QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

# VOLATILITY ESTIMATION IN MISSING AT RANDOM HIGH-FREQUENCY FINANCIAL TIME SERIES

BY

FERIEL ACHAIBOU

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters in Applied Statistics

June 2023

# COMMITTEE PAGE

The members of the Committee approve the Thesis of

Feriel Achaibou defended on 18/05/2023.

Dr. Mohamed Chaouch
CAS, Qatar University
Thesis Supervisor

Prof. Mhamed Mesfioui
University of Quebec at Trois Rivières, Canada
Committee Member

Prof. Lanouar Charfeddine
CBE, Qatar University
Committee Member

Approved:

Ahmed Elzatahry, Dean, College of Arts and Sciences

# ABSTRACT

Achaibou, Feriel, Masters : June : 2023, Masters in Applied Statistics

Title: Volatility estimation in missing at random high-frequency financial time series

Supervisor of Thesis: Dr. Mohamed Chaouch

CAS, Qatar University.

More than 15 years ago, the capital markets have seen significant development, introducing high-frequency trading and a shift of market towards high-frequency and algorithm trading. It was always believed that high-frequency trading and automated trading were source price shocks and rising of volatility. Therefore, more interest was recently given in modeling the volatility with high-frequency financial data. However, financial data can still be missing despite modern technology that allows data collection on a very fine time scale. Thus, this thesis focuses on the estimation of regression and volatility functions based on missing data using a nonparametric heteroscedastic regression model. A Nadaraya-Watson type estimator is used when the response variable is a real-valued random variable and subject to missing at random mechanism, while the predictor is a completely observed infinite-dimensional (functional) random variable. Based on the observed data, we first introduce a simplified, as well as inverse probability weighted, estimators. Second, these initial estimators are used to impute missing values and define estimators of the regression and volatility operators based on imputed data. Third, the performance of the proposed estimators is assessed using simulated data. Finally, an application to the estimation and forecasting of the daily volatility of Brent Oil Price returns conditionally to 1-minute frequency daily Natural Gas returns curves is also investigated.

# DEDICATION

*This work is wholeheartedly dedicated to my beloved parents and twin sisters, who always encouraged me and have continually provided their moral and emotional support throughout my life.*

# ACKNOWLEDGMENTS

First prise is to Allah Almighty, who gave me the strength and patience I needed to complete this thesis and enlightened me enough to continue this journey and reach this new success in my life.

Secondly, I would like to express my gratefulness to Dr. Mohamed Chaouch, my thesis supervisor, for his academic guidance, support and patience during my thesis work. I am also thankful to the committee members, Prof. Mhamed Mesfioui and Prof. Lanouar Charfeddine for their valuable feedback and suggestions that improved the quality of this thesis.

My acknowledgment would be incomplete if I did not thank my family. The prayers and benediction of my dearest parents, and the love and care of my spiritual twin sisters.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

## Problem statement

In ancient times, traders used manually follow up and post stock price on boards. However, the main drawbacks of this traditional method include slow trading activity, manually calculations and few orders are executed per day. Starting from nineties with the arrival of computers, the trading activity become more automated. The traders can follow the market dynamic and make the better decision. Nowadays, with the progress of computers and the increase of their computational capacity, financial institutions moved towards high frequency trading. This trading activity is mainly performed by computers and orders can be executed every millisecond. Despite its advantages in trading large volumes of securities and make profit from every very small price fluctuations, high frequency trading has been linked to increased market volatility and even market crashes. For instance, in 2010 the financial market experienced so called flash-crash a type of stock market crash which started at 2:32 pm and lasted for approximately 36 minutes. Therefore, stock indices, such as S&P500, Dow Jones or Nasdaq collapsed and rebounded very rapidly. From this perspective, understanding and modeling the financial market intraday volatility with high frequency financial data and if possible predict it would be of great interest to investors to take the right decisions. In addition, financial firms, that trade assets on high-frequency time scale, are not just interested in short-term forecasting of future values of financial assets, but also assess the risk associated to such predictions through the volatility components. Thus, risk analysis plays an important role in financial market. Financial risk could be defined as the risk of losing part or all of an investment. Regulators and owners of financial institutions routinely use risk measures to aid in decision making process. The purpose of risk measure is to

assess the level of risk in a portfolio. Higher risk investments should generate higher returns to offset the risk of losing money. [1] was the first to use volatility as proxy risk where the idea of risk optimisation was introduced. That is, for a particular return on a particular investment, investors need to minimize the volatility of those returns in order to maximize their utility so-called modern portfolio theory. Besides, [2] discussed several risk measures such as volatility, the most popular measure for financial risk. Investors are more interested in returns instead of prices because returns have statistical feature involving stationarity. For instance, if the price changes a lot over a given period time, then volatility is high. However, it is difficult to determine how high is the volatility. Because of latent nature of volatility, it must therefore be forecasted by a statistical model.

