Chapter 3

Smoothing Parameter Selection When the Errors are Correlated

Let us consider the following model:

$$Y_i = m(x_i) + \varepsilon_i, i = 1, \ldots, n \tag{3.1}$$

where the x_i's follow a fixed design with density f as defined earlier in (2.1). However, unlike previously assumed, the ε_i's are now correlated, that is, they follow a covariance stationary process with unknown covariance function $\gamma_n(k) = Cov(\varepsilon_i, \varepsilon_{i+k})$. This formulation of the covariance function allows the autocorrellation to vary with both the distance between the design points and the sample size. Here we define covariance stationary as $Cov(\varepsilon_i, \varepsilon_j) = Cov(\varepsilon_{i+k}, \varepsilon_{j+k})$ for any $k \geq 0$. For more details on such processes, see Brockwell and Davis (1987).

We surveyed, in the previous chapter, some automated smoothing parameter selectors which are known to perform well in the situation of independent observations. However, in the case of dependent observations, those smoothing parameter selectors will not produce good smoothing parameters.

In this chapter, we will first analyse how correlation in the data affects the appropriate amount of smoothing. Then, we will introduce four smoothing parameter selection methods for the case of dependent observations. These methods are intended to take into account the correlation and give smoothing parameters that are closer to the optimal smoothing parameter. The presented methods will be: "modified cross-validation" (MCV), "partitioned cross-validation" (PCV), "blockwise cross-validation" (BCV) and "time series cross-validation" (TSCV). For the sake of simplicity, we will present them in a kernel estimation context, but they can very well be generalized to other smoothing methods, such as splines.

3.1 Effect of Dependence on the Smoothing Parameter

Let us assume that we want to estimate the regression function m, which is assumed to be twice continuously differentiable and defined on $[0, 1]$. Furthermore, we would like to have a data-driven method that would allow us to chose an appropriate value for h, the smoothing parameter. We choose as our risk criterion the mean average squared error (MASE) defined as

$$MASE(h) = \frac{1}{n} \sum_{i=1}^{n} E(\hat{m}_h(x_i) - m(x_i))^2. \tag{3.2}$$

Under model (3.1) and for the Gasser-Müller kernel regression estimator, Hart (1991) showed that

$$MASE(h) \approx \frac{S(0)\nu_0(K^2)}{nh} + \frac{\nu_2^2(K)h^4}{4}\int_0^1 m''(x)^2 f(x)\,dx, \tag{3.3}$$

where $\nu_i(K)$ is as in section (2.1.1), K is a kernel of order 4 and S is the spectrum of the process $\{Y_i - m(x_i)\}$, i.e.,

$$S(\omega) = \gamma_n(0) + 2\sum_{i=1}^{\infty}\gamma_n(i)cos(\pi i\omega), \qquad 0 \le \omega \le 1. \tag{3.4}$$

We can see that the square of the estimator's bias (the second term of the right-hand side of (3.3)) is not affected by the correlation. However, the correlation does affect the variance and it does so in the same way that correlation affects the variance of a sample mean of a covariance stationary process. Indeed, Priestley (1981) showed that the sample mean ($\bar{Y} = \sum_{i=1}^{n} Y_i/n$) of a process with well-defined and bounded spectrum satisfies

$$Var(\bar{Y}) \sim \frac{S(0)}{n}.$$

One can show that the minimizer of (3.3) and, asymptotically, of (3.2) is

$$h_{opt} = \left[\frac{\nu_0(K^2)S(0)}{n\nu_2^2(K)\int_0^1\{m''(u)\}^2 f(u)\,du}\right]^{1/5}, \tag{3.5}$$

which is similar to equation (2.22) with σ^2 replaced by $S(0)$. So if $S(0) > \gamma_n(0) = Var(\varepsilon_i)$, meaning that the correlation is predominantly positive, then h_{opt} will be larger than it would be in the case where the data are independent. Thus, when errors are predominantly positively correlated, data should be oversmoothed. On the other hand, for $S(0) < \gamma_n(0)$, which is equivalent to predominant negative correlation, h_{opt} would be smaller than in the uncorrelated case. It is then obvious that correlation has a clear effect on the optimal smoothing parameter which leads us to consider techniques for smoothing parameter selection that take into account the fact that the data are correlated. For a more detailed description of the effect of dependence on smoothing parameter selection, the reader is refered to Hart and Wehrly (1986), Chiu (1989), Diggle and Hutchinson (1989) and Hart (1991).

