2011

Kernel Regression in the Presence of Correlated Errors

Kris De Brabanter, *KU Leuven*
Jos De Brabanter, *KU Leuven*
Johan A.K. Suykens, *KU Leuven*
Bart De Moor, *KU Leuven*

Journal of Machine Learning Research 12 (2011) 1955-1976 Submitted 2/11; Revised 4/11; Published 6/11

Kernel Regression in the Presence of Correlated Errors

Kris De Brabanter KRIS.DEBRABANTER@ESAT.KULEUVEN.BE
Department of Electrical Engineering SCD-SISTA
K.U. Leuven
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium

Jos De Brabanter JOS.DEBRABANTER@ESAT.KULEUVEN.BE
Departement Industrieel Ingenieur - E&A
KaHo Sint Lieven (Associatie K.U. Leuven)
G. Desmetstraat 1
B-9000 Gent, Belgium

Johan A.K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Bart De Moor BART.DEMOOR@ESAT.KULEUVEN.BE
Department of Electrical Engineering SCD-SISTA
K.U. Leuven
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium

Editor: Xiaotong Shen

Abstract

It is a well-known problem that obtaining a correct bandwidth and/or 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. Finally, some extensions are proposed to use the proposed criterion in support vector machines and least squares support vector machines for regression.

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

1. Introduction

Nonparametric regression is a very popular tool for data analysis because these techniques impose few assumptions about the shape of the mean function. Hence, they are extremely flexible tools for uncovering nonlinear relationships between variables. Given the data $\{(x_1,Y_1),\ldots,(x_n,Y_n)\}$ where $x_i \equiv i/n$ and $x \in [0,1]$ (fixed design). Then, the data can be written as

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

where $e_i = Y_i - m(x_i)$ satisfies $\mathbf{E}[e] = 0$ and $\mathbf{Var}[e] = \sigma^2 < \infty$. Thus Y_i can be considered as the sum of the value of the regression function at x_i and some error e_i with the expected value zero and the sequence $\{e_i\}$ 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.

Many techniques include a smoothing parameter and/or kernel bandwidth which controls the smoothness, bias and variance of the estimate. A vast number of techniques have been developed to determine suitable choices for these tuning parameters from data when the errors are independent and identically distributed (i.i.d.) with finite variance. More detailed information can be found in the books of Fan & Gijbels (1996), Davison & Hinkley (2003) and Konishi & Kitagawa (2008) and the article by Feng & Heiler (2009). However, all the previous techniques have been derived under the i.i.d. assumption. It has been shown that violating this assumption results in the break down of the above methods (Altman, 1990; Hermann, Gasser & Kneip, 1992; Opsomer, Wand & Yang, 2001; Lahiri, 2003). If the errors are positively (negatively) correlated, these methods will produce a small (large) bandwidth which results in a rough (oversmooth) estimate of the regression function. The focus of this paper 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 are extensively studied in Hart & Wehrly (1986), Hart (1991) and Park et al. (2006).

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 error process is said to be short-range dependent if for some $\tau > 0$, $\delta > 1$ and correlation function $\rho(\cdot)$, the spectral density $H(\omega) = \frac{\sigma^2}{2\pi}\sum_{k=-\infty}^{\infty}\rho(k)e^{-i\omega}$ of the errors satisfies (Cox, 1984)

$$H(\omega) \sim \tau\omega^{-(1-\delta)} \text{ as } \omega \to 0,$$

where $A \sim B$ denotes A is asymptotic equivalent to B. In that case, $\rho(j)$ is of order $|j|^{-\delta}$ (Adenstedt, 1974). In case of long-range dependence, the correlation decreases more slowly and regression estimation becomes even harder (Hall, Lahiri & Polzehl, 1995; Opsomer, Wand & Yang, 2001). Here, the decrease is of order $|j|^{-\delta}$ for $0 < \delta \leq 1$. Estimation under long-range dependence has attracted more and more attention in recent years. In many scientific research fields such as astronomy, chemistry, physics and signal processing, the observational errors sometimes reveal long-range dependence. Künsch, Beran & Hampel (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.*"

Further, since Kulkarni et al. (2002) have proven consistency for the data-dependent kernel estimators, that is, correlated errors and/or correlation among the independent variables, there is no need to alter the kernel smoother by adding constraints. Confirming their results, we show that the problem is due to the model selection criterion. In fact, we will show in Section 3 that there exists

a simple multiplicative relation between the bandwidth under correlation and the bandwidth under the i.i.d. assumption.

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 & Srivastava, 1990).

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 breakdown of commonly used automatic tuning parameter selection methods such as cross-validation (CV) or plug-in.

Data-driven bandwidth selectors tend to be "fooled" by the correlation, interpreting it as reflecting the regression relationship and variance function. So, the cyclical pattern in positively correlated errors is viewed as a high frequency regression relationship with small variance, and the bandwidth is set small enough to track the cycles resulting in an undersmoothed fitted regression curve. The alternating pattern above and below the true underlying function for negatively correlated errors is interpreted as a high variance, and the bandwidth is set high enough to smooth over the variability, producing an oversmoothed fitted regression curve.

The breakdown of automated methods, as well as a suitable solution, is illustrated by means of a simple example shown 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 Auto Regressive process of order 1, denoted by AR(1), $\text{corr}(e_i, e_j) = \exp(-\alpha|x_i - x_j|)$ (Fan & Yao, 2003). Figure 1 shows four local linear 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 3. Table 1 summarizes the bandwidths selected for the four data sets under both methods.

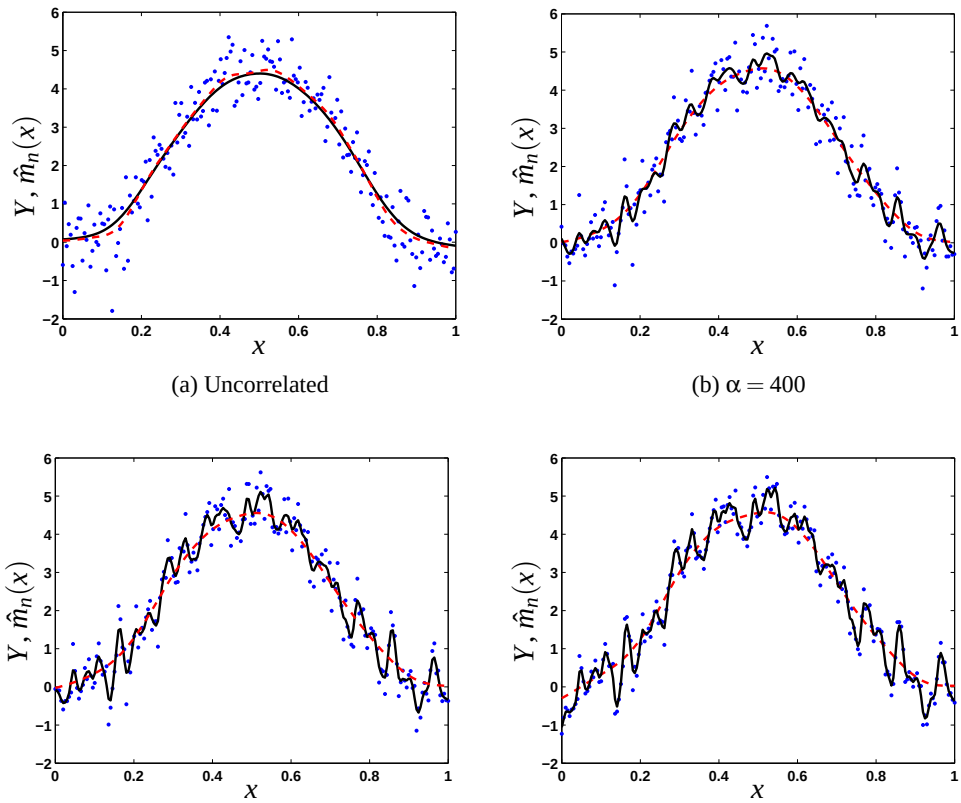
Table 1 and Figure 1 clearly show that when correlation increases, the bandwidth selected by CV becomes smaller and smaller, and the estimates become more undersmoothed. The bandwidths selected by CC-CV (explained in Section 3), a method that accounts for the presence of correlation, are much more stable and result in virtually the same estimate for all four cases. This type of undersmoothing behavior in the presence of positively correlated errors has been observed with most commonly used automated bandwidth selection methods (Altman, 1990; Hart, 1991; Opsomer, Wand & Yang, 2001; Kim et al., 2009).