## Volatility models

Let $(X_t, Y_t)_{t=1,\ldots,n}$ be a strictly stationary process with the same distribution as $(X, Y)$, where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$. Denote $m(x) := \mathbb{E}(Y|X = x)$ and $U^2(x) := \text{var}(Y|X = x)$ the regression function and the conditional variance, respectively. Suppose that the random sample $(X_t, Y_t)_{1\leq t\leq n}$ is generated by the following nonlinear heteroscedastic model:

$$Y_t = m(X_t) + U(X_t)\varepsilon_t, \qquad t = 1, \ldots, n, \tag{1.1}$$

where $\mathbb{E}(\varepsilon_t|X_t) = 0$ and $\text{var}(\varepsilon_t|X_t) = 1$. Several approaches to estimate the regression and the conditional variance functions were introduced in the literature. For instance, [3] discussed parametric methods to estimate the autoregression function and the volatility (see Chap.4) and nonparametric approaches to estimate the same quantities (see, Ch.8).

It is worth noting that model (1.1) encompasses several interesting volatility models. In the following, we discuss some specific models and the main existing contributions in the literature.

**(a) Parametric autoregressive models with ARCH/GARCH errors:**

If we consider $X_{t-1} \equiv (Y_{t-1}, \ldots, Y_{t-d})^\top$, then model (1.1) becomes

$$Y_t = m(Y_{t-1}, \ldots, Y_{t-d}) + U(Y_{t-1}, \ldots, Y_{t-d})\, \varepsilon_t. \tag{1.2}$$

Moreover, when $m(X_{t-1}) = 0$ and $U(X_{t-1}) \equiv U(X_{t-1}; \boldsymbol{\beta}) = \sqrt{1 + \sum_{j=1}^{d} \beta_j Y_{t-j}^2}$ one obtains the pure autoregressive conditional heteroscedasticity (ARCH) model introduced in [4].

An extension of Engle's model was considered in [5] where

$$m(X_{t-1}) \equiv m(X_{t-1}; \boldsymbol{\alpha}) = \sum_{j=1}^{q} \alpha_j Y_{t-j} \quad \text{and} \quad U(X_{t-1}; \boldsymbol{\beta}) = \sqrt{1 + \sum_{j=1}^{d} \beta_j Y_{t-j}^2}. \tag{1.3}$$

In such case the underling process $Y_t$ generated by the model (1.2) is an autoregressive process with ARCH errors. The study of the statistical properties of an AR-ARCH model was considered, for instance, in [6] and [7] when $q = d = 1$ (AR(1)-ARCH(1)) and [8] for higher orders of $q$ and $d$.

The AR-ARCH models presented above belong to the class of linear and parametric models for the volatility. Indeed, as one can see from (1.3), the shape of the volatility is identified through some unknown parameters which should be estimated. Moreover, common estimation techniques such as the maximum-likelihood method will require to impose a certain probability distribution on the errors.

In practice, before assuming that a financial time series is generated by an AR-

ARCH model, we have to use Goodness-of-fit (GOF) tests to assess the parametric form of the volatility (see [9], and [10] for more details). Moreover, assuming a specific probability distribution (in general Gaussian) on the $\varepsilon_t$ is a very restrictive condition which is in general violated when we model financial time series.

**(b) Nonparametric autoregressive models with ARCH errors:**

In order to avoid any miss-specification of the parametric form of the regression and volatility functions in (1.2), and to relaxe the assumption on the probability distribution of the innovations $\varepsilon_t$, nonparametric approaches represent a relevant alternative to the AR-ARCH models. In this case we do not impose any specific form on the functions $m(\cdot)$ and $U^2(\cdot)$. Only smoothness (regularity) conditions will be needed to achieve good statistical properties of the nonparametric estimators of the regression and the volatility functions.

