NONPARAMETRIC METHODS FOR
FUNCTIONAL REGRESSION WITH
MULTIPLE RESPONSES

Thesis submitted for the degree of

Doctor of Philosophy

at the

University of Leicester

by

Kurdistan M. Taher Omar

Department of Mathematics

University of Leicester

June 2019
Abstract

Nonparametric functional regression is of considerable importance due to its impact on the development of data analysis in a number of fields, least cost and saving time. In this thesis, we focus on nonparametric functional regression and its extensions, and its application to functional data.

We first review nonparametric functional regression, followed by a detailed discussion about model structures, semi-metrics and kernel functions. Secondly, we extend the independent response model to multivariate response variables with functional covariates. Our model uses the K-Nearest Neighbour function with automatic bandwidth selection by a cross-validation procedure, and where the closeness between functional data is measured via semi-metrics. Then, in the third topic, we use the principal component analysis to decorrelate multivariate response variables. After that, in the fourth topic, we add new results to the non-parametric functional regression when the covariate is functional and the response is multivariate in nature with different bandwidths for different responses, and where the correlation among different responses is taken into account with different bandwidths for different responses. Our model uses the kernel function with automatic bandwidth selection via a cross-validation procedure and semi-metrics as a measure of the proximity between functional data. Finally, we extend the univariate functional responses to the multivariate case and then take the correlation between different functional responses into account. The effectiveness of the proposed models is illustrated through simulated instances. The proposed methods are then applied to functional data and, through our numerical outcomes, we improve the results as compared with the various methods reported in the literature.

Keywords: Nonparametric functional regression, multivariate response, functional covariate, semi-metrics, functional data, principal component analysis, almost complete convergence, functional response, multivariate functional responses, covariance
Acknowledgements

It is pleasure to acknowledge the help and support I have received from many people during my PhD studies at the University of Leicester. I would like to take this opportunity to acknowledge their help and efforts.

Firstly, I would like to express my sincere gratitude to my supervisor, Assistant Professor Bo Wang, for his support, time and patience, who has also been my inspiration throughout my studies. I must say thanks for all of the help, encouragement and suggestions he has given, and which are by no means limited to my research but have also proven invaluable in life. This thesis draws on his experience, talents, knowledge and contributions.

I would like to thank all the respectable staff members of the Department of Mathematics for their help and making so many things easier. I extend my appreciation to my fellow research students in the Michael Atiyah offices.

Finally, a special thanks goes to my husband, Salam, who was always with me from the initial steps to the end of my PhD studies. Thank you for always believing in me, for your unconditional support, strength, looking after the kids and your love. My children Shayan, and Shydar, they are little boys but respect me when I study and we are at home, even on school holidays, and another point I want mentioned that after I had our baby, Rawezh, they looked after him as much as they could. I would like to show my thanks to my family, especially my parents. They deserve a tremendous amount of credit for this achievement as they have always been a source of inspiration and support throughout my life. Without their help, this thesis would not have been possible.
Contents

Abstract ii

Acknowledgements iii

List of Figures vi

List of Tables vii

Abbreviations viii

1 Introduction 1
  1.1 Introduction .......................... 1
  1.2 Multivariate Response with K-Nearest Neighbour model .... 3
  1.3 Principal Component Analysis in Nonparametric Functional Regression .......................... 8
  1.4 Correlation Between Different Components of the Responses in Functional Data .......................... 9
  1.5 Multivariate Functional Responses .......................... 10
  1.6 Contributions and Outline of the Thesis .......................... 13

2 Preliminaries 16
  2.1 Background .......................... 16
  2.2 Semi-metric .......................... 16
    2.2.1 Functional Principal Component Analysis .......................... 18
    2.2.2 Semi-metric Construct Based on Functional Principal Component Analysis (FPCA) .......................... 20
    2.2.3 Semi-metric Based on Derivative .......................... 22
  2.3 Local Weighting .......................... 24
    2.3.1 Real Case .......................... 24
    2.3.2 Multivariate Case .......................... 26
    2.3.3 Functional Case .......................... 27
    2.3.4 Local Weighting and Small Ball Probabilities .......................... 29
  2.4 Nonparametric Functional Kernel Regression .......................... 29
    2.4.1 The Estimate .......................... 30
List of Figures

2.1 Usual Symmetrical Kernels. ........................................... 26

3.1 A sample of 215 simulated curves. ................................. 38
3.2 The Spectrometric curves. ............................................ 41
3.3 The multivariate response model predictions for fat, water, and protein content from top to bottom, respectively. .................. 43
3.4 The independent response model predictions for fat, water, and protein content from top to bottom, respectively. .................. 44
3.5 The average monthly temperature for 35 Canadian weather stations. 45
3.6 The average monthly precipitation at 35 Canadian weather stations. 46
3.7 The prediction from multivariate response variables for the average monthly precipitation at 10 stations from the average monthly temperature curves. .......................... 48
3.8 The prediction from independent responses for average monthly precipitation at 10 stations from the average monthly temperature curves. 49

5.1 A sample of 215 simulated curves. ................................. 70
5.2 A sample of 100 simulated curves. ................................. 73
5.3 A sample of 215 simulated curves. ................................. 75

6.1 A sample of 200 simulated covariate curves. ................. 83
6.2 A sample of 200 simulated functional responses. .......... 84
6.3 Simulation study - three randomly chosen samples for each model with the true values. Top: functional response 1; bottom: functional response 2. The black curves (cross) represent the true values, and the blue (triangle), green curves (plus) and red (square) show the predicted curves by multivariate functional responses with correlation (MFRC) model, multivariate functional responses (MFR) method and individual functional responses (IFR) model, respectively. 86
6.4 The maximum and minimum temperatures from 1930 to 2015. Top panel: the monthly maximum temperature; bottom panel: the monthly minimum temperature. ............................... 87
6.5 Valley data - two randomly chosen samples for each functional responses (maximum temperature and minimum temperature) from top to bottom, respectively. The black curves (triangle) represent the true curves, the blue curves (circle), the green curves (diamond), and the red curves (square) show the predicted curves using the MFRC model, MFR method, and the IFR model, respectively. 88
List of Tables

3.1 The average MSEs for the two simulated samples data. ............. 40
3.2 The functional covariates and multivariate response variables. ... 42
3.3 The MSE for fat, water and protein. .................................. 43
3.4 The functional covariates and Multivariate response variables for 35 Canadian weather stations. ................................. 47
3.5 MSE for Canadian Weather Station. ................................. 48
4.1 The average RMSEs for the simulated sample data. ............. 56
4.2 The RMSE for the fat, water and protein. .......................... 57
4.3 The RMSE for the SOM and EC. ................................. 58
5.1 Average RMSE comparison between MRC, MRD and IR models based on ten replicated simulations. .......................... 72
5.2 Average RMSE comparison between MRC, MRD and IR models based on ten replicated simulations. .......................... 74
5.3 Average RMSE comparison between the MRC, MRD and IR models based on ten replicated simulations. .......................... 76
5.4 RMSE comparison between MRC, MRD and IR models. ........ 77
6.1 Simulation study - average of MSE for MRFC, MRF and IFR, based on 20 replications. ................................. 85
6.2 The average MSE by MRFC, MFR and IFR models for the UK weather data. ................................. 87
Abbreviations

*f.v.* = functional variable
*KNN* = K-Nearest Neighbour
*NW* = Nadayara-Watson
*FPCA* = Functional Principal Component Analysis
*Mul − R* = multivariate response variables
*Ind − R* = the independent response model
*MSE* = mean square error
*NFDA* = Nonparametric Functional Data Analysis
*Mul − RP* = Multivariate response variables with principal component analysis
*RMSE* = root mean square error
*SOM* = soil organic matter
*EC* = ergosterol concentration
*MRD* = multivariate response using different bandwidths
*MRC* = multivariate response with correlations and different bandwidths
*V* = covariance matrix
*MFR* = multivariate functional responses
*MFRC* = multivariate functional responses with correlation
*IFR* = individual functional response
*C* = generic finite real positive constants
*d(, ; )* = semi-metric on some functional space *E*
*E, F* = generic functional spaces
*E(Y)* = expectation of some r.r.v. *Y*
*E(Y|X = χ)* = conditional expectation of some r.r.v. *Y* given the f.r.v. *X*
*f.r.v.* = functional random variable
*h* = generic notation bandwidth *h = h(n)*
$K =$ generic notation for an asymmetrical kernel function

$O_{a.co.} =$ rate of almost complete convergence

$r,m =$ nonlinear regression operator

$X =$ generic real random variable (r.r.v.)

$\mathcal{X} =$ generic functional random variable

$\chi =$ generic (unrandom) functional element

$\mathcal{X}_i = i=1,\ldots,n$ sample of f.r.v.

$\chi_i = i=1,\ldots,n$ Statistical observations of the f.r.v $\mathcal{X}_i$

$R =$ set of real numbers without 0

$Z =$ set of integers without 0
To my husband and my children
Chapter 1

Introduction

1.1 Introduction

A well-known statistical issue contains in studying the relation between two variables (output and input) in order to predict one of them (the response variable) given a new value for the other one (the explanatory variable). Therefore, the regression problem has been studied widely for real or multivariate variables, which have been recently generalised to functional variables. Our research is devoted to the problem when the explanatory is a functional and the response is a multivariate variable or multivariate function. There are several methods to study prediction, and one of the most popular is the regression method, which is based on the conditional expectation.

Recently, functional variables have seen intensively development because of an increasing number of related problems in the various fields of applied science (biometric, chemometric, medicine, environmetrics, etc.), which collects data as curves (i.e., one observation contains at least one functional object, such as a curve). Measurement instruments have become increasingly powerful, and which allow the collection of data with great precision and at high frequency. Hence, technological developments that allow us to deal with a large set of variables (infinite dimensional data). Therefore, this type of data can only be analysed through special statistical structures or procedures, presenting the statistician with a particular challenge, for example where some random variable can be observed at several different times in the range \((t_{\text{min}}, t_{\text{max}})\). The random family \(X(t_j)_{j=1,...,J}\) can be expressed by an observation. One way to take this into account is to consider
the data to be observations of the continuous family $X = \{X(t); t \in (t_{\text{min}}, t_{\text{max}})\}$. However, other situations that can be measured that consider different wavelengths instead of points in time, such as the spectrometric curves. $X$ is referred to as a functional variable (f.v.) if it takes values in an infinite dimensional space, and an observation $\chi$ of $X$ is called functional data, where $\chi_1, ..., \chi_n$ is the observation of $n$ functional variables $X_1, ..., X_n$ which are identically distributed as $X$.

In the last few years, a number of statisticians have dealt with functional data due to the belief that traditional statistical methods are expensive and time consuming, whilst functional data offer a new modelling framework for prediction. In fact, if we consider a sample of infinity discretised curves, two important statistical issues are encountered. The first is related to the proportion between the size of the sample and the number of variables whilst the second is due to a possible correlation between variables, whereby the problem becomes ill-conditioned within the structure of a multivariate linear model. Therefore, this type of data is required to develop statistical models in order to properly take into account the functional framework.

For functional data analysis, two main streams of methodologies exist in the literature: functional parametric methods and functional nonparametric methods. Ramsay and Silverman (1997, 2005) studied the statistics of functional parametric models, where this area of statistics has become increasingly popular applying a number of vigorous illustrations of applications of functional data analysis using curves as data. Several different techniques can involve functional data (such as prediction, regression, and classification). Ramsay and Silverman (2007) define functional data in detail, as well as discuss a number of different examples and utilise linear regression and multiple regression (parametric model) to predict scalar responses from the functional covariates and functional responses from functional covariates (for example, predicting the average monthly precipitation curves from average monthly temperatures). Their study also contains a case studies and applied issues arising from collaborative research to explain how functional data analysis ideas can be applied in reality. Ferraty and Vieu (2003, 2004, 2006) first introduced nonparametric methods for functional data analysis with functional covariates and univariate scalar responses, which use semi-metrics as a measure of
proximity between curves.

The first nonparametric univariate regression estimator was proposed in the early 1960’s, where free-modelling (both a free distribution and in a free parameter) was the focus of the work by Mahalanobis (1961), who suggested a "distribution-free" regression analysis. Therefore, nonparametric statistics have been developed intensively in multivariate cases and functional models. Nonparametric prediction problems have been investigated both in real and multivariate cases. It is impossible to give a detailed description of the related index, but for further information the reader is referred to the work of Nadaraya (1964) and Watson (1964).

1.2 Multivariate Response with K-Nearest Neighbour model

In the case of multivariate response variables, this model has been studied for a number of decades using linear regression techniques; the problem has been considered by a number of authors, but the majority of such work has been restricted to the standard multivariate situation where both $X$ and $Y$ are real or multivariate. For instance, Liang et al. (1992) took more than one response to estimate equations for parameters in terms of mean and variance in the attempt to account for the associations between outcomes. Their study compared two generalised estimate equations for the parameter model: the first concentrated on the regression parameters, whilst the second simultaneously estimated the regression and association between parameters because the response vector can include repeated observations of one variable, as in longitudinal studies or genetic studies. Breiman and Friedman (1997) pointed out the prediction of several response variables from the same set of independent variables, where these same authors also discussed the advantage of correlation between the response variables to improve predictions. Harrar and Bathke (2008) defined the nonparametric model via unbalanced multivariate data and derived an associated asymptotic distribution, in which the number of factor levels tended to infinity; they also noted that their nonparametric model had better predictive power than its parametric counterparts. Crambes
et al. (2009) described functional linear regression, where scalar responses are modelled in terms of a dependence on covariate random functions, where it was shown that the convergence rates of the prediction errors relied on the smoothness of the slope function and the framework of the predictors. Lutz and Bühlmann (2006) explained the problem of multivariate responses in high-dimensional linear regression by using a boosting method based on squared error loss for multivariate data, which was then applied to multivariate linear regression with continuous responses and to vector autoregressive time series. Koch and Naito (2010) extended principal component regression in a number of ways: a comprehensive variable ranking was combined with a selection of the optimal number of components for principal component regression, which was further extended to regression with multivariate responses.

Over the last two decades, functional datasets have been undergone intensive development due to the collection of data as curves, surfaces or measurements that vary over continua. Recently, with the rapid improvements of the associated technologies, the quality of data collected has become more precise in various fields such as in the biometric, chemometric, medicine, environmetrics, etc. Ramsay and Silverman (1997, 2005) noted the field of the statistics of functional data had become increasingly popular, and further discussed a number of related case studies and applied issues Ramsay and Silverman (2007) that used linear regression and multiple regression (parametric model), which predict a scalar response variable from a functional covariate and a functional response from a functional covariate, respectively. Cardot et al. (2003) pointed to the regression analysis of functional data in the instance where the response is a real random variable and a square integrable covariate random function defined on some compact set, which also introduced a continuous model of the functional linear regression model with a scalar response; their study also mentioned a smooth version of functional principal components regression for which $L^2$ convergence can be achieved. The linear model for regression with functional data was also discussed by James (2002), Cai et al. (2006), Hall et al. (2007), and Ahmedou et al. (2016), whilst the functional response method with scalar predictors was studied by Faraway (1997) and Chion et al. (2003, 2004)). There are several examples of work in the literature
discussing nonparametric statistics, which can be found in a previous monograph for applied nonparametric regression by Härdle (1990), where kernel smoothers are investigated in great detail. Recently, statistical methods for longitudinal data were proposed for the functional linear model - see, for instance, Fan and Fan and Zhang (2000) and the application of their proposed model, as discussed by local polynomial smoothing, and who established a number of asymptotic results for local polynomial estimators.

Since the book published by Ferraty and Vieu (2006), the novelty of nonparametric functional data analysis has become increasingly apparent, and nonparametric functional regression methods have since been the main theme of a number of studies. The goal of their book was to study both the theory and implementation of these methods through different statistical problems containing prediction, time series, and classification. They also mentioned several different types of functional data which are applied to the different models nonparametrically, where three different types of semi-metrics are defined (as based on functional principal component analysis, derivatives, and partial least-squares) in detail for the measurement of the proximity between functional covariates. Another aim of their book was to present certain asymptotic results related to the nonparametric estimation of the three functional predictors (conditional expectation, conditional median and conditional mode), as linked to the issue of predicting real response variables given an explanatory variable which is allowed to be of infinite dimensions.

The existing literature contains a considerable number of theoretical and experimental studies with different models on nonparametric functional data when the response is independent response and covariate is functional. For instance, the functional Nadayara-Watson (NW) estimator model used kernel function with cross-validation procedure for choosing optimal bandwidth and semi-metrics for measure of proximity between curves by Ferraty and Vieu (2003, 2004, 2006), Ferraty et al. (2007), Preda (2007) and Rachdi and Vieu (2007) and their study contains some asymptotic results. Burba et al. (2009) described the functional k-nearest neighbour estimator with nonparametric functional regression where the asymptotic properties of the model were presented and it was noted that the rate
of convergence is similar to that of the Nadayara-Watson kernel estimator. Their study was applied to real data and stimulation data, from which it was reported that the k-nearest neighbour model can take into account the local structure of the data and give better predictions when the data are heterogeneously concentrated. Baillio and Grané (2009) presented a local linear regression with the nonparametric functional regression model in the context that the covariate is functional and the response is a scalar vector and the asymptotic assumption is considered to hold. Their most recent study was compared with a Nadayara-Watson kernel estimator and with linear regression via a Monte Carlo study and analysis of two real datasets and a simulation set, where in all cases the local linear regression performed better other models. For more information on the functional local linear estimator see, for example, Barrientos-Marin et al. (2010), and for the distance-based local linear estimator, for example, Boj et al. (2010). Ferraty et al. (2010) illustrated a bootstrap procedure to approximate the distribution in nonparametric functional regression, where both a naive and a wild bootstrap procedure are discussed, and their asymptotic validity demonstrated to confirm the procedure, which was applied to both simulated and real datasets. Geenens et al. (2011) developed the procedure of closeness between functional covariates via semi-norms, which is often suggested as a technical tool that can be used for dimension reduction purposes, in order to resolve otherwise intractable problems such as infinite dimensionality. Zhou and Lin (2016) pioneered the locally modelled regression estimator instead of the Nadaraya-Watson estimator for the nonparametric regression function of a scalar response on a functional covariate, and considered the asymptotic properties of this estimator for functional data, also they adapt the empirical likelihood method to construct the point-wise confidence intervals for the regression function and derive the Wilk’s phenomenon for the empirical likelihood inference.

Recently, multivariate regression modelling was also proposed for use with functional data analysis, for instance by Matsui et al. (2008) to illustrate the relationship between multiple scalar responses and functional predictors using Gaussian basis functions along with regularization, where the efficiency of the proposed method can be illustrated through Monte Carlo simulations and application to real data (spectrometric data). Chaouch et al. (2013) noted the problem of multivariate
response variables from functional covariates based on the \( L_1 \)-median regression estimation, and also proposed results by making the asymptotic assumption for the estimator, such as consistency, asymptotic normality, and evaluation of the bias expression. Their study also compared three approaches, namely the \( L_1 \)-median regression, the vector of marginal conditional median, and the non-functional multivariate median to predict random vectors. The nonparametric regression analysis for multivariate longitudinal data was studied by Xiang et al. (2013). In their study, the correlation between different components of the responses was considered, and which also used different bandwidths for different components of the responses. The model was based on local polynomial kernel smoothing, and where the theoretical part of the proposed model was studied.