3. New Developments in Kernel Regression with Correlated Errors

In this Section, we address how to deal with, in a simple but effective way, correlated errors using CV. We make a clear distinction between kernel methods requiring no positive definite kernel and kernel methods requiring a positive definite kernel. We will also show that the form of the kernel,

Correlation level	Autocorrelation	CV	CC-CV
Independent	0	0.09	0.09
$\alpha = 400$	0.14	0.034	0.12
$\alpha = 200$	0.37	0.0084	0.13
$\alpha = 100$	0.61	0.0072	0.13

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



(a) Uncorrelated

(b) $\alpha = 400$

(c) $\alpha = 200$

(d) $\alpha = 100$

Figure 1: Simulated data with four levels of AR(1) correlation, estimated with local linear regression; full line represents the estimates obtained with bandwidth selected by CV; dashed line represents the estimates obtained with bandwidth selected by our method.

based on the mean squared error, is very important when errors are correlated. This is in contrast with the i.i.d. case where the choice between the various kernels, based on the mean squared error, is not very crucial (Härdle, 1999). In what follows, the kernel K is assumed to be an isotropic kernel.

3.1 No Positive Definite Kernel Constraint

To estimate the unknown regression function m, consider the Nadaraya-Watson (NW) kernel estimator (Nadaraya, 1964; Watson, 1964) 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 kernel K. This kernel can be one of the following kernels: Epanechnikov, Gaussian, triangular, spline, etc. An optimal h can for example be found by minimizing the leave-one-out cross-validation (LCV) score function

$$\mathrm{LCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_n^{(-i)}(x_i;h)\right)^2, \tag{2}$$

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 suppressed. We can now state the following.

Lemma 2 *Assume that the errors are zero-mean, then the expected value of the LCV score function* (2) *is given by*

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

Proof: see Appendix A. ■

Note that the last term on the right-hand side in Lemma 2 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 $\mathbf{E}[\mathrm{LCV}(h)] \approx \sigma^2 + c/nh$ where $c < 0$ and c does not depend on the bandwidth. If the correlation is sufficiently strong and n sufficiently large, $\mathbf{E}[\mathrm{LCV}(h)]$ will be minimized at a value of h that is very near to zero. The latter corresponds to almost interpolating the data (see Figure 1). 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, Wand & Yang, 2001). The following theorem provides a simple but effective way to deal with correlated errors. 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 3 *Assume uniform equally spaced design,* $x \in [0,1]$, $\mathbf{E}[e] = 0$, $\mathbf{Cov}[e_i, e_{i+k}] = \mathbf{E}[e_i e_{i+k}] = \gamma_k$ *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.*

Assume further 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

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

Proof: see Appendix B. ■

Remark 4 *There are no major changes in the proof if we consider other smoothers such as Priestley-Chao and local linear regression. In fact, it is well-known that the local linear estimate is the local constant estimate (Nadaraya-Watson) plus a correction for local slope of the data and skewness of the data point under consideration. Following the steps of the proof of Theorem 3 for the correction factor will yield a similar result.*

From this result it is clear that, by taking a kernel satisfying the condition $K(0) = 0$, the correlation structure is removed without requiring any prior information about its structure and (3) reduces to

$$\mathbf{E}[\mathrm{LCV}(h)] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 + o(n^{-1}h^{-1}). \tag{4}$$

Therefore, it is natural to use a bandwidth selection criterion based on a kernel satisfying $K(0) = 0$, defined by

$$\hat{h}_b = \arg\min_{h \in Q_n} \mathrm{LCV}(h),$$

where Q_n is a finite set of parameters. 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. Such a kernel gives more weight to observations near to the point x of interest than those that are far from x. But at the same time it also reduces the weight of points which are too close to x. A major advantage of using a bandwidth selection criterion based on bimodal kernels is the fact that is more efficient in removing the correlation than leave-$(2l+1)$-out CV (Chu & Marron, 1991).

Definition 5 (Leave-$(2l+1)$-out CV) *Leave-$(2l+1)$-out CV or modified CV (MCV) is defined as*

$$\mathrm{MCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_n^{(-i)}(x_i)\right)^2, \tag{5}$$

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

Taking a bimodal kernel satisfying $K(0) = 0$ results in Equation (4) while leave-$(2l+1)$-out CV with unimodal kernel K, under the conditions of Theorem 3, yields

$$\mathbf{E}[\mathrm{MCV}(h)] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 - \frac{4K(0)}{nh - K(0)}\sum_{k=l+1}^{\infty}\gamma_k + o(n^{-1}h^{-1}).$$

The formula above clearly shows that leave-$(2l+1)$-out CV with unimodal kernel K cannot completely remove the correlation structure. Only the first l elements of the correlation are removed.

Another possibility of bandwidth selection under correlation, not based on bimodal kernels, is to estimate the covariance structure $\gamma_0, \gamma_1, \ldots$ in Equation (3). Although the usual residual-based estimators of the autocovariances $\hat{\gamma}_k$ are consistent, $\sum_{k=1}^{\infty} \hat{\gamma}_k$ is not a consistent estimator of $\sum_{k=1}^{\infty} \gamma_k$ (Simonoff, 1996). A first approach correcting for this, is to estimate $\sum_{k=1}^{\infty} \gamma_k$ by fitting a parametric model to the residuals (and thereby obtaining estimates of γ_k) and use these estimates in Equation (3) together with a univariate kernel. If the assumed parametric model is incorrect, these estimates can be far from the correct ones resulting in a poor choice of the bandwidth. However, Altman (1990) showed that, if the signal to noise ratio is small, this approach results in sufficiently good estimates of correlation for correcting the selection criteria. A second approach, proposed by Hart (1989, 1991), suggests estimating the covariance structure in the spectral domain via differencing the data at least twice. A third approach is to derive an asymptotic bias-variance decomposition under the correlated error assumption of the kernel smoother. In this way and under certain conditions on the correlation function, plug-ins can be derived taking the correlation into account, see Hermann, Gasser & Kneip (1992), Opsomer, Wand & Yang (2001), Hall & Van Keilegom (2003), Francisco-Fernández & Opsomer (2004) and Francisco-Fernández et al. (2005). More recently, Park et al. (2006) proposed to estimate the error correlation nonparametrically without prior knowledge of the correlation structure.