3.2 Smoothing Parameter Selection Methods for Dependent Observations

For dependent observations, Chu and Marron (1991a) showed that the cross-validated smoothing parameter does not converge to the optimal smoothing parameter, although it converges at the same rate as that given by Härdle, Hall and Marron (1988) for the case of independent observations.

One can easily see why CV yields such a poor estimator of the regression function when errors are correlated. The cross-validation method tries to find a good predictor of Y_i. The problem is that a good predictor of Y_i is often a poor estimate of $E(Y_i)$ when the data are correlated. The best mean squared error predictor of the data point Y_i is based on the remainder of the observations in the following way:

$$E(Y_i|Y_1,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_n) = m(x_i) + q(\varepsilon_1,\ldots,\varepsilon_{i-1},\varepsilon_{i+1},\ldots,\varepsilon_n) \quad (3.6)$$

for some function q.

In the case of independent data, $q \equiv 0$, and thus a good estimator of

$E(Y_i)$ is the same as a good predictor of Y_i. However, this is not generally the case when the data are correlated. Cross-validation tends to select a smaller smoothing parameter when data are positively correlated because a kernel regression smoother with a small smoothing parameter usually does a good job of estimating the predictor of Y_i which is not the same as $E(Y_i)$ in the case of dependent observations. Obviously, an adjustment to cross-validation is needed in order to alleviate the effect of dependence. Here are a few methods that take into account the correlation when deciding on the amount of smoothness needed.

3.2.1 Modified Cross-Validation

The modified cross-validation method (MCV), also called the "leave-($2l$+1)-out" method, was first introduced by Härdle and Vieu (1992) and further studied by Vieu and Hart (1990), Gyorfi, Härdle, Sarda and Vieu (1989) and Chu and Marron (1991a) who presented results of applications of the MCV in various settings, including mixing data.

Let us now introduce the "leave-($2l$+1)-out" smoothing parameter selector. For any $l \geq 0$ the smoothing parameter is obtained by minimizing the MCV criterion

$$MCV_l(h) = n^{-1} \sum_{i=1}^{n} \left(\hat{m}_h^{(i)}(x_i) - Y_i \right)^2 . \tag{3.7}$$

Here $\hat{m}_h^{(i)}(x_i)$ is the "leave-($2l$+1)-out" estimate of $m(x_i)$, that is, the observations (x_{i+j}, Y_{i+j}), $-l \leq j \leq l$ are left out in constructing the estimate. Note that when l is taken to be zero, MCV is simply ordinary cross-validation. In the case of the Nadaraya-Watson kernel estimator, $\hat{m}_h^{(i)}(x_i)$ is defined by

$$\hat{m}_h^{(i)}(x_i) = \frac{\sum_{j:|j-i|>l} K(x_i - x_j)Y_j}{\sum_{j:|j-i|>l} K(x_i - x_j)}. \tag{3.8}$$

The amount of dependence between $\hat{m}_h^{(i)}(x_i)$ and Y_i is gradually reduced as the value of l is increased. Chu and Marron (1991a) showed that whenever l is moderately large, MCV produces a smoothing parameter that is nearly asymptotically unbiased for the optimal smoothing parameter for estimating m. Marron (1987) suggested an approach for the choice of l that is based on the minimization of the mean squared error.

The results of Chu and Marron (1991a) were proven in the context of kernel regression estimators. We implemented this smoothing parameter selector and utilized it in the context of smoothing splines with model-based penalty. The results will be discussed in greater detail in the next chapter.

