New Developments in Kernel Regression with Correlated Errors

K. De Brabanter [*] **, J. De Brabanter** [*,**] **, I. Gijbels** [***]
J.A.K. Suykens [*] **, B. De Moor** [*]

[*] *K.U. Leuven, Department of Electrical Engineering ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
{kris.debrabanter,johan.suykens,bart.demoor}@esat.kuleuven.be

[**] *Hogeschool KaHo Sint-Lieven (Associatie K.U.Leuven), Departement Industrieel Ingenieur, B-9000 Gent*
jos.debrabanter@kahosl.be

[***] *K.U. Leuven, Department of Mathematics and Leuven Statistics Research Centre (LStat), Celestijnenlaan 200B, B-3001 Leuven*
irene.gijbels@wis.kuleuven.be

Abstract: It is a well-known problem that obtaining a correct bandwidth/smoothing parameter in nonparametric regression is difficult in the presence of correlated errors. There exist a wide variety of methods coping with this problem, but they all critically depend on a tuning procedure which requires accurate information about the correlation structure. We propose a bandwidth selection procedure based on bimodal kernels which successfully removes the correlation without requiring any prior knowledge about its structure and its parameters. Further, we show that the form of the kernel is very important when errors are correlated which is in contrast to the independent and identically distributed (i.i.d.) case.

Keywords: nonparametric regression, correlated errors, bimodal kernel, cross–validation

1. INTRODUCTION

In nonparametric regression problems, one is interested in estimating the mean function from a set of observations $(x_1, Y_1), \ldots, (x_n, Y_n)$, where the x_i can be either univariate or multivariate. Many methods are currently available, including kernel based methods, smoothing splines, wavelet and Fourier series expansions. The bulk of the literature in these areas has focused on the case where an unknown mean function is *masked* by a certain amount of white noise. The goal of regression is to *remove* the white noise and reveal the function. Suppose the noise is no longer white and instead contains a certain amount of structure in the form of correlation. The focus of this abstract is to look at the problem of estimating the mean function m in the presence of correlation, not that of estimating the correlation function itself. Approaches describing the estimation of the correlation function can be found in Hart and Wehrly (1986), Hart (1991) and Park et al. (2006). In this context we want to (i) explain some of the difficulties associated with the presence of correlation in nonparametric regression and (ii) discuss some recent developments in this area.

Suppose we want to recover the regression function from the following nonparametric regression model

$$Y_i = m(x_i) + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where m is an unknown, smooth function in which the design points are fixed and equally spaced i.e. $x_i \equiv i/n$. Also, we assume that $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma^2 < \infty$ and the sequence $\{e_i\}, i = 1, \ldots, n$ is a covariance stationary process.

Definition 1. (Covariance Stationarity). The sequence $\{e_i\}$ is covariance stationary if

- $\mathbf{E}[e_i] = \mu$ for all i
- $\mathbf{Cov}[e_i, e_{i-j}] = \mathbf{E}[(e_i - \mu)(e_{i-j} - \mu)] = \gamma_j$ for all i and any j.

It is well known when a nonparametric method is used to recover m, correlated errors trouble bandwidth selection severely. Bandwidth selection procedures designed for independent errors, such as cross–validation (CV) (Burman, 1989) and plug-ins (Fan and Gijbels, 1996), will suffer from significant bias. If the errors are positively (negatively) correlated, CV will produce a small (large) bandwidth which results in a rough (oversmooth) estimate of m.