Given $(X_t, Y_t)_{t=1,\ldots,n}$, a strict stationary process, several nonparametric estimators of $m(\cdot)$ and $U^2(\cdot)$ were proposed in the literature. For instance, [11] used Nadaraya-Watson type estimator to estimate the parameters in the nonlinear autoregressive model with ARCH errors given in (1.2). The estimator of the autoregressive part $m$ is defined, for any fixed x, as follows:

$$m_n(x) = \frac{\sum_{t=d}^{n-1} Y_{t+1} K\Big(\frac{x - X_t}{h_n}\Big)}{\sum_{t=d}^{n-1} K\Big(\frac{x - X_t}{h_n}\Big)}, \tag{1.4}$$

where, as defined above, $X_t \equiv (Y_t, \ldots, Y_{t-d+1})^\top$, is presented as a special case when the predicted values of the process $Y_t$ are obtained based on the past information $X_t$, $K$ is a kernel and $h_n$ is called bandwidth which is a sequence of positive numbers tending to zero as $n$ goes to infinity.

The nonparametric estimation of the regression function received a lot of interest among the statistics community. Asymptotic properties of such estimator, including consistency, asymptotic distribution, were discuss in [12], [13], [14], [15] among others. It has been proven in the cited literature that the choice of the kernel does not really affect the quality of the estimation. However, the bandwidth plays a crucial role in the estimation. The selection of the bandwidth can be obtained either by minimizing some risk measure such as the mean square error or numerically using cross-validation techniques.

Regarding the volatility part, observe that it can be estimated in two different ways. The first one consists in using the following simple decomposition of the conditional variance:

$$U^2(x) = \mathbb{E}(Y^2|X = x) - (m(x))^2. \tag{1.5}$$

Therefore, one may estimate $\widetilde{m}(x) \equiv \mathbb{E}(Y^2|X = x)$ nonparametrically using (1.4) and by replacing $Y_{t+1}$ by $Y_{t+1}^2$. Then a *difference based estimator* of $U^2(x)$ can be defined as

$$\widetilde{U}_n^2(x) = \widetilde{m}_n(x) - (m_n(x))^2. \tag{1.6}$$

The main drawback of the difference based estimator is that, in practice, it may lead to negative values of volatility.

The second method to estimate $U^2(x)$ is called *residual-based approach*. If the

regression function is know, one can see the conditional variance as:

$$U^2(x) = \mathbb{E}\Big((Y - m(X))^2 | X = x\Big), \tag{1.7}$$

which is a regression function of the squared residuals squared $(Y - m(X))^2$ of the predictor $X$. Therefore, a Nadaraya-Watson type estimator of $U^2(x)$ can be defined as follows:

$$U_n^2(x) = \frac{\sum_{t=d}^{n-1} \Big(Y_{t+1} - m_n(X_t)\Big)^2 K\Big(\frac{x - X_t}{h_n}\Big)}{\sum_{t=d}^{n-1} K\Big(\frac{x - X_t}{h_n}\Big)}, \tag{1.8}$$

where, as defined above, $X_t \equiv (Y_t, \ldots, Y_{t-d+1})^\top$. [11] investigated the local constant estimator of the regression and conditional variance given in (1.4) and (1.8), respectively, when data is supposed to be generated from model (1.2). He provided a uniform consistency rate for both estimators and established their asymptotic distributions. [15] introduced a local linear estimator of $m(\cdot)$ and $U^2(\cdot)$ and compared the efficiency of the residual-based and difference-based estimators of $U^2(x)$. [16] discussed the conditional variance estimation in heteroscedastic regression model where he introduced more efficient method than local linear regression estimation to ensure that the conditional variance estimator is always positive.

**(c) Stochastic volatility**

Another interesting reason to study model (1.1) is that it includes continuous-time stochastic models which are used to model diffusion processes. Several financial assets, say $X$, are modeled using diffusion processes solution of the following stochastic

differential equation:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \qquad t > 0, \tag{1.9}$$

where $W_t$ is a standard Brownian motion. The drift $\mu(\cdot)$ and the diffusion $\sigma^2(\cdot)$ are in general unknown functions. Several well-known models in financial econometrics, (including [17]; [18]; [19]; [20], among others), can be written under the form (1.9) with a specific form of drift and diffusion functions. In practice, a diffusion process $\{X_t\}$ cannot be observed continuously over time. It is rather observed at instants $\{t = i\Delta | i = 0, \ldots, n\}$, where $\Delta > 0$ is very small. For instance the series could be observed hourly, daily, weekly or monthly. High-frequency financial data are usually daily or intradaily series. Following Euler discretization scheme, one gets a discretized version of (1.9). That is

