Kernel Regression with Correlated Errors

K. De Brabanter [*] **, J. De Brabanter** [*,**] **, 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

Abstract: It is a well-known problem that obtaining a correct bandwidth 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. Since the errors cannot be observed, the latter is a hard goal to achieve. In this paper, we show the breakdown of several data-driven parameter selection procedures. We also develop a bandwidth selection procedure based on bimodal kernels which successfully removes the error correlation without requiring any prior knowledge about its structure. Some extensions are made to use such a criterion in least squares support vector machines for regression.

Keywords: nonparametric regression, correlated errors, short-range dependence, bimodal kernel, cross–validation

1. INTRODUCTION

In nonparametric regression problems, one is interested in estimating the mean function $\mathrm{E}[Y|X] = m(X)$ 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 the 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 paper is to look at the problem of estimating the mean function $m(\cdot)$ 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 a new development 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 (uniform) equally spaced i.e. $x_i \equiv i/n$. Also, we assume that $\mathrm{E}[e] = 0$ and is a covariance-stationary process. It is well known that when a nonparametric method is used to recover m, that 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; Opsomer et al., 2001), 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.

Another issue in this context is whether the errors are assumed to be short-range dependent, where the correlation decreases rapidly as the distance between two observations increases or long-range dependent. The latter makes regression estimation even harder and has become an active field of research. Künsch et al. (1993) made the following interesting remark: "*Perhaps most unbelievable to many is the observation that high-quality measurements series from astronomy, physics, chemistry, generally regarded as prototype of i.i.d. observations, are not independent but long-range correlated.*"

Note that in the parametric case, ordinary least squares estimators in the presence of autocorrelation are still linear-unbiased as well as consistent, but they are no longer efficient (i.e., minimum variance). As a result, the usual confidence intervals and the test hypotheses cannot be legitimately applied (Sen and Srivastava, 1990).

This paper is organized as follows: In Section 2 the practical difficulties associated with estimating m under model (1) are explained. In Section 3, some extensions of existing results as well as new developments are described. In Section 4 the proposed CV is adapted to least squares support vector machine for regression. Finally, in Section 5

the results of a simulation study justifying our findings are presented.

2. PROBLEMS WITH CORRELATION

Some quite fundamental problems occur when nonparametric regression is attempted in the presence of correlated errors. 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.

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 $m(x) = 300x^3(1-x)^3$, 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 $\sigma^2 = 0.3$ 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 Nadaraya-Watson kernel regression estimates for these data sets. For each data set, two bandwidth selection methods were used: standard CV and a correlation-corrected CV (CC-CV) which is further discussed in Section 4. 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 CV becomes smaller and smaller, and the estimates become progressively more undersmoothed. The bandwidths selected by CC-CV, a method that accounts for the presence of correlation, 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).

Table 1. Summary of bandwidth selection for simulated data in Figure 1

Correlation level	Autocorrelation	CV	CC-CV
Independent	0	0.081	0.072
$\alpha = 400$	0.14	0.012	0.071
$\alpha = 200$	0.37	0.008	0.075
$\alpha = 100$	0.61	0.006	0.071

3. BANDWIDTH SELECTION WITH BIMODAL KERNELS

To estimate the function m consider the Nadaraya-Watson kernel estimator defined as

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

where h is the bandwidth of the kernel K. This kernel can be one of the following kernels: Epanechnikov, Gaussian, box, triangle,... In this section, we address the problem of finding the bandwidth h when the errors are correlated.

