Brachmann, Klaus

Working Paper

Choosing the optimal bandwidth in case of correlated data

Discussion papers in statistics and econometrics, No. 8/95

Provided in cooperation with:
Universität zu Köln

DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 8/95

Choosing the optimal bandwidth in case of correlated data

by

Klaus Brachmann*

November 1995

Summary: In case of estimating growth curves nonparametrically one faces the fact that the data driven bandwidth selectors published in standard textbooks mostly choose bandwidths much too low. This is due to the positive autocorrelation observed in growth data. This paper introduces an easy way to incorporate this effect in the known concept of penalizing functions.

*) Seminar für Wirtschafts– und Sozialstatistik, Universität zu Köln, Albertus–Magnus–Platz, 50923 Köln, Deutschland; Tel: +49/221/470 4129, Fax: +49/221/470 2816, e–mail: brach@wiso.uni-koeln.de

1 Introduction

Nonparametric regression is a flexible tool for estimating the underlying true but unknown regression function from noisy data. Unlike the parametric methods it does not require any assumptions other than the smoothness of the curve. But although the method is called nonparametric there is a very important parameter to be estimated. As this paper tackles the widely known kernel method, this parameter is to be characterized as the window width since kernel estimators are locally weighted averages of the response variables where the weights assigned are derived from the kernel function. The width of the window containing positive weights is determined by the bandwidth parameter.

The magnitude of bandwidth controls the smoothness of the estimated regression function. If the bandwidth is too large, the regression function is oversmoothed (i.e. the estimated function may miss some important local properties of the true function) whereas in case of too low a bandwidth the regression function is undersmoothed (i.e. the estimated function may have some features which are the result of some irregularities and are in no way systematic). Hence, choosing an optimal value of the bandwidth is the main problem in the field of nonparametric regression. This holds true since the other variable component, the kernel function itself does not have any significiant influence in the performance of the estimator (see e.g. Härdle (1990a)).

For independent observations there are several methods to be found in standard textbooks (see e.g. Härdle (1990b)). These are mostly designed to minimize a quadratic error measure like the mean average square error (MASE). But, as Härdle and Marron (1986) proved, the choice of the error measure to be minimized is not critical in the sense that

$$\sup_{h \in H_n} \left| \frac{d(h) - MISE(h)}{MISE(h)} \right| \xrightarrow{a.s.} 0 \qquad (n \to \infty)$$

where $H_n = [n^{\delta-1}, n^{-\delta}], 0 < \delta < \frac{1}{2}$ and $d(h)$ is one of the widely known error measures $ASE(h)$, $MASE(h)$ and $ISE(h)$. h is the bandwidth parameter and n the number of observations.

Two main methods have been proven very useful in the case of independent observations, the crossvalidation method and the more flexible concept of penalizing functions which even includes measures known from time series analysis. This paper presents a generalization of this concept of penalizing functions to the case of correlated errors which will be observed when working with growth data. It will be shown how the conventional concept fails in this case and how this failure may be corrected. A small simulation study will support these results.

In the literature there are already some solutions to special problems connected with correlated data. Müller and Stadtmüller (1988) worked out a test for serial correlation which heavily depends upon the assumption that with increasing n in a fixed interval and because of the smoothness of the underlying true regression function the influence of this regression function vanishes asymptotically and therefore can be viewed as a nuisance regression function. Further, they do not offer a recipe how to choose the optimal bandwidth. Hart and Wehrly (1986) have developed a way to calculate the optimal bandwidth in case of correlated data using observations from several experimental units. Chu and Marron (1991) compare the modified crossvalidation (MCV) with the partitioned crossvalidation (PGCV) introduced by Marron (1987). Altman (1990) corrects for the correlation by using the empirical autocorrelation function, a method which is also pursued in this paper in the context of penalizing functions. Chu (1992) offers a bandwidth selector derived by minimzing the $MASE(h)$ and using the empirical autocorrelation function.

In the next section the conventional concept of penalizing function is presented and it is shown why and how this concept fails in case of correlated data. The third section contains the generalization of this concept which corrects for the above failure. The paper closes with a short simulation study which shows (in the case of a growth function) the superiority of the new bandwidth selector.