$$X_{t+\Delta} - X_t = \mu(X_t)\Delta + \sigma(X_t)\Delta^{1/2}\varepsilon_t, \tag{1.10}$$

where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed standard normal random variables. Taking $Y_t = X_{t+\Delta} - X_t$, $\mu(X_t)\Delta = m(X_t)$, and $\sigma(X_t)\Delta^{1/2} = U(X_t)$, model (1.10) can be viewed as a special case of model (1.1). [17] and [21] studied nonparametric estimation of $\mu$ and $\sigma$.

Particular cases discussed in **(a)**, **(b)** and **(c)** clearly show that it is important to study a general heteroscedastic regression model of the form (1.1). In this thesis we are interested in extending the model (1.1) to the case where $Y \in \mathbb{R}$ and the predictor $X$ is a functional random variable. That is $X \in \mathcal{E}$ where $\mathcal{E}$ is an infinite-dimensional space endowed with a certain semi-metric. Moreover, we suppose that the response variable

$Y$ is subject to a missing at random mechanism. Note that whenever $Y$ is completely observed real random variable and $X \in \mathcal{E}$, [22] studied nonparametric estimation of $m(\cdot)$ and $U^2(\cdot)$, the regression and conditional variance operators and investigated their asymptotic properties.

Moreover, all the above mentioned references assume that the time series is completely observed. In practice, financial time series may be subject of a missing at random mechanism (see [23]). Therefore estimators given in (1.4) and (1.8) cannot be used to predict future values of financial assets or quantify the risk assigned to such prediction through the volatility component. Recently, when $x \in \mathbb{R}$, [24] considered heteroscedastic regression model with fixed design ($X$ is not random) and used local polynomial method to estimate conditional variance function with correlated errors and missing at random response. Four nonparametric estimators of the conditional variance function were proposed and authors conclude that imputed estimator provides better results than simplified one.

## Thesis outline

The organization of this thesis is as follows. Chapter 1 describes the problem statement, reviews different types of volatility models introduced in the literature and the contribution of this thesis. Chapter 2 discusses the relevant literature which is used for the objectives of this thesis. The estimation of the regression and conditional variance for both cases complete and missing data is discussed in Chapter 3 . The imputed estimators of the conditional variance are obtained in Chapter 4 based on imputation approaches including regression imputation and inverse probability weighting. Chapter 5 reports the simulation results, while in Chapter 6 shows an application of the methodology on

high-frequency financial data. Finally, the main findings of this thesis are summarised in Chapter 7 and some open research ideas are given for further investigations in the future.

## Contribution of the Thesis

This thesis deals with nonparametric estimation of the regression and conditional variance operators in a nonlinear heteroscedastic functional regression model. It supposes that the response $Y$ is a real-valued random variable subject to a missing at random mechanism. However, the covariate $X$ is functional in nature taking value in an infinite dimensional space endowed with a certain semi-metric and completely observed. This thesis can be seen as an extension of several recent contributions in functional data analysis:

- In [22] a heteroscedastic functional regression model was considered when the data are completely observed. We extend the estimators in [22] to the missing at random case.

- In [25] an inverse probability weighting kernel-based estimator of the regression operator was investigated when the response variable is missing at random and a homoscedastic functional regression model is considered. Under the same model, [26] studied the simplified estimator of the regression operator. In this thesis, we extend both results in the case of heteroscedastic functional regression model with MAR response.

- In [24] a fixed design heteroscedastic regression model with correlated errors was considered, and based on local polynomial regression, four nonparametric estimators of the conditional variance were proposed when there are missing responses.

In this thesis, we extend the work in [24] to heteroscedastic functional regression model by generalizing the Nadaraya-Watson type estimator of the conditonal variance taking into consideration the missing at random response.