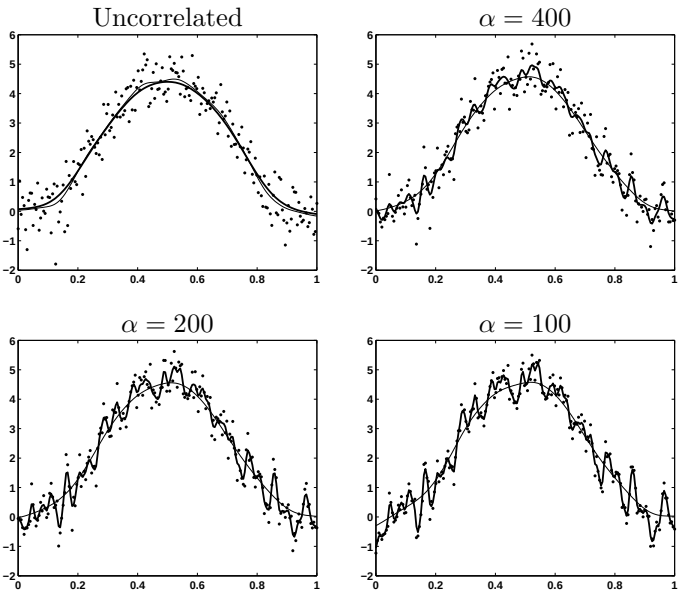

Fig. 1. Simulated data with four levels of AR(1) correlation, estimated with Nadaraya-Watson kernel regression; (bold line) represents estimate obtained with bandwidth selected by CV; (thin line) estimate obtained with bandwidth selected by our method.

Some automated bandwidth selection procedures are based on the minimization of the residual sum of squares given by

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

Taking expectations of (2) results in Lemma 1. This lemma provides insights in why a bimodal kernel is useful in removing the error correlation without having prior knowledge about its structure. Figure 2 illustrates the bimodal kernel $K_{bimod} = 630(4x^2-1)^2x^4 I_{[-1/2,1/2]}(x)$ which will be used in the remaining of the paper.

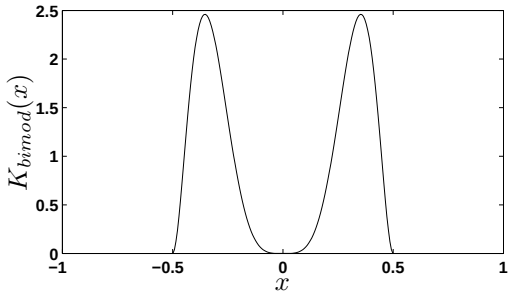

Fig. 2. Illustration of the bimodal kernel used in the paper $K_{bimod} = 630(4x^2-1)^2x^4 I_{[-1/2,1/2]}(x)$.

Lemma 1. Assume the errors are zero-mean, then the expected value of the residual sum of squares (2) is given by

$$\text{E}[\text{RSS}(h)] = \text{E}[\text{MASE}(h)] + \gamma_0 - \frac{2}{n}\sum_{i=1}^{n} \text{Cov}(\hat{m}(x_i), e_i), \tag{3}$$

where $\text{MASE}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - \hat{m}(x_i)\right)^2$ *is the mean average squared error and* $\gamma_0 = \text{E}(e_i e_i), i = 1, \ldots, n$.

Proof: see Appendix A. ■

Lemma 1 shows why CV breaks down in the presence of correlated errors. In this case, the last term in (3) will be positive (negative) for positively (negatively) correlated errors. Hence, the CV will be seriously biased if this is not taken into account. For positively correlated errors, the last term will start canceling the other two resulting in bandwidths which tend to zero for increasing correlation. On the other hand, for negatively correlated errors, it will bias the CV producing larger bandwidths.

Theorem 1. Assume uniform equally spaced design and conditions (A1)–(A3) are satisfied.

(A1) *Assume K is Lipschitz continuous at $x = 0$.*
(A2) *Assume that the covariance is decreasing according to $\gamma_j \sim j^{-a}$ for some $a > 2$.*
(A3) *Assume that the errors are zero-mean covariance stationary process.*

Then, the expected value of the residual sum of squares, in case of the Nadaraya-Watson kernel smoother, is given by

$$\mathrm{E}[\mathrm{RSS}(h)] = \mathrm{E}[\mathrm{MASE}(h)] + \gamma_0 - 2\frac{K(0)}{S(K)}\left[\gamma_0 + 2\sum_{p=1}^{\infty}\gamma_p\right] + o(n^{-1}h^{-1}),$$

where $S(K) = nK(0) + 2\sum_{p=1}^{n-1}(n-p)K\left(\frac{p}{nh}\right)$, $\gamma_p = \mathrm{Cov}[e_{i+p}, e_i], i = 1, \ldots, n$ and $p = 0, \pm 1, \pm 2, \ldots$.