Here, we develop a new method of nonparametric functional data analysis when the covariate \( X \) takes values in some infinite dimensional space and where the response \( Y \) is the multivariate response. For spectrometric instances, many authors have taken fat content as a scalar response and then made predictions from curves, so in our work we will also take water and protein content as scalar responses and their content predicted separately. To the best of our knowledge, the multivariate response vector problem with nonparametric conditional expectation has not been previously reported in the literature. We also note that it would be preferable to predict all components of a vector of random variables simultaneously in order to take into account any correlation between them, rather than predicting each component separately. For example, in real spectrometric data the estimation of the three variables (fat, water, and protein contents) together represents an important research issue in chemistry in terms of both cost and time savings. Notice that the fat, water and protein content of each piece of meat are strongly correlated. Therefore, it is more convenient to predict these variables simultaneously, rather than predicting each of them separately. On the other hand, for the Canadian Weather Station data, with temperature curves as the covariate function and average monthly precipitation as the multivariate response vectors, its should be noted that the values for average monthly precipitation are correlated, so once again it is more convenient to predict these variables simultaneously rather than individually. The performance of the proposed model is compared with the
independent response model from the functional covariates in the nonparametric approach.

1.3 Principal Component Analysis in Nonparametric Functional Regression

Principal Component Analysis is considered a useful tool in multivariate situations for illustrating data in a reduced dimensional space. Recently, principal component analysis has seen use in several different statistical approaches such as being extended to functional data and to decorrelate multivariate response variables. Daultrey (1976) pioneered the principal component analysis is transformation manner for a set of data. Chatfield and Collins (1980) indicated the relationships between a set of $p$ correlated variables could be transferred to a new set of uncorrelated variables called principal components, where the new variables are linear combinations of the original variables and are derived in decreasing order of significance such that, for instance, the first principal component accounts for as much of the variation in the original data as possible. Critchley (1985) illustrated principal component analysis with linear regression and the theoretical influence function developed in their study. Abdi and Williams (2010) and Bro and Smilde (2014) noted the idea of principal component analysis in many different areas, especially the chemometric.

In recent years, multivariate regression modelling was also proposed for use in functional data analysis as mentioned in the previous section, for example by Matsui et al. (2008), and Chaouch et al. (2013). Wang et al. (2017) proposed the Gaussian process regression with multivariate response and the use of principal component analysis for the decorrelation of multivariate responses with functional covariates and multivariate covariate variables. Their study used two different kinds of semi-metric, as based on the functional principal component analysis, and on derivatives. Their study improved the results obtained from simulated and real data examples (spectrometric data and soil data).
According to Omar and Wang (2017) extended the independent response model to the multivariate response model with a functional covariate in nonparametric functional data analysis. In their model, they took the correlation between different components of the response variables into account and used also different bandwidths for different components of the response variables. In this study, a new model is proposed to deal with multivariate responses and functional covariates. We use principal component analysis to decorrelate the multivariate responses or each component of the responses will be a linear combination of other components of the responses. Furthermore, we use the KNN method for independent prediction, and a cross-validation procedure is used to obtain the optimal bandwidth. In the KNN model, we utilize two different kinds of semi-metric (one based on the derivative and one built on the functional principal component analysis) as a measure of the closeness between the functional covariates.

We add some new results to those previously found for nonparametric functional regression when the covariate $X$ are functional and for a response $Y$ use principal component analysis to decorrelate the multivariate response (or each component of the response will be a linear combination of other components of the response). As we know, the multivariate response vectors problem with principal component analysis in nonparametric functional regression has not been previously reported in the literature. The performance of the proposed method is compared with the independent response from functional covariates in the nonparametric approaches.

1.4 Correlation Between Different Components of the Responses in Functional Data

Functional data concerns data which are collected as curves, surfaces or measurements that vary over a continuum. With the rapid advances in technology, functional data have become increasingly prevalent in a large number of fields, such as the environment, medicine, finance, industry and social sciences.
Nonparametric modelling has also been proposed for other types of functional data analysis. For instance, Wang and Chen (2015) proposed the formulation of the covariance function for the multi-response Gaussian process regression, and the presented method is able to learn from the data dependencies between different outputs. The excellence of the multi-response Gaussian process regression model over the independent Gaussian process regression is illustrated through a numerical example. Xiang et al. (2013) studied multivariate nonparametric regression analysis in the context of longitudinal data. In their study, they used different bandwidths for different components of the responses and took the correlation between different components of response variables into account. They applied their model to both real and simulated data. In section 1.2, we noted that the model directly extended from an independent response variable to multivariate response variables in the functional data analysis.

We have developed two new solutions to address the functional regression problem with functional covariates and a multivariate response variable. The first solution is to directly extend the nonparametric method for univariate response to a multivariate response with different bandwidths for different responses. In the second solution, the correlation between different responses is taken into account with different bandwidths for different responses and incorporated into the model. Our methods utilize the kernel function with automatic bandwidth selection via a cross-validation procedure and semi-metrics as a measure of the proximity between functional data. The rate at which almost complete convergence of the methods was achieved had been studied, and the effectiveness of the proposed methods is demonstrated through several simulation studies and a real data example (spectrometric data).

1.5 Multivariate Functional Responses

In recent years, interest has grown in the problem of functional regression due to the sophistication of recent technological advances regarding the collection and storage of data as curves. This field of study began with the popular monograph by Donoho and Johnstone (1994) and Ramsay and Silverman (1997, 2005) that
gives an itemized demonstration of the functional linear model. The existing literature includes the instance of the functional linear model with scalar response and functional predictors (Cardot et al. (2003), James (2002), Müller et al. (2005), Cai et al. (2006), Hall et al. (2007), Ahmedou et al. (2016)) and the functional response model with scalar predictors approach (see, for instance, Faraway (1997), Chiou et al. (2003, 2004)).

Ramsay and Dalzell (1991) introduced a functional response model with a functional covariate, and the study displays how the theory of $L - splines$ can support generalisations of linear modelling and principal components analysis to samples drawn from random functions, research into which has been further expanded by a number of researchers (see for example, Cuevas et al. (2002), Aguilera et al. (2008), Crambes et al. (2013), Horváth et al. (2009)). Ramsay and Silverman (2005) predicted average monthly precipitation curves from average monthly temperatures curves, and Antoch et al. (2008) used historical data to predict future hourly electricity consumption. Kadri et al. (2010) proposed nonlinear functional regression whereby the responses are functions and the input attributes, and developed the kernel Hilbert space approach and demonstrated the basic properties of kernel Hilbert space, proving the theorem for this setting. Chiou et al. (2016) proposed a multivariate functional linear regression approach to predict multivariate functional data in the instance when both the response and the predictor contain multivariate random functions. Their study also explains the multivariate functional linear regression model, connects with the multivariate functional principal component analysis study, which takes the preference of cross-correlation between component functions within the multivariate response and predictor variables, respectively.

In previous sections, we mentioned the novelty of nonparametric functional data analysis when the covariate is functional and the response is real or multivariate.

Recently, the functional response model was also suggested for nonparametric functional data analysis. For instance, Lian (2007) presented an expansion of reproducing kernel Hilbert space theory which supplied a new structure for analysing functional responses from functional covariate with nonlinear regression
models, and used a cross-validation procedure to choose automatic smoothing parameter estimation. Shi et al. (2007) noted the use of a Gaussian process functional regression method to model functional response curves with a set of functional covariates. Their study addressed two issues through their model: modelling a non-linear and nonparametric regression relationship and modelling covariance and mean structures simultaneously, to confirm their hypothesis, which was applied to both real and simulated data and showed that the model provided excellent results for the purposes of curve fitting and prediction. Shi and Wang (2008) proposed Gaussian process functional regression with a functional response and functional covariate, where their study focussed on the problem of clustering functional relationships between response functions and covariate curves. Lian et al. (2011) investigated the rates of strong (almost sure) convergence of functional k-nearest neighbour regression estimates with functional responses. Lian et al. (2012) illustrated the almost sure convergence rates of both the Nadaraya-Watson and nearest neighbour estimators in a unified style when both predictors and responses are functions but in two different spaces, the semi-metric space and a separable Hilbert space, respectively. Ferraty et al. (2011) noted a number of asymptotic properties (uniform, almost complete convergence rate) for the kernel regression estimator when both the covariate and response variables are functional. Ferraty et al. (2012) illustrated the formulae for asymptotic bias and variance in the instance when the response and the covariate were both functions. Their approach studied both a naive and a wild componentwise bootstrap procedure, and their asymptotic validity was demonstrated.

We consider the nonparametric methods for functional regression when the explanatory is functional and the response is a multivariate function, which, to the best of our knowledge, have not been previously addressed in the literature. We consider two methods: firstly, the method for univariate functional response described by Ferraty et al. (2012), which is directly extended to multivariate functional responses; we then propose a new method whereby the correlation between different functional responses is taken into account. The functional kernel with automatic bandwidth selection is utilized by the cross-validation procedure along
with a semi-metric built on the functional principal component analysis as a measure of the proximity between curves. The usefulness of the proposed models is illustrated through simulated and real data examples.

1.6 Contributions and Outline of the Thesis

The main topic of this thesis is to develop nonparametric functional regression methods and then apply them to functional data and, subsequently, the related theories in the instance when the covariate is a function and the response is a multivariate response variable, and where the correlation between different components of the responses is taken into account through the different bandwidths for different components of the responses. The main contributions of this thesis are discussed in Chapters 3, 4, 5, and 6 in more detail. This thesis contains three papers, where two revised manuscripts and one submission have been achieved during the course of this thesis:


The remainder of the thesis is organized as follows.

In Chapter 2, we briefly present a number of fundamental ideas about nonparametric functional regression, discuss two different forms of semi-metric (based on derivatives and built on functional principal component analysis), and local weights in the real, multivariate and functional cases. The estimating nonlinear regression model is by the kernel functional estimator, and choosing the optimal bandwidth is by cross-validation procedure.

In Chapter 3, we extend the independent response variable model to the multivariate response variable method in nonparametric functional regression, estimating the nonlinear model via the functional k-nearest neighbour estimator and choosing the bandwidth corresponding to the optimal number of neighbours via a cross-validation procedure. In nonparametric functional regression, the closeness
between functional covariates was calculated via semi-metrics (as based on derivatives, and built on functional principal component analysis). The multivariate response variable model is then applied to two types of functional data (spectrometric data and Canadian Weather Station) and a set of simulated data. We compare the predictive performance of the proposed model via the independent response variable method by calculating the mean square error between the true values and the predictor values.

In Chapter 4, a new model is presented to deal with correlated multivariate response variables with functional covariates. We use principal component analysis to decorrelate the multivariate response, which then enables us to model each principal component independently via nonparametric functional regression. In the nonparametric functional regression model, the closeness between functional covariates is measured via semi-metrics (as based on derivatives and on functional principal component analysis). The model presented uses the k-nearest neighbour method with automatic bandwidth selection via a cross-validation procedure. The performance of the proposed method is compared with that of the independent response, where the response variables are modelled independently and without considering their correlation. The effectiveness of this functional regression model is verified through instances of simulated and real data (spectrometric data and Soil data).

In Chapter 5, we present two models to the functional regression problem. The first method is to directly extend the nonparametric method for the univariate response to multivariate responses with different bandwidths for different responses. In the second model, the correlation between different components of the responses is taken into account using different bandwidths for different responses which are incorporated into the model. Our methods use the kernel function with automatic bandwidth selection via a cross-validation procedure, using semi-metrics as a measure of the proximity between functional data. Two kinds of semi-metrics were used in this chapter (as based on derivatives and on functional principal component analysis). The asymptotic properties of the estimators are studied, and the effectiveness of the proposed methods is demonstrated through several examples of simulated and real data. A similar study to that presented
in this chapter is given in a paper published in the journal COMMUNICATION IN STATISTICS-THEORY AND METHODS.

Chapter 6, Ferraty et al. (2012) illustrated a nonparametric regression method for cases in which the response and the covariate are both functional. In this chapter, we study the functional regression with multivariate functional responses. We first extend the method from the univariate functional response to the multivariate functional responses model, and then propose a new method whereby the correlation between different functional responses is taken into account. The functional kernel with automatic bandwidth selection is utilized by the cross-validation procedure, which uses a semi-metric built on the functional principal component analysis as a measure of the proximity between curves. The utility of the proposed models is illustrated through simulated and a real data examples of the UK weather, in which their performances are compared. It is identified that the multivariate functional responses model with correlation, which takes advantage of the correlation between different functional responses, and improves predictive accuracy in terms of mean square error. A similar study to that presented in this chapter was presented at a conference, the Research Students Conference 2018, at the University of Sheffield, UK.

Some important discussion and future work are offered in Chapter 7.
Chapter 2

Preliminaries

2.1 Background

Utilizing functional data requires one to answer important statistical questions. Actually, the larger the space \( E \) from which the variable takes its values, the sparser the resultant data. In the case of functional data, we know that \( E \) is an infinite dimensional space. Therefore, in this chapter, we will review some important essential problems relating to infinite dimensional data. In particular, the sparseness concept is strongly related to the method used to measure the proximity of curves; here, we used a semi-metric to approach this problem.

2.2 Semi-metric

We use a semi-metric in functional data instead of a general metric because the metric spaces can otherwise become too restrictive; semi-metric spaces are better adapted than metric spaces to these kinds of situation. In some cases, measuring the closeness of two curves requires the use of a classical norm, especially in a finite dimensional space, but this type of measure is not generally important except in several practical situations.

Ferraty and Vieu (2003, 2006) illustrate that the Euclidean norm \( \| \cdot \| \) is the most popular in \( \mathbb{R}^p \), which is based on the squares of the components of any vector. According to Ferraty and Vieu (2006), \( \| \cdot \| \) is defined as semi-norm on some space \( F \):

1. \( \forall (\lambda, x) \in \mathbb{R} \times F, \| \lambda x \| = |\lambda| \| x \| \),
2. \( \forall (x, y) \in F \times F \times F, \quad \| x + y \| \leq \| x \| + \| y \|. \)

If we let \( x = (x_1, \ldots, x_p)^T \) be a vector in \( \mathbb{R}^p \), then the classical formula for the Euclidean norm is defined by:

\[
\| x \|^2 = \sum_{j=1}^{p} (x_j)^2 = x^T x.
\]

From this formula it can be inferred that a family of norms are based on the Euclidean norm by utilizing different definite positive matrices \( M \) in the following manner:

\[
\| x \|^2_M = (x)^T M x.
\]

According to Ferraty and Vieu (2006),  is defined as a semi-metric on the sample space \( F \):

1. \( \forall x \in F, d(x, x) = 0. \)

2. \( \forall (x, y, z) \in F \times F \times F, \quad d(x, y) \leq d(x, z) + d(y, z). \)

Notice that unlike metrics the semi-metrics allow to have \( d(x, y) = 0 \) for \( x \) not equal to \( y \).

Now, in the functional case the choice of the semi-metric becomes crucial.

Ferraty and Vieu (2006) pointed out the different kinds of semi-metrics. Thus, initially, statisticians used Principal Component Analysis as the measure of the proximity between two curves or subjects being calculated by means of the classical \( l_2 \)-metric, which is defined by:

\[
\sqrt{\left( \int (\chi_i(t) - \chi_i'(t))^2 dt \right)},
\]

for all observed curves \( \chi_i \) and \( \chi_i' \).

Nevertheless, another type of semi-metric used as a measure of proximity between curves which is based on the second derivative:

\[
\sqrt{\left( \int (\chi_i^{(2)}(t) - \chi_i'^{(2)}(t))^2 dt \right)}.
\]
Thus, one advantage of this particular semi-metric approach is that it can detect more structure in the variable than the $l_2$-metric Principal Component Analysis.

Because of the different types of datasets and various statistical models available, from a more practical perspective we require several kinds of semi-metrics to measure proximity, and functional datasets are well adapted to measure closeness between curves via different types of semi-metrics. In this chapter, we focus on two different families of semi-metrics, semi-metric based on Functional Principal Component Analysis and semi-metric built on derivative, which is based on the shape of the curve.

### 2.2.1 Functional Principal Component Analysis

With today's rapid technological advances, technologies which allow datasets to be recorded are becoming increasingly common, often being applied to financial variables. Frequently, there is a potential to reduce the number of dimensions to a problem while still retaining much of the information available in the original data. Principal component analysis is probably the best known technique and most frequently used to reduce the dimensionality of multivariate data. Finding linear combinations, called principal components, via principal component analysis can reduce the dimensionality, that successfully have maximum variance for the data, see for instance Jolliffe (2011).

Several decades ago, the majority of statisticians used principal component analysis for money different solutions was then extended to functional principal component analysis because of different types of data such as functional data and longitudinal data. Locantore et al. (1999) using principal component analysis with functional data which images is consider, and the objects are summarized by feature vectors. For examples of functional principal component analysis techniques applied to financial time series, see for example, Ingrassia and Costanzo (2005, 2006). Viviani et al. (2005) applied principal component analysis to functional magnetic resonance imaging (fMRI) data based on nonparametric functional data analysis, and showed that functional principal component analysis is more effective than its standard counterpart. Their study also discussed the rationale and advantages of the present model compared to other models, such as clustering...
or independent component analysis. Functional principal component analysis can also be applied to longitudinal data, see Greven et al. (2011) proposed an estimation procedure that is based on principal components.

Rice and Silverman (1991) pointed out a model of smoothed principal components analysis of functional data such as estimating the mean function in a nonparametric manner under the assumption that this function is smooth, and suggest a variation on the usual method of cross-validation to chose the degree of smoothing. Silverman (1995) extended the idea of functional principal component analysis to deal with data that is a hybrid of functional and parametric effects. His approach emphasized the situation where it is helpful to support that the method for the observed data contains parametric effects that are specific to each observation. His methodology was applied to Canadian weather data. According to Silverman et al. (1996), an attractive model that incorporates smoothing of the functional principal component analysis of functional data replaces the usual $L^2$ – orthonormality constraint on the principal components by orthonormality with respect to an inner product. Their study also proposed some theoretical properties of the model, such as its estimates can be shown to be consistent under appropriate conditions, and asymptotic expansion styles are used to investigate their bias and variance properties, and cross-validation is discussed as the choice of smoothing parameter. Boente and Fraiman (2000) presented kernel-based estimates of the functional principal components when the data are a continuous channel of stochastic processes, and noted that the strong consistency and the asymptotic distribution are derived under mild conditions. Yao and Lee (2006) developed an iterative technique which reduces the dependence on the repeated measurements made for the same subject, and their study shows that, after iteration, the resulting data can be theoretically shown to be asymptotically equivalent to sets of independent data. Hall and Hosseini-Nasab (2006) discuss the properties of functional principal component analysis, which reduces the dimensionality of functional data analysis to a finite level, and points to the most significant components of the data. Their study also suggests bootstrap models for building simultaneous confidence regions for an infinite number of eigenvalues, and also for individual eigenvalues and eigenvectors. Di et al. (2009) illustrated multilevel
functional principal component analysis and applied it to the Sleep Heart Health Study. Górecki and Krzyśko (2012) proposed a functional principal component analysis based on the principal components found for vector data.