# CHAPTER 2: LITIRATURE REVIEW

## Nonparametric regression estimation: finite dimensional case

In statistics is is very common that we want to understand how a response variable $Y$ is concomitant to a certain predictor $X$. Regression models represent one of the most powerful tools to understand the relationship between these random variables. In this section we suppose that the data generating model is written as follows:

$$Y = m(X) + \texttt{error}, \tag{2.1}$$

where $m : \mathbb{R} \to \mathbb{R}$ is an unknown function to be estimated and the `error` term is centered conditionally to $X$. Therefore, one can show that, when $X = x$, $m(x) = \mathbb{E}\left(Y|X = x\right)$.

In what follows we briefly remind the reader how to estimate $m(x)$ nonparametrically using an i.i.d random sample $(X_i, Y_i)_{i=1,\dots,n}$ copies of $(X, Y) \in \mathbb{R} \times \mathbb{R}$. Note that the regression function or the conditional mean $m(\cdot)$ can be rewritten as

$$m(x) = \mathbb{E}(Y|X = x) = \int y f_{Y|X=x}(y)dy = \int \frac{y f(x, y)}{f(x)}dy, \tag{2.2}$$

where $f(x, y)$ is the joint density function of $(X, Y)$ and $f(x)$ is the marginal density of $X$. A plug-in estimator of $m(x)$ can be obtained by replacing each unknown quantity in (2.2) by its empirical version. Therefore, a Nadaraya-Watson type estimator of $m(x)$, at any fixed point $x$, can be written as follows:

$$m_n(x) = \int \frac{y f_n(x, y)}{f_n(x)}dy, \tag{2.3}$$

where $f_n(x,y)$ and $f_n(x)$ are the nonparametric estimators of $f(x,y)$ and $f(x)$, respectively. These estimators are defined as follows:

$$f_n(x,y) = \frac{1}{nh^2}\sum_{t=1}^{n} K\Big(\frac{x-X_t}{h}\Big)K\Big(\frac{y-Y_t}{h}\Big) \text{ and } f_n(x) = \frac{1}{nh}\sum_{t=1}^{n} K\Big(\frac{x-X_t}{h}\Big). \quad (2.4)$$

After substituting (2.4) in (2.3), a Nadaraya-Watson type estimator of the regression function is obtained for any fixed $x$, as follows

$$m_n(x) = \frac{\displaystyle\sum_{t=1}^{n} Y_t K\Big(\frac{x-X_t}{h_n}\Big)}{\displaystyle\sum_{t=1}^{n} K\Big(\frac{x-X_t}{h_n}\Big)}. \quad (2.5)$$

where, $K$ is the kernel and $h := h_n$ is the bandwidth which is a sequence of positive numbers tending to zero as $n$ goes to infinity. Note that closer the observation $X_t$ to the fixed point $x$ where $m(x)$ is estimated, higher will be the weight assigned to corresponding $Y_t$.

Observe that the Nadaraya-Watson type regression estimator can be adapted to the multivariate case, where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$, and defined as follows:

$$m_n(x) = \frac{\displaystyle\sum_{t=1}^{n} Y_t K\left(\frac{||x-X_t||}{h}\right)}{\displaystyle\sum_{t=1}^{n} K\left(\frac{||x-X_t||}{h}\right)}, \quad (2.6)$$

where, $||\cdot||$ denotes the Euclidean norm and defined, for any $u = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$, as $\|u\| = (u_1^2 + \cdots + u_d^2)^{1/2}$.

Alternatively, one can consider the so-called the kernel product estimator of the regression function where, for any $u \in \mathbb{R}^d$, $K(u) = \prod_{j=1}^{d} K_j(u_j)$. For simplicity, we consider here the same kernel $K_j$ for any $j = 1, \ldots, d$. Therefore, one can define $m_n(x)$

as follows:

$$m_n(x) = \frac{\sum_{t=1}^{n}\prod_{j=1}^{d} Y_t K\left(\frac{x_j - X_{t,j}}{h}\right)}{\sum_{t=1}^{n}\prod_{j=1}^{d} K\left(\frac{x_j - X_{t,j}}{h}\right)}. \tag{2.7}$$

Note that the kernel product estimator is easy to calculate compared to one given in (2.6). However, it does not preserve the dependence structure in the vector $X$ as is the case in (2.6). The statistical properties of such estimator were discussed in the literature. For instance [27] and [28] investigated the strong consistency and the asymptotic normality of $m_n(x)$ for $\alpha$-mixing processes. In addition, the uniform consistency of the estimator for $\phi$-mixing process were discussed in [29],[30] and [13].