Proof: see Appendix B. ■

It is clear that, by taking a kernel satisfying $K(0) = 0$, the complete correlation structure is removed without requiring any prior information on the structure. Hence, here we propose a bandwidth selector, based on such a kernel, defined by

$$\hat{h}_b = \arg\min_h \mathrm{RSS}(h).$$

Another possibility, not based on bimodal kernels, is to estimate the covariance structure $\gamma_0, \gamma_1, \ldots$ This approach is extensively studied in Hart (1991) and Park et al. (2006).

Notice that if K is a symmetric probability density function, then $K(0) = 0$ implies that K is not unimodal. In this case, it is natural to use bimodal kernels. Such a kernel gives more weight to observations near to the point x of interest than those that are far from x. But in the same time it also reduces the weight of points which are too close to x. In fact, using such a kernel is equivalent with the leave-$(2l+1)$-out version of CV proposed by Chu and Marron (1991).

However, one drawback of using bimodal kernels to estimate m is that the estimate $\hat{m}$ will suffer from increased mean squared error. It can be shown under certain conditions that

$$\mathrm{MASE}(h_M) = cG_K^{2/5}n^{-4/5} + o(n^{-4/5}),$$

where h_M denotes the bandwidth minimizing the mean average squared error, c depends neither on the bandwidth nor on the kernel K and

$$G_K = \left(\int K(u)^2\,du\right)^2 \int u^2K(u)\,du.$$

Using the Epanechnikov kernel $K_{epa} = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$, which is optimal in the L_2 sense, gives $G_{K_{epa}} = 0.072$. On the other hand, using the following bimodal kernel $K_{bimod} = 630(4x^2-1)^2x^4I_{[-1/2,1/2]}(x)$ results in $G_{K_{bimod}} = 0.374$. Reducing this effect can be done by finding a bimodal kernel that makes $\mathrm{MASE}(h_M)$ as small as possible. However, Kim et al. (2009) pointed out that no such kernels exist in the class of smooth bimodal kernels.

4. TOWARDS AN LS-SVM APPROACH WITH CC-CV

Before explaining the CC-CV algorithm, we briefly sketch the least squares support vector machine (LS-SVM) for regression. Further information regarding this topic can be found in Suykens et al. (2002).

4.1 LS-SVM for Regression

In the primal weight space the following optimization problem can be formulated

$$\min_{w,b,e} \mathcal{J}(w,e) = \tfrac{1}{2}w^Tw + \tfrac{\lambda}{2}\sum_{i=1}^{n} e_i^2 \quad \text{s.t.} \quad Y_i = w^T\varphi(x_i) + b + e_i, \quad i = 1, \ldots, n, \tag{4}$$

By using Lagrange multipliers, the solution of (4) can be obtained by taking the Karush-Kuhn-Tucker (KKT) (Bertsekas, 1982) conditions for optimality. The result is given by the following linear system in the dual variables α

$$\left(\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\lambda}I_n \end{array}\right)\left(\frac{b}{\alpha}\right) = \left(\frac{0}{Y}\right),$$

with $Y = (Y_1, \ldots, Y_n)^T$, $1_n = (1, \ldots, 1)^T$, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ and $\Omega_{il} = \varphi(x_i)^T\varphi(x_l) = \mathcal{K}(x_i, x_l)$ for $i, l = 1, \ldots, n$ with $\mathcal{K}(x_i, x_l)$ positive definite. Based on Mercer's theorem, the resulting LS-SVM model for function estimation becomes

$$\hat{m}(x) = \sum_{i=1}^{n}\hat{\alpha}_i\mathcal{K}(x, x_i) + \hat{b}. \tag{5}$$

