Chapter 6

Conclusion and Discussion

The motivation of this work comes from an environmental problem. Indeed, estimating the underlying trend in air pollution data is complex as one has to consider the serial correlation of the time series while keeping in mind the usual seasonal behaviour of such data. Usual time series methods appear to take correlation into account but lack the flexibility of nonparametric methods in terms of tracking finer features of the data.

In Chapter 3, we described five smoothing parameter selection techniques used in nonparametric kernel regression which, when applied in situations where the regression errors are correlated, are designed to perform more effectively than smoothing parameter selection techniques used in the case of independent errors. We modified two of these techniques, namely the "modified cross-validation" (MCV_l) and the "blockwise cross-validation" (BCV) methods, and used them in the context of smoothing splines with model based penalty.

The performance of these two smoothing parameter selection techniques was investigated, in Chapter 4, via a simulation study. Using simulated data sets to which correlated errors were added and for which the underlying functions were known, we could evaluate the extent to which the effect of the correlation was alleviated, that is, evaluate the accuracy with which the function was estimated in the presence of correlated errors. Our two techniques were compared to ordinary cross-validation and to three commonly used parametric methods.

Overall the performance of MCV_l in the simulations is superior to any other methods that we considered. Furthermore, for best results, an optimal value of l should be chosen. Indeed, an appropriate choice of l would lead to a smoothing parameter that efficiently alleviates the effect of correlation and yields an estimate of the regression function that is more accurate. On the other hand, if the value of l is chosen to be too small compared to the optimal one, then the obtained smoothing parameter would, for positively correlated errors, tend to yield an estimate that is more "wiggly", hence leading to interpolation. Conversely, the use of an excessively large smoothing parameter has not been studied in depth, thus we can not assess its effect in this work. We proposed two techniques to select l: the first one uses the order of the estimated autoregressive model and the second one considers the magnitude of the estimates of $\mathrm{Cor}(\varepsilon_i, \varepsilon_{i+j})$. However, to properly assess their performance, further work needs to be done.

The BCV method did not perform very well. Indeed, by the way it minimizes its criterion, BCV is highly influenced by the amount of noise in the data and by its periodicity. Also, one has to correctly choose the number of cycles for BCV. The choice of these cycles is usually made according to the seasonality of the observed natural process. However, the regularity of these cycles plays an important role in the performance of BCV. Indeed, in the case where the data are periodic with similar cycles, BCV will select an excessively small smoothing parameter, thus yielding an estimate of the regression function that is interpolating the data. Conversely, if the data are not periodic, BCV will have a tendancy to oversmooth the data as it selects a smoothing parameter that is too large.

In Chapter 5, we applied MCV_l and BCV to daily and monthly ground level ozone concentration. Once again, MCV_l performed well compared to the other methods of analysis. BCV, which was highly affected by the periodicity of the monthly data and by the high level of noise in the daily data, did not yield satisfying results. For a few selected sites, we considered $\mathrm{MCV}_{\hat{p}}$ where $\hat{p}$ is the estimate of the autoregressive order of the residuals. For both data sets, the selection rule of $l = \hat{p}$ worked well. Indeed, the estimate obtained with $\mathrm{MCV}_{\hat{p}}$ appears, for all selected sites, to give the best fit among all methods used.

As usual, there is further work to be done. The smoothing parameter selection techniques were considered in the context of no covariate. However, the importance of covariates such as temperature, humidity, season, etc., is undeniable. Therefore, modifying our techniques to apply them to regression of higher dimensional data might increase the accuracy of the regression function estimation. Furthermore, studying the effect of different error structures, such as those that are heteroscedastic and otherwise nonstationary, could be helpful to improve the understanding of the effect of correlation. Finally, and as mentioned before, further investigation on the choice of l in MCV_l is needed. We believe that the range and the magnitude of the correlation should be used to select an optimal value of l.