3.2 Positive Definite Kernel Constraint

Methods like support vector machines (SVM) (Vapnik, 1999) and least squares support vector machines (LS-SVM) (Suykens et al., 2002) require a positive (semi) definite kernel (see Appendix C for more details on LS-SVM for regression). However, the following theorem reveals why a bimodal kernel $\tilde{K}$ cannot be directly applied in these methods.

Theorem 6 *A bimodal kernel $\tilde{K}$, satisfying $\tilde{K}(0) = 0$, is never positive (semi) definite.*

Proof: see Appendix D. ■

Consequently, the previous strategy of using bimodal kernels cannot directly be applied to SVM and LS-SVM. A possible way to circumvent this obstacle, is to use the bandwidth $\hat{h}_b$, obtained from the bimodal kernel, as a pilot bandwidth selector for other data-driven selection procedures such as leave-$(2l+1)$-out CV or block bootstrap bandwidth selector (Hall, Lahiri & Polzehl, 1995). Since the block bootstrap in Hall, Lahiri & Polzehl (1995) is based on two smoothers, that is, one is used to compute centered residuals and the other generates bootstrap data, the procedure is computationally costly. Therefore, we will use leave-$(2l+1)$-out CV or MCV which has a lower computational cost. A crucial parameter to be estimated in MCV, see also Chu & Marron (1991), is l. Indeed, the amount of dependence between $\hat{m}_n(x_k)$ and Y_k is reduced as l increases.

A similar problem arises in block bootstrap where the accuracy of the method critically depends on the block size that is supplied by the user. The orders of magnitude of the optimal block sizes are known in some inference problems (see Künsch, 1989; Hall, Horowitz & Jing, 1995; Lahiri, 1999; Bühlmann & Künsch, 1999). However, the leading terms of these optimal block sizes depend on various population characteristics in an intricate manner, making it difficult to estimate these parameters in practice. Recently, Lahiri et al. (2007) proposed a nonparametric plug-in principle to determine the block size.

For $l = 0$, MCV is ordinary CV or leave-one-out CV. One possible method to select a value for l is to use $\hat{h}_b$ as pilot bandwidth selector. Define a bimodal kernel $\tilde{K}$ and assume $\hat{h}_b$ is available, then

one can calculate

$$\hat{m}_n(x) = \sum_{i=1}^{n} \frac{\tilde{K}\left(\frac{x-x_i}{\hat{h}_b}\right) Y_i}{\sum_{j=1}^{n} \tilde{K}\left(\frac{x-x_j}{\hat{h}_b}\right)}. \tag{6}$$

From this result, the residuals are obtained by

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

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

$$|r_q| = \left| \frac{\sum_{i=1}^{n-q} \hat{e}_i \hat{e}_{i+q}}{\sum_{i=1}^{n} \hat{e}_i^2} \right| \leq \frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}, \tag{7}$$

where Φ^{-1} denotes the quantile function of the standard normal distribution and α is the significance level, say 5%. Observe that Equation (7) is based on the fact that r_q is asymptotically normal distributed under the centered i.i.d. error assumption (Kendall, Stuart & Ord, 1983) and hence provides an approximate $100(1-\alpha)\%$ confidence interval for the autocorrelation. The reason why Equation (7) can be legitimately applied is motivated by combining the theoretical results of Kim et al. (2004) and Park et al. (2006) stating that

$$\frac{1}{n-q} \sum_{i=1}^{n-q} \hat{e}_i \hat{e}_{i+q} = \frac{1}{n-q} \sum_{i=1}^{n-q} e_i e_{i+q} + O(n^{-4/5}).$$

Once l is selected, the tuning parameters of SVM or LS-SVM can be determined by using leave-$(2l+1)$-out CV combined with a positive definite kernel, for example, Gaussian kernel. We then call Correlation-Corrected CV (CC-CV) the combination of finding l via bimodal kernels and using the obtained l in leave-$(2l+1)$-out CV. Algorithm 1 summarizes the CC-CV procedure for LS-SVM. This procedure can also be applied to SVM for regression.

Algorithm 1 Correlation-Corrected CV for LS-SVM Regression

1: Determine $\hat{h}_b$ in Equation (6) with a bimodal kernel by means of LCV
2: Calculate l satisfying Equation (7)
3: Determine both tuning parameters for LS-SVM by means of leave-$(2l+1)$-out CV Equation (5) and a positive definite unimodal kernel.

3.3 Drawback of Using 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 squared error (MSE). The following theorem indicates the asymptotic behavior of the MSE of $\hat{m}_n(x)$ when the errors are covariance stationary.

Theorem 7 (Simonoff, 1996) *Let Equation* (1) *hold and assume that m has two continuous derivatives. Assume also that* $\mathbf{Cov}[e_i, e_{i+k}] = \gamma_k$ *for all k, where* $\gamma_0 = \sigma^2 < \infty$ *and* $\sum_{k=1}^{\infty} k|\gamma_k| < \infty$. *Now, as*

$n \to \infty$ and $h \to 0$, the following statement holds uniformly in $x \in (h, 1-h)$ for the Mean Integrated Squared Error (MISE)

$$\mathrm{MISE}(\hat{m}_n) = \frac{\mu_2^2(K) h^4 \int (m''(x))^2 \, dx}{4} + \frac{R(K)[\sigma^2 + 2\sum_{k=1}^{\infty} \gamma_k]}{nh} + o(h^4 + n^{-1}h^{-1}),$$

where $\mu_2(K) = \int u^2 K(u)\, du$ and $R(K) = \int K^2(u)\, du$.

An asymptotic optimal constant or global bandwidth $\hat{h}_{\mathrm{AMISE}}$, for $m''(x) \neq 0$, is the minimizer of the Asymptotic MISE (AMISE)

$$\mathrm{AMISE}(\hat{m}_n) = \frac{\mu_2^2(K) h^4 \int (m''(x))^2 \, dx}{4} + \frac{R(K)[\sigma^2 + 2\sum_{k=1}^{\infty} \gamma_k]}{nh},$$

w.r.t. to the bandwidth, yielding

$$\hat{h}_{\mathrm{AMISE}} = \left[\frac{R(K)[\sigma^2 + 2\sum_{k=1}^{\infty} \gamma_k]}{\mu_2^2(K) \int (m''(x))^2 \, dx} \right]^{1/5} n^{-1/5}. \tag{8}$$

We see that $\hat{h}_{\mathrm{AMISE}}$ is at least as big as the bandwidth for i.i.d data $\hat{h}_0$ if $\gamma_k \geq 0$ for all $k \geq 1$. The following corollary shows that there is a simple multiplicative relationship between the asymptotic optimal bandwidth for dependent data $\hat{h}_{\mathrm{AMISE}}$ and bandwidth for independent data $\hat{h}_0$.