4.2 Correlation-Corrected Cross–Validation

Theorem 1 stated that a bimodal kernel is well suited to automatically remove the correlation structure. So, in theory we could plug such a kernel into the LS-SVM. However, this method requires a positive definite kernel function and the bimodal kernel does not fulfill this condition. Therefore, this kernel cannot immediately be used but its bandwidth $\hat{h}_b$ can serve as a pilot bandwidth selector for other data-driven selection procedures such as leave-$(2l+1)$-out CV or block bootstrap bandwidth selector (Hall et al., 1995). In this paper we choose the leave-$(2l+1)$-out CV. The leave-$(2l+1)$-out CV can be defined as follows:

Definition 1. The leave-$(2l+1)$-out version of CV (LCV) is defined as

$$LCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{(-i)}(x_i) - Y_i\right)^2. \tag{6}$$

Here $\hat{m}_{(-i)}(x_i)$ is the leave-$(2l+1)$-out version of $m(x_i)$, that is, the observations (x_{i+j}, Y_{i+j}), $-l \le j \le l$, are left out to estimate $\hat{m}(x_i)$.

A crucial parameter to be estimated, see also Chu and Marron (1991), in this procedure is l. Indeed, the amount of dependence between $\hat{m}(x_i)$ and Y_i is reduced as l increases. For $l = 0$, LCV is ordinary CV. One possible method to select a value for l is to use $\hat{h}_b$ as pilot bandwidth selector. Define a bimodal kernel K_{bimod} and $\hat{h}_b$, then it is possible to calculate

$$\hat{m}(x) = \sum_{i=1}^{n} \frac{K_{bimod}(\frac{x-x_i}{\hat{h}_b})Y_i}{\sum_{j=1}^{n} K_{bimod}(\frac{x-x_j}{\hat{h}_b})}. \quad (7)$$

From this result, the residuals are obtained by

$$\hat{e}_i = Y_i - \hat{m}(x_i), \text{ for } i = 1, \ldots, n$$

and choose l to be the smallest $q \geq 1$ such that

$$|r_q| = \left| \frac{\sum_{i=1}^{n-q} \hat{e}_i \hat{e}_{q+i}}{\sum_{i=1}^{n} \hat{e}_i^2} \right| \leq \frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}, \quad (8)$$

where Φ^{-1} denotes the quantile function of the standard normal distribution and α is the significance level, say 5%. Observe that (8) is based on the fact that r_q is asymptotically normal distributed under the centered i.i.d. error assumption (Kendall et al., 1983) and hence provides an approximate $100(1-\alpha)\%$ confidence interval for the autocorrelation.

Once we have selected l, $\hat{h}_{LCV}$ and $\hat{\lambda}_{LCV}$ can be determined by using leave-$(2l+1)$-out CV combined with (5) based on a positive definite unimodal kernel $\mathcal{K}$. Algorithm 1 summarizes the complete CC-CV method based on LS-SVM. Note that this algorithm is applicable to a wide range of smoothers other than LS-SVM such as Nadaraya-Watson kernel regression (see example in Section 2), local polynomial regression, Priestley-Chao kernel estimator, Gasser-Müller kernel estimator, support vector machines,... These types of smoothers can be simply plugged in step 3 of Algorithm 1.

Algorithm 1 Correlation-Corrected CV for LS-SVM

1: Determine $\hat{h}_b$ in (7) with K a bimodal kernel by means of any CV procedure
2: Calculate l satisfying (8)
3: Determine $(\hat{h}_{LCV}, \hat{\lambda}_{LCV})$ for LS-SVM (5) by means of leave-$(2l+1)$-out CV (6) and a positive definite unimodal kernel $\mathcal{K}$, e.g. Gaussian kernel.

