Chapter 1

Introduction

Estimating the underlying signal, often called the regression function, of a set of noisy data is a very common task that statisticians encounter and it has two main purposes. Firstly, it provides a way of analysing and presenting the relationship between the design variable and the response variable; secondly, it allows the prediction of observations that have not yet been made. In parametric regression, the underlying signal is assumed to follow a given parametric form. For instance, in the case of linear regression, the regression function is assumed to belong to the family of linear functions. Oftentimes, nonparametric methods of estimation such as kernel smoothers or smoothing splines are desirable, because they are extremely flexible as they do not restrict the model to a defined family of functions (see Green and Silverman (1994), Eubank (1988)). The use of an initial nonparametric estimate may very well suggest a parametric model (like a linear model or low order polynomial model) but it will at least give the data a chance to express themselves more freely. However, the use of nonparametric regression involves the choice of a *smoothing parameter* which controls the balance between goodness of fit to the data and smoothness of the regression function. This choice can be made subjectively but automated methods of selection are often preferred.

Due to their dependence on the cyclic nature of solar radiation, tides, meteorological processes, economical activities, etc., many observed time series exhibit a periodic pattern along with serial correlation. Let us consider ground-level ozone (O_3) concentration. O_3 concentration has a strong yearly cycle as it highly depends on solar radiation. It is also serially correlated. We would like to estimate the underlying trend of ground-level ozone concentration using a nonparametric regression technique as we want our method of estimation to be flexible. However, since we strongly suspect that the underlying signal is periodic and of the form $\{sin, cos\}$, we would like to use that knowledge. Therefore, we will use smoothing splines with model based penalty. This nonparametric method, which will be described in Chapter 3, allows us to specify a default parametric family. Also, we will need to choose a value of the smoothing parameter and, in order to avoid subjectivity, we would like to use an automated selection method. Many automatic methods of choosing the smoothing parameter exist for the case where the regression errors (or noise) are assumed to be independent, namely via *plug-in* (see Ruppert, Sheather and Wand (1995)), *cross-validation* and *generalized cross-validation* (see Stone (1974) and Craven and Wahba (1979)). However, if the observations are dependent, the smoothing parameter selectors designed for independent observations will produce poor results. In fact, many authors such as Opsomer (1996), Altman (1990) and Chu and Marron (1991a) showed, in the context of kernel regression, that the effect of ignoring even minor amounts of error correlation when selecting a smoothing parameter could potentially be severe. This effect is also present in smoothing spline regression, as a simulated example readily shows in Figure 1.1: the data are generated by a $\{sin, cos\}$ trend (discussed in greater detail in Chapter 4) and are overlaid with errors that follow an AR(1) model with autoregressive coefficient $\phi = 0.5$. Ordinary cross-validation, which performs extremely well when the errors are uncorrelated, selects a smoothing parameter that is far too small and results in an overly "wiggly" estimated regression function. The other estimate displayed in Figure 1.1 uses a larger smoothing parameter selected automatically by the MCV_1 method studied in this thesis, and is much closer to the true mean function of the data.

For kernel regression, many authors have tackled the problems associated with correlated regression errors and have proposed methods to solve them. Hart and Wehrly (1986) proved that, under certain conditions, correlation may lead to an inconsistent estimator of the regression function. Several other authors have shown that correlation causes smoothing parameter selection techniques designed for independent observations to converge to the "wrong" smoothing parameter, and have proposed various methods to remedy the situation. Chiu (1989) and Hart (1991,1994) assumed a parametric form for the correlation structure and adjusted some smoothing parameter selection

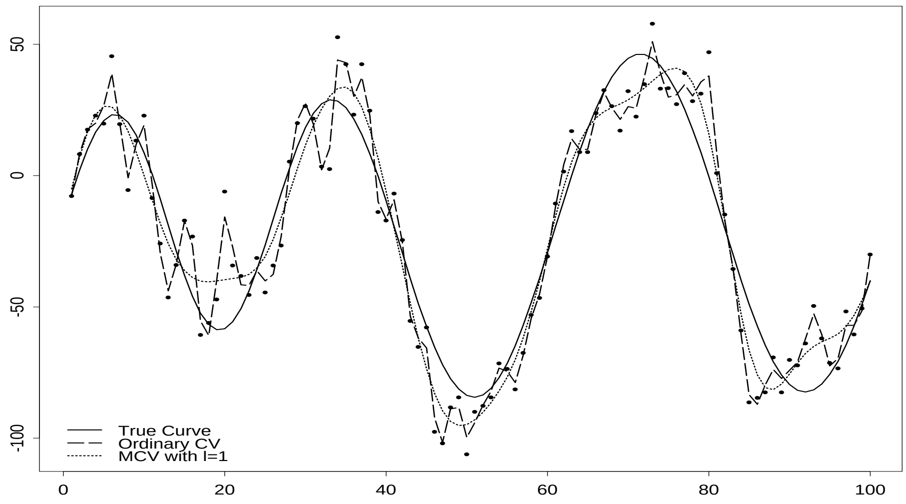

Figure 1.1: *Estimated Mean Functions using CV and* MCV_1.

techniques. Altman (1990) estimated both the regression and correlation function nonparametrically and proposed a plug-in smoothing parameter method. Opsomer (1996) also proposed a plug-in method and it will be discussed later in Chapter 3. Chu and Marron (1991a) bypassed the problem of estimating the correlation function and adjusted the cross-validation criterion to obtain an improved bandwidth selector.

However, most authors have worked in a kernel estimation setting. We propose, in this thesis, to modify some of the ideas used in the cross-validation criterion in the kernel regression context, and to apply them to smoothing splines with model based penalty. Chapter 2 surveys the nonparametric methods most commonly used for regression function estimation. Chapter 3 discusses a few smoothing parameter selection techniques that account for correlation in the regression errors. In Chapter 4, we will analyse the performance of the developed smoothing parameter selection methods by applying them to a simulation study and in Chapter 5, using the same smoothing parameter selectors, we will analyse a data set of daily and monthly ground-level O_3 concentrations.