{-1} - I]\bar{m}_A$ is a global (weighted) average of ε_i.

We now come to the proofs of (6.8) - (6.10). Claim (6.8) immediately follows from (6.10). For claim (6.10) it suffices to show

$$\max\{\|(\widehat{H} - H)(I - \pi)f\|_\infty : \|f\|_2 \leq 1\} = o_p(1). \tag{6.23}$$

This can be seen by a similar treatment of the matrices $(I - \widehat{H}\pi)^{-1}$ and $(I - H\pi)$ as above. Claim (6.23) follows immediately from (A4').

It remains to show (6.9). This claim follows from (6.10) and

$$\begin{aligned}
&\max\{\|Rf\|_\infty : \|f\|^2 \leq 1\} \\
&\leq \max\{\max_{x\in[0,1]^p}[\int \|R^2(y,x)\|^2\ dx]^{1/2}\ \|f\| : \|f\|^2 \leq 1\} \\
&\leq \{\max_{x\in[0,1]^p}[\int \|R^2(y,x)\|^2\ dx]^{1/2} \\
&\leq \max_{x,y\in[0,1]^p} \|R(y,x)\| \\
&< +\infty
\end{aligned}$$

Proof of Theorem 4. The proofs follows along the same lines of the proofs of Theorems 1 and 2. Compare also the proofs of Theorems 1' - 4' in Mammen et al. (1999).

References.

Arellano, M. (2003). *Panel Data Econometrics*, Oxford University Press: Oxford.

Arellano, M. and Honoré, B. E. (2001). Panel data models: Some recent developments. In J. J. Heckman and E. Leamer (eds.), *Handbook of Econometrics* 5, ch. 53, 3229-3296. North Holland.

Baltagi, B. H. (1995). *Econometric Analysis of Panel Data*, John Wiley: New York.

Baltagi, B. H., Griffin, J. M. and Xiong, W. (2000). To pool or not to pool: Homogenous versus heterogenous estimators applied to cigarette demand. *The Review of Economics and Statistics* **82**, 117 - 126.

Baltagi, B. H. and Li, Q. (2002). On instrumental variable estimation of semiparametric dynamic panel data models. *Economic Letters* **76**, 1-9.

Chang, Y. (2004). Bootstrap unit root test in panels with cross-sectional dependency. *Journal of Econometrics* **120**, 263-293.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710-723.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series*, Springer-Verlag: New York.

Fu, B., Li, W-K. and Fung, W-K. (2002). Testing model adequacy for dynamic panel data with intercorrelation. *Biometrika* **89**, 591-601.

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

Hjellvik, V., Chen, R. and Tjøstheim, D. (2004). Nonparametric estimation and testing in panels of intercorrelated time series. *Journal of Time Series Analysis* **25**, 831-872.

Hjellvik, V. and Tjøstheim, D. (1999). Modelling panels of intercorrelated autoregressive time series. *Biometrika* **86**, 573-590.

Honoré, B. E. and Lewbel, A. (2002). Semiparametric binary choice panel data models without strictly exogenous regressors. *Econometrica* **70**, 2053-2063.

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge University Press: Cambridge.

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

Linton, O., Mammen, E. (2005). Estimating semiparametric ARCH (∞) models by kernel smoothing methods. *Econometrika* **73**, 771-836.

Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-101.

Luchka, A. Y. (1965). *The Method of Averaging Functional Corrections: Theory and Applications*, Academic Press: New York and London.

Mammen, E., Linton, O., Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statist.* **27**, 1443 - 1490.

Mátyás, L. and Sevestre, P. (1992). *The Econometrics of Panel Data*, Kluwer Academic: Dordrecht.

Nielsen, J. P. and Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society, Ser. B* **67**, 43-61.

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

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* **25**, 186-211.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continous data. *Journal of Econometrics* **119**, 99-130.

Sperlich, S., Tjøstheim, D. and Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory* **18**, 197-251.

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

Wooldridge, J. (2005a). Fixed effects and related estimation for correlated random coefficient and treatment effect panel data models. Forthcoming *Review of Economics and Statistics*.

Wooldridge, J. (2005b). Simple solutions to the initial conditions problem for dynamic nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**, 39-54.