5. SIMULATIONS

In this section we will compare the finite sample performance of the CC-CV, see Algorithm 1, with the classical leave-one-out CV (LOO-CV) based on a unimodal kernel. The used kernel functions are $630(4x^2-1)^2x^4 I_{[-1/2,1/2]}(x)$ and $\exp(-x^2)$ for the bimodal and unimodal kernel respectively. The LS-SVM smoother is taken in both cases. An overall comparison is made among $\hat{h}_{LCV}$ and $\hat{h}_{LOO}$ where $\hat{h}_{LCV}$ denotes the bandwidth of the CC-CV method and $\hat{h}_{LOO}$ the bandwidth of the unimodal kernel tuned with LOO-CV. The performance measure is taken to be the MASE(h, γ) for both bandwidths.

The sample size is set to $n = 200$ and the regression function is $m(x) = 300x^3(1-x)^3$ for $0 \leq x \leq 1$. We consider two types of noise models: (i) an AR(5) process $e_j = \sum_{l=1}^{5} \phi_l e_{j-l} + \sqrt{1-\phi_1^2} Z_j$ where Z_j are i.i.d. normal random variables with variance $\sigma^2 = 0.5$ and zero mean. The errors e_j for $j = 1, \ldots, 5$ are standard normal random variables. The AR(5) parameters are set to $[\phi_1, \phi_2, \phi_3, \phi_4, \phi_5] = [0.7, -0.5, 0.4, -0.3, 0.2]$. (ii) m-dependent models $e_i = r_0\delta_i + r_1\delta_{i-1}$ with $m = 1$ where δ_i is i.i.d. standard normal random variable, $r_0 = \frac{\sqrt{1+2\gamma}+\sqrt{1-2\gamma}}{2}$ and $r_1 = \frac{\sqrt{1+2\gamma}-\sqrt{1-2\gamma}}{2}$ for $\gamma = 1/2$.

Table 2 summarizes the average of the regularization parameters, bandwidths and MASE(h, γ) for 50 runs for both noise models. By looking at the MASE(h, γ) it is clear that the tuning parameters obtained by CC-CV result into better estimates. Also notice the small bandwidths and larger regularization constants found by LOO-CV for both noise models. This provides clear evidence that the kernel smoother is trying to model the noise instead of the true underlying function. These findings are also valid if one uses generalized CV or v-fold CV. Figure 3 shows typical results of the regression estimates for both noise models.

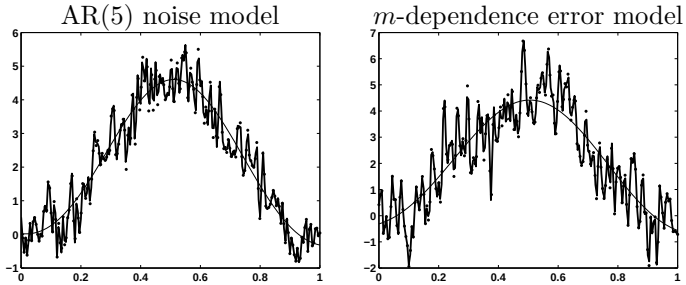

Fig. 3. Typical results of the regression estimates for both noise models. The bold line represents the estimate with tuning parameters determined by LOO-CV and the thin line is the estimate based on the CC-CV tuning parameters.

Table 2. Average of the regularization parameters, bandwidths and MASE for 50 runs for both noise models

	AR(5)		m-dependence models	
	LOO-CV	CC-CV	LOO-CV	CC-CV
$\hat{\lambda}$	224.69	2.28	1.03×10^5	6.96
$\hat{h}$	0.027	1.06	0.03	1.89
MASE$(\hat{h}, \hat{\gamma})$	0.36	0.021	0.89	0.04

Figure 4 and Figure 5 show the CV surfaces for both methods on the AR(5) noise model. These plots clearly demonstrate the shift of the tuning parameters. A cross section for both tuning parameters is provided below each surface plot. Also notice that the surface of the CC-CV tends to be flatter than LOO-CV and so it is harder to minimize numerically (Hall et al., 1995). Because of this extra difficulty, we used a state-of-the-art fast global optimization technique called Coupled Simulated Annealing with variance control (Xavier de Souza et al., 2009) in all the examples.

