12

Application to Continuous Time Processes Prediction

In Chapter 11, the functional nonparametric methodology was shown to have appealing theoretical supports for dependent statistical samples. The aim of this chapter is to show how this methodology can be used in practical situations for analysing time series. After a short discussion in Section 12.1 on how nonparametric finite dimensional statistics are used in the standard literature to treat time series, Section 12.2 explains how time series analysis can be viewed as specific functional nonparametric problems for dependent data for which all the methodology described in Chapter 11 will apply directly. Then we will see in Section 12.3 that, despite their rather technical look, these nonparametric functional methods are easy to implement. To emphasize this point, a real dataset will be quickly treated in Section 12.4. Finally, the source codes, functional datasets, descriptions of the $R/S+$ routines, guidelines and examples of use are detailed in the companion website *http://www.lsp.ups-tlse.fr/staph/npfda*.

12.1 Time Series and Nonparametric Statistics

The statistical analysis of some time series $\{Z_t, t \in \mathbb{R}\}$ is always linked with models and methods involving dependent data. Let us look for instance at the standard case when the process has been observed until time T and when the problem is to predict some future value Z_{T+s} of the process. Usually, the process is observed at a grid of N discretized times, and the observations are denoted by $\{Z_1, \dots Z_N\}$. The first step for predicting future values is to decide how much information has to be taken into account from the past?

The simpler situation consists in predicting the future just by taking into account one single past value. This is usually done by constructing some two-dimensional statistical sample of size $n = N - s$:

$$X_i = Z_i \text{ and } Y_i = Z_{i+s},\ i = 1, \dots, N - s, \tag{12.1}$$

in such a way that the problem turns to be a standard prediction problem of a real valued response Y given some real explanatory variable X. The only additional difficulty comes from the obvious necessity for allowing dependence structure in the statistical sample (X_i, Y_i). This approach can be used for many different statistical purposes and with various nonparametric estimates. It can lead to appealing results in practical situations as shown in several different real data studies that have been performed in the statistical literature (see for instance [HV92], [CD93], [H96], [R97], [GSY03], or [Co04] for the treatment of several time series coming from various fields of applied statistics).

Of course, this univariate modelling of the explanatory variable can be too restrictive to take into account sufficient information in the past of the series. To bypass this problem one could think in terms of constructing some $(p+1)$-dimensional statistical sample (of size $n = N - s - p + 1$) in the following manner:

$$\mathbf{X_i} = (Z_{i-p+1}, \ldots, Z_i) \text{ and } Y_i = Z_{i+s}, \; i = p, \ldots, N - s, \tag{12.2}$$

in such a way that the problem turns to be a standard prediction problem of a real valued response Y given some p-dimensional explanatory variable $\mathbf{X}$. Once again, as before when $p = 1$, it is necessary to attack this regression problem by allowing some dependence into the statistical sample $(\mathbf{X_i}, Y_i)$. Indeed, nonparametric approach to such a multidimensional prediction problem suffers from the curse of dimensionality (see discussions and references in Sections 3 and 13.5). Because of this curse of dimensionality the question of the choice of the order p turns to be a crucial one (see for instance [V95], [V02], [GQV802], [TA94] and [AT90] for an unexhaustive list of recent approaches to this question and for more references). From a practical point of view, most people prefer the use of semi-parametric modelling in order to reduce the effects of the dimension. It is out of the scope of this book to discuss in detail these semi-parametric and/or reduction dimension modelling approaches. For that, and to stay inside within the most recent references, we could encourage the reader to look at the advances provided by [GT04], [G99], [AR99], [G98], as well as at the general discussions presented by [HMSW04], [FY03], [HLG00] and [Gh99].

Finally, if one wishes to minimize the modelling errors by staying in pure nonparametric framework, it seems that there is a trade-off to balance between taking too many explanatory past values of the series (but with bad influence on the statistical performance of the estimates) and insuring good behaviour of the estimates (but by reducing the information from the past). We will see in the next section that the functional methodology is one way to answer this question.