2 The conventional concept of penalizing functions

First it is considered the equally spaced fixed design

$$Y_i = m(x_i) + \varepsilon_i$$

with support on $[0;1]$ where m is a smooth unknown regression function and the ε_i are iid errors. To estimate the regression function m, we consider the widely known kernel estimator as introduced by Nadaraya (1964) and Watson (1964):

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) Y_i}{\sum_{i=1}^{n} K_h(x - x_i)}$$

$K_h(x - x_i)$ is the simplified notation for $\frac{1}{h} K\left(\frac{x - x_i}{h}\right)$, where $K(u)$ is the known kernel function. It is easily seen that $\hat{m}(x)$ belongs to the class of linear estimators as it can be written as a weighted sum of the observations:

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} W_i(x)Y_i \quad \text{with} \quad W_i(x) = \frac{K_h(x-x_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i)}$$

As stated in the introduction the optimal bandwidth may be estimated by minimizing a quadratic error measure like the $ASE(h)$, which can be decomposed as following:

$$\begin{aligned} ASE(h) &= \frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - \hat{m}(x_i)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} m^2(x_i) + \frac{1}{n}\sum_{i=1}^{n}\hat{m}^2(x_i) - \frac{2}{n}\sum_{i=1}^{n} m(x_i)\hat{m}(x_i) \end{aligned}$$

Obviously only the third term is of practical interest since the first term is independent of h and the second term can be directly computed as it does not contain any unknown elements.

Estimating the third term may be simply done by substituting the unknown regression function $m(x_i)$ by the observable value Y_i. If this is done throughout the whole expression for $ASE(h)$, one obtains the following substitution estimate:

$$p(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(x_i)\right)^2$$

Minimization of this term must produce a value of h which leads to pure interpolation since in this case $p(h) = 0$. In other words $p(h)$ is a biased estimator of $ASE(h)$, which is easily seen by the following decomposition:

$$\begin{aligned} p(h) &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}(x_i)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + ASE(h) - \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{m}(x_i) - m(x_i)\right) \end{aligned}$$

As the first term is again independent of h the usefulness of $p(h)$ as an estimator of $ASE(h)$ depends heavily on the magnitude of the last term. Unfortunately this is not of negligible size:

$$\begin{aligned} E\left[-\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{m}(x_i - m(x_i)\right)\right] &= -\frac{2}{n}\sum_{i=1}^{n} E\left[\varepsilon_i\left(\frac{1}{n}\sum_{i=1}^{n} W_j(x_i)Y_j - m(x_i)\right)\right] \\ &= -\frac{2}{n}\sum_{i=1}^{n} E\left[\varepsilon_i\left(\frac{1}{n}\sum_{i=1}^{n} W_j(x_i)(m(x_j) + \varepsilon_j) - m(x_i)\right)\right] \\ &= -\frac{2}{n}\sum_{i=1}^{n} E\left[\varepsilon_i\right]\left(\frac{1}{n}\sum_{i=1}^{n} W_j(x_i)m(x_j) - m(x_i)\right) \\ &= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_j(x_i)E\left[\varepsilon_i\varepsilon_j\right] \end{aligned}$$

The conventional concept of penalizing functions now makes the assumption that the errors ε_i are independent random variables with expectation zero and variance $\sigma^2(x_i)$. This reduces the above expression in the following way:

$$\begin{aligned} E\left[-\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\hat{m}(x_i)-m(x_i)\right)\right] &= -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}W_j(x_i)E\left[\varepsilon_i\varepsilon_j\right] \\ &= -\frac{2}{n^2}\sum_{i=1}^{n}W_i(x_i)E\left[\varepsilon_i^2\right] \\ &= -\frac{2}{n^2}\sum_{i=1}^{n}W_i(x_i)\sigma^2(x_i) \end{aligned}$$

As this term does not vanish, the substitution estimator has to be adjusted by a correcting term which penalizes too small bandwidths. This leads to the penalizing function estimator:

$$G(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{m}(x_i)\right)^2\Xi\left(\sum_{i=1}^{n}W_i(x_i)\right)$$

Any function $\Xi(u)$ which satisfies the condition

$$\Xi(u) = 1 + 2u + O(u^2) \qquad (u \to 0)$$

may serve as a correcting term. The major examples are shown in table 1.