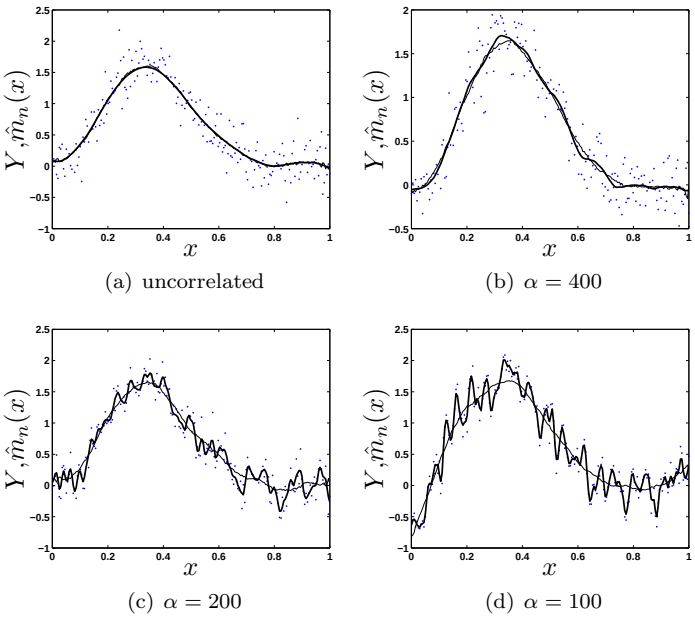
This abstract is organized as follows: In Section 2 the practical difficulties associated with estimating m under model (1) are explained. In Section 3, some recent developments are described. Finally, in Section 4, the capability of the method is illustrated on a real life data set.

2. THE EFFECTS OF CORRELATION

For all nonparametric regression techniques, the shape and the smoothness of the estimated function depends on a large extent on the specific value(s) chosen for the kernel bandwidth (and/or regularization parameter). In

order to avoid selecting values for these parameters by trial and error, several data-driven methods are developed. However, the presence of correlation between the errors, if ignored, causes the commonly used automatic tuning parameter selection methods such as cross–validation (CV) or plug-in, to break down (Opsomer et al., 2001; Francisco-Fernández et al., 2005).

The breakdown of automated methods, as well as a suitable solution, is illustrated by means of a simple example in Figure 1. For 200 equally spaced observations and a polynomial mean function $m(x) = 500x^3(1-x)^6$, four progressively more correlated sets of errors were generated from the same vector of independent noise and added to the mean function. The errors are normally distributed with variance $\sigma^2 = 0.05$ and correlation following an AR(1) process, $\text{corr}(e_i, e_j) = \exp(-\alpha|x_i - x_j|)$ (Fan and Yao, 2003). Figure 1 shows four local linear regression estimates (Fan and Gijbels, 1996) for these data sets. For each data set, the bandwidth was selected according to leave-one-out CV (LOO-CV) with a unimodal kernel K (Epanechnikov) and a kernel satisfying $K(0) = 0$. Table 1 summarizes the bandwidths selected for the four data sets under both methods. Table 1 and Figure 1 clearly show that as the correlation increases, the bandwidth selected by LOO-CV based on a unimodal kernel K becomes smaller and smaller, and the estimates become progressively more undersmoothed. The bandwidths selected by LOO-CV based on a kernel K satisfying $K(0) = 0$ are much more stable and result in virtually the same estimate for all four cases. This type of undersmoothing behavior in the presence of correlated errors has been observed with most commonly used automated bandwidth selection methods (Altman, 1990; Hart, 1991; Opsomer et al., 2001).



(a) uncorrelated (b) $\alpha = 400$ (c) $\alpha = 200$ (d) $\alpha = 100$

Fig. 1. Simulated data with four levels of AR(1) correlation, estimated with local linear regression; (bold line) represents estimate obtained with bandwidth selected by leave-one-out CV; (thin line) estimate selected by leave-one-out CV based on a kernel satisfying $K(0) = 0$.

Table 1. Bandwidth selection for simulated data in Figure 1 based on LOO-CV with unimodal kernel (Epa) and kernel satisfying $K(0) = 0$

Correlation level	Autocorrelation	Epa	$K(0) = 0$
Independent	0	0.11	0.10
$\alpha = 400$	0.14	0.09	0.10
$\alpha = 200$	0.37	0.04	0.11
$\alpha = 100$	0.61	0.02	0.12

3. RECENT DEVELOPMENTS

3.1 Towards Bimodal Kernels

For our theoretical results, consider the Nadaraya-Watson (NW) kernel estimator to estimate the unknown regression function m defined as

$$\hat{m}_n(x) = \sum_{i=1}^{n} \frac{K(\frac{x-x_i}{h})Y_i}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})},$$

where h is the bandwidth of the (unimodal) kernel K. The bandwidth h can for example be selected by minimizing the leave-one-out cross-validation (LOO-CV) score function

$$\text{LOO–CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_n^{(-i)}(x_i;h)\right)^2, \qquad (2)$$

where $\hat{m}_n^{(-i)}(x_i;h)$ denotes the leave-one-out estimator where point i is left out from the training. For notational ease, the dependence on the bandwidth h will be further suppressed. We can now state the following.