**Remark 1** (**On the principle of local weighting).** *The local weighting techniques are very well adapted to nonparametric estimation. One of the most common approaches among theses local weighting techniques in finite dimensional case is the kernel one. Kernel local weighting is based on a kernel function $K$ and bandwidth $h$. In finite dimensional case, if $x$ is fixed point, then the kernel smoothing transforms the observation $X_t$ into $\Delta_t$ as follows such that $\Delta_t = \frac{1}{h}K\left(\frac{x - X_t}{h}\right)$, $t = 1, \ldots, n$. The kernel density estimator is defined as $f(x) = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{h}K\left(\frac{x - X_t}{h}\right)$. The basic idea behind the local weighting around a fixed $x$ is to assign a weight to each observation $X_t$ based on the similarity between $x$ and $X_t$. The more $X_t$ is close to $x$, the higher is the weighting.*

In practice, there are different choices of possible kernels that listed below and displayed in Figure 2.1.

`Gaussian Kernel:` $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$

`Epanechnikov or Quadratic Kernel:` $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{\{|u|\leq 1\}}$

`Uniform or Rectangular Kernel:` $K(u) = \frac{1}{2}\mathbb{1}_{(|u|\leq 1)}$

`Triangular Kernel:` $K(u) = (1 - |u|)\mathbb{1}_{(|u|\leq 1)}$

`Biweight Kernel:` $K(u) = \frac{15}{16}(1 - u^2)^2\mathbb{1}_{(|u|\leq 1)}$

`Cosine Kernel:` $K(u) = \frac{1}{2}(1 + \cos(\pi u))\mathbb{1}_{(|u|\leq 1)}$

`Optcosine Kernel:` $K(u) = \frac{\pi}{4}\cos(\frac{\pi}{2}u)\mathbb{1}_{(|u|\leq 1)}$

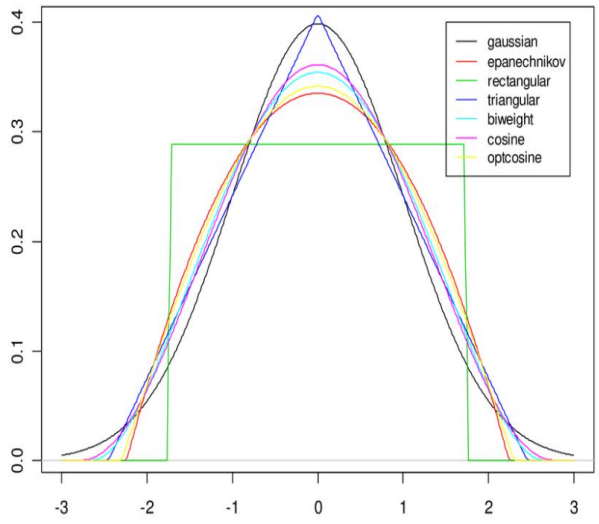



Figure 2.1. Different type of usual kernels used in nonparametric estimation.

**How to choose the bandwidth?** The accuracy of the kernel smoothing method mainly depends on the smoothing parameter $h$. It is well-know in nonparametric estimation that the choice of the kernel is not a determinant factor of the quality of estimation even though it has been proven that the Epanechnikov gives slightly better results. In contrast, the choice of the smoothing parameter is crucial theoretically as well as in practice. Indeed, from a theoretical point of view, it has been proven (see Theorem 3.1 in Bosq (1998), page 70) that the “optimal" mean square convergence rate of the kernel-type estimator of the regression function is of order $n^{-4/(d+4)}$ for a bandwidth $h_n = c_n n^{-1/(d+4)}$, where

$d$ is the dimension of the predictor $X$. Thus, a good choice of the bandwidth allows to reach some optimal convergence rates of the estimators. From a practical point of view, the optimal bandwidth is usually selected based on the minimization of a certain risk measure. For instance, if we can explicitly find an analytical expression of the asymptotic mean square error (AMSE) or the asymptotic integrated mean square error, as a function of $h$, then one can find the analytical expression of the corresponding bandwidth minimizing such criterion. In general this approach will lead to a value of bandwidth which is also depending on some unknown parameters that should be estimated as well. The second approach, called cross-validation, is purely numerical which consists in choosing the optimal bandwidth that minimizes the sum of the square prediction errors. For illustration, the Figure 2.2 displays an example of the effect of varying the bandwidth. A smoothing parameter that is too large can obscure the characteristics of the distribution (e.g. oversmoothed curve). However, a too small value of $h$ can overemphasize the variability (e.g. undersmoothed curve). The decision how much is too smooth is important in nonparametric estimation, hence the selection of the smoothing parameter is well known to be a challenging task.