### 2.2.2 Semi-metric Construct Based on Functional Principal Component Analysis (FPCA)

Historically, principal component analysis was found to be a useful tool for displaying data in a reduced dimensional space in multivariate data. Because of development of collecting data in large sets and other statistical purposes, principal component analysis was extended to functional principal component analysis for functional data in order to measure the proximity between curves in a reduced dimensionality; for further discussion in this regard, see Dauxois et al. (1982), Castro et al. (1986), Ferraty and Vieu (2003, 2006), and Febrero-Bande et al. (2012).

If we suppose $\mathcal{X}(t)$ to be a functional data, we assume that $E \int \mathcal{X}^2(t)dt$ is finite (see, for instance, Ferraty and Vieu (2003, 2006)). The functional principal component analysis of a functional random variable $\mathcal{X}$ allows the following expansion of $\mathcal{X}$ to be obtained:

$$\mathcal{X} = \sum_{k=1}^{\infty} \left( \int \mathcal{X}(t)\nu_k(t) dt \right) \nu_k,$$

$\nu_1, \nu_2, \ldots$, being the orthonormal eigenfunctions of the covariance operator discussed by Silverman et al. (1996), and Ferraty and Vieu (2003, 2006).

$$\Gamma_{\mathcal{X}}(s,t) = E(\mathcal{X}(s)\mathcal{X}(t)),$$

is associated with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$. Now, let

$$\tilde{\mathcal{X}}^{(q)} = \sum_{k=1}^{q} \left( \int \mathcal{X}(t)\nu_k(t) dt \right) \nu_k,$$

be a truncated text of the previous expansion of $\mathcal{X}(t)$. This truncated version minimizes $E \left( \int (\mathcal{X}(t) - P_q\mathcal{X}(t))^2 dt \right)$ over all projections $P_q$ of $\mathcal{X}$ into $q$-dimensional
spaces. Therefore, a parameterized class of semi-norms from the classical $L_2$-norm is defined in the following way:

$$\|\chi\|^{PCA}_q = \sqrt{\int (\hat{\chi}^{(q)}(t))^2 dt} = \left[ \sum_{k=1}^{q} \left( \int \chi(t) \nu_k(t) dt \right)^2 \right]^{\frac{1}{2}},$$

then the parameterized family of semi-metrics comes from the previous expansion:

$$d_q^{PCA}(X_i, \chi) = \left[ \sum_{k=1}^{q} \left( \int [X_i(t) - \chi(t)] \nu_k(t) dt \right)^2 \right]^{\frac{1}{2}}.$$

In this case, $q$ is a tuning parameter and not really a smoothing parameter. Note that in practice, $\Gamma_{X}$ and $\nu_k$'s are unknown, but the covariance operator can be approximated using empirical methods.

$$\Gamma^n_X(s, t) = \frac{1}{n} \sum_{i=1}^{n} X_i(s) X_i(t),$$

and the eigenfunctions of $\Gamma^n_X$ are consistent estimators of those $\Gamma_X$ see Ferraty and Vieu (2006).

In fact, we never observe $\{X_i = \{\chi_i(t); t \in T\}\}_{i=1,...,n}$ but only a discretized version $x_i = (\chi_i(t_1), ..., \chi_i(t_J))^T_{i=1,...,n}$ (note that this is straightforward when the data are balanced). Therefore, from a practical point of view, according to Castro et al. (1986), we can approximate the integral in the following way:

$$\int [X_i(t) - \chi(t)] \nu_k(t) dt \approx \sum_{j=1}^{J} \omega_j \mid X_i(t_j) - \chi(t_j) \mid \nu_k(t_j),$$

where $\omega_1, ..., \omega_J$ are quadrature weights which define the approximate integration. To fix ideas, note that a standard choice could be $\omega_j = t_j - t_{j-1}$. Hence, if we have two discretized curves $x_i$ and $x_{i'}$, then the quantity $d_q^{PCA}(X_i, X_{i'})$ can be approximated by:

$$d^{PCA}_q(x_i, x_{i'}) = \left[ \sum_{k=1}^{q} \left( \sum_{j=1}^{J} \omega_j (\chi_i(t_j) - \chi_{i'}(t_j)) [v_k]_j \right)^2 \right]^{\frac{1}{2}},$$

where $v_1, v_2, ..., $ are the $W$-orthogonal eigenvectors of the covariance matrix ($W = diag(\omega_1, ..., \omega_J)$).
\[ \Gamma^n W = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T W, \]

associated with the eigenvalues \( \lambda_{1,n} \geq \lambda_{2,n} \geq \ldots \). Note that as soon as the grid \((t_1, \ldots, t_J)\) is sufficiently fine, then \( d_{q}^{PCA}(x_i, x_{i'}) \) is close to \( d_{q}^{PCA}(\chi_i, \chi_{i'}) \).

Therefore, this kind of semi-metric has both advantages and disadvantages. One of the disadvantages is that it can only be used with balanced data. On the other hand, this type of semi-metric is advantageous because it can be utilized even if the curve is rough.

### 2.2.3 Semi-metric Based on Derivative

The second semi-metric to construct a parameterized family between curves considers a distance between one among their derivatives. Consider the following semi-metric of two observed curves \( \chi_i \) and \( \chi_{i'} \):

\[
d_{q}^{\text{deriv}}(\chi_i, \chi_{i'}) = \int \left( \chi_i^{(q)}(t) - \chi_{i'}^{(q)}(t) \right)^2 dt
\]

where \( \chi^{(q)} \) denotes the \( q \)th derivatives \( \chi \); note that \( d_{0}^{\text{deriv}}(\chi, 0) \) is the classical \( l^2 \)-norm of \( \chi \) (see for instance, Ferraty and Vieu (2003, 2006), and Febrero-Bande et al. (2012)). Also the authors noted that the computation of successive derivatives is sensitive numerically, and thus a B-Spline approximation should be used to address this numerical problem for each curve so that the successive derivatives are directly computed by differentiating their analytic forms many times.

In this case, we use a \( B \)-spline basis to approximate the discretized curve \( x_i = (\chi_i(t_1), \ldots, \chi_i(t_J))^T \). Let \( \{B_1, B_2, \ldots, B_B\} \) be a \( B \)-spline basis (a spline of order \( n \) is a piecewise polynomial function of degree \( n - 1 \) in a variable \( x \)) (for more detail about the monograph of \( B \)-splines, see Härdle (1990), Ramsay and Silverman (2005), Cardot et al. (2003), and Ferraty and Vieu (2006)):

\[
\hat{\chi}_i(.) = \sum_{b=1}^{B} \hat{\beta}_b B_b(.),
\]
where

\[
\hat{\beta}_i = (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{iB}) = \arg \inf_{(\alpha_1, \ldots, \alpha_B) \in \mathbb{R}^B} \sum_{j=1}^{p} \left( \chi_i(t_j) - \sum_{b=1}^{B} \alpha_b B_b(t_j) \right)^2.
\]

In this formula, the \(B_b\)'s are well known, so the successive derivative can be computed and the approximated curves can be easily differentiated:

\[
\hat{\chi}_i^{(q)}(\cdot) = \sum_{b=1}^{B} \hat{\beta}_b B_b^{(q)}(\cdot),
\]

where the analytic expression of \(B_b^{(q)}(\cdot)\) is known. Now, since \(x_i\) and \(x_i'\) are two discretized curves, then the semi-metric between them can be expressed as follows:

\[
d_{\text{deriv}}(x_i, x_i') = \sqrt{\int \left( \hat{\chi}_i^{(q)}(t) - \hat{\chi}_i'^{(q)}(t) \right)^2 dt}.
\]

To calculate the previous formula, let:

\[
f(t) = \left( \hat{\chi}_i^{(q)}(t) - \hat{\chi}_i'^{(q)}(t) \right)^2.
\]

According to Ferraty and Vieu (2003, 2006), in order to calculate this integral using the Gauss method, this method can be applied to evaluate the integrand \(f\) at any point. Indeed, the Gauss method proposes the following approximation:

\[
\int_{a}^{b} f(t) dt \sim \frac{b-a}{2} \sum_{i=1}^{K} \omega_k f \left( \frac{b - a}{2} + \frac{b - a}{2} \delta_k \right),
\]

where the weights \(\omega_k\) and the real \(\delta_k\) are tabulated, so that the accuracy of this numerical method comes from the fact that it is exact for any polynomial of degree \(\leq 2K - 1\).

We can use this class of semi-metric even in the context of unbalanced data (i.e., when the curves are not necessarily observed at the same points), where one advantage of the semi-metric is that due to the curves being replaced by their \(B\)-spline expansions. Therefore, in practice this kind of semi-metric will be well adapted and used in the presence of smooth curves.
2.3 Local Weighting

This section is directly related to the section on semi-metric modelling because local weighting requires certain topological methods to measure the proximity between curves. Obviously, local studies need to have some topological methods at hand to measure the proximity between curves, hence this section is directly connected with the semi-metric modelling discussed in the previous section.

In the community of nonparametricians, the idea of local weighting methods is very popular, especially in the context of finite dimensional data because the local weight techniques are well adapted to nonparametric models. The kernel is the most common approach to local weighting in finite dimensional cases.

In this section, we expand on how the kernel’s smoothing notions can be adapted to infinite dimensional variables and also discuss the kernel method, as well as how it is applied to finite dimensional data. First, we define the kernel local weighting in real and a multivariate cases and then extend this definition to functional data.

2.3.1 Real Case

Kernel local weighting is built on the kernel function denoted by \( K \) and a smoothing parameter (which is called the bandwidth \( h \)); then, if \( x \) is a fixed real number, the kernel local weighting transforms \( n \) real random variables (r.r.v.) \( X_1, X_2, ..., X_n \) into \( \Delta_1, \Delta_2, ..., \Delta_n \) as follows:

\[
\Delta_i = \Delta_i(x, h, K) = \frac{1}{h} K \left( \frac{x - X_i}{h} \right).
\]

The aim of the local weighting around \( x \) is to trail each real random variable \( X_i \) with a weight, taking into account the distance between \( x \) and \( X_i \); if \( x \) is close to \( X_i \), then the higher the weighting (see, for example, Härdle (1990) and Ferraty and Vieu (2006)).

Before continuing, we should define what a kernel function actually is in this situation. Indeed, there are several different kinds of kernel function, where we can consider any density function to a kernel but even unnecessary positive
functions can be pointed by Gasser and Müller (1979). There is a large body of
the literature that considers this field (see, for instance, Marron et al. (1988) and
Bosq (1993)). Figure 2.1 displays several different types of functional kernel which
are analytically defined as below:

a) Rectangular kernel:

\[ K(x) = \frac{1}{2} 1_{[-1,1]}(x). \]

b) Triangle kernel:

\[ K(x) = (x + 1) 1_{[-1,0]}(x) + (1 - x) 1_{[0,1]}(x). \]

c) Quadratic kernel:

\[ K(x) = \frac{3}{4} (1 - x^2) 1_{[-1,1]}(x). \]

d) Gaussian kernel:

\[ K(x) = \frac{1}{\sqrt{2\pi}} exp\left( -\frac{x^2}{2} \right). \]

For a more precise idea of kernel local weighting, we will use the rectangular
kernel as an example and rewrite \( \Delta_i \)'s as below:

\[ \Delta_i = \frac{1}{h} 1_{[x-h,x+h]}(X_i). \]

In this case, the local weighting feature occurs and the real random variable inside
the range \([x-h, x+h]\) is one; otherwise, it is discounted. Also, the normalization
\( \frac{1}{h} \) is proportional to the size of the set \([x-h, x+h]\) by (Härdle (1990), Ferraty
and Vieu (2003, 2004, 2006), and Rachdi and Vieu (2007)). Therefore, these ideas
are not only true for the Box kernel but for all types of kernel.
2.3.2 Multivariate Case

From the real kernel, local weighting can be extended to the multivariate case. In this situation, it is observed that $n$ random vectors $X_1, X_2, ..., X_n$ are valued in $\mathbb{R}^P$. First, we have to define the multivariate kernel $K^*$, which is a function of $\mathbb{R}^p$ to $\mathbb{R}$. The first way to do this is to define $K^*$ as the product of $p$ real kernel functions $K_1 = K_2 = ... = K_p$:

$$\forall u = (u_1, ..., u_p)^t \in \mathbb{R}^p, K^*(u) = K_1(u_1) \times ... \times K_p(u_p).$$

The second way combines a real kernel function $H$ with a norm $\| \cdot \|$ in $\mathbb{R}^p$ as follows:
\[ \forall u \in \mathbb{R}^p, K^*(u) = K(\| u \|). \]

If \( K_1 = K_2 = \ldots = K_p = 1_{[-1,1]} \) and if \( \| . \| \) is the supremum norm, both studies coincide by taking \( K = 1_{[0,1]} \) and \( \| u \| \) as a positive quantity (see for example, Ferraty and Vieu (2006)).

To define multivariate kernel local weighting in a similar manner to the real kernel local weighting, let \( X \) be a vector of \( \mathbb{R}^p \) and transform the \( n \) random vectors \( X_1, X_2, \ldots, X_n \) into \( n \) variables \( \Delta_1, \ldots, \Delta_n \):

\[
\Delta_i = \frac{1}{h^p} K^* \left( \frac{x - X_i}{h} \right),
\]

where \( \Delta_i \) are local weighting transformations of the variables \( X_i \), if \( X_i \) is out of some neighbourhood of \( x \) then \( \Delta_i = 0 \), and the normalization \( \frac{1}{h^p} \) is proportional to the volume of the set \( X_i \) (for more detail see, for example, Härdle (1990), Ferraty and Vieu (2006), and Rachdi and Vieu (2007)).

### 2.3.3 Functional Case

The main difference between the two previous cases is the introduction of the kernel local weighting in the functional case. In this situation, let \( \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n \) be \( n \) functional random variables (\( f.r.v. \)) valued in \( E \). Indeed, a function is extended from multivariate kernel local weighting notions. In this case, we would transform the \( n \) functional random variables \( \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n \) into the quantities:

\[
\frac{1}{V(h)} K \left( \frac{d(\chi, \mathcal{X}_i)}{h} \right),
\]

where \( d \) is a semi-metric of \( E \), and \( K \) is a real kernel. In this situation \( V(h) \) would be the volume of:

\[
B(\chi, h) = \{ \chi' \in E, d(\chi, \chi') \leq h \}.
\]
This is a ball centred at $\chi$ with a radius $h$ described by Ferraty and Vieu (2004, 2006). In this case, we need to have at hand a measure of $E$. The user is free to choose any particular measure, using the probability distribution of the functional random variables to establish normalization in this situation, where the functional kernel local weighting variables are defined by:

$$\Delta_i = \frac{K\left(\frac{d(x,X_i)}{h}\right)}{E\left(K\left(\frac{d(x,X_i)}{h}\right)\right)}.$$  \hfill (2.1)

Ferraty and Vieu (2006) illustrated that this class expands from the multivariate case, so that for some constant $C$ that depends on $K$ and on the norm $\| \cdot \|$ used in $R^p$:

$$E[K(\| x - X_i \| /h)] \sim Cf(x)h^p,$$

as long as $X_i$ has a density of $f$ with respect to the Lebesgue measure, which is continuous and such that $f(x) > 0$. Therefore, it is clear that (2.1) is an extension of the multivariate kernel local weighting in the functional framework. Ferraty and Vieu (2006) defined two different kinds of kernels for weighting functional variables.

1. A function $K$ from $R$ into $R^+$ such that $\int K = 1$ is called a kernel of kind $I$ if there exist two real constants $0 < C_1 < C_2 < \infty$ such that:

$$C_1 1_{[0,1]} \leq K \leq C_2 1_{[0,1]},$$

where this type of kernel contains the usual discontinuous kernels such as the asymmetrical rectangle (box) kernel.

2. A function $K$ from $R$ into $R^+$ such that $\int K = 1$ is called a kernel of kind $II$ if its support is $[0,1]$ and if its derivative $K'$ exists on $[0,1]$ and satisfies two real constants $-\infty < C_1 < C_2 < 0$:

$$C_2 \leq K' \leq C_1.$$
where this definition considers the standard asymmetrical continuous case such as the triangle, quadratic, etc.

2.3.4 Local Weighting and Small Ball Probabilities

We can now build the relation between local weighting and the concept of small ball probabilities. Consider the simplest kernel among those of type 1, namely the asymmetrical box kernel. Let $X$ be a functional response variable valued in space $E$ and $\chi$ again be a fixed element of space $E$. We can then write:

$$E \left( 1_{[0,1]} \left( \frac{d(\chi, X)}{h} \right) \right) = E(1_{B(\chi,h)}(X)) = P(X \in B(\chi,h)).$$

Where $d$ is a semi-metric and $h$ is the optimal bandwidth. According to (Ferraty and Vieu (2006), Ferraty et al. (2007), Burba et al. (2009), Ferraty et al. (2011), Ferraty et al. (2012), and Ciollaro (2014)), $B(\chi,h)$ is a small ball and $P(X \in B(\chi,h))$ is a probability of a small ball when we take the number of curves $n$ to be very large, and the bandwidth $h$ tends towards zero. Then, for all positive real $h$, we use the formula:

$$\varphi(\chi)(h) = P(X \in B(\chi,h)).$$

This concept of small ball probabilities will play a major role in both the theoretical and practical points of view because the notion is strongly related to the semi-metric $d$, and further that the choice of semi-metric will be a crucial point in obtaining greater accuracy. According to Ferraty and Vieu (2006), consider two types of kernel for weighting functional variables.

2.4 Nonparametric Functional Kernel Regression

Let $(X_i, Y_i)_{i=1,2,...,n}$ be $n$ pairs of independent and identically distributed $(X,Y)$ and valued in $E \times R$, where $(E,d)$ is a semi-metric space (i.e., $X$ is a functional random variable and $d$ is a semi-metric). In this chapter, we wish to review some problems connecting functional data. The prediction of a scalar response variable $Y$ given the functional variable $X$, more precisely, leads to a focus on conditional
expectations as well as the main goal of estimating the regression of the nonlinear operator \( r \) of \( Y \) on \( \mathcal{X} \), defined by:

\[
  r(\chi) = E(Y \mid \mathcal{X} = \chi).
\] (2.2)

This nonlinear operator provides information about \( \mathcal{X} \) and \( Y \), which is useful for predicting \( y \) given \( \chi \) (see, for example, Ferraty and Vieu (2003, 2004, 2006), Rachdi and Vieu (2007), Ferraty et al. (2007), and Burba et al. (2009)). A prediction is obtained directly from the regression operator by:

\[
  \hat{y} = \hat{r}(\chi),
\]

where \( \hat{y} \) is the predicted value and \( \hat{r} \) is the estimate of \( r \). Therefore, the predictor is based on the estimation of a number of (nonlinear) operators (i.e., the regression operator \( r \)).