Table 1: Types of Correcting Terms

name	$\Xi(u)$
Shibata	$1+2u$
Generalized CV	$\frac{1}{(1-u)^2}$
AIC	$\exp(2u)$
FPE	$\frac{1+u}{1-u}$
Rice	$\frac{1}{1-2u}$

The penalizing function estimator now decomposes (ignoring terms of higher order) as follows:

$$\begin{aligned} G(h) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varepsilon_i^2+\left(m(x_i)-\hat{m}(x_i)\right)^2+2\varepsilon_i\left(m(x_i)-\hat{m}(x_i)\right)\right]\cdot\left[1+\frac{2}{n}W_i(x_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2+ASE(h)+\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(m(x_i)-\hat{m}(x_i)\right)+\frac{2}{n^2}\sum_{i=1}^{n}\varepsilon_i^2W_i(x_i) \end{aligned}$$

One can easily verify that the expectation of the third term is exactly the negative expected value of the last term which makes $G(h)$ an unbiased estimator of $ASE(h)$.

This of course only holds true under the crucial assumption of uncorrelated errors which means in this case

$$\sum_{i=1}^{n}\sum_{j=1}^{n} W_j(x_i)E\left[\varepsilon_i\varepsilon_j\right] = \sum_{i=1}^{n} W_i(x_i)E\left[\varepsilon_i^2\right] .$$

The question now is what happens if this assumption does not hold. Heuristically the following conclusions may be drawn:

- If the correlations are predominantly positive (i.e. $\sum_{i=1}^{n}\sum_{j=1}^{n} W_j(x_i)E\left[\varepsilon_i\varepsilon_j\right] > \sum_{i=1}^{n} W_i(x_i)E\left[\varepsilon_i^2\right]$) then the penalizing term is too small to compensate for the bias and therefore the penalizing of too small values for the bandwidth fails which leads to undersmoothing.

- If the correlations are predominantly negative (i.e. $\sum_{i=1}^{n}\sum_{j=1}^{n} W_j(x_i)E\left[\varepsilon_i\varepsilon_j\right] < \sum_{i=1}^{n} W_i(x_i)E\left[\varepsilon_i^2\right]$) then the penalizing term is too big which leads to an overpenalizing of too small bandwidths and thus produces an oversmoothed estimate.

3 Generalized concept of penalizing functions

As the bias of the substitution estimator contains the correlation structure of the errors the penalizing term should contain this correlation structure, too. As one can easily see the bias extends in the correlated case to

$$Bias(p) = -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_j(x_i)\gamma(|i-j|) ,$$

where the correlated case is assumed to be characterized by

$$E\left[\varepsilon_i\varepsilon_j\right] = \sigma^2\rho(|i-j|)$$

and where $\gamma(|i-j|)$ is the autocovariance function ($\rho(|i-j|)$ the corresponding autocorrelation function) depending only upon the distance of the points x_i and x_j which is in the relevant case of equidistant time points equivalent to the dependence upon the distance of the points i and j.

The penalizing functions to be constructed have to estimate this bias. Thereby the Generalized concept should mimic the traditional concept insofar to further use the functions from table 1 (especially the first–order Tayler expansion should be the same: $\Xi(u) = 1 + 2u + O(u^2) \quad (u \to 0)$). The autocorrelation function may now be included to construct the generalized penalizing function selector:

$$\tilde{G}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(x_i))^2\,\Xi\left(\frac{1}{n}\sum_{j=1}^{n}W_j(x_i)\rho(|i-j|)\right)$$

Decomposing $\tilde{G}(h)$ it can now be shown that the penalizing term will correct for this bias:

$$\begin{aligned}\tilde{G}(h) &= \frac{1}{n}\sum_{i=1}^{n}\left[\varepsilon_i + (m(x_i) - \hat{m}(x_i))^2 + 2\varepsilon_i\,(m(x_i) - \hat{m}(x_i))\right] \\ &\quad \cdot \left[1 + \frac{2}{n}\sum_{j=1}^{n}W_j(x_i)\rho(|i-j|)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + ASE(h) + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\,(m(x_i) - \hat{m}(x_i)) + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_i^2 W_j(x_i)\rho(|i-j|)\end{aligned}$$

As in the uncorrelated case with the conventional concept of penalizing function the expectation of the last term in the decomposition of $\tilde{G}(h)$ is exactly the bias of the pure substitution estimator.