3.2.2 Partitioned Cross-Validation

We have now seen one way of overcoming the effect of dependence on smoothing parameter selection. However other possibilities exist. One of them, proposed by Marron (1987) in the context of density estimation, is called "partitioned cross-validation" (PCV). Chu and Marron (1991a) later used PCV for kernel regression estimation.

Although the rate of convergence of this smoothing parameter is faster than that for MCV, the resulting smoothing parameter is not very good. Indeed, Chu and Marron (1991a) proved that the asymptotic mean of this smoothing parameter is far from the optimal smoothing parameter. In fact, they showed that the limiting distribution of the smoothing parameter is centered at the optimal smoothing parameter for the case of independent observations, which is different from the true optimal smoothing parameter. Thus, the resulting smoothing parameter suffers from asymptotic bias and leads to poor results in practice.

The idea behind partitioned cross-validation is to split the observations into k subgroups. This is done by taking observations $\{1, 1+k, 1+2k, \ldots\}$ to build the first subgroup, $\{2, 2+k, 2+2k, \ldots\}$ for the second subgroup and so on. Then, we calculate the ordinary cross-validation score $CV_{0,j}(h)$ for the j^{th} subgroup of observations separately, $j = 1, 2, \ldots, k$. We then minimize the average of these k cross-validation scores to find the optimal partitioned cross-validated smoothing parameter. In the case where n is not a multiple of k, we then take observations $1, 2, \ldots, k \cdot [n/k]$ and we discard the others. The notation $[a]$ denotes the largest integer which is less than or equal to a. When $k = 1$, partitioned cross-validation is simply ordinary cross-validation. For correlated data, as long as the value of k is sufficiently large, the errors within each subgroup will be essentially independent.

3.2.3 Blockwise Cross-Validation

A quite common problem seen in applications is that of taking repeated measurements on the same experimental subject or unit. Such data are often used to make inferences about a population. A model for such a problem can be

$$Y_j(x_{ij}) = m(x_{ij}) + \varepsilon_j(x_{ij}), \qquad j = 1, \ldots, M \;\; i = 1, \ldots, n_j; \tag{3.9}$$

where m, assumed to be smooth, is the population mean and $\varepsilon_j(x_{ij})$; $i = 1, \ldots, n_j$ follow a causal ARMA process. This leads to having M independent samples of respective size n_j of the same process with mean m. We do not assume that each sample is measured at the same time points nor do we assume that we have the same amount of data points for each sample. Therefore we use the notation x_{ij} which corresponds to the i^{th} time point for the j^{th} sample. Also, it is assumed that the errors ε_j between samples are independent.

Now the question that arises is how to analyse such data. Based on the assumption that m is smooth, we can use nonparametric methods such as kernel smoothers or smoothing splines. But how do we decide on the amount of smoothing? The most popular technique of smoothing parameter selection, cross-validation, is ineffective for this case since, as shown at the beginning of this chapter, observations are correlated. Wehrly and Hart (1988) and Rice and Silverman (1991) proposed a version of CV based on the idea of predicting data of one of the M experimental units by using the other $M-1$ units. It is done for each of the units and the M predicting errors are then averaged. The optimal smoothing parameter is the one that minimizes the mean prediction error. This method was called "unitwise" or "blockwise" cross-validation since one unit or block is left out and then predicted. Then, the blockwise cross-validation criterion is defined as

$$BCV(h) = \frac{1}{N} \sum_{j=1}^{M} \sum_{i=1}^{n_j} \left(Y_j(x_{ij}) - \hat{m}_h^j(x_{ij}) \right)^2, \tag{3.10}$$

where $N = \sum_{j=1}^{M} n_j$ and $\hat{m}_h^j(x_{ij})$ is the estimate of $m(x_{ij})$ computed with all the data except for the j^{th} unit: $\{Y_j(x_{1j}), \ldots, Y_j(x_{n_j j})\}$. The regression function is then estimated by $\hat{m}_{h_{opt}}$, where h_{opt} minimizes $BCV(h)$ and all N observations are treated as a data set from one individual.