Corollary 8 *Assume the conditions of Theorem 7 hold, then*

$$\hat{h}_{\mathrm{AMISE}} = \left[1 + 2 \sum_{k=1}^{\infty} \rho(k) \right]^{1/5} \hat{h}_0, \tag{9}$$

where $\hat{h}_{\mathrm{AMISE}}$ is the asymptotic MISE optimal bandwidth for dependent data, $\hat{h}_0$ is the asymptotic optimal bandwidth for independent data and $\rho(k)$ denotes the autocorrelation function at lag k, that is, $\rho(k) = \gamma_k/\sigma^2 = \mathbf{E}[e_i e_{i+k}]/\sigma^2$.

Proof: see Appendix E. ■

Thus, if the data are positively autocorrelated ($\rho(k) \geq 0 \quad \forall k$), the optimal bandwidth under correlation is larger than that for independent data. Unfortunately, Equation (9) is quite hard to use in practice since it requires knowledge about the correlation structure and an estimate of the bandwidth $\hat{h}_0$ under the i.i.d. assumption, given correlated data. By taking $\hat{h}_{\mathrm{AMISE}}$ as in Equation (8), the corresponding asymptotic MISE is equal to

$$\mathrm{AMISE}(\hat{m}_n) = c D_K^{2/5} n^{-4/5},$$

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

$$D_K = \mu_2(K) R(K)^2 = \left(\int u^2 K(u)\, du \right) \left(\int K^2(u)\, du \right)^2. \tag{10}$$

It is obvious that one wants to minimize Equation (10) with respect to the kernel function K. This leads to the well-known Epanechnikov kernel K_{epa}. However, adding the constraint $K(0) = 0$ (see Theorem 3) to the minimization of Equation (10) would lead to the following optimal kernel

$$K^{\star}(u) = \begin{cases} K_{\text{epa}}(u), & \text{if } u \neq 0; \\ 0, & \text{if } u = 0. \end{cases}$$

Certainly, this kernel violates assumption (C1) in Theorem 3. In fact, an optimal kernel does not exist in the class of kernels satisfying assumption (C1) and $K(0) = 0$. To illustrate this, note that there exist a sequence of kernels $\{K_{\text{epa}}(u,\varepsilon)\}_{\varepsilon \in]0,1[}$, indexed by ε, such that $K_{\text{epa}}(u)$ converges to $K^{\star}(u)$ and the value of $\int K_{\text{epa}}(u,\varepsilon)^2 du$ decreases to $\int K^{\star}(u)^2 du$ as ε tends to zero. Since an optimal kernel in this class cannot be found, we have to be satisfied with a so-called ε-optimal class of bimodal kernels $\tilde{K}_{\varepsilon}(u)$, with $0 < \varepsilon < 1$, defined as

$$\tilde{K}_{\varepsilon}(u) = \frac{4}{4 - 3\varepsilon - \varepsilon^3} \begin{cases} \frac{3}{4}(1-u^2) I_{\{|u| \leq 1\}}, & |u| \geq \varepsilon; \\ \frac{3}{4}\frac{1-\varepsilon^2}{\varepsilon}|u|, & |u| < \varepsilon. \end{cases}$$

For $\varepsilon = 0$, we define $\tilde{K}_{\varepsilon}(u) = K_{\text{epa}}(u)$. Table 2 displays several possible bimodal kernel functions with their respective D_K value compared to the Epanechnikov kernel. Although it is possible to express the D_K value for $\tilde{K}_{\varepsilon}(u)$ as a function of ε, we do not include it in Table 2 but instead, we graphically illustrate the dependence of D_K on ε in Figure 2a. An illustration of the ε-optimal class of bimodal kernels is shown in Figure 2b.

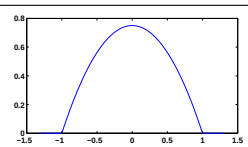
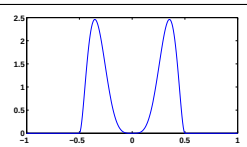
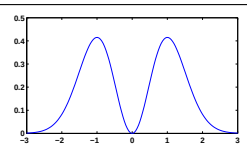
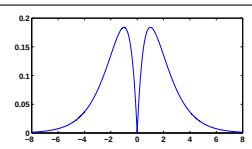
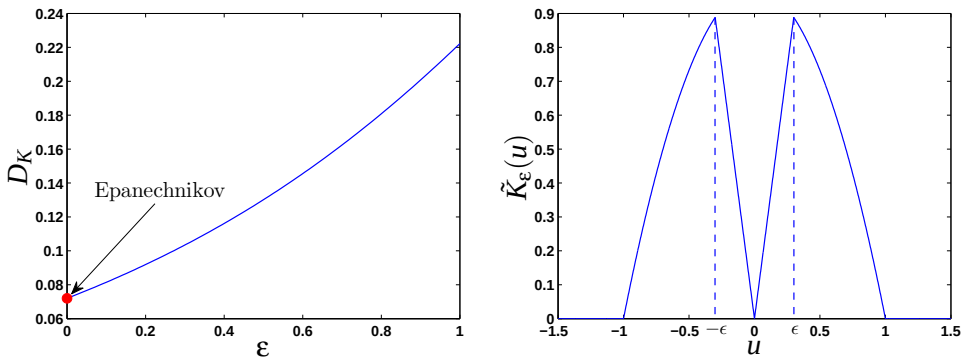
	kernel function	**Illustration**	$\mathbf{D_K}$
K_{epa}	$\frac{3}{4}(1-u^2)I_{\{\lvert u\rvert \leq 1\}}$		0.072
$\tilde{K}_1$	$630(4u^2-1)^2 u^4 I_{\{\lvert u\rvert \leq 1/2\}}$		0.374
$\tilde{K}_2$	$\frac{2}{\sqrt{\pi}} u^2 \exp(-u^2)$		0.134
$\tilde{K}_3$	$\frac{1}{2}\lvert u\rvert \exp(-\lvert u\rvert)$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

Table 2: Kernel functions with illustrations and their respective D_K value compared to the Epanechnikov kernel. I_A denotes the indicator function of an event A.

Remark 9 *We do not consider ε as a tuning parameter but the user can set its value. By doing this one should be aware of two aspects. First, one should choose the value of ε so that its D_K*



(a) (b)

Figure 2: (a) D_K as a function of ε for the ε-optimal class of kernels. The dot on the left side marks the Epanechnikov kernel; (b) Illustration of the ε-optimal class of kernels for $\varepsilon = 0.3$.

value is lower than the D_K value of kernel $\tilde{K}_3$. This is fulfilled when $\varepsilon < 0.2$. Second, by choosing ε extremely small (but not zero) some numerical difficulties may arise. We have experimented with several values of ε and we concluded that the value taken in the remaining of the paper, that is, $\varepsilon = 0.1$ is small enough and it does not show any numerical problems. In theory, there is indeed a difference between kernel $\tilde{K}_3$ and the ε-optimal class of bimodal kernels. However, in practice the difference is rather small. One can compare it with the i.i.d. case where the Epanechnikov kernel is the optimal kernel, but in practice the difference with say a Gaussian kernel is negligible.