Unfortunately, the true autocorrelation function will be rarely known so that it has to be estimated. As in Altman (1990) and in Chu (1992) the autocorrelation function may be estimated nonparametrically:

$$\hat{\rho}_h(s) = \frac{\sum\limits_{i=[nh/2]}^{n+1-[nh/2]-s}\hat{\varepsilon}_i\hat{\varepsilon}_{i+s}}{\sum\limits_{i=[nh/2]}^{n+1-[nh/2]-s}\hat{\varepsilon}_i^2}$$

$\hat{\varepsilon}_i = Y_i - \hat{m}_h(x_i)$ are the estimated residuals. The limits of the above sums are formulated to avoid the usual bias in nonparametric regression. So the $nh/2$ observations near the border of the support $[0;1]$ are just left out. The estimated residuals are the result of a separate nonparametric regression. This leads to another problem: Estimating the autocorrelation function is done using nonparametric regression which involves fixing another optimal bandwidth. Fortunately the choice of bandwidth is not so critical in this stage as the simulation in the next chapter and other works (e.g. Altman (1990)) show.

Another possible way is to suppose a specific parametric model for the error structure and then to estimate the correlations by usual parametric methods. This approach has been followed by Truong (1991). This of course leads to a general discussion about the comparison between parametric and nonparametric methods. A completely nonparametric method should not contain any parametric assumptions, so the path used in this paper is to be preferred.

The influence of the bandwidth h used to construct the autocorrelation function is explained heuristically as follows:

- Using too high a bandwidth will lead to a too smooth estimate and thus some parts of the systematic component will be found in the estimated residuals. This will increase the influence of the assumed autocorrelation and thus the correction used above will overpenalize smaller bandwidths which results in generally too large values for $\hat{h}$.
- Using too low a bandwidth will lead to a wiggly estimate (in the extreme case even to interpolation) and thus the estimated residuals will underestimate the real stochastic influence so that the possible autocorrelation will be underestimated. This of course leads to less penalization of too low bandwidths which results in generally too low values for $\hat{h}$.

4 Simulation Study

Throughout the whole section the true function $m(x)$ (normally unknown) will be considered to be a logistic function

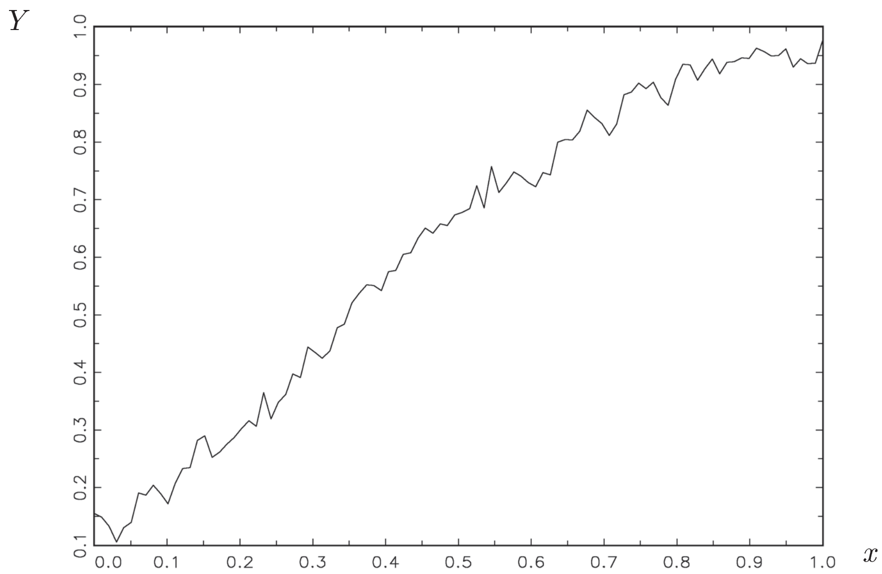
$$m(x) = \frac{Y^*}{1 + \exp\{a - b * t\}}$$

with the parameters fixed at $Y^* = 1$, $a = 1,8$ and $b = 5$. The errors are generated from a Gaussian AR(1)–process. As the number of observations in growth curve analysis mostly does not exceed $n = 100$ this was also the length of the generated times series. In each simulation study there are $N = 500$ times series generated to perform the MC–simulation.

The kernel function used was the Epanechnikov–kernel ($K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$) since it can be easily shown that this is an optimal kernel (see standard textbooks). Anyway the use of the kernel function is not as critical as the choice of the bandwidth.