We also implemented a modification of this method and we applied it to our simulation study (Chapter 4) and data set (Chapter 5).

3.2.4 Time Series Cross-Validation

Based on the *prequential* method of Dawid (1984), "time series" cross-validation (TSCV) was proposed by Hart (1994). The prequential approach is based on the concept of choosing a parametric model that will perform well in predicting future observations. It is similar to cross-validation but instead of using all the data other than Y_i, it uses only the previous data $Y_1, \ldots, Y_{i-1}$ to predict. Since forecasting is so important in the time series context, prequential analysis becomes very appealing in that situation. So like the prequential approach, TSCV has roots in giving a good one-step ahead predictor. However, it can be modified to be more suitable for estimating the regression function. Details on how to modify this technique appear at the end of this section.

Now, let us assume that we believe that the parametric probability model M_θ would yield good prediction for the process $\{\varepsilon_i = Y_i - m(x_i)\}$. Here θ represents the vector of the unknown parameters of the error model M_θ. Consider the following predictor

$$\hat{Y}_i(h, \theta) = \tilde{m}_h^i(x_i) + c_\theta(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_{i-1}), \tag{3.11}$$

where $\tilde{m}_h^i(x_i)$ is an estimate of $m(x_i)$, based on a smoothing parameter h, that uses only $Y_1, \ldots, Y_{i-1}$, $\hat{\varepsilon}_j = Y_j - \tilde{m}_h^j(x_j), j = 1, \ldots, i-1$ and $c_\theta(\varepsilon_1, \ldots, \varepsilon_{i-1})$ is the optimal mean squared error predictor of ε_i given $\varepsilon_1, \ldots, \varepsilon_{i-1}$ and assuming that the probability model is M_θ. Since TSCV only uses data to the left of x_i, Hart (1994) suggests that one uses a boundary-type kernel when estimating $m(x_i)$.

Then, the TSCV criterion is defined as

$$T(h, \theta) = \sum_{i=2}^{n} (\hat{Y}_i(h, \theta) - Y_i)^2. \tag{3.12}$$

The TSCV method consists of choosing h and θ that minimize (3.12). However, with certain models for the errors such as autoregressive ones, it is possible to find a closed form, for each h, of the value $\hat{\theta}(h)$ that minimizes $T(h, \theta)$ with respect to θ. In those cases, one obtains a curve $T(h) = T(h, \hat{\theta}(h))$ that can be analyzed in order to find the optimal time series cross-validated smoothing parameter, $\hat{h}_{TSCV}$.

This method yields a smoothing parameter that is evidently intended for a good one-step ahead predictor based on the past data. With this estimate, one obtains, not only the predictors, but also estimates of the trend and the covariance function. However, if the mean function (trend) is the main interest of the data analyst, then one would want to have an estimate that uses data on both sides of x_i. In that case, Hart (1994) proposed using a smoothing parameter "equivalent" to $\hat{h}_{TSCV}$ to estimate m. If the regression function is estimated using a kernel K, then the corrected smoothing parameter is $C_{K,L} \cdot \hat{h}_{TSCV}$, where the term $C_{K,L}$ is

$$C_{K,L} = \left\{ \frac{\nu_0(K^2)\nu_2^2(L)}{\nu_0(L^2)\nu_2^2(K)} \right\}^{1/5} \tag{3.13}$$

and L is the boundary-type kernel that was used in the estimation of $\tilde{m}^i(x_i)$.

3.2.5 Plug-in Method

Let us now introduce a plug-in method called *Correlation Direct Plug-in* (CDPI) developped by Opsomer (1996). This method is based on the *Direct Plug-in* method seen in section 2.1.4 but is intended for local polynomial regression when the errors are correlated. Early work on the plug-in method in the context of correlated errors includes Altman (1990).