Lemma 1. (Hart, 1991; De Brabanter et al., 2010). Assume that the errors are zero-mean and ignoring boundary effects, then the expected value of the LOO-CV score function (2) is given by

$$\mathbf{E}[\text{LOO–CV}(h)] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 - \frac{2}{n}\sum_{i=1}^{n}\mathbf{Cov}\left[\hat{m}_n^{(-i)}(x_i), e_i\right].$$

Note that the last term on the right-hand side in Lemma 1 is in addition to the correlation already included in the first term. Hart (1991) shows, if $n \to \infty$, $nh \to \infty$, $nh^5 \to 0$ and for positively correlated errors, that $\mathbf{E}[\text{LOO–CV}(h)] \approx \sigma^2 + c/nh$ where $c < 0$ and c does not depend on the bandwidth. If the correlation is sufficiently strong and n sufficiently large, $\mathbf{E}[\text{LOO–CV}(h)]$ will be minimized at a value of h that is very near to zero. This result does not only hold for leave-one-out cross-validation but also for Mallow's criterion (Chiu, 1989) and plug-in based techniques (Opsomer et al., 2001). The following theorem provides evidence why a kernel satisfying $K(0) = 0$ efficiently removes the correlation structure without requiring prior knowledge about its structure. Notice that if K is a symmetric probability density function, then $K(0) = 0$ implies that K is not unimodal. Hence, it is obvious to use bimodal kernels. In what follows we will use the following notation

$$k(u) = \int_{-\infty}^{\infty} K(y)e^{-iuy}\,dy$$

for the Fourier Transform of the kernel function K.

Theorem 1. (De Brabanter et al., 2011). Assume the model structure in (1), $\mathbf{E}[e] = 0$, $\mathbf{Cov}[e_i, e_{i+k}] = \mathbf{E}[e_i e_{i+k}] = \gamma_k$ and $\gamma_k \sim k^{-a}$ for some $a > 2$. Assume that

(C1) K is Lipschitz continuous at $x = 0$;
(C2) $\int K(u)\,du = 1, \lim_{|u|\to\infty} |uK(u)| = 0, \int |K(u)|\,du < \infty, \sup_u |K(u)| < \infty$;
(C3) $\int |k(u)|\,du < \infty$ and K is symmetric.

Further, assume that boundary effects are ignored and that $h \to 0$ as $n \to \infty$ such that $nh^2 \to \infty$, then for the NW smoother it follows that

$$\mathbf{E}[\text{LOO–CV}(h)] = \frac{1}{n}\,\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 - \frac{4K(0)}{nh - K(0)}\sum_{k=1}^{\infty}\gamma_k + o(n^{-1}h^{-1}). \tag{3}$$

3.2 Drawback of Bimodal Kernels

Although bimodal kernels are very effective in removing the correlation structure, they have an inherent drawback. When using bimodal kernels to estimate the regression function m, the estimate $\hat{m}_n$ will suffer from increased mean integrated squared error (MISE). It can be shown that, for NW, the asymptotic MISE is equal to

$$\text{AMISE}(\hat{m}_n) = cD_K^{2/5}n^{-4/5},$$

where c depends neither on the bandwidth nor on the kernel K and

$$D_K = \left(\int u^2 K(u)\,du\right)\left(\int K^2(u)\,du\right)^2. \tag{4}$$

It is obvious that one wants to minimize (4) with respect to the kernel function K. This leads to the well-known Epanechnikov kernel. However, adding the constraint $K(0) = 0$ to the minimization of (4) would lead lead to a discontinuous kernel in zero and hence violating condition (C1) of Theorem 1. In fact, using arguments from Kim et al. (2009) and De Brabanter et al. (2011) it can be shown that an optimal kernel does not exist in the class of kernels satisfying assumption (C1) and $K(0) = 0$. Table 2 displays several possible bimodal kernel functions with their respective D_K value compared to the Epanechnikov kernel. On the other hand, we can create an

Table 2. Bimodal kernel functions and their respective D_K value compared to the Epanechnikov (unimodal) kernel. I_A denotes the indicator function of an event A.