4. Simulations

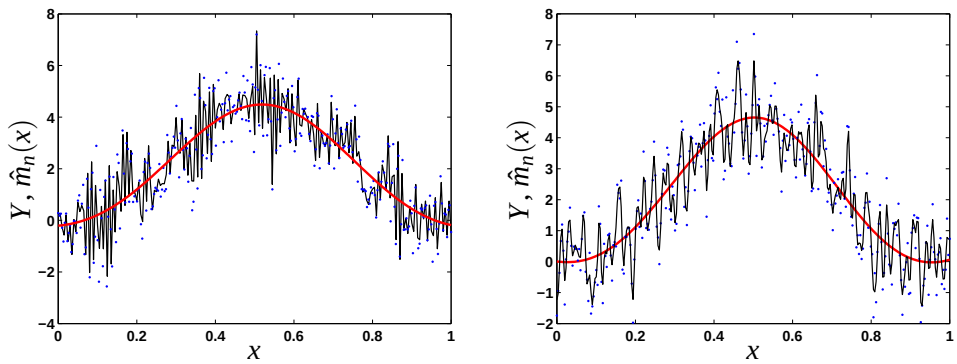
In this Section, we illustrate the capability of the proposed method on several toy examples corrupted with different noise models as well as a real data set.

4.1 CC-CV vs. LCV with Different Noise Models

In a first example, we compare the finite sample performance of CC-CV (with $\tilde{K}_\varepsilon$ and $\varepsilon = 0.1$ in the first step and the Gaussian kernel in the second step) to the classical leave-one-out CV (LCV) based on the Epanechnikov (unimodal) kernel in the presence of correlation. Consider the following function $m(x) = 300x^3(1-x)^3$ for $0 \le x \le 1$. The sample size is set to $n = 200$. 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 data is generated according to Equation (1). 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, $r_0 = \frac{\sqrt{1+2\nu}+\sqrt{1-2\nu}}{2}$ and $r_1 = \frac{\sqrt{1+2\nu}-\sqrt{1-2\nu}}{2}$ for $\nu = 1/2$.

Figure 3 shows typical results of LS-SVM regression estimates for both noise models. Table 3 summarizes the average of the regularization parameters $\hat{\gamma}$, bandwidths $\hat{h}$ and asymptotic squared error, defined as $\text{ASE} = \frac{1}{n}\sum_{i=1}^{n}(m(x_i) - \hat{m}_n(x_i))^2$, for 200 runs for both noise models. By looking at the average ASE, it is clear that the tuning parameters obtained by CC-CV result into better

estimates which are not influenced by the correlation. Also notice the small bandwidths and larger regularization constants found by LCV 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 4 and Figure 5 show the CV surfaces



(a) AR(5) (b) m-dependence models

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

	AR(5)		m-dependence models	
	LCV	CC-CV	LCV	CC-CV
av. $\hat{\gamma}$	226.24	2.27	1.05×10^5	6.87
av. $\hat{h}$	0.014	1.01	0.023	1.88
av. ASE	0.39 (2.9×10^{-2})	0.019 (9.9×10^{-4})	0.90 (8.2×10^{-2})	0.038 (1.4×10^{-3})

Table 3: Average of the regularization parameters $\hat{\gamma}$, bandwidths $\hat{h}$ and average ASE for 200 runs for both noise models. The standard deviation is given between parenthesis.

for both model selection methods on the AR(5) noise model corresponding to the model selection of the estimate in Figure 3(a). These plots clearly demonstrate the shift of the tuning parameters. A cross section for both tuning parameters is provided below each surface plot. Also note that the surface of the CC-CV tends to be flatter than LCV and so it is harder to minimize numerically (see Hall, Lahiri & Polzehl, 1995).

4.2 Evolution of the Bandwidth Under Correlation

Consider the same function as in the previous simulation and let $n = 400$. 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 average ASE for both methods. The results are summarized in Table 4. We used the $\tilde{K}_\varepsilon$ kernel with $\varepsilon = 0.1$ in the first step and the Gaussian kernel in the second step for CC-CV and the Gaussian kernel for classical leave-one-out CV (LCV). The 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

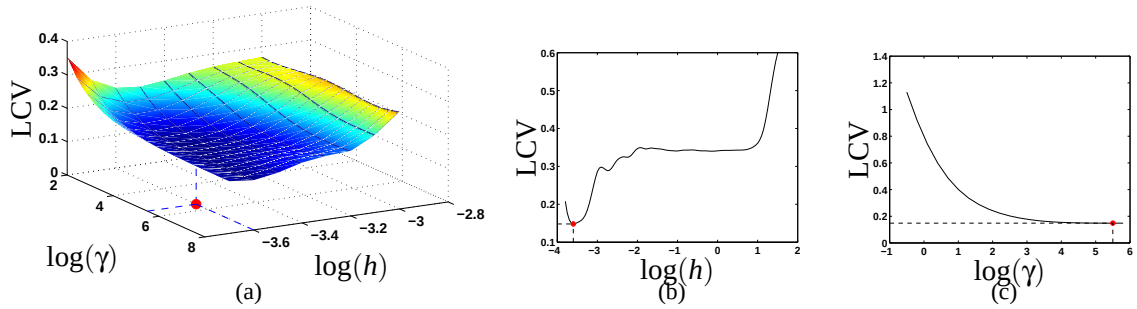


Figure 4: (a) CV surface for LCV; (b) cross sectional view of $\log(h)$ for fixed $\log(\gamma) = 5.5$; (c) cross sectional view of $\log(\gamma)$ for fixed $\log(h) = -3.6$. The dot indicates the minimum of the cost function. This corresponds to the model selection of the wiggly estimate in Figure 3(a).

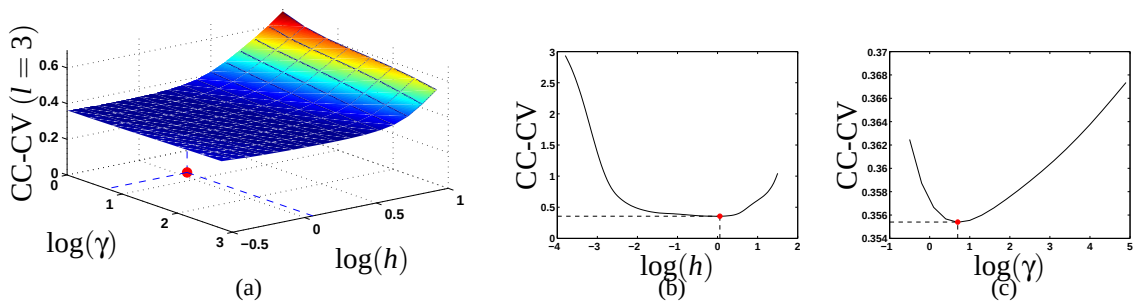


Figure 5: (a) CV surface for CC-CV; (b) cross sectional view of $\log(h)$ for fixed $\log(\gamma) = 0.82$; (c) cross sectional view of $\log(\gamma)$ for fixed $\log(h) = 0.06$. The dot indicates the minimum of the cost function. This corresponds to the model selection of the smooth estimate in Figure 3(a).

outperforms the classical LCV for positively correlated errors, that is, $\phi > 0$. The method is capable of producing good bandwidths which do not tend to very small values as in the LCV case.