Opsomer (1996) developped this method for the random design case but the results also hold for the nonrandom fixed design case. Opsomer (1996) defined the covariance of the residuals as $Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \rho_n(X_i - X_j)$ where ρ_n is a stationary, symmetric and continuous correlation function which can depend on the sample size. To relate ρ_n to our γ_n, consider the case with $X_i = i/n$. Then, $\gamma_n(k) = \sigma^2 \rho_n(k/n)$. In order to ensure convergence of the estimator, Opsomer imposed the following set of conditions on ρ_n.

1. ρ_n is differentiable, $\int n|\rho_n(t-x)|\,dt = O(1)$, $\int n\rho_n(t-x)^2\,dt = o(1)$ for all x,

2. $\exists \xi > 0$: $\int |\rho_n(t)| I_{(|t/h| > \xi)}\,dt = o(\int |\rho_n(t)|\,dt)$, where $h = h_n \to 0$ and $nh \to \infty$ as $n \to \infty$ This means that the dependence is short range, that is, correlation between observations sufficiently far apart can be ignored.

Opsomer (1996) showed that the smoothing parameter that minimizes the mean average squared error of the estimate of the regression function is of the form

$$h_{opt} = \left(\frac{\nu_0(K^2) I C_n}{n \nu_2^2(K) \theta_{22}} \right)^{1/5} \tag{3.14}$$

where K is a kernel function satisfying assumption 2 of Theorem 1. IC_n is called the integrated covariance function and $\theta_{ij} = E(m^{(i)}(X)m^{(j)}(X))$. These two quantities are unknown. Therefore one needs to estimate them.

The estimation of IC_n, when the observations are equally spaced, can be done in one simple step by estimating the spectral density at 0. However, when the observations are not equally spaced, this method can not be directly applied. Opsomer (1996) proposed using a technique based on binning approximations to increase the efficiency of local polynomial regression (see Fan and Marron (1994) and Turlach and Wand (1996)). A binned dataset with approximately the same correlation structure as the original data is constructed and spectral theory is used to estimate IC_n. The estimation of IC_n also requires the separate estimation of the variance parameter σ^2. Opsomer (1996) suggests that one uses a generalization of the variance estimator of Opsomer and Ruppert (1998) in order to get $\hat{\sigma}^2$.

As an estimator of θ_{22}, $\hat{\theta}_{22} = n^{-1}\sum_{i=1}^{n}(\hat{m}''(x_i))^2$, is proposed with $\hat{m}''$ the estimate of the the second derivative gotten from a local polynomial estimator of degree $p = 3$. Opsomer (1996) showed that the asymptotically optimal smoothing parameter to estimate θ_{22} is of the form

$$h_{opt}^{1} = \left(C_\theta \frac{12\nu_0(K_{(2,3)}^2)IC_n}{|\theta_{24}|\nu_4(K_{(2,3)})} \right)^{1/7}, \tag{3.15}$$

where $K_{(r,3)}$ is as defined in (2.16) and

$$C_\theta = \begin{cases} 1 & if \quad \theta_{24} \le 0 \\ \frac{5}{2} & if \quad \theta_{24} > 0 \end{cases}.$$

Then, one simply needs, in order to get $\hat{h}_{opt}$, to follow these steps.

1. "Pilot" estimators of θ_{24} and σ^2 are computed from a piecewise quartic fit, where the number of pieces is chosen by minimizing Mallows C_p;

2. a pilot estimator for IC_n is computed and is used to obtain the nonparametric estimate $\hat{IC}_n$ as described in section 3 of Opsomer (1996);

3. use θ_{24} and $\hat{IC}_n$ in (3.15) and use the estimated h^1_{opt} to find $\hat{\theta}_{22}$;

4. $\hat{h}_{opt}$ is obtained by plugging $\hat{\theta}_{22}$ and $\hat{IC}_n$ in equation (3.14).