	kernel function	$\mathbf{D_K}$
Epa	$\frac{3}{4}(1-u^2)I_{\{\lvert u\rvert\le 1\}}$	0.072
$\tilde{K}_1$	$630(4u^2-1)^2u^4 I_{\{\lvert u\rvert\le 1/2\}}$	0.374
$\tilde{K}_2$	$\frac{2}{\sqrt{\pi}}u^2\exp(-u^2)$	0.134
$\tilde{K}_3$	$\frac{1}{2}\lvert u\rvert\exp(-\lvert u\rvert)$	0.093

ϵ-optimal class of bimodal kernels $\tilde{K}_\epsilon(u)$, with $0 < \epsilon < 1$, defined as (De Brabanter et al., 2011)

$$\tilde{K}_\epsilon(u) = \frac{4}{4-3\epsilon-\epsilon^3}\begin{cases}\frac{3}{4}(1-u^2)I_{\{|u|\le 1\}}, & |u| \ge \epsilon;\\ \frac{3}{4}\frac{1-\epsilon^2}{\epsilon}|u|, & |u| < \epsilon.\end{cases} \tag{5}$$

which approximates the optimal value as $\epsilon \to 0$.

4. REAL LIFE EXAMPLE

We apply the proposed method to a time series of the Beveridge index of wheat prices from the year 1500 to 1869. These data are an annual index of prices at which wheat was sold in European markets. The data used for analysis are the natural logarithms of the Beveridge indices. This transformation is done to correct for heteroscedasticity in the original series. The result is shown in Figure 2 for the NW with Epanechnikov kernel. It is clear that this estimate is very rough. The estimate based on a bimodal kernel $\tilde{K}_\epsilon$ with $\epsilon = 0.05$ (bandwidth obtained with LOO-CV) produces a smooth regression fit.

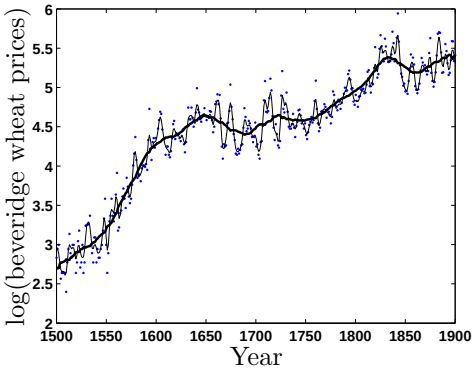


Fig. 2. Difference in NW regression estimates for leave-one-out CV based on Epanechnikov kernel (thin line) and bimodal kernel $\tilde{K}_\epsilon$ with $\epsilon = 0.05$ (bold line).

REFERENCES

Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *J. American Statistical Association*, 85(411), 749–759.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3), 503–514.

Chiu, S.T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statist. Probab. Lett.*, 8(4), 347–354.

De Brabanter, K., De Brabanter, J., Suykens, J., and De Moor, B. (2010). Kernel regression with correlated errors. *Proc. of the 11th International Symposium on Computer Applications in Biotechnology*, pp. 13–18.

De Brabanter, K., De Brabanter, J., Suykens, J., and De Moor, B. (2011). Kernel regression in the presence of correlated errors. Technical Report 11–35, K.U.Leuven, `ftp://ftp.esat.kuleuven.be/pub/SISTA//kdebaban/11-35.pdf`.

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

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer.

Francisco-Fernández, M., Opsomer, J., and Vilar-Fernández, J. (2005). A plug-in bandwidth selector for local polynomial regression estimator with correlated errors. *J. Nonparametr. Stat.*, 18(1–2), 127–151.

Hart, J. (1991). Kernel regression estimation with time series errors. *J. Royal Statist. Soc. B*, 53(1), 173–187.

Hart, J. and Wehrly, T. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.*, 81(396), 1080–1088.

Kim, T.Y., Park, B.U., Moon, M.S., and Kim, C. (2009). Using bimodal kernel inference in nonparametric regression with correlated errors. *J. Multivariate Analysis*, 100, 1487–1497.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, 16(2), 134–153.

Park, B., Lee, Y., Kim, T., and Park, C. (2006). A simple estimator of error correlation in non-parametric regression models. *Scandinavian Journal of Statistics*, 33(3), 451–462.