In general, the regularization parameter obtained by LCV is larger than the one from CC - CV. However, the latter is not theoretically verified 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.

4.3 Comparison of Different Bimodal Kernels

Consider a polynomial mean function $m(x_k) = 300x_k^3(1-x_k)^3$, $k = 1,\ldots,400$, where the errors are normally distributed with variance $\sigma^2 = 0.1$ and correlation following an AR(1) process, $\mathrm{corr}(e_i, e_j) = \exp(-150|x_i - x_j|)$. The simulation shows the difference in regression estimates (Nadaraya-Watson) based on kernels $\tilde{K}_1$, $\tilde{K}_3$ and $\tilde{K}_\varepsilon$ with $\varepsilon = 0.1$, see Figure 6a and 6b respectively. Due to the larger D_K value of $\tilde{K}_1$, the estimate tends to be more wiggly compared to kernel $\tilde{K}_3$. The difference be-

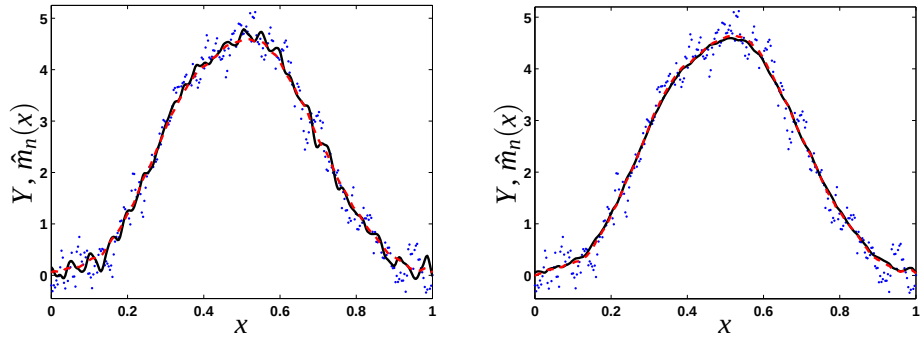
ϕ	LCV			CC-CV		
	$\hat{\gamma}$	$\hat{h}$	av. ASE	$\hat{\gamma}$	$\hat{h}$	av. ASE
-0.95	14.75	1.48	0.0017	7.65	1.43	0.0019
-0.9	11.48	1.47	0.0017	14.58	1.18	0.0021
-0.8	7.52	1.39	0.0021	18.12	1.15	0.0031
-0.7	2.89	1.51	0.0024	6.23	1.21	0.0030
-0.6	28.78	1.52	0.0030	5.48	1.62	0.0033
-0.5	42.58	1.71	0.0031	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

Table 4: Average of the regularization parameters, bandwidths and average ASE for 100 runs for the AR(1) process with varying parameter ϕ

tween the regression estimate based on $\tilde{K}_3$ and $\tilde{K}_\varepsilon$ with $\varepsilon = 0.1$ is very small and almost cannot be seen on Figure 6b. For illustration purposes we did not visualize the result based on kernel $\tilde{K}_2$. For the sake of comparison between regression estimates based on $\tilde{K}_1$, $\tilde{K}_2$,$\tilde{K}_3$ and $\tilde{K}_\varepsilon$ with $\varepsilon = 0.1$, we show the corresponding asymptotic squared error (ASE) in Figure 7 based on 100 simulations with the data generation process described as above. The boxplot shows that the kernel $\tilde{K}_\varepsilon$ with $\varepsilon = 0.1$ outperforms the other three.

4.4 Real Life Data Set

We apply the proposed method to a time series of the Beveridge (1921) index of wheat prices from the year 1500 to 1869 (Anderson, 1971). 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 (no other preprocessing was performed). The result is shown in Figure 8 for LS-SVM with Gaussian kernel. It is clear that the estimate based on classical leave-one-out CV (assumption of no correlation) is very rough. The proposed CC-CV method produces a smooth regression fit. The selected parameters $(\hat{\gamma}, \hat{h})$ for LS-SVM are $(15.61, 29.27)$ and $(96.91, 1.55)$ obtained by CC-CV and LCV respectively.



(a) (b)

Figure 6: Difference in the regression estimate (Nadaraya-Watson) (a) based on kernel $\tilde{K}_1$ (full line) and $\tilde{K}_3$ (dashed line). Due to the larger D_K value of $\tilde{K}_1$, the estimate tends to be more wiggly compared to $\tilde{K}_3$; (b) based on kernel $\tilde{K}_3$ (full line) and ε-optimal kernel with $\varepsilon = 0.1$ (dashed line).

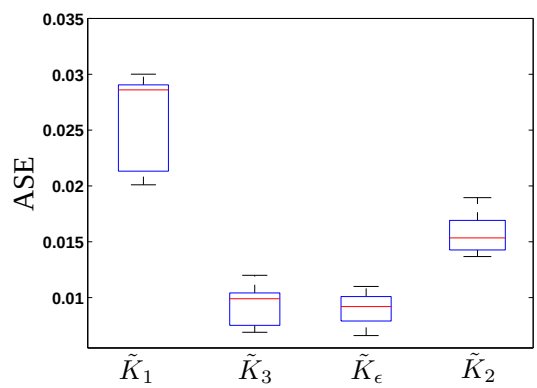


Figure 7: Boxplot of the asymptotic squared errors for the regression estimates based on bimodal kernels $\tilde{K}_1$, $\tilde{K}_2$, $\tilde{K}_3$ and $\tilde{K}_\varepsilon$ with $\varepsilon = 0.1$.

5. Conclusion

We have introduced a new type of cross-validation procedure, based on bimodal kernels, in order to automatically remove the error correlation without requiring any prior knowledge about its structure. We have shown that the form of the kernel is very important when errors are correlated. This in contrast with the i.i.d. case where the choice between the various kernels on the basis of the mean squared error is not very important. As a consequence of the bimodal kernel choice the estimate suffers from increased mean squared error. Since an optimal bimodal kernel (in mean squared error sense) cannot be found we have proposed a ε-optimal class of bimodal kernels. Further, we have used the bandwidth of the bimodal kernel as pilot bandwidth selector for leave-$(2l+1)$-out cross-validation. By taking this extra step, methods that require a positive definite kernel (SVM and LS-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.

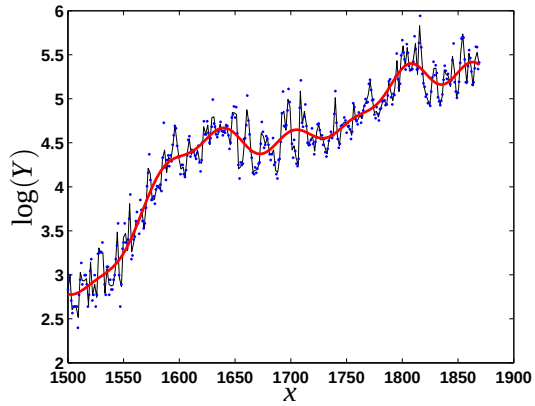


Figure 8: Difference in regression estimates (LS-SVM) for standard leave-one-out CV (thin line) and the proposed method (bold line).