First the influence of the bandwidth h used by the preceding nonparametric regression will be analyzed. Therefore the ASE–optimal bandwidth is computed and relative to this value h is determined. Three situations are examined (the situation of figure 1 is given):

Figure 1: *Logistic function with* $Y^* = 1$, $a = 1,8$, $b = 5$ *overlayed by a Gaussian AR(1)–process with* $\sigma = 0,02$ *and* $\rho = 0,5$



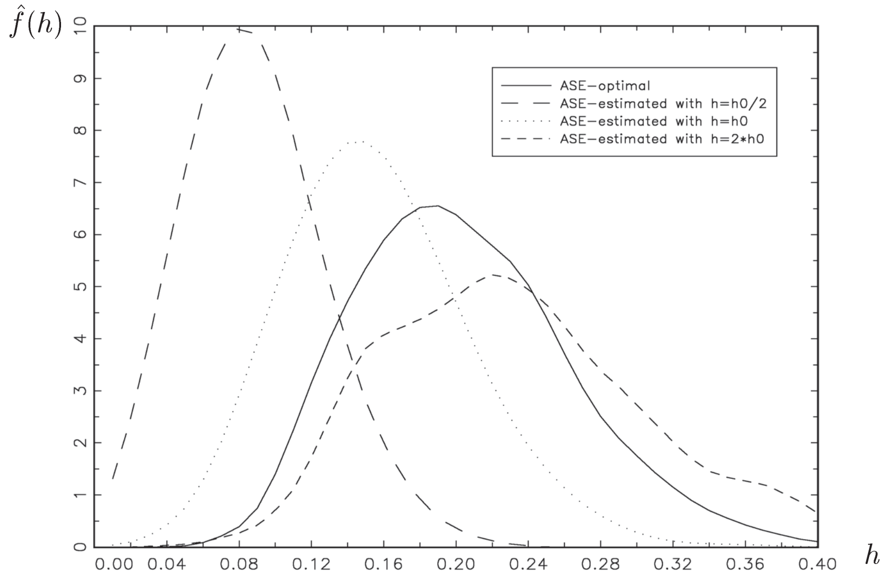
1. $h = h_{opt}/2$

2. $h = h_{opt}$

3. $h = 2 \cdot h_{opt}$

Using these values the optimal bandwidth $\hat{h}$ is estimated. Since it is more important to examine the difference in ASE, also the relative values of the ASE ($ASE(\hat{h})/ASE(h_{opt})$) in each of the three cases are calculated. Figure 2 shows the nonparametric density estimate $\hat{f}(h)$ of the MC–distribution of the resulting optimal values of h in the true case (direct ASE–minimized) and the three cases to be examined. It can be easily seen that the success of the generalized concept of penalizing function depends severely upon the choice of bandwidth in estimating the autocorrelation function. Choosing too low a bandwidth results in much too low estimated values for h. Even values which are exactly identical with the optimal bandwidth lead to a too low bandwidth. Doubling the value leads, as expected, to a too big value for the estimated bandwidth. But the bias is much smaller than in the first two cases. Thus oversmoothing is much less critical than undersmoothing. This is supported by Figure 3 which shows the nonparametric density estimate of the MC–distribution of the relative ASE–values at the estimated values for h

Figure 2: *Distribution of the true and the estimated optimal values for h*



$(ASE_{rel} = ASE(h_{opt})/ASE(\hat{h}))$.

The second part of the simulation study shows the superiority of the proposed method over the conventional concept. Further, the consequences of using the empirical autocorrelation function are examined by comparing with the situation when the true but normally unkown autocorrelation function is used. It will be seen that the empirical autocorrelation function leads indeed to acceptable results.

The data are generated according to the first simulation study with varying values for the autocorrelation coefficient ρ and for the standard deviation σ. For all the combinations the relative values of the ASE are calculated for the following three different cases:

- Generalized concept using the empirical autocorrelation function (ASE_1)
- Generalized concept using the true autocorrelation function (ASE_2)
- Conventional concept (ASE_3)

It is easy to see that the relative ASE–values have a minimum of 1 which indicates that the used concept leads to an ASE–optimal choice of h. Increasing values show worse performance. Table 2 shows the medians of all the different cases (for each case there were 500 simulations done). Although the median is a useful measure it does not contain