### 2.4.1 The Estimate

To predict the value of the scalar response given the explanatory functional variable, estimating \( r \) is an important issue. Therefore, we need to discern how to make functional nonparametric estimates that are adaptable to such types of statistical model, and thus we need to find a model to estimate \( r \). Consequently, the previous section explained that the kernel estimators are good applicants through which to obtain a local weighting in the functional setting. Nevertheless, it is an extremely sensitive task because \( r \) is a nonlinear operator from \( E \) to \( R \), and so the functional kernel regression estimator is:

\[
  \hat{r}(\chi) = \frac{\sum_{i=1}^{n} Y_i K \left( h^{-1} d(\chi, \mathcal{X}_i) \right)}{\sum_{i=1}^{n} K \left( h^{-1} d(\chi, \mathcal{X}_i) \right)},
\] (2.3)

where \( K \) is a kernel function and \( h \) (depending on \( n \)) is a positive real bandwidth (Ferraty and Vieu (2003, 2004, 2006), Rachdi and Vieu (2007), Ferraty et al. (2007), Burba et al. (2009), and Akbar and Ullah (2011)). It is also an adoption of the familiar Nadaraya-Watson estimate by Nadaraya (1964) and Watson (1964), as well as by Härdle (1990) who introduced it for finite dimensional nonparametric
regression. The only difference comes from the semi-metric which measures the proximity between the functional objects or curves.

For more simplicity and to see how such an estimator works, let us consider the following formula:

\[ \omega_{i,h}(\chi) = \frac{K(h^{-1}d(\chi, \mathcal{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(\chi, \mathcal{X}_i))}. \]

Therefore, it is straightforward to rewrite the kernel estimator (2.3) as:

\[ \hat{r}(\chi) = \sum_{i=1}^{n} \omega_{i,h}(\chi) Y_i. \quad (2.4) \]

Here, if \( \mathcal{X}_i \) is close to \( \chi \), then more weight is placed on \( Y_i \). Also, we have \( \omega_{i,h}(\chi) = 0 \), when \( d(\chi, \mathcal{X}_i) > h \). The estimator \( \hat{r}(\chi) \) is only taken into account within \( y_i \)'s, for which the corresponding \( \mathcal{X}_i \)'s are distant from \( \chi \) at most \( h \). In this case, \( h \) controls the number of terms in the weighted average, and thus the parameter \( h \) plays an important role. In other words, the larger the number of terms in the sum when \( h \) is larger, the less sensitive \( \hat{r}(\chi) \) is with respect to small variations in the \( Y_i \)'s. \( h \) has a smoothing effect, and in this case is called the smooth parameter for infinite dimensional data (see, for example, Ferraty and Vieu (2003, 2006)) and also the multivariate case by Härdle (1990)).

### 2.4.2 Prediction via Regression

We previously mentioned the prediction via regression, which requires some attention in order to compute estimators and choose the smoothest parameter or select the bandwidth. Therefore, in this chapter, we focus on revising the functional nonparametric regression method because it is the most popular form used in statistics. Accordingly, there are different kinds of kernel estimators with several automatic choices of smoothing parameter.

Here we recall the problem of prediction, which relates to the observation of \((\mathcal{X}_i, y_i)_{i=1,...,n}\) independent and identically distributed: \( \mathcal{X}_i = \{\chi_i(t_1), ..., \chi_i(t_J)\} \) is the discretized version of the curve \( \chi_i = \{\chi_i(t); t \in T\} \) measured at \( J \) points \( t_1, t_2, ..., t_J \), whereas the scalar response values are obtained by \( y_i \)'s. In addition,
the measure of proximity between the observed curves $\chi_i$ and $\chi_i'$ is denoted by any type of semi-metric, as previously described. Therefore, when predicting the scalar responses from the curves, we first consider the functional version of the Nadaraya-Watson kernel-type estimator defined in (2.3).

According to Ferraty and Vieu (2004, 2006), Ferraty et al. (2007), and Burba et al. (2009), two approaches are obtained, the first that is dependent on the smoothing parameter being the bandwidth $h$, which is a real positive number that attains the prediction at an observed curve $\chi_i'$ by structuring a weighted average of the $y_i$'s, for which the corresponding $X'$ is such that the measure $d(\chi_i, X_i')$ is smaller than $h$ in the kernel method. The second approach is to change $h$ to the number $K$ (i.e., the number of neighbours), which is also a smoothing parameter and takes its values from a discrete set to calculate the average for this case using the method K-Nearest Neighbour. Härdle (1990) states that in the multivariate random vector, the K-Nearest Neighbour estimate is a weighted average with a varying Neighbour, while the kernel estimate can be defined as a weighted average of the response variables for fixed neighbours around $X'$.

Therefore, there are several different types of functional kernel estimator which are based on a selection of smooth parameters that can be either fixed or automatic in nature. The $R$ procedures used for all kinds of predictions are available at http://www./sp.ups-tlse.fr/staph/npfda.

In this work, we focus on two types to estimate the different regression operators: the first is based on the kernel estimator and is called $\text{funopare.kernel.cv}$, we can found in http://www./sp.ups-tlse.fr/staph/npfda, whilst the second utilizes the KNN model and is called $\text{funopare.knn.gcv}$, and which are described in more detail in chapter 3. The first uses an automatic bandwidth selection as provided by a cross-validation procedure, as proposed on at http://www./sp.ups-tlse.fr/staph/npfda. The main purpose is to compute the formula:

$$R_{CV}^{kernel} = \frac{\sum_{i=1}^{n} y_i K(d(\chi, X_i)/h_{opt})}{\sum_{i=1}^{n} K(d(\chi, X_i)/h_{opt})},$$

where the pairs $(X_i, y_i)_{i=1,2,...,n}$ are independent and identically distributed, and where $h_{opt}$ (the optimal bandwidth of the kernel estimator) is obtained by a cross-validation framework:
\[ h_{opt} = \arg\min_{h} CV(h), \]

where

\[ CV(h) = \sum_{i=1}^{n} \left( y_i - R_{(-i)}^{\text{kernel}}(\chi_i) \right)^2, \]

with

\[ R_{(-i)}^{\text{kernel}}(\chi) = \frac{\sum_{j=1,j\neq i}^{n} y_j K(d(\chi, \chi_i)/h)}{\sum_{j=1,j\neq i}^{n} K(d(\chi, \chi_i)/h)}. \]

The semi-metric \( d(.,.) \) and the kernel function \( K(.) \) have been fixed by the partitioner (see, for example, Ferraty and Vieu (2004, 2006), and Burba et al. (2009)).
Chapter 3

K-Nearest Neighbour model with multivariate response in functional nonparametric regression

In recent years the problem of functional regression has become a subject of growing interest, due to the sophistication in recent technological advances regarding collecting and storing data curves. In this chapter, we propose a new model for nonparametric functional regression analysis with conditional expectation in the context of response as a multivariate variable while the covariates take values in some infinite dimensional space. We use the formula of the Nadaraya-Watson estimator (k-nearest neighbour (KNN)) for prediction with two kinds of semi-metrics as a measure of proximity between curves (semi-metric based on the derivatives and semi-metric built on the Functional Principal Component analysis). The present model is more convenient for the prediction of the components of a vector of random variables together rather than for predicting each of them separately because of saving time and least cost. The performance of this model is then evaluated by calculating mean square prediction errors which are then compared to the independent response model. The model is illustrated by both a simulation study and analysis of two real data sets from functional data analysis (Spectrometric Data and Canadian Weather Stations).

The chapter is organized as follows. Section 3.1 introduces the multivariate response variable method with K-Nearest Neighbour estimator. Section 3.2 contains the simulation study. Two examples of treatment of real data are presented.
in Section 3.3. Finally, some concluding remarks will be given in Section 3.4.

3.1 Methodology

In the past decades, nonparametric functional regression has been developed as an effective statistical method for nonlinear regression. The novelty of functional regression for linear model is pointed by Ramsay and Silverman (2005), also case studies and applied issues are discussed by Ramsay and Silverman (2007) with linear regression and multiple regression (parametric model). The regression model has for many decades formed part of the interest in nonparametric functional data analysis, and has been the subject of various studies. For example, Ferraty and Vieu (2003, 2004, 2006), Ferraty et al. (2007), Rachdi and Vieu (2007), illustrated the nonparametric functional regression when the response is scalar variable and the covariate is function. In their studies, they used semi-metric as a measure of proximity between curves, and cross-validation procedure used to obtain the optimal bandwidth. Burba et al. (2009) proposed the functional k-nearest neighbour estimator with nonparametric functional regression with scalar responses, in which the explanatory variable is valued in some abstract semi-metric functional space and asymptotic properties of the model are studied and noted that the rate is similar to the rate of convergence of the Nadaraya-Watson kernel estimator. Geenens et al. (2011) illustrate the nonparametric functional regression when the covariate is function and the response is scalar variable and used the semi-norm as a measure of proximity between curves. Consider the nonlinear regression model:

\[ Y = m(\mathcal{X}) + \text{error}, \]

where \( Y \) is a response variable, \( \mathcal{X} \) is a functional random variable (i.e., \( \mathcal{X} \) takes values in some infinite-dimensional space) and the nonparametric feature of the model arises from the fact that the only restrictions on \( m \) are smoothness restrictions. In this chapter, we use KNN model to estimate \( m(\cdot) \).
3.1.1 Multivariate response in functional nonparametric regression

Let \((X_i, Y_i)_{i=1,...,n}\) be \(n\) pairs that are independently and identically distributed as \((X, Y)\) and valued in \(f \times \mathbb{R}^q\), where \(f\) is an infinite-dimensional space equipped with a semi-metric \(d(\cdot, \cdot)\), and \(Y_i = (y_{i1}, ..., y_{iq})^t\) are the \(q\)-dimensional response variables. We use the general frame of the functional nonparametric regression:

\[
Y_i = m(X_i) + \varepsilon_i, \quad i = 1, ..., n 
\]  

(3.2)

where \(\varepsilon_i\) is independent random error with \(E[\varepsilon/X] = 0\). Then, the subject we want to estimate is the non-linear operator, \(m(.) = E[Y/X = .]\). Thus, the formula of the KNN estimator (the main idea of KNN is to replace the parameter \(h\) with \(h_k\) which is the bandwidth allowing us to take the account \(k\) terms in the weighted average see Ferraty and Vieu (2006)) can be written as:

\[
\hat{m}^{KNN} GCV(\chi) = \sum_{i=1}^{n} Y_i \omega_i(\chi). 
\]  

(3.3)

where

\[
\omega_i(\chi) = \frac{K \left( d(\chi, X_i) / h_{kopt}(\chi) \right)}{\sum_{i=1}^{n} K \left( d(\chi, X_i) / h_{kopt}(\chi) \right)}. 
\]

where \(K(\cdot)\) is a symmetrical kernel and \(h_{kopt}(\chi)\) is the bandwidth corresponding to the optimal number of neighbours obtained by a cross-validation procedure:

\[
k_{opt} = \arg \min_k GCV(k). 
\]

where

\[
GCV(k) = \sum_{i=1}^{n} \left( y_i - m_{KNN}^{(i)}(X_i) \right)^2 
\]

with

\[
m_{KNN}^{(i)}(\chi) = \frac{\sum_{j=1, j \neq i}^{n} y_j K \left( d(X_j, \chi) / h_k \right)}{\sum_{j=1, j \neq i}^{n} K \left( d(X_j, \chi) / h_k \right)}. 
\]
And $h_k$ is the bandwidth such that

$$\text{card}\{i : d(\chi, \chi_i) < h_k\} = k$$

Utilizing the same number of neighbours at any curves: $h_{k_{opt}}$ depends clearly relies on $\chi$ (the bandwidth $h_{k_{opt}}$ is such that only the $k_{opt}$-nearest neighbours of $\chi$ are taken into account) but $k_{opt}$ is the same of any curve $\chi$. Therefore, the user has to settle the semi-metric $(d(.,.))$ and the kernel function $K(.)$ (see for instance, Ferraty and Vieu (2004, 2006), and Burba et al. (2009)).

According to Ferraty and Vieu (2004, 2006), semi-metrics can be used as a closeness measure for functional data, in present model we used two kinds of semi-metric (semi-metric based on derivative and semi-metric build on functional component analysis). Our aim for the proposed method (predict the components of the response vectors simultaneously) is doubtless less expensive (in terms of time and cost).

### 3.2 Simulation study

The aim of this section is to illustrate the theoretical result through the simulated study. We consider the nonparametric functional regression:

$$Y_i = m(\chi_i) + \varepsilon_i \quad i = 1, ..., n = 215$$

First of all, we generate the simulate curves:

$$\chi_i(t_j) = \cos(t_j) + a_i(t_j - 0.5)^2 + b_i, \quad i = 1, ..., n$$

where $0 = t_1 < t_2 < ... < t_{100} = 1$ are equispaced points, and $a_i, b_i$ are independently drawn from a normal distribution $a_i \sim N(0, 1)$ and $b_i \sim N(0, 1)$. Figure 3.1 presents the 215 simulated curves from one replication.

Once the curves are defined, we simulate a nonparametric function regression method to calculate the outputs.

We considered two functional operators for building the regression operator $m$, which are expressed as
Chapter 3  K-Nearest Neighbour model with multivariate response  

Figure 3.1: A sample of 215 simulated curves.

\[
\begin{align*}
    m_1(X_i(t)) &= \int_0^\pi X'_i(t) dt, \\
    m_2(X_i(t)) &= \int_0^\pi |X'_i(t)| \log|X'_i(t)| dt,
\end{align*}
\]

where \( X'_i(t) \) is the first derivative of \( X(t) \). We compute the corresponding response variables:

\[
Y_{1i} = m_1(X_i) + \varepsilon_{1i}, \quad i = 1, 2, ..., n,
\]

\[
Y_{2i} = m_2(X_i) + \varepsilon_{2i}, \quad i = 1, 2, ..., n.
\]

We generate the errors \( \varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}) \sim N(0, \Sigma) \), \( \Sigma = \begin{bmatrix} (\sigma_1)^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & (\sigma_2)^2 \end{bmatrix} \), where \( \Sigma \) is the covariance matrix.

We take three different cases for correlations between the two response variables, \( \rho = 0.9 \) or 0.2 or 0.0, respectively, and use \( \sigma_1, \sigma_2 = 1, 2 \). We take two different sample spaces \( (n = 215 \text{ and } n = 400) \), which are divided into sub-sample; that is, samples of size \( n = 160 \) and \( n = 300 \), from which all the estimates are computed, and \( n = 55 \) and \( n = 100 \) for testing, which are used to check the
In both models the multivariate response variables (Mul-R) model and the independent response model (Ind-R) that are considered, the quadratic kernel function and semi-metric based on the second derivative is used as a measure of proximity between curves because the curves are smooth. For the measure of prediction efficiency, we use the mean square error (MSE) between the predicted values and the true regression values for the 55 and 100 test samples. We calculate the MSE for the performance of the first and second samples as below

\[ MSE = \frac{1}{55} \sum_{i=161}^{215} (y_i - \hat{y}_i)^2, \]

and

\[ MSE = \frac{1}{100} \sum_{i=301}^{400} (y_i - \hat{y}_i)^2. \]

The execution of the suggested model, multivariate response variables (Mul-R) model, is compared with that of independent response model (Ind-R) where the two responses are defined independently and without considering the correlation between them. The simulated procedure is repeated 20 times, and the average of the MSEs is presented in Table 3.1. From Table 3.1, it can be seen that in all cases, the multivariate method improves the precision of the prediction compared with modelling each response independently (Ind-R). Therefore, it is obvious that from Table 3.1 that the multivariate model is more convenient for prediction than the independent model even though there is no correlation between the components of the responses.

### 3.3 Real Data Application

In this section, we apply the proposed model to two real data sets, Tecator real data and Canadian Weather station.

**Tecator data.** This kind of data is very popular among the community of nonparametricians because chemometrics was the point behind the development of the functional nonparametric methodology, and therefore several applications have been based on this type of data using different methods (see, for instance, Ferraty
Table 3.1: The average MSEs for the two simulated samples data.

<table>
<thead>
<tr>
<th>n</th>
<th>case</th>
<th>components</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>215</td>
<td>I</td>
<td>1</td>
<td>0.9933 1.0049</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>2</td>
<td>3.8281 3.8732</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1.0149 1.0163</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>2</td>
<td>4.4200 4.4209</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>2</td>
<td>0.9600 0.9823</td>
</tr>
<tr>
<td>400</td>
<td>I</td>
<td>1</td>
<td>1.0313 1.0351</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>2</td>
<td>4.0715 4.0725</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1.0539 1.0569</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>2</td>
<td>4.0680 4.0789</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>2</td>
<td>1.0747 1.0822</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0788 4.0818</td>
</tr>
</tbody>
</table>

and Vieu (2003, 2006), Ferraty et al. (2007), Burba et al. (2009) and Wang et al. (2017)). A brief description of spectrometric data comes from the quality control problem, and can be found at http://lib.stat.cmu.edu/datasets/tector. These data concern a sample of finely chopped pieces of meat. More precisely, for each meat sample, the data consist of a 100 channel spectrum of absorbances, where the absorbance is the $-\log_{10}$ of the transmittance measured by the spectrometer. Figure 3.2 represents the associated spectrometric curves.

One of the main aims of spectrometric analysis is to allow for the discovery of the proportion of specific chemical content because the analysis by chemical processing would take a considerable amount of time and be extremely expensive. This example is solved (see, for example, Ferraty and Vieu (2006) and Shang (2013)), when the response is still real and the covariate is function, so we use the same model and the same case for predicting fat, water and protein content from curves in a separate manner. Indeed, the correlation coefficients between the three variables (fat, water, and protein content) are given by $\rho_{\text{fat,water}} = 0.988$, $\rho_{\text{fat,protein}} = -0.86$ and $\rho_{\text{water,protein}} = 0.82$. As can be seen, because the fat, water, and protein content in meat are strongly correlated, it will be more convenient to
predict these variables simultaneously rather than predicting each one separately. In this study, we describe the data for multivariate response variables in this manner, since for each subject \( i \) (among 215 pieces of finely chopped meat) we observe one spectrometric discretized curve \( (x_i) \) that corresponds to the absorbance measured on a grid of 100 wavelengths (i.e., \( x_i = (\chi_i(\lambda_1), ..., \chi_i(\lambda_{100})) \)). Furthermore, we have at hand the fat, water, and protein content \( Y_i \) for each unit \( i \), as obtained by chemical analysis, hence the pairs \((x_i, y_i)_{1, ..., n}\), which are organized in Table 3.2.

The three last columns involve the multivariate response values and the first 100 columns correspond to the 100 channel spectrum. Given a new spectrometric curve \( x \), the goal is to predict multivariate response variables. Therefore, in order to highlight the performance of functional nonparametric prediction regression, the original sample is split into two sub-samples. The first one includes the first 160 units \((x_i, y_i)_{i=1,2,...,160}\) and is referred to as the 'learning sample'; at this point,
we use both the $x_i$'s and the responding $y_i$'s that permits us to build the functional kernel estimators with optimal smoothing parameters. The latter contains the last 55 units, and is referred to as the 'testing sample', which is beneficial for getting predictions and measuring their accuracy. In the same style as the simulation study, the mean square error (MSE) is used as the measure of the performance of the model, as

$$MSE = \frac{1}{55} \sum_{i=161}^{215} (y_i - \hat{y}_i)^2.$$  

The function is to predict the fat, water and protein contents from the spectroscopic curves.