Acknowledgments

The authors would like to thank Prof. László Györfi and Prof. Irène Gijbels for their constructive comments which improved the results of the paper.
Research supported by Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/post-doc & fellow grants; Flemish Government: FWO: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare, Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011), IBBT, EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), Contract Research: AMINAL, Other: Helmholtz, viCERP, ACCM. BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. JS is a professor at the Katholieke Universiteit Leuven, Belgium.

Appendix A. Proof of Lemma 2

We first rewrite the LCV score function in a more workable form. Since $Y_i = m(x_i) + e_i$

$$\begin{aligned}
\mathrm{LCV}(h) &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}_{(-i)}(x_i)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[m^2(x_i) + 2m(x_i)e_i + e_i^2 - 2Y_i\hat{m}_n^{(-i)}(x_i) + \left(\hat{m}_n^{(-i)}(x_i)\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[m(x_i) - \hat{m}_n^{(-i)}(x_i)\right]^2 + \frac{1}{n}\sum_{i=1}^{n} e_i^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\left[m(x_i) - \hat{m}_n^{(-i)}(x_i)\right] e_i.
\end{aligned}$$

Taking expectations, yields

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

Appendix B. Proof of Theorem 3

Consider only the last term of the expected LCV (Lemma 2), that is,

$$A(h) = -\frac{2}{n}\sum_{i=1}^{n}\mathbf{Cov}\left[\hat{m}_n^{(-i)}(x_i), e_i\right].$$

Plugging in the Nadaraya-Watson kernel smoother for $\hat{m}_n^{(-i)}(x_i)$ in the term above yields

$$A(h) = -\frac{2}{n}\sum_{i=1}^{n}\mathbf{Cov}\left[\sum_{j\neq i}^{n}\frac{K\left(\frac{x_i-x_j}{h}\right)Y_j}{\sum_{l\neq i}^{n}K\left(\frac{x_i-x_l}{h}\right)}, e_i\right].$$

By using the linearity of the expectation operator, $Y_j = m(x_j) + e_j$ and $\mathbf{E}[e] = 0$ it follows that

$$\begin{aligned} A(h) &= -\frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbf{E}\left[\frac{K\left(\frac{x_i-x_j}{h}\right)Y_j}{\sum_{j\neq i}^{n}K\left(\frac{x_i-x_l}{h}\right)}e_i\right] \\ &= -\frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{K\left(\frac{x_i-x_j}{h}\right)}{\sum_{j\neq i}^{n}K\left(\frac{x_i-x_l}{h}\right)}\mathbf{E}\left[e_i e_j\right]. \end{aligned}$$

By slightly rewriting the denominator and using the covariance stationary property of the errors (Definition 1), the above equation can be written as

$$A(h) = -\frac{2}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{K\left(\frac{x_i-x_j}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{x_i-x_l}{h}\right) - K(0)}\gamma_{|i-j|}. \tag{11}$$

Let f denote the design density. The first term of the denominator can be written as

$$\begin{aligned} \sum_{j=1}^{n}K\left(\frac{x_i-x_l}{h}\right) &= nh\hat{f}(x_i) \\ &= nhf(x_i) + nh(\hat{f}(x_i) - f(x_i)). \end{aligned}$$

If conditions (C2) and (C3) are fulfilled, f is uniform continuous and $h \to \infty$ as $n \to \infty$ such that $nh^2 \to \infty$, then

$$|\hat{f}(x_i) - f(x_i)| \leq \sup_{x_i}|\hat{f}(x_i) - f(x_i)| \xrightarrow{P} 0 \quad \text{as } n \to \infty,$$

due to the uniform weak consistency of the kernel density estimator (Parzen, 1962). $\xrightarrow{P}$ denotes convergence in probability. Hence, for $n \to \infty$, the following approximation is valid

$$nh\hat{f}(x_i) \approx nhf(x_i).$$

Further, by grouping terms together and using the fact that $x_i \equiv i/n$ (uniform equispaced design) and assume without loss of generality that $x \in [0,1]$, Equation (11) can be written as

$$\begin{aligned} A(h) &= -\frac{2}{n}\sum_{i=1}^{n}\frac{1}{nhf(x_i)-K(0)}\sum_{j\neq i}^{n}K\left(\frac{x_i-x_j}{h}\right)\gamma_{|i-j|} \\ &= -\frac{4}{nh-K(0)}\sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k. \end{aligned}$$

Next, we show that $\sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k = K(0)\sum_{k=1}^{\infty}\gamma_k + o(n^{-1}h^{-1})$ for $n\to\infty$. Since the kernel $K \geq 0$ is Lipschitz continuous at $x=0$

$$[K(0)+C_2x]_+ \leq K(x) \leq K(0)+C_1x,$$

where $[z]_+ = \max(z,0)$. Then, for $K(0)\geq 0$ and $C_1 > C_2$, we establish the following upperbound

$$\begin{aligned} \sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k &\leq \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\left(K(0)+C_1\frac{k}{nh}\right)\gamma_k \\ &\leq \sum_{k=1}^{n-1}K(0)\gamma_k + \sum_{k=1}^{n-1}C_1\frac{k}{nh}\gamma_k. \end{aligned}$$

Then, for $n\to\infty$ and using $\gamma_k \sim k^{-a}$ for $a>2$,

$$C_1\sum_{k=1}^{n-1}\frac{k}{nh}\gamma_k = C_1\sum_{k=1}^{n-1}\frac{k^{1-a}}{nh} = o(n^{-1}h^{-1}).$$

Hence,

$$\sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k \leq K(0)\sum_{k=1}^{\infty}\gamma_k + o(n^{-1}h^{-1}).$$

For the construction of the lower bound, assume first that $C_2<0$ and $K(0)\geq 0$ then

$$\sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k \geq \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\left[K(0)+C_2\frac{k}{nh}\right]_+\gamma_k.$$

Since $C_2<0$, it follows that $k \leq \frac{K(0)}{-C_2}nh$ and therefore

$$\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\left[K(0)+C_2\frac{k}{nh}\right]_+\gamma_k = \sum_{k=1}^{\min\left(n-1,\frac{K(0)}{-C_2}nh\right)}\left(1-\frac{k}{n}\right)\left(K(0)+C_2\frac{k}{nh}\right)\gamma_k.$$

Analogous to deriving the upper bound, we obtain for $n\to\infty$

$$\sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k \geq K(0)\sum_{k=1}^{\infty}\gamma_k + o(n^{-1}h^{-1}).$$

In the second case, that is, $C_2>0$, the same lower bound can be obtained. Finally, from the upper and lower bound, for $n\to\infty$, yields

$$\sum_{k=1}^{n-1}\left(\frac{n-k}{n}\right)K\left(\frac{k}{nh}\right)\gamma_k = K(0)\sum_{k=1}^{\infty}\gamma_k + o(n^{-1}h^{-1}).$$