### Functional data analysis

While classical statistics refers to the analysis of random univariate, vectors, and matrices, functional data analysis (FDA) deals with the analysis and theory of random functions. FDA is a new field in statistics which aims to analyze curve-type data. This area's history is much older and go back to [31] and [32]. Data from many fields come to us through a process that is naturally described as functional including chemometric data, speech recognition data, and electricity consumption data, etc.

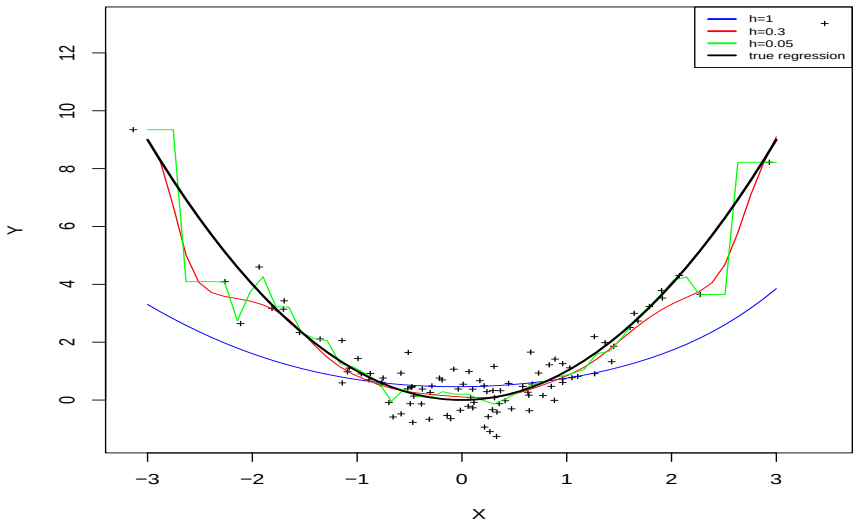



Figure 2.2. Example of effect of the bandwidth selection on the regression estimation.

In the last decade, FDA received special interest among the statistical community where several statistical approaches, studied in multivariate statistics framework, were generalized when functional/curve-type data are available. Several monographs were published to discuss FDA statistical methods. For instance [33] discussed parametric models for functional. In [34] more focus was given to nonparametric approaches. The last methods are doubly infinite in nature because they suppose that the regression function is not defined through some parameters, belongs to an infinite dimensional family and the predictor is an infinite-dimensional random variable (see, e.g. [35]).

Formally, from now on, we will denote $\mathcal{E}$ an infinite dimensional space where the functional random variable takes values. The larger is the space, the sparser are the data. An interesting question is that about the sparseness of the data in high-dimensional space when we work with functional data? Clearly, the sparseness concept is strongly linked to method used to measure closeness between data. Closeness measure between mathematical objects is an important concept in several statistical analysis methods.

In many situations, a classical norm can be used to measure the proximity of two objects. For instance, the usual Euclidean norm, which is based on sum of squares of the components of any vector, is the one of the most widely used in finite dimensional euclidean space ($\mathcal{E} = \mathbb{R}^d, d \geq 1$). In finite dimensional space, there is an equivalence in all norms, the choice of norm is therefore not crucial. However, in the functional data context, it is necessary to approach the issue differently since the equivalence between norms fails.

*Semi-metric choice and measure of similarity between curves*

Suppose that $Z_1 := (Z_1^1, Z_1^2, \ldots, Z_1^d)^\top$ and $Z_2 := (Z_2^1, Z_2^2, \ldots, Z_2^d)^\top$ are two $d$-dimensional random vectors. The similarity between $Z_1$ and $Z_2$ can be quantified using the Euclidean distance is defined as:

$$\|Z_1 - Z_2\|_E = \left((Z_1^1 - Z_2^1)^2 + \cdots + (Z_1^d - Z_2^d)^2\right)^{1/2}.$$