For the independent response model, we run the `funopare.knn.gev` function in R program for prediction; this function is available on the website of Nonparametric Functional Data Analysis (NFDA) at [http://lib.stat.cmu.edu/datasets/tector](http://lib.stat.cmu.edu/datasets/tector). Because of the smoothing data, we use the semi-metric based on derivatives, but after trying various derivatives, the second derivative was found to be the best for analysing both independent response model and multivariate response vectors. For each model, mean square error was now used as an evaluation criterion, as shown in Table 3.3.

We can deduce from Table 3.3 that the multivariate response model (Mul-R) significantly improves the prediction precision for the fat and protein compared to the independent response model (Ind-R) which is studied in the literature. The prediction for the amount of water by the multivariate response (Mul-R) method

<table>
<thead>
<tr>
<th>Row</th>
<th>Col 1</th>
<th>Col j</th>
<th>Col 100</th>
<th>Col 101</th>
<th>Col 102</th>
<th>Col 103</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_1(\lambda_1)$</td>
<td>...</td>
<td>$x_1(\lambda_j)$</td>
<td>...</td>
<td>$x_1(\lambda_{100})$</td>
<td>$y_{11}$</td>
</tr>
<tr>
<td>2</td>
<td>$x_2(\lambda_1)$</td>
<td>...</td>
<td>$x_2(\lambda_j)$</td>
<td>...</td>
<td>$x_2(\lambda_{100})$</td>
<td>$y_{21}$</td>
</tr>
<tr>
<td>i</td>
<td>$x_i(\lambda_1)$</td>
<td>...</td>
<td>$x_i(\lambda_j)$</td>
<td>...</td>
<td>$x_i(\lambda_{100})$</td>
<td>$y_{i1}$</td>
</tr>
<tr>
<td>215</td>
<td>$x_{215}(\lambda_1)$</td>
<td>...</td>
<td>$x_{215}(\lambda_j)$</td>
<td>...</td>
<td>$x_{215}(\lambda_{100})$</td>
<td>$y_{2151}$</td>
</tr>
</tbody>
</table>

Table 3.2: The functional covariates and multivariate response variables.
was worse than for the independent response (Ind-R) model. As we see in Figure 3.3 and Figure 3.4, the predictions are shown as dashed lines while the solid line is the true values for fat, water, and protein content from top to bottom, respectively, as obtained using the multivariate response (Mul-R) and independent response (Ind-R) models.

<table>
<thead>
<tr>
<th>Responses</th>
<th>Mul-R</th>
<th>Ind-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat</td>
<td>1.798</td>
<td>1.885</td>
</tr>
<tr>
<td>Water</td>
<td>2.418</td>
<td>2.298</td>
</tr>
<tr>
<td>Protein</td>
<td>1.793</td>
<td>1.847</td>
</tr>
</tbody>
</table>

Table 3.3: The MSE for fat, water and protein.

Figure 3.3: The multivariate response model predictions for fat, water, and protein content from top to bottom, respectively.

**Canadian Weather Stations** This dataset was studied by Ramsay and Silverman (2005); Rachdi and Vieu (2007).
Chapter 3 *K-Nearest Neighbour model with multivariate response*

Figure 3.4: The independent response model predictions for fat, water, and protein content from top to bottom, respectively.

The dataset was obtained from Canadian Weather Stations, and found in the functional data analysis package in R. In this example, we take only the average monthly precipitation and the average monthly temperature for 35 Canadian weather stations.

Figure 3.5 shows different stations, where each curve represents average monthly temperature of a single station. The data consists of 12 observations; one station occurs clearly as a discretized curve, so we have 35 curves representing 35 stations. Therefore, each station is a continuous curve because of the fineness of the grid of the data.

Figure 3.6 illustrates the average monthly precipitation of the 35 stations, with each curve representing the one station.
Therefore, each station $i$ (among 35 stations) has one discretized curve $x_i$ which corresponds to the mean monthly temperature measured at 12 time points (i.e. $x_i = (\chi_i(\lambda_1), \chi_i(\lambda_2), ..., \chi_i(\lambda_{12}))$. More precisely, the prediction of each station $i$ average monthly precipitation contains $Y_i$ from Canadian Weather Stations. Hence, the data contain the pairs $(x_i, y_i)_{i=1,2,...,35}$, which are organized in Table 3.4. The first 12 columns correspond to the 12 average monthly temperatures as a functional covariate, and the last 12 columns average monthly precipitation is multivariate response values. Given a new mean monthly temperature curve $x$, our aim is to predict the average monthly precipitation (twelve months) from average monthly temperature. Therefore, in order to highlight the performance of functional nonparametric prediction regression, we use the same case study as the first example which split the original sample into two sub-samples. The first is the learning sample, which includes the first 25 stations $((x_i, y_i)_{i=1,2,...,25})$, as well as in this point both use the $x_i$’s and the responding $y_i$’s that permit the
building of the functional kernel estimators with optimal smoothing parameters. The second sample contains the last 10 stations, and forms the testing sample. It is beneficial for getting predictions and measuring their accuracy. Furthermore, the correlation coefficients between all 12 months (Jan., Feb.,..., and Dec.) are positive and strongly correlated, so it is clearly preferable to predict these vectors simultaneously rather than on the basis of individual months. In this instance, we have also calculated the mean square error for the measure of performance for both models.

\[ MSE = \frac{1}{10} \sum_{i=26}^{35} (y_i - \hat{y}_i)^2. \]

We run the `funopare.knn.gcv` function from R program are available at http://www.sp.ups-tlse.fr/staph/npfda for the Ind-R model.

In both models the multivariate response variables (Mul-R) model and the independent response model (Ind-R) that are considered, the quadratic kernel
function and the semi-metric based on the functional principal component analysis for the measure of proximity between curves with $q = 1$. The mean square error of the proposed multivariate response (Mul-R) model was compared with the independent response (Ind-R) model, as reported in Table 3.5.

It can be seen in Table 3.5, in comparison with the independent response model, that the multivariate response model significantly improves the prediction accuracy across all components of the responses of the average monthly precipitation from the average monthly temperature. We can suggest, then, that when the correlation between the responses are strong so the multivariate response model will give us better results than the independent response model.

As we can see in Figure 3.7 and Figure 3.8, the dashed line is the true values for the 10 stations that we selected from four regions (Atlantic, Continental, Pacific, and Arctic) from the Canadian Weather Station dataset, while the predictions when using the multivariate response variables and Independently responses are shown as solid lines.

<table>
<thead>
<tr>
<th>Row 1</th>
<th>$\chi_1(\lambda_1)$</th>
<th>...</th>
<th>$\chi_1(\lambda_j)$</th>
<th>...</th>
<th>$\chi_1(\lambda_{12})$</th>
<th>Col 12</th>
<th>Col 13</th>
<th>Col 14</th>
<th>...</th>
<th>Col 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 2</td>
<td>$\chi_2(\lambda_1)$</td>
<td>...</td>
<td>$\chi_2(\lambda_j)$</td>
<td>...</td>
<td>$\chi_2(\lambda_{12})$</td>
<td>Col 12</td>
<td>Col 13</td>
<td>Col 14</td>
<td>...</td>
<td>Col 24</td>
</tr>
<tr>
<td>Row $i$</td>
<td>$\chi_i(\lambda_1)$</td>
<td>...</td>
<td>$\chi_i(\lambda_j)$</td>
<td>...</td>
<td>$\chi_i(\lambda_{12})$</td>
<td>Col 12</td>
<td>Col 13</td>
<td>Col 14</td>
<td>...</td>
<td>Col 24</td>
</tr>
<tr>
<td>Row 35</td>
<td>$\chi_{35}(\lambda_1)$</td>
<td>...</td>
<td>$\chi_{35}(\lambda_j)$</td>
<td>...</td>
<td>$\chi_{35}(\lambda_{12})$</td>
<td>Col 12</td>
<td>Col 13</td>
<td>Col 14</td>
<td>...</td>
<td>Col 24</td>
</tr>
</tbody>
</table>

Table 3.4: The functional covariates and Multivariate response variables for 35 Canadian weather stations.
<table>
<thead>
<tr>
<th>Responses</th>
<th>MSE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mul-R</td>
<td>Ind-R</td>
<td></td>
</tr>
<tr>
<td>January</td>
<td>2.61</td>
<td>3.08</td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>2.03</td>
<td>2.55</td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>0.97</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td>April</td>
<td>1.02</td>
<td>1.58</td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>0.68</td>
<td>1.12</td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>0.96</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>1.10</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>August</td>
<td>1.09</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td>September</td>
<td>3.35</td>
<td>4.03</td>
<td></td>
</tr>
<tr>
<td>October</td>
<td>7.43</td>
<td>8.51</td>
<td></td>
</tr>
<tr>
<td>November</td>
<td>4.51</td>
<td>5.09</td>
<td></td>
</tr>
<tr>
<td>December</td>
<td>3.15</td>
<td>3.71</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5: MSE for Canadian Weather Station.

Figure 3.7: The prediction from multivariate response variables for the average monthly precipitation at 10 stations from the average monthly temperature curves.
3.4 Conclusion

This chapter has proposed a new model for analysing a nonparametric regression where the response is multivariate and the covariates is functional. Under the proposed method, we show that the K-nearest neighbour estimator provides good predictions when compared with the results obtained from the independent response model. The measure of proximity between the functional covariates is via the semi-metrics (semi-metric based on derivatives and semi-metric build on functional principal component analysis) and choosing the optimal bandwidth is via cross-validation procedure, as pointed out by Ferraty and Vieu (2006). The utility of the multivariate response variables model and independent response variable model is illustrated through some numerical instances (simulation data and
two real data examples). Our numerical results proceed reasonably well, by presented model (multivariate response). In future work, we will attempt to study nonparametric functional regression when the covariate is a function and take the correlation between different components of the responses into account and the different bandwidth for the different component of the responses. Applying another procedure for choosing optimal bandwidth such as the Bayesian bandwidth estimation procedure as proposed by Shang (2013). Moreover, applying the proposed methodologies to other functional regression estimators, such as the kernel estimator (Ferraty and Vieu (2003, 2006), Ferraty et al. (2007)) and functional local linear kernel estimator by Benhenni et al. (2007). Finally, an asymptotic properties of the proposed model (multivariate response variables) is the good topic for future study.
Chapter 4

Principal Component Analysis in Nonparametric Functional Regression

Today’s the rapid advances in technology have allowed datasets to be recorded with great accuracy and at high frequency, and often include measurements of a large number of variables (high dimensions). Frequently, it is possible to reduce the number of dimensions while still retaining much of the information available in the original data. Principal component analysis is probably the best known technique used for such purposes, and is most commonly used to reduce the dimensionality of multivariate data. By finding linear combinations, called the principal components, principal component analysis can reduce the dimensionality whilst successfully maintaining maximum variance in the data (see, for instance, Jolliffe (2011), and Bro and Smilde (2014)).

In this chapter, a new method is proposed to deal with nonparametric functional data analysis with conditional expectation in the context of the covariate being a function and where the principal component analysis is used to decorrelate the multivariate response variables. We use the Nadaraya-Watson estimator (k-nearest neighbour (KNN)) with a cross-validation procedure to choose the optimal bandwidth, for further prediction with two types of semi-metrics in terms of measuring the proximity between curves: a semi-metric based on derivatives, and a semi-metric built on functional principal component analysis. The performance of this model was then evaluated by calculating the root mean square errors which were then compared to the independent response model studied in the literature. The model is illustrated by both a simulation study and analysis of two
real datasets (Spectrometric data and Soil data).

The remainder of the chapter is organized as follows. Section 4.1 contains the models that use principal component analysis to decorrelate the multivariate response variables for estimation. Section 4.2 presents the efficiency of the suggested model through a simulation. Section 4.3 applies the method to two real data examples, and the chapter is then concluded by Section 4.4.

4.1 Methodology to Decorrelate Multivariate Response Variables via Principal Component Analysis

Let $Y = (y_1, ..., y_m)^t$, be a multivariate response in $R^m$, and $(X(t))$ be the functional covariate. Consider the issue of nonlinear regression:

$$Y = m(X(t)) + \varepsilon$$ (4.1)

Given n observations $(X_1, Y_1), ..., (X_n, Y_n)$, let $\hat{\mu}$ and $\hat{\Sigma}$ be the sample mean and the sample covariance matrix of $Y = (Y_1, ..., Y_n)^t = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix}$ respectively, and have the eigenvalue-(normalized) eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_m, e_m)$ where $\lambda_1 \geq \lambda_2 \geq ... \lambda_m \geq 0$. Then, via the principal component analysis, the principal scores are given by:

$$Y = (Y - \eta)v,$$ (4.2)

where $v = (e_1, ..., e_m)$ and $\eta = (\hat{\mu}, ..., \hat{\mu})^t$ is an $n \times m$ matrix. Denote $Y_l = (Y_{1l}, Y_{2l}, ..., Y_{nl})^t$, the lth column of $Y$, then $Y_1, ..., Y_m$ are samples of $m$ uncorrelated random variables. The nonparametric regression function $m(.)$ can then be represented by the relationship between $Y_{il}$ and $X_i(t)$, that is, for $l = 1, ..., m$ and $i = 1, ..., n$

$$Y_{il} = m_l(X_i(t)) + e_{il}$$ (4.3)
where \( e_i \sim N(0, \sigma^2_l) \). By the \( K - NN \) method, we estimate \( m_l(\cdot) \) by:

\[
m_l(\chi) = \frac{\sum_{i=1}^{n} \gamma_i K \left( d(\chi, \chi_i) / h_{kopt}(\chi) \right)}{\sum_{i=1}^{n} K \left( d(\chi, \chi_i) / h_{kopt}(\chi) \right)}.
\]

where \( K(.) \) is a symmetrical kernel and \( h_{kopt}(\chi) \) is the bandwidth corresponding to the optimal number of neighbours obtained via a cross-validation procedure:

\[
k_{opt} = \arg\min_k GCV(k)
\]

And \( h_k \) is the bandwidth such that

\[
\text{card}\{i : d(\chi, \chi_i) < h_k\} = k
\]

where

\[
GCV(k) = \sum_{i=1}^{n} \left( \gamma_i - m_{KNN}^{(-i)}(\chi_i) \right)^2
\]

with:

\[
m_{KNN}^{(-i)}(\chi) = \frac{\sum_{j=1,j\neq i}^{n} \gamma_j K \left( d(\chi_j, \chi) / h_k \right)}{\sum_{j=1,j\neq i}^{n} K \left( d(\chi_j, \chi) / h_k \right)}.
\]

Using the same number of neighbours for any curve provides a global choice, and \( h_{kopt}(\chi) \) clearly relies on \( \chi \) (the bandwidth \( h_{kopt} \) is such that only the \( k_{opt} \)-nearest neighbours of \( \chi \) are taken into account) but \( k_{opt} \) is the same for any curve \( \chi \).

Therefore, the user has to fix the semi-metric \( (d(\cdot, \cdot)) \) and the kernel function \( K(\cdot) \) (see, for instance, Ferraty and Vieu (2004, 2006), and Burba et al. (2009)).

According to Ferraty and Vieu (2004, 2006), the semi-metrics can be used as a measure of the closeness of functional data, and several different semi-metrics have been defined in the literature. In this chapter, we employed two different kinds of semi-metrics in our numerical examples, which are built on derivative and functional principal component analysis (FPCA), as discussed in detail in chapter 2. Let \( \mathcal{X}_1, ..., \mathcal{X}_n \) be a sample of the curves and \( \mathcal{X} = \{\mathcal{X}_i(t); t \in \tau\} \).

The semi-metric built on FPCA is defined as:

\[
d_q^{FPCA}(\mathcal{X}_i, \mathcal{X}_j) = \sqrt{\sum_{k=1}^{q} \left( \int (\chi_i(t) - \chi_j(t)) v_k(t) \right)^2 dt},
\]
where $\nu_1, ..., \nu_q$ are the orthonormal eigenfunctions of the covariance operator $\Gamma_X(s,t) = E(X(s)X(t))$ associated with the largest $q$ eigenvalues. This type of semi-metric is adapted to rough curves, see Ferraty and Vieu (2006).

The semi-metric based on derivatives is defined as:

$$d^{\text{deriv}}_q(X_i, X_j) = \sqrt{\int \left( X_i^{(q)}(t) - X_j^{(q)}(t) \right)^2 dt}.$$ 

where $X^{(q)}$ is the $q$th derivative of $X$ with respect to $t$, which in practice is calculated utilizing the B-spline approximation of the curves. The derivative semi-metric is suitable for smooth curves, see Ferraty and Vieu (2006) for further details.

Let $X^*$ be a test point and $Y^*$ be the corresponding response point. Therefore, the predictive means of the scores then can be obtained by nonparametric functional regression and presented by $\hat{Y}_l^*$ for $l = 1, ..., m$. Thus the predictive mean of the $m$-dimensional response $Y^*$ is given by:

$$\mathbb{E}(Y^*) = \hat{\mu} + \nu \Upsilon^*,$$

where $\Upsilon^* = (\hat{\Upsilon}_1^*, ..., \hat{\Upsilon}_m^*)^t$.

### 4.2 Simulation Study

The aim of this section is to illustrate the theoretical result through further simulation. We consider the nonparametric functional regression:

$$Y_i = m(X_i) + \varepsilon_i \quad i = 1, ..., n = 215$$

First of all, we generate the simulate curves:

$$X_i(t_j) = \cos(t_j) + a_i(t_j - 0.5)^2 + b_i, \quad i = 1, ..., n$$

where $0 = t_1 < t_2 < ... < t_{100} = 1$ are equispaced points, and $a_i, b_i$ are independently drawn from the normal distribution, $a_i \sim N(0, 1)$ and $b_i \sim N(0, 1)$. Figure 3.1 in chapter 3 presents the 215 simulated curves from one replication.
Once the curves are defined, we simulate a nonparametric functional regression method to calculate the outputs. We consider two functional regression operators with components \( m_1, m_2 \) which are expressed as:

\[
\begin{align*}
    m_1(x_i(t)) &= \int_0^\pi x_i'(t) \, dt, \\
    m_2(x_i(t)) &= \int_0^\pi |x_i'(t)| \log |x_i'(t)| \, dt,
\end{align*}
\]

where \( x_i'(t) \) is the first derivative of \( x(t) \). We calculate the corresponding response variables:

\[
Y_{1i} = m_1(x_i) + \varepsilon_{1i}, \quad i = 1, 2, \ldots, n,
\]

\[
Y_{2i} = m_2(x_i) + \varepsilon_{2i}, \quad i = 1, 2, \ldots, n.
\]

We generate the error \( \varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}) \sim N(0, \Sigma) \), where

\[
\Sigma = \begin{bmatrix}
    (\sigma_1)^2 & \rho \sigma_1 \sigma_2 \\
    \rho \sigma_1 \sigma_2 & (\sigma_2)^2
\end{bmatrix}.
\]

We take two different cases for correlations between the two response variables, \( \rho = (0.9 \text{ or } 0.1) \), respectively, and use \( \sigma_1, \sigma_2 = 1, 2 \). We split the simulated sample \( n = 215 \) into two sets: the training sample of size \( n = 160 \), from which all the estimates are computed, and \( n = 55 \) to test the sample, which is used to check the performance of the predictions made by the model.