In a second example we take the same function $m(x)$ and $n = 400$. Further, the noise error model is taken to be an AR(1) process with varying parameter $\phi = -0.95, -0.9, \ldots, 0.9, 0.95$. For each ϕ, 100 replications of size n were made to report the average regularization parameter, bandwidth and MASE for both methods. The

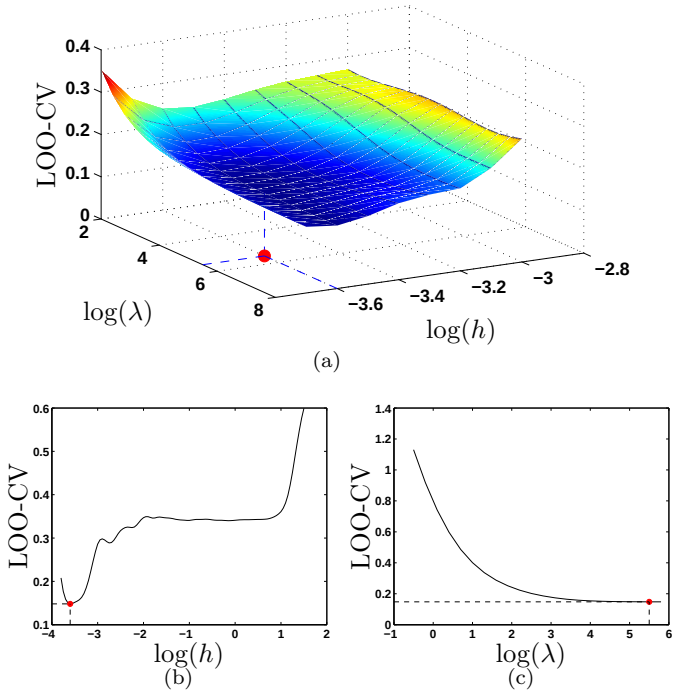

Fig. 4. (a) CV surface for LOO-CV; (b) cross sectional view of $\log(h)$ for fixed $\log(\lambda) = 5.5$; (c) cross sectional view of $\log(\lambda)$ for fixed $\log(h) = -3.6$. The dot indicates the minimum of the cost function obtained by Coupled Simulated Annealing. These results correspond with the first column of Table 2.

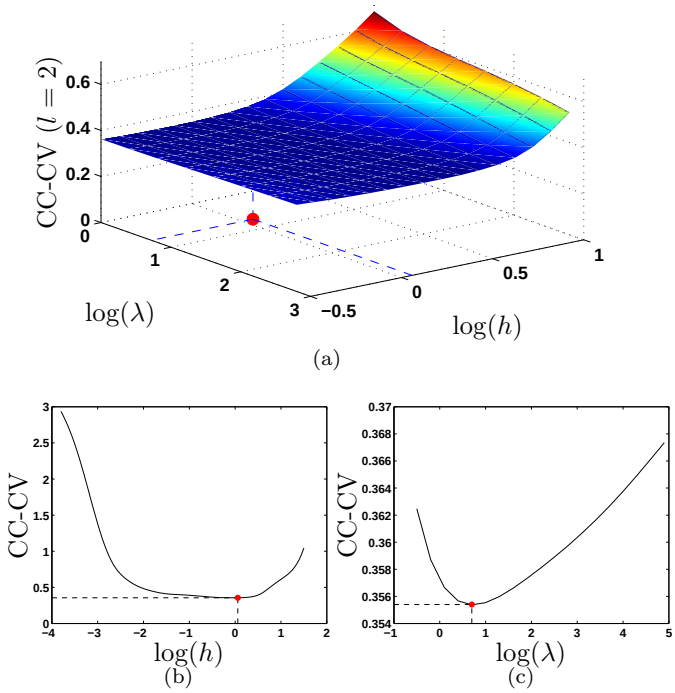

Fig. 5. (a) CV surface for CC-CV; (b) cross sectional view of $\log(h)$ for fixed $\log(\lambda) = 0.82$; (c) cross sectional view of $\log(\lambda)$ for fixed $\log(h) = 0.06$. The dot indicates the minimum of the cost function obtained by Coupled Simulated Annealing. These results correspond with the second column of Table 2.