Figure 3: *Distribution of the relative ASE–values*

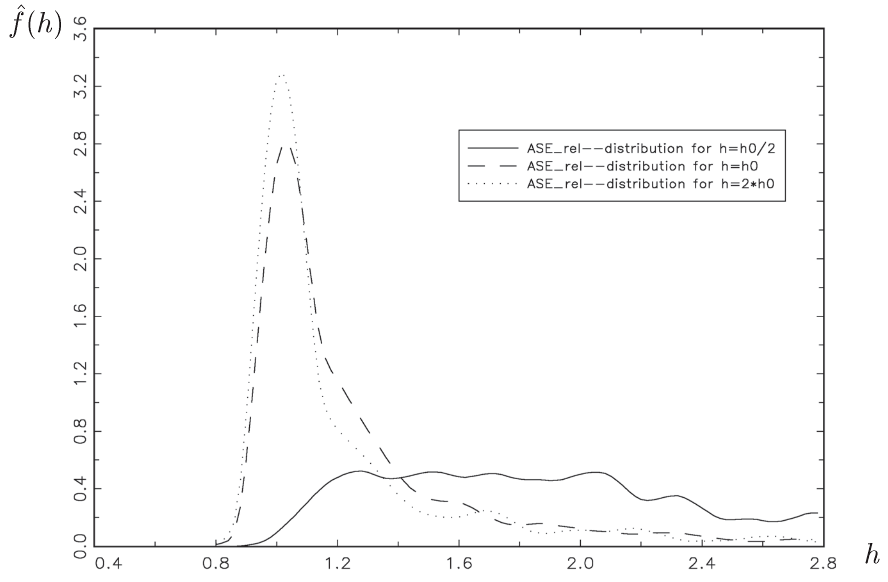


Table 2: Performance of the proposed concept

		$\sigma = 0,1$			$\sigma = 0,02$		
		ASE_1	ASE_2	ASE_3	ASE_1	ASE_2	ASE_3
	-0,9	1,0154	1,0272	4,5913	1,0301	1,0408	3,7726
	-0,6	1,0137	1,0120	1,7006	1,0194	1,0252	1,6410
	-0,4	1,0248	1,0334	1,1989	1,0164	1,0285	1,2301
	-0,3	1,0325	1,0414	1,1415	1,0158	1,0383	1,1330
	-0,2	1,0332	1,0627	1,1129	1,0185	1,0379	1,0684
	-0,1	1,0328	1,0736	1,0731	1,0149	1,0405	1,0511
ρ	0	1,0608	1,0804	1,0804	1,0151	1,0581	1,0581
	0,1	1,0429	1,0942	1,1067	1,1210	1,0502	1,0696
	0,2	1,0557	1,1155	1,1542	1,0806	1,0715	1,2088
	0,3	1,0753	1,1260	1,6069	1,0500	1,0655	1,7678
	0,4	1,0756	1,1347	2,8296	1,0423	1,0654	2,4843
	0,6	1,1444	1,1361	5,2338	1,0368	1,0842	2,8105
	0,9	1,2384	1,2820	3,0399	1,0610	1,1292	1,8477

the complete information of the MC–distribution in each of the above cases. In order to compare the performance of the different concepts more accurately figures 4 and 5 show in analogy to figure 3 the nonparametric density estimate of the simulated distributions in the case of $\sigma = 0,1$ and varying values for ρ (starting with $\rho = -0,9$ up to $\rho = 0,9$ using the same steps as in table 2). For the uncorrelated case $\rho = 0$ a graphical presentation is omitted since all three concepts work well (ASE_2 and ASE_3 are even equal in this case). The graphs are in some way misleading since they show positive densities in areas below 1. This is the result of the chosen bandwidth ($h = 0,1$ in this case) for the nonparametric density estimation.

One can easily see that the use of the empirical autocorrelation function is indeed acceptable since the distributions of ASE_1 and ASE_2 are in fact quite similar. The dotted lines which represent the traditional concept without taking into account the autocorrelation show that an increasing absolute value of ρ leads to a worse performance, whereas the value of ρ does not have any influence on the distribution of the relative ASE when using the proposed concept of the generalized penalizing functions so that this concept is applicable without regard of the underlying correlation structure.

Figure 4: *Distribution of the relative ASE–values in case of negative autocorrelation. ASE_1 is represented by the solid, ASE_2 by the dashed and ASE_3 by the dotted line.*