In both models (multivariate response variables with principal component analysis to decorrelate responses (Mul-RP) and the independent response model (Ind-R)) that are considered, the quadratic kernel function and semi-metric based on the second derivative is used as measures of the proximity between the curves because the curves are smooth. As a measure of prediction efficiency, we use the root mean square error (RMSE) between the predicted values and the true regression values for the 55 test samples.

The achievement of the suggested model Mul-RP is compared with that of Independent response (Ind-R) model where the two responses are defined independently and without considering the correlation between them. The simulation procedure is repeated 20 times and the average of the root mean square error is presented in Table 4.1. From Table 4.1, it can be seen that in both cases the (Mul-RP) method improves the prediction precision compared with the model.
each response predicts independently (Ind-R). Therefore, it is obvious from Table 4.1 that the present (Mul-RP) model is more covenant for prediction than the the independent model, even through the correlation between the components of the responses is small.

### 4.3 Real Data Application

Now we apply the presented model to two real datasets, Tecator data and soil data.

**Tecator data.** This kind of data is very popular among the community of nonparametricians because several applications have been performed on it by different methods (see, for instance, Ferraty and Vieu (2003, 2006), Ferraty et al. (2007), Burba et al. (2009) and Wang et al. (2017)). Then, a brief characterization of spectrometric data from the quality control problem will be undertaken, and can be found at [http://lib.stat.cmu.edu/datasets/tecatort](http://lib.stat.cmu.edu/datasets/tecatort).

The aim of spectrometric analysis is to allow for the detection of the proportion of specific chemical content because the analysis by chemistry processing would take a long time and be much more expensive. This example is solved using nonparametric functional regression (see, for instance, Ferraty and Vieu (2006) and Shang (2013)), when the response is a scalar value and the covariate is a function. In the previous chapter, we mentioned the correlation coefficients between three variables (fat, water, and protein content) are given by \( \rho_{\text{fat,water}} = 0.988 \), \( \rho_{\text{fat,protein}} = -0.86 \) and \( \rho_{\text{water,protein}} = 0.82 \). We notice that fat, water, and protein content in meat are strongly correlated. As we mentioned in the previous chapter, the three last columns contain the multivariate response values and the first 100 columns correspond to the 100 channel spectrum. Given a new spectrometric

<table>
<thead>
<tr>
<th>Case</th>
<th>Components</th>
<th>RMSE Mul-RP</th>
<th>RMSE Ind-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>1.0866</td>
<td>1.0948</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.6675</td>
<td>3.7156</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>1.0534</td>
<td>1.0521</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.2935</td>
<td>4.3312</td>
</tr>
</tbody>
</table>

**Table 4.1:** The average RMSEs for the simulated sample data.
Table 4.2: The RMSE for the fat, water and protein.

<table>
<thead>
<tr>
<th>Responses</th>
<th>RMSE</th>
<th>Mul-RP</th>
<th>Ind-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat</td>
<td>1.819</td>
<td>1.999</td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>1.937</td>
<td>2.120</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>1.532</td>
<td>1.630</td>
<td></td>
</tr>
</tbody>
</table>

curve \( x \), the goal is to predict the response. Therefore, in order to highlight the performance of functional nonparametric prediction regression, the original sample is split into two subsamples. The first sample contains of the first 160 units, and is called learning sample, and which permits one to build the functional kernel estimators with optimal smoothing parameters. The second includes the last 55 units, and is called the testing sample, which is beneficial for obtaining predictions and measuring their accuracy. In the same manner as the simulation study, the RMSE is a measure of the performance of the model as

\[
RMSE = \left( \frac{1}{55} \sum_{i=161}^{215} (y_i - \hat{y}_i)^2 \right)^{1/2}.
\]

The subject is to predict fat, water and protein content from the spectrometric curves.

(Multivariate response variables with principal component analysis to decorrelate responses (Mul-RP) and the independent response model (Ind-R)) we run the `funopare.knn.gcv` function as an R procedure to allowing prediction of the independent response model, where this function can found on the Nonparametric Functional Data Analysis (NFDA) website. Because of the type of data, we use the semi-metric built on derivatives, but after trying certain derivatives, the second derivative was found to be the best for the analysis of both the independent response (Ind-R) model and the multivariate response variables with principal component analysis to decorrelate responses (Mul-RP) method. To compare the efficiency of the models, we take randomly 10 times each time 55 curves for testing then take the average of the 10 repetitions, as reported in Table 4.2, whilst the remaining data were used for model training. Table 4.2 illustrate that the Mul-RP model significantly improves the predicted precisions for the fat, water and protein determinations compared to the Ind-R model.

Soil data. Originally, this data set was analysed by Rinnan and Rinnan (2007). They proposed that fluorescence, and especially near infrared reflectance
spectroscopy, have the capability to reveal changes in the properties of the highly organic arctic soil, as shown by the principal component analysis separation of the spectra measured to give the soil data from different climate change manipulation treatments. Subsequently, Wang et al. (2017) took samples of these data (organic matter (SOM) and ergosterol concentration (EC)) and used Gaussian process regression when the response was multivariate and the covariate was functional and multivariate. In their study, principal component analysis was used to decorrelate the multivariate response variables which then enabled them to model each principal component independently by Gaussian process regression. Furthermore, semimetrics were used to measure the proximity between curves (semi-metric based on derivatives, and semi-metric built on functional principal component analysis).

The soil samples were obtained from a long-term field experiment at a subarctic fell in Abisko, northern Sweden. We have 108 samples, recorded over a the wavelength range of 400-2500 nm (visible and near infrared spectrum) which was scanned at 2 nm intervals with an INR spectrophotometer. Fluorescence excitation-emission matrices (EEMs) were recorded with a spectrofluorometer for more detail, see Rinnan and Rinnan (2007) and Wang et al. (2017). Two component values, soil organic matter (SOM) was weighted as loss on ignition at 550 °C, and ergosterol concentration (EC) was defined through HPLC. As the functional covariates were smooth, the semi-metric built on second derivative was adopted in our example. To compare the effectiveness, leave-one-out cross-validation was undertaken, that is, each of the 108 samples was left as test data while the remaining data were utilized for method training. The root mean square error is calculated as a measure of efficiency of the comparison of two models (multivariate response variables with principal component analysis to decorrelate responses (Mul-RP) and independent response model (Ind-R)), as reported in Table 4.3. It can be seen that the Mul-RP model presented significantly improves the efficiency

<table>
<thead>
<tr>
<th>Responses</th>
<th>Mul-RP</th>
<th>Ind-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td>1.092</td>
<td>3.749</td>
</tr>
<tr>
<td>EC</td>
<td>6.146</td>
<td>46.005</td>
</tr>
</tbody>
</table>

Table 4.3: The RMSE for the SOM and EC.
of the prediction for both SOM and EC in comparison with the Ind-R method.

4.4 Conclusion

This chapter has proposed a new method for analysing nonparametric regressions where the covariate is a function and the response is multivariate. We used principal component analysis to decorrelate the multivariate response. Under the model presented, we showed that the K-nearest neighbour estimator supplied good predictions when compared with the results obtained from the independent response method. The measure of proximity between the functional covariates was obtained via the semi-metrics (the semi-metric based on functional component analysis, and the semi-metric built on derivatives) and where the bandwidth was chosen by a cross-validation procedure. The use of the multivariate response variables with principal component analysis to decorrelate responses (Mul-RP) model and the independent response (Ind-R) model was illustrated through a number of numerical simulations and real data (Tecator real data and soil data). The theoretical part of the proposed model represents an excellent topic for future study.
Chapter 5

Nonparametric regression method with functional covariates and multivariate response

The majority context of this chapter is the paper published in the journal of COMMUNICATION IN STATISTICS-THEORY AND METHODS. Nonparametric regression methods have been widely studied in functional regression analysis in the context of functional covariates and univariate response, but it is not the case for functional covariates with multivariate response. In this chapter, we present two solutions for the latter functional regression problem. The first solution is to directly extend the nonparametric method for univariate response to multivariate response with different bandwidths for different responses. In the second solution, the correlation among different responses is taken into account with different bandwidths for different responses and incorporated into the model. Our methods utilize the kernel function with an automatic bandwidth selection by cross-validation procedure and semi-metrics used as a measure of the proximity between functional data. Two kinds of semi-metrics used in this chapter (semi-metric based on derivatives and semi-metric based on the functional principal component analysis). The asymptotic properties of the estimators are studied, and the effectiveness of the proposed methods are demonstrated through several simulation studies and a real data example.

The rest of the chapter is organised as follows. Section 5.1 presents the models and the estimators. Section 5.2, includes some theoretical properties such
as the rate of almost complete convergence are discussed. In Section 5.3, the effectiveness of the proposed methods are illustrated through several simulation studies and a real data example. Section 5.4 concludes the chapter.

5.1 Models and Estimation

Let \((X_i, Y_i)_{i=1,...,n}\) be \(n\) pairs of samples independently and identically distributed as \((X, Y)\) and valued in \(f \times \mathbb{R}^q\), where \(f\) is an infinite-dimensional space equipped with a semi-metric \(d(\cdot, \cdot)\). \(Y_i = (y_{i1},...,y_{iq})^t\) is the \(q\)-dimensional response variable and \(X_i\) is the functional predictor. The general framework of the functional regression can be defined by:

\[
Y_i = r(X_i) + \varepsilon_i, \quad i = 1,...,n, \tag{5.1}
\]

where \(r(\cdot)\) is an unknown nonlinear operator, and \(\varepsilon_i\) is independent random errors with \(E[\varepsilon_i|X_i] = 0\). Therefore, the objective is to estimate the nonlinear operator \(r(\chi) = E[Y|X = \chi]\) for a given value \(X = \chi \in f\). Three parameters should be settled in this point: \(K(\cdot)\) the kernel, the bandwidth \(h\), and the semi-metric.

Choice of the kernel: \(K\) is the kernel function. There are several possible kinds of density function, chosen to be a symmetric density function with support \([-1,1]\). In the nonparametric estimation, it is a greed that the exact form of the kernel function does not greatly impact the estimate with regards to the choice of the bandwidth. In this setting, we can choose any kind of kernel functions; for instance a quadratic kernel, Epanechnikov kernel, indicator kernel, and integrated triangle kernel.

Choice of the bandwidth \(h\): In practice, choosing the bandwidth is critical to achieving accurate results, so there several different procedures for choosing the bandwidth, such as Bowman (1984) who apply the likelihood cross validation to obtain the optimal bandwidth for univariate kernel error density estimator, Shang (2013) studied Bayesian bandwidth and (Ferraty and Vieu (2003, 2004, 2006), Ferraty et al. (2007), Burba et al. (2009)) used cross validation for choosing the optimal bandwidth. In this chapter, we used the cross validation to choice different bandwidth for different components of the response variable.
By the nonparametric kernel method, the regression estimator can be expressed as:

$$\hat{r}(\chi) = \sum_{i=1}^{n} \omega_i(\chi)Y_i,$$

where $\omega_i(\chi)$ is the weights.

In this chapter, we present two methods for the estimation of the weights $\omega_i(\chi)$. Firstly, we extend the method for univariate scalar response case to the multivariate case but using different bandwidths for different components of the response. This method is denoted by MRD and given as follows:

$$\omega_i(\chi) = \left(\sum_{i=1}^{n} \tilde{\omega}_i\right)^{-1/2} \left(\sum_{i=1}^{n} \tilde{\omega}_i\right)^{-1/2},$$

where

$$\tilde{\omega}_i = \begin{bmatrix} a_{11}^i & 0 & 0 & \ldots & 0 \\ 0 & a_{22}^i & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & a_{qq}^i \end{bmatrix},$$

and

$$a_{11}^i = K(h_1^{-1}d(\chi, \mathcal{X}_i))$$
$$a_{22}^i = K(h_2^{-1}d(\chi, \mathcal{X}_i))$$
$$\vdots$$
$$a_{qq}^i = K(h_q^{-1}d(\chi, \mathcal{X}_i)).$$

Here $K(\cdot)$ is the kernel function, and $h_1, h_2, \ldots, h_q$ are the $q$th bandwidths for different components of the response variables.

In the second model (denoted by MRC), we take the correlations between different components of the response variable into account and with use different bandwidths for different components of the response variables.
Chapter 5 Multivariate response

Since \((X_i, Y_i), i = 1, ..., n\), are independent and identically distributed, we can define

\[
V = \begin{bmatrix}
V_{11} & V_{12} & \cdots & V_{1q} \\
V_{21} & V_{22} & \cdots & V_{2q} \\
\vdots & \vdots & \ddots & \vdots \\
V_{q1} & V_{q2} & \cdots & V_{qq}
\end{bmatrix} = \text{Cov}(Y_i | X_i).
\]

Therefore, \(V\) is the \(q \times q\) covariance matrix of \(Y_i\), where the diagonal elements give the variances of each component of the response and the off-diagonal elements give the covariances among different components of the response.

Now let \(\hat{V}\) be an estimate of \(V\), i.e. the sample covariance matrix of \(Y_i\), and define

\[
K_{iH} = \text{diag}\{K(h_l^{-1} d(X_i, \chi))\}, \quad l = 1, 2, ..., q, \]

and

\[
\hat{\omega}_i = (K_{iH}^{-1/2} \hat{V} K_{iH}^{-1/2})^{-1}.
\]

Then in the second method the weights \(\omega_i(\chi)\) are given by

\[
\omega_i(\chi) = \left(\sum_{i=1}^{n} W_i\right)^{-1/2} \hat{\omega}_i \left(\sum_{i=1}^{n} W_i\right)^{-1/2},
\]

where \(W_i\) is a diagonal matrix containing the row sums of the matrix \(\hat{\omega}_i\).

It is noted that we need a \(q\)-dimensional bandwidth vector \(H = (h_1, \ldots, h_q)\) to allow different degrees of smoothing in different components of the response. In practice, the bandwidth is critical to achieving accurate results. Ferraty and Vieu (2004, 2006), Ferraty et al. (2007), and Burba et al. (2009) use the cross-validation procedure for selecting bandwidths. Similarly, we can also select bandwidth by cross-validation procedure. Let

\[
CV_l(h_l) = \frac{1}{n} \sum_{i=1}^{n} (y_{il} - \hat{r}_{l,-i}(X_i))^2, \quad l = 1, ..., q
\]

and

\[
CV(H) = CV_1(h_1) + CV_2(h_2) + \ldots + CV_q(h_q),
\]
where \( \hat{r}_{l-i}(\cdot) \) is the estimate of \( r_l(\cdot) \) without the \( i \)th sample. Therefore, the optimal bandwidths can be determined by minimizing \( CV(H) \).

### 5.2 Asymptotic properties

Let \( x \) be a given point in \( f \), and denote by \( B(x, h) \) the closed ball of centre \( x \) and radius \( h \), namely:

\[
B(x, h) = \{ x' \in f : d(x, x') \leq h \}.
\]

The model requires that the probability of \( X \) is such that there exists a non-decreasing function \( \phi_x \) such that:

(H1) \( \exists (C_1, C_2), \forall x \in f, \forall \varepsilon > 0, \]

\[
0 < C_1 \phi_x(\varepsilon) \leq P(X \in B(x, \varepsilon)) \leq C_2 \phi_x(\varepsilon) < \infty.
\]

And the joint distribution of \((X, Y)\) needs to satisfy:

(H2) \( \exists C_3, \forall r > 1, E(\|Y\|^r |X) < C_3 r! < \infty. \)

(H3) \( \exists C_4, \exists b > 0, \exists \gamma > 0, \forall x, x' \in f, \]

\[
\|r(x) - r(x')\| \leq C_4 d^b(x, x').
\]

In addition, we also need the following technical conditions on the kernel function and the bandwidth.

(H4) The kernel function has to be such that:

(i) \( K \) is a bounded and Lipschitz continuous function with support \([0, 1]\),

and if \( K(1) = 0 \) it has to fulfill, together with \( \phi_x(\cdot) \), the conditions:

(ii) \( \exists (C_5, C_6) > 0, \) such that \(-\infty < C_5 \leq K'_{\text{ih}} \leq C_6 < 0. \)

(iii) \( \exists C_7 > 0, \exists \gamma_0 > 0, \forall \gamma < \gamma_0, \]

\[
\int_0^\gamma \phi_x(u) du > C_7 \gamma \phi_x(\gamma).
\]
(H5) The bandwidth $h$ is a positive sequence such that:

$$\lim_{n \to \infty} h = 0 \quad \text{and} \quad \lim_{n \to \infty} \frac{\log n}{n \phi_x(h)} = 0.$$ 

It is noted that this set of assumptions are different from the case of functional covariate with univariate response and when both explanatory and response variables are functional, because of the different bandwidths for the different components of the response variable.

The rate of convergence of $\hat{r}(x)$ is stated in Theorem 5.1. As mentioned before, this is the first outcome of this type in a nonparametric functional regression setting when the covariate is functional while the response is multivariate with different bandwidths for each component.

Result B1: Define

$$K_{iH} = \text{diag}\{K(h_l^{-1}d(x_i, x)), \quad l = 1, 2, ..., q\},$$

then, $\exists(C_8, C_9) > 0$ such that

$$C_8 \phi_x(h) \leq E[\|K_{iH}\|] \leq C_9 \phi_x(h).$$

This result is apparent when $K(1) > 0$ and can be extended to the continuous kernel, satisfying $H4$ as shown in Lemma 4.4 in Ferraty and Vieu (2006).

**Theorem 5.1.** Under the hypotheses $H1 - H5$, we have:

$$\|\hat{r}(x) - r(x)\| = O(h^b) + O_{a.co.}\left(\sqrt{\frac{\log n}{n \phi_x(h)}}\right).$$

*Proof.* For $i = 1, ..., n$, we define

$$\Delta_i = \frac{(K_{iH}^{-1/2}\hat{\nu}K_{iH}^{-1/2})^{-1}}{E\|K_{iH}^{-1/2}\hat{\nu}K_{iH}^{-1/2}\|^{-1}}.$$

Note that $H1$, $H4$ and Result B1, ensure that $E\|K_{iH}^{-1/2}\hat{\nu}K_{iH}^{-1/2}\|^{-1} > 0$. 
Let \( \hat{r}_1(x) \) and \( \hat{r}_2(x) \) be the following quantities:

\[
\hat{r}_1(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i
\]

and

\[
\hat{r}_2(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \Delta_i.
\]

Then \( \hat{r}(x) = \hat{r}_2(x)\hat{r}_1^{-1}(x) \). Our proof depends on the decomposition:

\[
\hat{r}(x) - r(x) = \left\{ (\hat{r}_2(x) - E\hat{r}_2(x)) - (r(x) - E\hat{r}_2(x)) \right\} \hat{r}_1^{-1}(x)
- r(x)\left\{ \hat{r}_1(x) - 1 \right\} \hat{r}_1^{-1}(x).
\]

Theorem 5.1 will be true providing both the following Lemma 5.2 and Lemma 5.3 can be proved. However, the second part of Lemma 5.3 is addressed directly by using part (i) of Lemma 5.3 when \( Y_i = 1 \) in combination with part (i) of Proposition A.6 in Ferraty and Vieu (2006).