results are summarized in Table 3. These results indicate that the CC-CV method is indeed capable of finding good tuning parameters in the presence of correlated errors. The CC-CV method outperforms the classical LOO-CV for positively correlated errors, i.e. $\phi > 0$. The method is capable of producing good bandwidths which do not tend to very small values as in the LOO-CV case. In general, the regularization parameter obtained by LOO-CV is larger than the one from CC-CV. However, the latter is not theoretically verified by the author and serves only as a heuristic.

On the other hand, for negatively correlated errors ($\phi < 0$), both methods perform equally well. The reason why the effects from correlated errors is more outspoken for positive ϕ than for negative ϕ might be related to the fact that negatively correlated errors are seemingly hard to differentiate form i.i.d. errors in practice.

Table 3. Average of the regularization parameters, bandwidths and MASE for 50 runs for the AR(1) process with varying parameter ϕ

ϕ	LOO-CV			CC-CV		
	$\hat{\gamma}$	$\hat{h}$	MASE	$\hat{\gamma}$	$\hat{h}$	MASE
-0.95	14.75	1.48	0.0017	7.65	1.43	0.0019
-0.9	11.48	1.47	0.0017	14.58	1.18	0.0021
-0.8	7.52	1.39	0.0021	18.12	1.15	0.0031
-0.7	2.89	1.51	0.0024	6.23	1.21	0.0030
-0.6	28.78	1.52	0.0030	5.48	1.62	0.0033
-0.5	42.58	1.71	0.0031	87.85	1.75	0.0048
-0.4	39.15	1.55	0.0052	39.02	1.43	0.0060
-0.3	72.91	1.68	0.0055	19.76	1.54	0.0061
-0.2	98.12	1.75	0.0061	99.56	1.96	0.0069
-0.1	60.56	1.81	0.0069	101.1	1.89	0.0070
0	102.5	1.45	0.0091	158.4	1.89	0.0092
0.1	1251	1.22	0.0138	209.2	1.88	0.0105
0.2	1893	0.98	0.0482	224.6	1.65	0.0160
0.3	1535	0.66	0.11	5.18	1.86	0.0161
0.4	482.3	0.12	0.25	667.5	1.68	0.023
0.5	2598	0.04	0.33	541.8	1.82	0.033
0.6	230.1	0.03	0.36	986.9	1.85	0.036
0.7	9785	0.03	0.41	12.58	1.68	0.052
0.8	612.1	0.03	0.45	1531	1.53	0.069
0.9	448.8	0.02	0.51	145.12	1.35	0.095
0.95	878.4	0.01	0.66	96.5	1.19	0.13

6. CONCLUSION

We have introduced a type of CV procedure (CC-CV), based on bimodal kernels, in order to automatically remove the error correlation without requiring any prior knowledge about its structure. Since the estimate suffers from increased mean squared error, due to the bimodal kernel, we have used the bandwidth of the bimodal kernel as pilot bandwidth selector for leave-$(2l+1)$-out CV. By taking this extra step, methods like LS-SVM and SVM can be equipped with this technique of handling data in the presence of correlated errors since they require a positive definite kernel. Also other kernel methods which do not require positive definite kernels can benefit from

the proposed method. Finally, we have demonstrated the capability of the method by means of toy examples.

The authors would like to thank Prof. László Györfi for his constructive comments.

ACKNOWLEDGEMENTS

Research supported by: Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06 l, G.0321.06, G.0302.07, G.0320.08, G.0558.08, G.0557.08, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite+; Helmholtz: viCERP. Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, 2007-2011); EU: ERNSI.

REFERENCES

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

Bertsekas, D.P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press.

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.

Chu, C.K. and Marron, J.S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, 19(4), 1906–1918.

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.

Hall, P., Lahiri, S.N., and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *Ann. Statist.*, 23(6), 1921–1936.