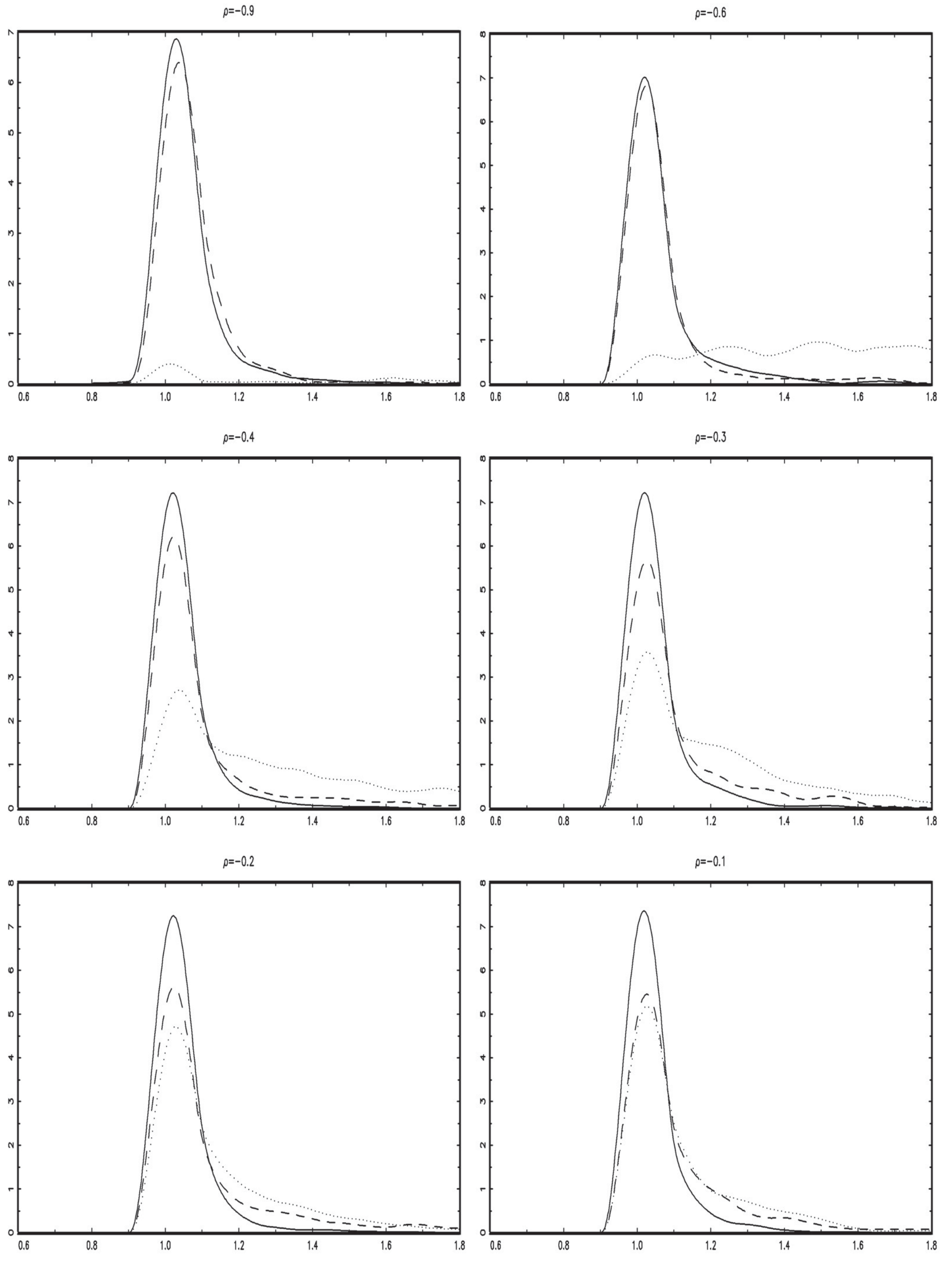
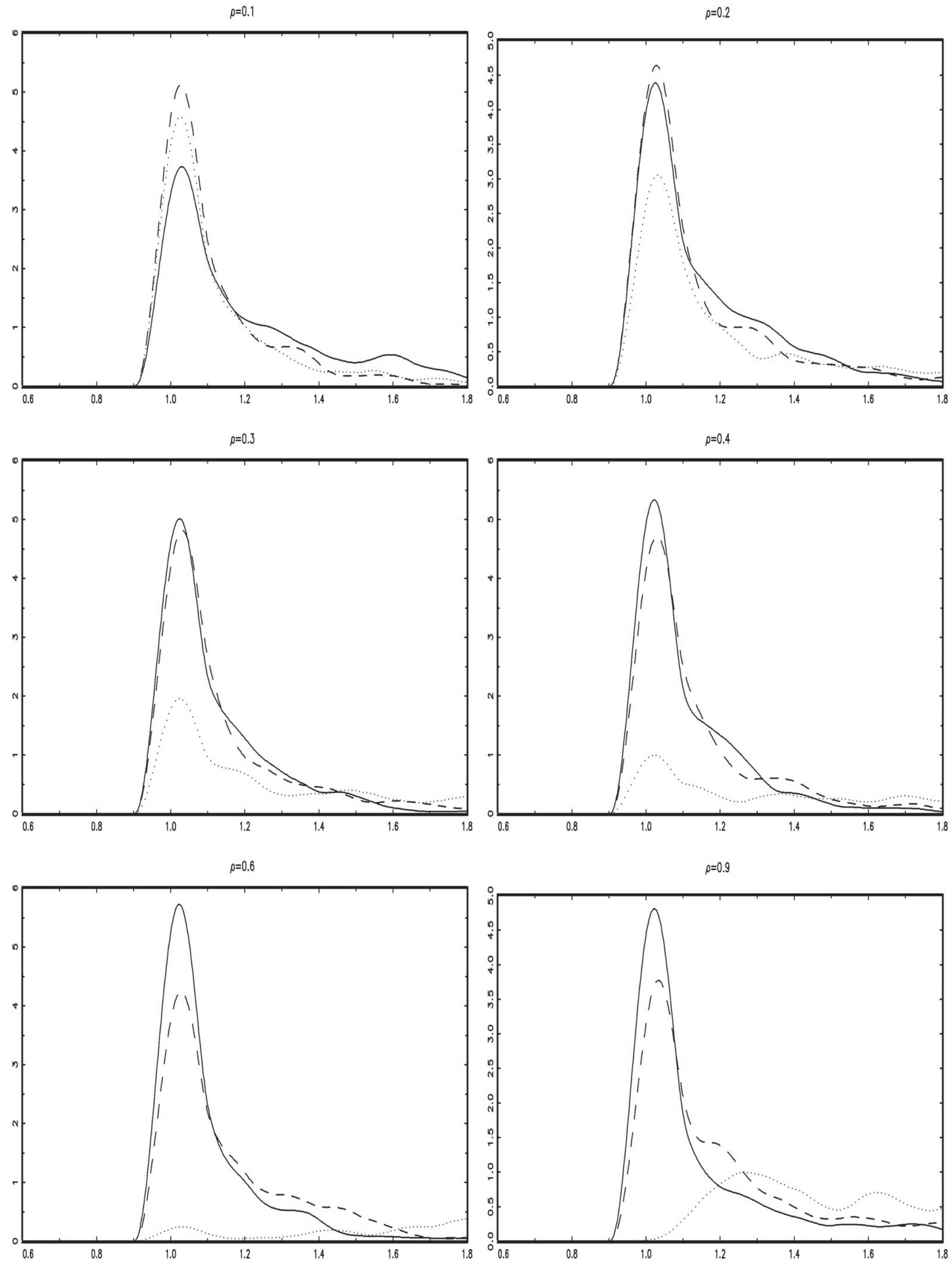