**Lemma 5.2.** Under \( H_3, H_4, \) and \( H_5 \) we have:

\[
E\|r(x) - E\hat{r}_2(x)\| = O(h^b).
\]

**Proof.** Model (5.1) allows us to directly write:

\[
\|r(x) - E\hat{r}_2(x)\| = \|r(x) - E(Y_1 \Delta_1)\|
= \|r(x) - E[E(Y_1 \Delta_1 | X_1)]\|
= \|r(x) - E[E(Y_1 | X_1) \Delta_1]\|
= \|r(x) - E(r(X_1) \Delta_1)\|
\leq E[\|r(x) - r(X_1)\|\|\Delta_1\|],
\]

and by using \( H_3 \), this becomes:

\[
\|r(x) - E\hat{r}_2(x)\| \leq E[d^b(x, X_1)\|\Delta_1\|].
\]
Therefore, with the hypothesis \( H_1 \) and Result B1, and since \( \|\Delta_1\| = 1, \exists C > 0 \), such that

\[
\|r(x) - E\hat{r}_2(x)\| \leq C h^b.
\]

The above inequality yields the proof, since \( C \) does not depend on \( x \). \( \square \)

**Lemma 5.3.** We have:

(i) Under assumptions \( H_1 - H_5 \) we have:

\[
\|\hat{r}_2(x) - E\hat{r}_2(x)\| = O_{a.co.}\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right).
\]

(ii) Under assumptions \( H_1, H_4, \) and \( H_5 \), we have:

\[
\|\hat{r}_1(x) - 1\| = O_{a.co.}\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right).
\]

**Proof.** The idea of this proof is based on using a Bernstein-type exponential inequality. In fact:

\[
P\left(\|\hat{r}_2(x) - E\hat{r}_2(x)\| > \epsilon\right) = P\left(\frac{1}{n}\|\sum_{i=1}^{n}(Y_i\Delta_i - E(Y_i\Delta_i))\| > \epsilon\right),
\]

and we have to offer that it exists \( \epsilon_0 > 0 \) such that:

\[
\sum_{n \in \mathbb{N}^*} P\left(\frac{1}{n}\|\sum_{i=1}^{n}(Y_i\Delta_i - E(Y_i\Delta_i))\| > \epsilon_0\sqrt{\frac{\log n}{n\phi_x(h)}}\right) < \infty.
\]

Therefore, applying Corollary A.8 – ii in Ferraty and Vieu (2006) gives the exponential inequality with \( Z_i = Y_i\Delta_i - EY_1\Delta_1 \). We first need to show that:

\[
\exists C_{10} > 0, \forall m = 2, 3, ..., \|E(Y_1\Delta_1 - EY_1\Delta_1)\|^m \leq C_{10}\phi_x(h)^{-m+1}. \quad (5.3)
\]

- We first prove that for \( m \geq 2 \):

\[
E\|Y_1\Delta_1\|^m = O(\phi_x(h)^{-m+1}). \quad (5.4)
\]
For this, we write:

\[ E\|Y_1\Delta_1\|^m \leq E[E[\|Y_1\|^m|X_1]\|\Delta_1\|^m]. \]

Clearly, we get from the assumption \( H2 \) that:

\[ E\|Y_1\|^m = E[E[\|Y_1\|^m|X]] < Cm! < \infty, \]

which implies that:

\[ E\|Y_1\Delta_1\|^m \leq E[E[\|Y_1\|^m|X_1]\|\Delta_1\|^m] \leq Cm!E\|\Delta_1\|^m. \]

By applying \( H4 \) and Result B1, we get:

\[ \frac{C_5}{\phi_x(h)^{m-1}} \leq E\|\Delta_1\|^m \leq \frac{C_6}{\phi_x(h)^{m-1}}, \]

then, we get:

\[ E\|Y_1\Delta_1\|^m = O(\phi_x(h)^{-m+1}). \quad (5.5) \]

- Furthermore, we utilize the Newton's binomial expansion and get:

\[ \|Y_1\Delta_1 - E(Y_1\Delta_1)\|^m = \sum_{k=0}^{m} C_{k,m} \|Y_1\Delta_1\|^k \|E[Y_1\Delta_1]\|^{m-k}(-1)^{m-k}, \]

where

\[ C_{k,m} = \frac{m!}{k!(m-k)!}, \]

which implies that

\[ E\|Y_1\Delta_1 - E(Y_1\Delta_1)\|^m \leq C \sum_{k=0}^{m} C_{k,m} E\|Y_1\Delta_1\|^k \|r(x)\|^{m-k} \leq C \max_{k=0,1,2,\ldots,m} k!\phi_x(h)^{1-k}, \]
where $C$ is a real positive constant. Because $\phi_x(h)$ tends to zero when $n$ goes to infinity, it becomes the case that:

$$E\|Y_1 \Delta_1 - E(Y_1 \Delta_1)\|^m = O((\phi_x(h))^{1-m}).$$

Therefore, we can apply Corollary A.8 – ii in Ferraty and Vieu (2006) with $a^2 = \phi_x(h)^{-1}$.

Subsequently, we have $u_n = (a^2 \log n / n) = \log n / (n \phi_x(h)) \to 0$ as $n \to \infty$, by using assumptions $H4$ and $H5$.

And the second part of the theorem is directly derived from part one by taking $Y_i = 1$, and we get the result:

$$\|\hat{r}_1(x) - 1\| = O_{a.c.} \left( \sqrt{\frac{\log x}{n \phi_x(h)}} \right).$$

\[\square\]

Theorem 5.1 offers that the estimator $\hat{r}(x)$ is consistent in terms of almost complete convergence. It also states of pointwise almost complete convergence. As discussed in Ferraty and Vieu (2006), this mode of convergence is stronger than almost sure convergence and convergence in probability, and the links of almost complete convergence with other modes of convergence are also shown in the above reference.

### 5.3 Numerical examples

The main goal of this section is to illustrate the usefulness of the methodologies through simulated data and real data examples (Spectrometric data).

#### 5.3.1 Simulation studies

**Example 1.** We first consider the following regression model:

$$Y_i = r(X_i) + \varepsilon_i, \quad i = 1, 2, ..., n = 215.$$
The functional predictor is generated by

\[ \mathcal{X}_i(t_j) = a_i \cos(2t_j) + b_i \sin(4t_j) + c_i(t_j^2 - \pi t_j + \frac{2}{9} \pi^2), \quad (5.6) \]

where 0 = t_1 < t_2... < t_{100} = \pi are equispaced points, a_i, b_i and c_i are independently drawn from a uniform distribution on [0, 1], and n represents the sample size. Fig. 5.1 shows the simulated 215 curves for one replication.

![Figure 5.1: A sample of 215 simulated curves.](image)

The multivariate response is simulated via the following steps.

1. Construct two regression function operators \( r_1 \) and \( r_2 \), each for one output.

The two nonlinear operators \( r_1 \) and \( r_2 \) are defined as follows:

\[
\begin{align*}
    r_1(\mathcal{X}_i) &= \int_0^\pi t \cos(t)(\mathcal{X}'_i(t))^2 dt + \eta_i + \gamma_i, \\
    r_2(\mathcal{X}_i) &= \int_0^\pi [t \cos(t)(4b_i \cos(4t) - a_i \sin(t)) + c_i(2t - \pi)^2] dt + \eta_i + \gamma_i + \omega_i,
\end{align*}
\]

where \( \mathcal{X}'(t) \) represents the first derivative of \( \mathcal{X}(t) \), \( \eta_i \) is a real-valued continuous variable following a standard normal distribution, \( \gamma_i \) is a discrete-valued
variable drawn from a Bernoulli distribution, and \( \omega_i \) is a real-valued continuous variable from an exponential distribution with rate parameter 1.

2. Generate two sets of random errors \( \{ \varepsilon_{1i} \}_{i=1}^{n} \) and \( \{ \varepsilon_{2i} \}_{i=1}^{n} \):

\[
\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N_2 \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} (\sigma_1)^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & (\sigma_2)^2 \end{bmatrix} \right),
\]

where \( \sigma_1 = 1 \), \( \sigma_2 = 2 \), and we consider two cases for \( \rho \): \( \rho = 0.9 \) (Case I) and \( \rho = 0.1 \) (Case II).

3. Compute the corresponding response variables:

\[
Y_{1i} = r_1(\mathcal{X}_i) + \varepsilon_{1i}, \quad Y_{2i} = r_2(\mathcal{X}_i) + \varepsilon_{2i}, \quad i = 1, 2, \ldots, n.
\]

We divide the 215 samples into two subsets: the first 160 are used for building the model, and the remaining 55 are used for prediction to assess the performance of the methods. We compare the two proposed models the multivariate with correlations between different components of the response variable into account and use different bandwidths for different components (MRC) model and multivariate with using different bandwidths for different components of the response (MRD) model with the independent response (IR) model where the two responses are modelled independently, based on the criterion of the root mean square prediction error defined as

\[
\text{RMSPE} = \sqrt{MSE}, \quad MSE = \frac{1}{55} \sum_{i=161}^{215} (y_i - \hat{y}_i)^2.
\]

The optimal bandwidths in the multivariate with correlations between different components of the response variable into account and use different bandwidths for different components (MRC) model, multivariate with using different bandwidths for different components of the response (MRD) model and the independent response (IR) model are determined by using a cross-validation procedure, that is, by minimising \( CV(H) = CV_1(h_1) + CV_2(h_2) \). While this minimisation is time-consuming, we see in our simulation studies that the optimal bandwidths for
MRC, MRD and IR are quite close to each other, therefore, to simplify the computation we instead use a two-step CV procedure as done in Xiang et al. (2013): we first determine the individual bandwidths \((h_{1,0}, h_{2,0})\) independently, then determine the two bandwidths by minimising \(CV(H)\) in a small neighbourhood of \((h_{1,0}, h_{2,0})\). In all the models considered, the Epanechnikov kernel function and semi-metric based on the second derivative \((B=2)\) are used for measure of proximity between curves. The above process is repeated 10 times the average of root mean square error is shown in Table 5.1. From Table 5.1, it can be seen that in case I, the multivariate with correlations between different components of the response variable into account and use different bandwidths for different components (MRC) model performs better than the multivariate with using different bandwidths for different components of the response (MRD) model and the independent response (IR) method, while the multivariate with using different bandwidths for different components of the response MRD model performs slightly better than the independent response IR model. On the other hand, neither MRC nor MRD are better than the IR method for Case II. This means that when the correlation is strong among different components of the response, the multivariate models (the multivariate with correlations between different components of the response variable into account and use different bandwidths for different components (MRC) model and multivariate with using different bandwidths for different components of the response (MRD) model) perform better than the independent (IR) model, even when the covariance is ignored in the method.

<table>
<thead>
<tr>
<th>Case</th>
<th>Components</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MRC</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1.0041</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.9678</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>1.0257</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.1027</td>
</tr>
</tbody>
</table>

Table 5.1: Average RMSE comparison between MRC, MRD and IR models based on ten replicated simulations.
Example 2. In the second example we generate the functional predictor by
\[ X_i(t_j) = a_i(t_j - 0.5)^2 + b_i, \quad i = 1, 2, \ldots, n = 100, \]  
(5.7)
where \(0 = t_1 < t_2 < \ldots < t_{100} = 1\) are equispaced points, \(a_i\) and \(b_i\) are independently drawn from a uniform distribution on \([0, 1]\), and \(n\) represents the sample size. Figure 5.2 shows the simulated 100 curves from one replication.

![Figure 5.2: A sample of 100 simulated curves.](image)

The regression functional operators \(r_1\) and \(r_2\) are constructed by
\[
\begin{align*}
    r_1(X_i) &= \int_0^{\pi/2} X_i'(t) dt, \\
    r_2(X_i) &= \int_0^{\pi/2} |X_i'(t)| \log |X_i'(t)| dt.
\end{align*}
\]
The other settings including the random errors are the same as in Example 1.

We also divide the samples into two subsets: the training set contains the first 75 units and the testing set includes the remaining 25 units. We compare the two proposed models, the the multivariate with correlations between different components of the response variable into account and use different bandwidths for
different components (MRC) model and multivariate with using different bandwidths for different components of the response (MRD) model with the independent response (IR) model. The bandwidths are determined in the same way as before. The results based on 10 replications are reported in Table 5.2. These outcomes confirm the conclusion of Example 1 that when the correlation is strong among different components of the response, the multivariate methods perform better than the independent model, whilst all the three methods give essentially the same results when the correlation is weak.

<table>
<thead>
<tr>
<th>Case</th>
<th>Components</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MRC</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>0.9900</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.9750</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>1.0150</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.9047</td>
</tr>
</tbody>
</table>

Table 5.2: Average RMSE comparison between MRC, MRD and IR models based on ten replicated simulations.

Example 3. We now consider another simulated example where the time points are randomly collected and the error term follows a multivariate $t$-distribution.

We generate the functional predictor by

$$X_i(t_j) = a_i \cos(t) + b_i \sin(t) + W_i, \quad i = 1, 2, ..., 215, \quad (5.8)$$

where $0 = t_1 < t_2 < ... < t_{100} = \pi$ are uniformly sampled in $[0, \pi]$, $a_i$, $b_i$ and $W_i$ are independently drawn from a uniform distribution on $[0.2, 1.5]$. Figure 5.3 shows the simulated 215 curves from one replication.
Chapter 5 *Multivariate response*

![Figure 5.3: A sample of 215 simulated curves.](image)

The two regression functional operators $r_1$ and $r_2$ are defined as: for $i = 1, \ldots, 215$

\[
\begin{align*}
    r_1(X_i) &= \frac{1}{100} \int_0^\pi X'_i(t)X_i(t) dt, \\
    r_2(X_i) &= \frac{1}{100} \int_0^\pi X''_i(t)X_i(t) dt,
\end{align*}
\]

where $X''(t)$ denotes the first derivative of $X(t)$. The response variables are given by

\[
Y_{1i} = r_1(X_i) + \varepsilon_{1i}, \quad Y_{2i} = r_2(X_i) + \varepsilon_{2i}, \quad i = 1, 2, \ldots, 215,
\]

where the error term $(\varepsilon_{1i}, \varepsilon_{2i})^t$ follows a bivariate $t$-distribution with the degrees of freedom $\nu = 3$ and the covariance matrix

\[
\frac{\nu}{\nu - 2} \begin{pmatrix}
    (\sigma_1)^2 & \rho \sigma_1 \sigma_2 \\
    \rho \sigma_1 \sigma_2 & (\sigma_2)^2
\end{pmatrix},
\]

where $\sigma_1 = 1$, $\sigma_2 = 2$, and we consider two cases for $\rho$: $\rho = 0.9$ (Case I) and $\rho = 0.2$ (Case II).
The same experiment procedure as in the first instance is then conducted, and the average RMSEs based on 10 repetitions by the three models are reported in Table 5.3. The results further confirm the conclusion of the previous two examples, that is, when the correlation is strong the multivariate with correlations between different components of the response variable into account and use different bandwidths for different components (MRC) model significantly improves the prediction accuracy compared with the multivariate with using different bandwidths for different components of the response (MRD) and the independent response (IR) models, whilst all the three methods have similar achievement when the correlation is weak.

<table>
<thead>
<tr>
<th>Case</th>
<th>Components</th>
<th>RMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MRC</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1.6195</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.0334</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>1.4706</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.0142</td>
</tr>
</tbody>
</table>

Table 5.3: Average RMSE comparison between the MRC, MRD and IR models based on ten replicated simulations.

5.3.2 Real data example

We now demonstrate the effectiveness of the suggested models using a real data example - the spectrometric data.

Spectrometric data is a well known example in functional data analysis, and several applications have been undertaken using different models; see, for example, Ferraty and Vieu (2003, 2006), Ferraty et al. (2007), Benhenni et al. (2007), Burba et al. (2009) and Wang et al. (2017). The data comes from quality control, and can be found at http://lib.stat.cmu.edu/datasets/tecat or. It contains a sample of finely chopped pieces of meat. For each meat sample, the data consists of a 100-channel absorbance spectrum with a wavelength range between 850-1050nm (see Ferraty and Vieu (2006) for more details). The aim of spectrometric analysis
is to determine the proportion of specific chemicals’ content because the analysis by chemistry programming would take a lot more time and be more expensive. For this particular data, we want to predict fat, water and protein content simultaneously from spectrometric curves.

In this analysis, the response variable $Y_i$ are the percentages of fat, water and protein content in each piece of meat and the predictor is the spectrometric curves. We split the original sample into two subsets: the learning sample includes the first 160 units, and the testing sample consists of the last 55 units. The sample correlations between fat and water, between fat and protein and between water and protein are -0.9881, -0.8604 and 0.8145, respectively. The measure of performance for different models is achieved by calculating the root mean square prediction error in the testing sample. The results are reported in Table 5.4.

From Table 5.4, it can be clearly seen that the two new multivariate methods (the multivariate with correlations between different components of the response variable into account and use different bandwidths for different components (MRC) model and multivariate with using different bandwidths for different components of the response (MRD) model) perform better than the independent method. Compared to the independent method, their root mean square error values are smaller across two components of the response variable (fat and water) whilst all the three models give quite similar results for protein. This suggests that the multivariate methods are more accurate by modelling all the response components simultaneously.

<table>
<thead>
<tr>
<th>Components</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRC</td>
</tr>
<tr>
<td>Fat</td>
<td>2.144</td>
</tr>
<tr>
<td>Water</td>
<td>1.957</td>
</tr>
<tr>
<td>Protein</td>
<td>1.603</td>
</tr>
</tbody>
</table>

Table 5.4: RMSE comparison between MRC, MRD and IR models.
5.4 Conclusion

This chapter proposes two new models for nonparametric functional regression when the covariate is functional and the response is multivariate. The first solution is to directly extend the nonparametric method for univariate response to multivariate response with different bandwidths for different responses, and in the second model the correlation among different responses is taken into account. We used kernel function and cross-validation procedure for getting bandwidth and two different types of semi-metrics (semi-metric based on derivatives and semi-metric built on the functional principle component analysis) are used for measure of proximity between curves. The rate of almost complete convergence is presented under certain conditions. The numerical instances presented in the chapter suggest that our models can perform well, especially when the correlation between different components of the response is strong.

It is noted that when we choose different bandwidths for different components of the response by a cross-validation procedure, our numerical results perform reasonably well. Xiang et al. (2013) mentioned that the bandwidths chosen by this approach usually have a large variability. Therefore, the study is needed to provide a structure for a simple procedure to choose the bandwidths, and our method may not be particularly suitable for high-dimensional multivariate functional data because of the complexity in choosing different bandwidths for different components. Furthermore, applying the proposed methodologies to other functional regression estimators, such as the $k$-nearest neighbour kernel estimator (Burba et al. (2009)) and functional local linear kernel estimator (Benhenni et al. (2007)).
Chapter 6

Nonparametric method for functional regression with multivariate functional responses

Ferraty et al. (2012) proposed a nonparametric regression model for cases in which the response and the covariate are both functional. In this chapter, we study the functional regression with multivariate functional responses. We first extend the method from univariate functional response to multivariate functional responses model (MFR), and then propose a new method where the correlation among different functional responses is taken into account (MFRC). The functional kernel with an automatic bandwidth selection is utilized by the cross-validation procedure along with a semi-metric build on the functional principal component analysis for the measure of the proximity between curves. The usefulness of the proposed models is illustrated through a simulation data and a real data example in the UK weather and their performances are compared. It is identified that the multivariate functional responses model with correlation (MFRC), which takes advantage of correlation between different functional responses, improves the accuracy of prediction in terms of mean square error.