Appendix C. Least Squares Support Vector Machines for Regression

Given a training set defined as $\mathcal{D}_n = \{(x_k, Y_k) : x_k \in \mathbb{R}^d, Y_k \in \mathbb{R}; k = 1, \ldots, n\}$. Then least squares support vector machines for regression are formulated as follows (Suykens et al., 2002)

$$\begin{aligned} \min_{w,b,e} \mathcal{J}(w,e) &= \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{k=1}^{n} e_k^2 \\ \text{s.t.} \quad Y_k &= w^T \varphi(x_k) + b + e_k, \quad k = 1, \ldots, n, \end{aligned} \tag{12}$$

where $e_k \in \mathbb{R}$ are assumed to be i.i.d. random errors with zero mean and finite variance, $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the feature map to the high dimensional feature space (possibly infinite dimensional) and $w \in \mathbb{R}^{n_h}$, $b \in \mathbb{R}$. The cost function $\mathcal{J}$ consists of a residual sum of squares (RSS) fitting error and a regularization term (with regularization parameter γ) corresponding to ridge regression in the feature space with additional bias term.

However, one does not need to evaluate w and φ explicitly. By using Lagrange multipliers, the solution of Equation (12) can be obtained by taking the Karush-Kuhn-Tucker (KKT) conditions for optimality. The result is given by the following linear system in the dual variables α

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

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

$$\hat{m}_n(x) = \sum_{k=1}^{n} \hat{\alpha}_k K(x, x_k) + \hat{b},$$

where $K(\cdot, \cdot)$ is an appropriately chosen positive definite kernel. In this paper we choose K to be the Gaussian kernel, that is, $K(x_k, x_l) = (2\pi)^{-d/2} \exp\left(\frac{-\|x_k - x_l\|^2}{2h^2} \right)$.

Appendix D. Proof of Theorem 6

We split up the proof in two parts, that is, for positive definite and positive semi-definite kernels. The statement will be proven by contradiction.

- Suppose there exists a positive definite bimodal kernel $\tilde{K}$. This leads to a positive definite kernel matrix Ω. Then, all eigenvalues of Ω are strictly positive and hence the trace of Ω is always larger than zero. However, this is in contradiction with the fact that Ω has all zeros on its main diagonal. Consequently, a positive definite bimodal kernel $\tilde{K}$ cannot exist.

- Suppose there exists a positive semi-definite bimodal kernel $\tilde{K}$. Then, at least one eigenvalue of the matrix Ω is equal to zero (the rest of the eigenvalues is strictly positive). We have now two possibilities, that is, some eigenvalues are equal to zero and all eigenvalues are equal to zero. In the first case, the trace of the matrix Ω is larger than zero and we have again a contradiction. In the second case, the trace of the matrix Ω is equal to zero and also the determinant of Ω equals zero (since all eigenvalues are equal to zero). But the determinant can never be zero since there is no linear dependence between the rows or columns (there is a zero in each row or column).

Appendix E. Proof of Corollary 8

From Equation (8) it follows that

$$\begin{aligned}
\hat{h}_{\text{AMISE}} &= \left[\frac{R(K)\sigma^2}{n\mu_2^2(K)\int(m''(x))^2\,dx} + \frac{2R(K)\sum_{k=1}^{\infty}\gamma_k}{n\mu_2^2(K)\int(m''(x))^2\,dx}\right]^{1/5} \\
&= \left[\hat{h}_0^5 + \frac{\sigma^2 R(K)}{n\mu_2^2(K)\int(m''(x))^2\,dx}\,\frac{2\sum_{k=1}^{\infty}\gamma_k}{\sigma^2}\right]^{1/5} \\
&= \left[1+2\sum_{k=1}^{\infty}\rho(k)\right]^{1/5}\hat{h}_0.
\end{aligned}$$

References

R.K. Adenstedt. On large sample estimation for the mean of a stationary sequence. *Ann. Statist.*, 2(6):1095–1107, 1974.

N.S. Altman. Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, 85(411):749–759, 1990.

T.W. Anderson. *The Statistical Analysis of Time Series*. Wiley, New York, 1971.

P. Bühlmann and H.R. Künsch. Block length selection in the bootstrap for time series. *Computational Statistics & Data Analysis*, 31(3):295310, 1999

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

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

D.R. Cox. Long-range dependence: a review. In *Proceedings 50th Anniversary Conference. Statistics: An Appraisal.* pages 55–74, Iowa State Univ. Press.

A.C Davison and D.V. Hinkley. *Bootstrap Methods and their Application* (reprinted with corrections). Cambridge University Press, 2003.

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

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

Y. Feng and S. Heiler. A simple bootstrap bandwidth selector for local polynomial fitting. *J. Stat. Comput. Simul.*, 79(12):1425–1439, 2009.

M. Francisco-Fernández and J.D. Opsomer. Smoothing parameter selection methods for nonparametric regression with spacially correlated errors. *Canad. J. Statist.*, 33(2):279–295, 2004.

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

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

P. Hall, J.L. Horowitz, and B.-Y. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561574, 1995

P. Hall and I. Van Keilegom. Using difference-based methods for inference in nonparametric regression with time-series errors. *J. Roy. Statist. Assoc. Ser. B Stat. Methodol.*, 65(2):443–456, 2003.

W. Härdle. *Applied Nonparametric Regression* (Reprinted). Cambridge University Press, 1999.

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

J.D. Hart. Differencing as an approximate de-trending device. *Stoch. Processes Appl.*, 31(2):251–259, 1989.

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

E. Hermann, T. Gasser and A. Kneip. Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, 79(4):783–795, 1992.

M.G. Kendall, A. Stuart and J. K. Ord. *The Advanced Theory of Statistics, vol. 3, Design and Analysis, and Time-Series* (4th ed.). Griffin, London, 1983.

T.Y. Kim, D. Kim, B.U. Park and D.G. Simpson. Nonparametric detection of correlated errors. *Biometrika*, 91(2):491–496, 2004.

T.Y. Kim, B.U. Park, M.S. Moon and C. Kim. Using bimodal kernel inference in nonparametric regression with correlated errors. *J. Multivariate Anal.*, 100(7):1487–1497, 2009.

S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer, 2008.

S.R. Kulkarni, S.E. Posner and S. Sandilya. Data-dependent k_n-NN and kernel estimators consistent for arbitrary processes. *IEEE Trans. Inform. Theory*, 48(10):2785–2788, 2002.

H. Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17(3):12171261, 1989.

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

S.N. Lahiri. Theoretical comparisons of block bootstrap methods. *Ann. Statist.*, 27(1):386404, 1999.

S.N. Lahiri. *Resampling Methods for Dependent Data*. Springer, 2003.

S.N. Lahiri, K. Furukawa, and Y.-D. Lee. A nonparametric plug-in rule for selecting optimal block lengths for block bootstrap methods. *Statistical Methodology*, 4(3):292321, 2007.

E.A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.

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

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

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):10651076, 1962

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

J.S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

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

V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1999.

G.S. Watson. Smooth regression analysis. *Sankhyā Ser. A*, 26(4):359–372, 1964.