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

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

Kendall, M.G., Stuart, A., and Ord, J.K. (1983). *The Advanced Theory of Statistics*, volume 3. Griffin, London, 4th edition.

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.

Künsch, H., Beran, J., and Hampel, F. (1993). Contrasts under long-range correlations. *Ann. Statist.*, 21(2), 943–964.

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

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

Sen, A. and Srivastava, M. (1990). *Regression Analysis: Theory, Methods and Applications*. Springer-Verlag.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Xavier de Souza, S., Suykens, J.A.K., Vandewalle, J., and Bollé, D. (2009). Coupled simulated annealing. *IEEE Transactions on Systems, Man and Cybernetics - Part B*, in press.

Appendix A. PROOF OF LEMMA 1

Since $Y_i = m(x_i)+e_i$ and denote $m(x_i) = m_i, \hat{m}(x_i) = \hat{m}_i$, we can write the following

$$\begin{aligned}
\mathrm{RSS}(h) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(m_i^2 + 2m_i e_i + e_i^2 - 2Y_i\hat{m}_i + \hat{m}_i^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}(m_i - \hat{m}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}e_i^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}(\hat{m}_i - m_i)(m_i - Y_i).
\end{aligned}$$

Taking expectations and using the zero-mean error assumption gives

$$\begin{aligned}
\mathrm{E}[\mathrm{RSS}(h)] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(m_i - \hat{m}_i)^2\right] + \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}e_i^2\right] \\
&\quad + \mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}(\hat{m}_i - m_i)(m_i - Y_i)\right] \\
&= \mathrm{E}[\mathrm{MASE}(h)] + \gamma_0 \\
&\quad + \frac{2}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}(\hat{m}_i - m_i)(m_i - Y_i)\right],
\end{aligned}$$

where $\gamma_p = \mathrm{Cov}[e_{i+p}, e_i], p = 0, \pm1, \pm2, \ldots$. For ease of notation set $A(h) = \frac{2}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}(\hat{m}_i - m_i)(m_i - Y_i)\right]$. By noting that $Y_i = m_i + e_i$ and using the zero-mean error assumption, we obtain

$$\begin{aligned}
A(h) &= -\frac{2}{n}\,\mathrm{E}\left[\sum_{i=1}^{n}(\hat{m}_i - m_i)e_i\right] \\
&= -\frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(\hat{m}_i, e_i),
\end{aligned}$$

which completes the proof.

Appendix B. PROOF OF THEOREM 1

Plugging in the Nadaraya-Watson kernel regression estimator for $\hat{m}_i$ in the last term of Lemma 1 results in

$$A(h) = -\frac{2}{n}\sum_{i=1}^{n}\left(\mathrm{E}\left[\sum_{l=1}^{n}\frac{K(\frac{x_i - x_l}{h})e_l}{\sum_{j=1}^{n}K(\frac{x_i - x_j}{h})}e_i\right]\right).$$

The above expression reduces to

$$\begin{aligned}
A(h) &= -\frac{2}{n}\sum_{i=1}^{n}\sum_{l=1}^{n}\mathrm{E}\left[\frac{K(\frac{x_i - x_l}{h})}{\sum_{j=1}^{n}K(\frac{x_i - x_j}{h})}e_l e_i\right] \\
&= -\frac{2}{n}\frac{1}{S(K)}\left[nK(0)\gamma_0 + 2n\sum_{p=1}^{n-1}\left(\frac{n-p}{n}\right)K\left(\frac{p}{nh}\right)\gamma_p\right],
\end{aligned}$$

where $S(K) = nK(0) + 2\sum_{p=1}^{n-1}(n-p)K\left(\frac{p}{nh}\right)$. Using condition (A1) and $\gamma_p \sim p^{-a}$ for some $a > 2$, we obtain

$$A(h) = -2\frac{K(0)}{S(K)}\left[\gamma_0 + 2\sum_{p=1}^{\infty}\gamma_p\right] + o(n^{-1}h^{-1}).$$