The chapter is organized as follows. Section 6.1 presents the models and the methods of estimation. Section 6.2 illustrates the efficiency of the suggested models through a simulation example. Section 6.3 applies the models to a real data example for UK historic weather data. Finally, a general conclusion and future plans are provided in Section 6.4.
6.1 Construction of the estimators

Suppose \((X, Y)\) is a pair of random functions in \(E \times E'\), where \(E\) is an infinite dimensional space equipped with a semi-metric \(d(\cdot, \cdot)\). Let \((X_i, Y_i)_{i=1,...,n}\) be a sample of \(n\) pairs, independent and identically distributed as \((X, Y)\). Following the general framework of functional nonparametric regression:

\[
Y_i = m(X_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{6.1}
\]

where \(Y_i = (y_{i1}, ..., y_{iq})^t\), \(m(X_i) = (m_1(X_i), m_2(X_i), ..., m_q(X_i))^t\), and \(\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iq})^t\) is a \(q\)-dimensional random error with \(E[\varepsilon_i | X_i] = 0\). Therefore, the object to be estimated is the nonlinear operator \(m(\chi) = E[Y | X = \chi]\) for a given \(X = \chi \in E\).

By the nonparametric kernel method, \(m(\cdot)\) can be estimated by

\[
\hat{m}(\chi) = \frac{1}{n} \sum_{i=1}^{n} \omega_i(\chi) Y_i, \tag{6.2}
\]

where \(\omega_i(\chi)\) is the weights.

In this chapter, we present two models for the estimation of \(\omega_i(\chi)\). Firstly, we extend the method for univariate functional response proposed in Ferraty et al. (2012) to the multivariate functional response case and define the following multivariate functional responses model (MFR):

\[
\omega_i(\chi) = \frac{\hat{\omega}_i(\chi)}{\sum_{i=1}^{n} \hat{\omega}_i(\chi)},
\]

with

\[
\hat{\omega}_i(\chi) = K(h^{-1}d(X_i, \chi)),
\]

where \(K(\cdot)\) is a kernel function, \(h\) is the bandwidth, and \(d\) is the semi-metric defined on \(E\). Therefore, \(m(\cdot)\) can be estimated by MFR as:

\[
\hat{m}(\chi) = \frac{\sum_{i=1}^{n} K(h^{-1}d(X_i, \chi)) Y_i}{\sum_{i=1}^{n} K(h^{-1}d(X_i, \chi))}. \tag{6.3}
\]

For the second model (MFRC), we consider the discrete version of the problem and the correlations between different functional responses are taken
into account. Let $\mathbf{Y}_{ip} = (y_{i1p}, y_{i2p}, ..., y_{ijp})$ be the vector of $J$ observations on the $p$th dimension of the functional response for the $i$th pair, and define $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, ..., \mathbf{Y}_{iq})$, $i = 1, 2, ..., n$. Since $(X_i, Y_i)$, $i = 1, ..., n$, are independent and identically distributed, we denote

$$\mathbf{\nu} = [\nu_{lm}]_{Jq \times Jq} = \text{Cov}(\mathbf{Y}_i | X_i).$$

Thus $\mathbf{\nu}$ is the covariance matrix of size $Jq \times Jq$ and consists of $q$ submatrices of size $J \times J$. The diagonal sub-matrices provide the covariances within each dimension of the functional responses, and the off-diagonal sub-matrices give the covariances among different dimensions of the functional responses.

Let

$$K_i = \begin{pmatrix} K(h^{-1}d(X_i, \chi)) & 0 & \cdots & 0 \\ 0 & K(h^{-1}d(X_i, \chi)) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K(h^{-1}d(X_i, \chi)) \end{pmatrix}_{Jq \times Jq},$$

$$\hat{\omega}_i(\chi) = \left( K_i^{-1/2} \mathbf{\nu} K_i^{-1/2} \right)^{-1},$$

and $W_i(\chi)$ be a diagonal matrix containing the row sums of the matrix $\hat{\omega}_i(\chi)$. The weight $\omega_i(\chi)$ is then defined as follows:

$$\omega_i(\chi) = \left( \sum_{i=1}^{n} W_i(\chi) \right)^{-1/2} \hat{\omega}_i(\chi) \left( \sum_{i=1}^{n} W_i(\chi) \right)^{-1/2},$$

and the estimated functional response is given by

$$\hat{\mathbf{m}}(\chi) = \sum_{i=1}^{n} \omega_i(\chi) \mathbf{Y}_i. \quad (6.4)$$

It is noted that if $\mathbf{\nu}$ is an identity matrix, the estimator (6.4) is equivalent to the vectorized version of (6.3).

The semi-metric $d(\cdot, \cdot)$ measures the closeness between curves in functional data analysis. Ferraty and Vieu (2006) introduces three families of semi-metrics,
which are based on functional principal component analysis, on derivatives and on partial least squares (PLS), respectively. In our numerical experiments the semi-metric based on functional principal component analysis is adopted. The bandwidth $h$ is determined by cross-validation procedure.

6.2 Simulation study

The prime purpose of this section is to illustrate the suggested methodology through a simulated example. We consider the following regression model:

$$Y_i(t_j) = m(X_i)(t_j) + \varepsilon_{ij}, \quad i = 1, 2, ..., n = 200,$$

where $X_1, X_2, ..., X_n$ are $n = 200$ functional explanatory curves, such that

$$X_i(t_j) = a_i \exp(w_i t_j) + b_i \cos(w_i t_j), \quad j = 1, ..., p,$$

where, $0 = t_1 < t_2 < ... < t_p = \pi$ are $p = 10$ equally spaced points, and $a_i, b_i$ (respectively $w_i$) are independently drawn from uniform distribution on $[0.2, 1.5]$ (respectively $[0.2, 2.5]$). Figure 6.1 displays the simulated 200 curves for one replication.

We construct a two-dimensional functional regression operator with components $m_1(\cdot)$ and $m_2(\cdot)$ as follows: for all $i = 1, 2, ...200$ and $j = 1, ..., p$,

$$\begin{cases} 
m_1(X_i)(t_j) = \int_0^{t_j} X'_i(u)X_i(u)du, \\
m_2(X_i)(t_j) = \int_0^{t_j} X_i(u)^2du,
\end{cases}$$

where $X''(t)$ is the first derivative of $X(t)$.

The random errors $\varepsilon_{ij}$ are independent samples from a bivariate normal distribution:

$$\varepsilon_{ij} \sim N_2 \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \right)$$

with $\sigma_1 = 1.5$ and $\sigma_2 = 0.64$. We consider two different values for $\rho$: $\rho = 0.9$ (Case I) and $\rho = 0.1$ (Case II), to represent strong and weak correlations between the
two functional responses. Figure 6.2 demonstrates the two functional responses from one replication.

We consider three methods for this example. Besides the proposed the multivariate functional responses (MFR) model and the correlations between different functional responses (MFRC) model, we also apply the univariate method of Ferraty et al. (2012) to each dimension of the functional response to obtain estimators of the individual components of \( m(\cdot) \), denoted as individual functional response (IFR) model. In all of these models, the quadratic kernel function and the semi-metric based on the functional principal component analysis are utilized with \( q = 1 \). The cross-validation procedure is adopted for selecting the bandwidths.

The original sample is split into two subsets: the first 150 is the training sample and used for building the model; the remaining 50 is the testing sample and used for predictions. To measure the prediction accuracy, the mean square
error (MSE) is computed as:

\[ MSE = \frac{1}{50} \frac{1}{10} \sum_{j=1}^{10} \sum_{i=151}^{200} (m_{ij} - \hat{y}_{ij})^2, \]

where \( m_{ij} \) and \( \hat{y}_{ij} \) are the true functional regression values and their estimates, respectively.
### Table 6.1: Simulation study - average of MSE for MRF C, MRF and IFR, based on 20 replications.

<table>
<thead>
<tr>
<th>Case</th>
<th>fun. resp.</th>
<th>MRC</th>
<th>MRD</th>
<th>IFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>0.8243</td>
<td>1.4371</td>
<td>1.4123</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.3931</td>
<td>6.9640</td>
<td>6.9333</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>0.6229</td>
<td>1.3723</td>
<td>1.1682</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.5070</td>
<td>7.9542</td>
<td>7.8177</td>
</tr>
</tbody>
</table>

The above process is repeated 20 times, and Table 6.1 presents the averages of the mean square errors of the three models for two different cases. It can be seen in Table 6.1 that the multivariate functional responses with correlation (MFR C) model is significantly better than both the individual functional responses (IFR) and multivariate functional responses (MFR) models in both cases, which implies that accommodating the correlations among the different functional responses improves the prediction accuracy whenever the correlations between functional responses are strong or weak. It is noted that the results obtained from MFR model is worse than the individual functional responses (IFR) model, which may be due to the fact that MFR model uses the same bandwidth for both dimensions.

Figure 6.3 presents three randomly chosen samples from one replication of the study.

### 6.3 Application to the UK weather data

This section evaluates the proposed methods by applying them to a real data example. The data contains the monthly maximum and minimum temperatures at Valley Station in the UK between the years 1931 and 2015. In Figure 6.4, the top plot represents the monthly maximum temperature and the bottom plot depicts the monthly minimum temperature.

For this example, the temperatures in the first half year of each year are used as the functional predictor while the ones in the second half year are treated as the functional responses. The data is randomly split into two parts: one part of
Figure 6.3: Simulation study - three randomly chosen samples for each model with the true values. Top: functional response 1; bottom: functional response 2. The black curves (cross) represent the true values, and the blue (triangle), green curves (plus) and red (square) show the predicted curves by multivariate functional responses with correlation (MFRC) model, multivariate functional responses (MFR) method and individual functional responses (IFR) model, respectively.

65 years is used for training dataset and the other of 20 years for testing dataset. The mean square errors (MSE) between the true values and the predicted values for all three methods are reported in Table 6.2 for both functional responses.
Figure 6.4: The maximum and minimum temperatures from 1930 to 2015. Top panel: the monthly maximum temperature; bottom panel: the monthly minimum temperature.

<table>
<thead>
<tr>
<th>Functional responses</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFRC</td>
</tr>
<tr>
<td>Maximum temperature</td>
<td>1.2553</td>
</tr>
<tr>
<td>Minimum temperature</td>
<td>1.0946</td>
</tr>
</tbody>
</table>

Table 6.2: The average MSE by MFRC, MFR and IFR models for the UK weather data.

It can be observed that the multivariate functional responses with correlation (MFRC) model has the smallest mean square error (MSE) than the multivariate functional responses (MFR) model and individual functional responses (IFR) method, and there are considerable reductions in the percentage of MSE by MFRC model compared to the MFR method and the IFR model. It indicates that the MFRC has better prediction accuracy. These testing results show that taking
account of the correlation between the two functional responses is of significant importance for prediction accuracy.

Figure 6.5 explains two test samples, which are randomly chosen from each functional responses (maximum temperature and minimum temperature) from top to bottom, respectively.

6.4 Conclusion

This chapter contributes to the development of new methods for constructing nonparametric estimator of functional regression, when the covariate is functional and the response is multivariate functional. Two methods are considered: one is
a direct extension of the method for univariate functional response (Ferraty et al. (2012)) to multivariate functional responses and the other is a new method where the correlation among different functional responses is taken into account. The numerical examples show that the method in which the correlations between the functional responses are taken into account has superiority in terms of prediction accuracy, even though the correlations between functional responses are weak.

The bandwidth is automatically selected using the cross-validation procedure. Hall and Robinson (2009) suggested that the bandwidths chosen by the cross-validation approach usually have a large variability. However, in our numerical examples this method executes well in implementation. In the proposed models, the same bandwidth is used for all dimension of the functional response. It may be more sensible to consider different bandwidths for different dimensions to allow different degrees of smoothing, or to study the Bayesian bandwidth estimation procedure as proposed by Shang (2013). Furthermore, the semi-metric used in this paper is based on the functional principal component analysis. It is worth investigating how to choose the optimal semi-metric in practice, which is still an open problem Ferraty and Vieu (2006).
Chapter 7

Conclusions and future work

7.1 Conclusions

In this thesis, we introduce the framework of a nonparametric functional regression and its extensions, including the prediction of multivariate response variables (Mul-R) from a functional covariate, principal component analysis to decorrelate multivariate response variables (Mul-RP) with functional covariate, multivariate response variables with different bandwidths for different components of the response (MRD), and taking the correlation between different components of the response into account with different bandwidths for different components of the response (MRC), and the multivariate functional responses (MRF), and also take the correlation between different functional response into account (MFRC) from a functional covariate. By applying all these methods to functional data and simulation datasets, nonparametric functional regression and its extensions show the powerful ability and utility of this method in different fields. This thesis is divided into four main parts.

7.2 Multivariate Response Variable

In chapter 3, we extended the nonparametric functional regression to the situation where the response is multivariate response variables and the covariate takes values from certain infinite dimensional spaces. We used the K-nearest neighbour estimator for predicting nonlinear functional regression. We selected the optimal bandwidth using a cross-validation procedure. Under the model presented, we
show that the K-nearest neighbour estimator provides good predictions when compared with the outcomes obtained from the independent response method, which is studied in the literature. The measure of proximity between the functional covariates is achieved via the semi-metrics, namely the semi-metric based on functional principal component analysis and that built on the derivatives, as pointed out by Ferraty and Vieu (2006). The effectiveness of the proposed model was then evaluated by calculating mean square prediction errors and then comparing these to those obtained from the independent method. The utility of the multivariate response variable (Mul-R) and independent response variables (Ind-R) models were illustrated through numerical examples of real data (spectrometric data, and Canadian weather station data) and simulated data. We found that our estimator is well adapted to the prediction of multivariate response variables for several reasons, such as saving in time and reduced cost it represents. There are several ways in which the presented methodologies can be extended, and we will briefly indicate some of these possibilities. Apply the presented model to other functional data, such as the ozone level data studied by Quintela-del Rı et al. (2011). Consider other functional regression, such as the functional Nadaraya-Watson (NW) estimator model, the functional local linear estimator, and distance-based local linear estimator (Ferraty and Vieu (2003, 2004, 2006), Ferraty et al. (2007), Burba et al. (2009), Barrientos-Marin et al. (2010), and Boj et al. (2010). Also, in any future work, we will attempt to study nonparametric functional regression when the covariate is a function and take the correlation between different components of the responses into account.

7.3 Principal Component Analysis

In chapter 4, we introduced a new model with which to analyse nonparametric functional regression where the covariates include functional variables and a multidimensional response. In this chapter, we treat with principal component analysis as a tool to decorrelate multivariate response variables. We also use the K-nearest neighbour function to predict variables with a cross-validation procedure in order to choose the optimal bandwidth. The closeness between the functional
covariates is measured by two different kinds of semi-metrics, the first of which is based on the derivatives and the second of which is built on the functional principal component analysis discussed by Ferraty and Vieu (2006). The usefulness of the proposed model over the independent method is demonstrated through the results obtained from examples of real data and the simulation instance.

The nonparametric functional model incorporates a functional covariate. The principal component analysis style used to deal with multivariate response variables avoids the widely known difficulty in nonparametric regression model for multiple responses, that is, to formulate a covariance function that characterizes not only the correlation between data points, but also the correlation between responses. In this chapter, we suppose that the dimension of the response variable is low so that all the principal component analyses are utilized in the subsequent regression method. However, if the response variable has a very high dimensionality it is pointless to build a nonparametric functional regression method for each of the principal components, we can instead choose to model a smaller number of components which demonstrate at least 95% of the total variation or by setting this number by cross validation, which was pioneered by Wang et al. (2017). Furthermore, in the future, the model could be extended to other functional regression models such as the functional local linear kernel estimator, and the semi-functional partial linear regression method. The theoretical part of the proposed model also represents a good subject for future study.

7.4 Taking Different Bandwidths for Different Responses and Taking the Correlation

Chapter 5 focuses on a nonparametric kernel function for regression estimator with a multivariate response. This chapter presents two new methods for nonparametric functional regression when the covariate is functional and the response is multivariate. The first method is to directly extend the nonparametric model for independent response to a multivariate response variable with different bandwidths for different components of the responses, and in the second solution the correlation between different components of the responses is taken into account with
different bandwidths for different components of the responses and incorporated into the model. This chapter contains the theoretical study, the rate of almost complete convergence is presented under certain conditions (when the covariate includes a function and the response is multivariate with different bandwidths for different component of the responses, and the correlation between different components of the responses is taken into account). We use the kernel function for estimation with a cross-validation procedure to choose the optimal bandwidth. Two types of semi-metrics are used to measure the closeness between functional covariates (semi-metric based on the derivatives, and a semi-metric built on the functional principal component analysis). The numerical examples presented in the study suggest that our methods can be successfully implemented, especially when the correlations between different components of the response variable is strong. This means that the correlation between different components of the response and the different bandwidths for each component of the response positively affect the outcomes.

It was noted that when we chose different bandwidths for different components of the response via the cross-validation process, our numerical outputs executed reasonably well. Xiang et al. (2013) noted that the bandwidths chosen by cross-validation procedures usually have quite a large variability. Therefore, the approach needs to provide the structure for a simple procedure with which to choose the bandwidths, and our model may not be particularly suitable for high-dimensional multivariate functional data because of the complexity in choosing different bandwidths for different components of the response. Therefore, we can extend the proposed methodologies by applying them to other bandwidth estimation procedures, such as the Bayesian bandwidth studied by Shang (2013). Furthermore, applying the proposed method to other functional regression estimators, such as the $k$-nearest neighbour kernel estimator model (Burba et al. (2009)) and the functional local linear kernel estimator method (Benhenni et al. (2007)), is also a candidate for future work.
7.5 Multivariate Functional Responses

Chapter 6 extended the nonparametric functional regression to the situation where the response is a multivariate functional response and the covariate is a functional or multivariate functional. Two models are observed: one is a direct extension of the method for independent functional response (Ferraty et al. (2012)) to a multivariate functional response and the other is a new model where the correlation among different functional responses is taken into account. We use both methods in the kernel function for estimation purposes and the cross-validation procedure to choose the optimal bandwidth. The semi-metric based on the functional principal component analysis is used as a measure of the proximity between curves, the results of which were compared with the independent functional response which is commonly studied in the literature where, significantly, the MRC model shows outstanding performance in both the simulated and real examples. The numerical instances show that the model in which the correlations between the functional responses are taken into account is superior in terms of prediction accuracy, even though the correlations between the functional responses are themselves quite weak.

The optimal bandwidth is automatically chosen using the cross-validation procedure. Hall and Robinson (2009) noted that the optimal bandwidths selected by the cross-validation procedure usually have a wide variability. Therefore, for future study, we can choose the optimal bandwidth using other procedures such as the Bayesian bandwidth pioneered by Shang (2013). While, in our numerical examples this model performed well in application. In the methods presented, the same bandwidth is used for all dimensions of the functional response. It may be more reasonable to consider different bandwidths for different dimensions of the functional response to allow different degrees of smoothing. Moreover, the semi-metric utilized in this chapter is built on the functional principal component analysis. It is worth investigating how to select the optimal semi-metric in practice, which is still an open problem Ferraty and Vieu (2006).
Bibliography