Figure 5: *Distribution of the relative ASE–values in case of positive autocorrelation. ASE_1 is represented by the solid, ASE_2 by the dashed and ASE_3 by the dotted line.*



References

Altman, N.S. (1990): Kernel Smoothing of data with correlated errors. Journal of the American Statistical Association 85, 749–759.

Chu C.-K., J.S. Marron (1991): Comparison of two bandwidth selectors with dependent errors. The Annals of Statistics 19, 1906–1918.

Chu, C.K. (1992): Double Smoothing bandwidth selection with dependent errors. Chinese Journal of Mathematics 20,223–240.

Härdle, W., J.S. Marron (1986): Random approximations to an error criterion of nonparametric statistics. Journal of Multivariate Analysis, 20, 91–113.

Härdle, W. (1990a): Applied Nonparametric Regression.

Härdle, W. (1990b): Smoothing Techniques.

Hart, J.D., T.E. Wehrly (1986): Kernel Regression Estimation using repeated measurements data. Journal of the American Statistical Association, 81, 1080–1088.

Marron, J.S. (1987): Partitioned cross–validation. Econometric Review 6, 271–284.

Müller, H.-G., U. Stadtmüller (1988): Detecting Dependencies in smooth regression models. Biometrika, 75, 639–650.

Nadaraya, E.A. (1964): On Estimating Regression. Theory of Probability and its Applications 10, 186-190.

Truong, Y.K. (1991): Nonparametric curve estimation with time series errors. Journal of Statistical Planning and Inference 28, 167–183.

Watson, G.S. (1964): Smooth regression analysis. Sankya Series A 26, 359–372.