Now suppose that $X_1$ and $X_2$ are elements in a certain functional space, take for instance $\mathcal{E} = L^2(T)$, the space of square integrable functions on the interval $T$ endowed with its $L^2$-norm. A natural extension of the Euclidean distance to measure the similarity between $X_1$ and $X_2$ could be defined as:

$$\|X_1 - X_2\|_2 = \left(\int_T (X_1(t) - X_2(t))^2 \, dt\right)^{1/2}. \tag{2.8}$$

Figure 2.3 displays two curves with very similar shape but with different magnitude. The general appearance tells that these curves are “similar". However, the calculation of the $L^2$-distance comes to calculate the area between the curves $X_1$ and

$X_2$ which in such case will be large. Thus, one can wrongly conclude that $X_1$ and $X_2$ are not similar.

Note that if the magnitude of the predictor $X$ plays an important role in explaining the response variable $X$ then using the $L^2$-distance could be of interest. However, if the shape the $X$ is more relevant than its magnitude in explaining $Y$ then one cannot use the $L^2$-distance as a measure of similarity between curves. In such case one may use as a measure of proximity between curves the norm of the *difference between the first derivative* of $X_1$ and $X_2$ (when $X_1$ and $X_2$ are differentiable). That is

$$d(X_1, X_2) = \left( \int_T \left( X_1'(t) - X_2'(t) \right)^2 dt \right)^{1/2}. \tag{2.9}$$

Note that, in contrast to the $L^2$-distance, the distance based on the first derivative in (2.9) between $X_1$ and $X_2$ in Figure 2.3 would be $d(X_1, X_2) = 0$.

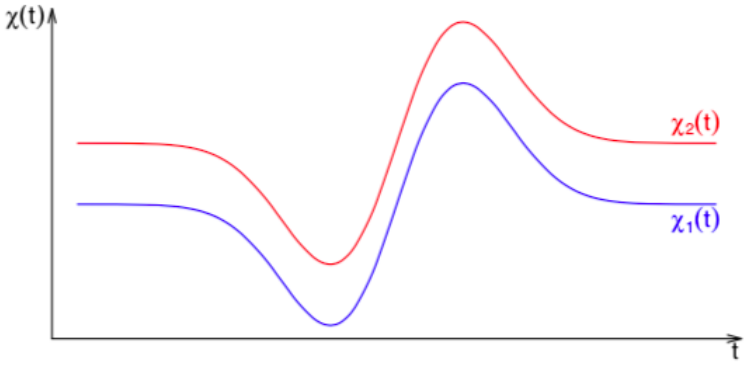



Figure 2.3. Example of curves.

Now we discuss another example that shows how important the selection of the appropriate semi-metric to measure the similarity between curves. Figure 2.4 displays three different curves, say $X_1, X_2$ and $X_3$.

A first comparison, based on the overall trend, between the three curves reveals that $X_1$ and $X_2$ are similar and completely different from $X_3$. However, $\|X_1 - X_2\|_2$ is

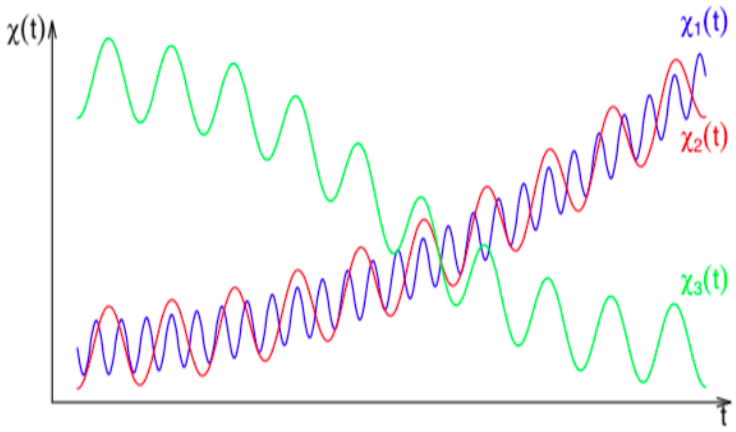



Figure 2.4. Example of curves.

certainly not small since $X_1$ and $X_2$ represent several differences in terms of amplitude and phase of the respective oscillations. If we suppose that the oscillations are just noise and that the general trend of the curves is most relevant to explain the variability in the response $Y$, one can think of using a semi-metric based on the projection of the curves on a suitable polynomial basis in order to have $d(X_1, X_2) \approx 0$, $d(X_1, X_3) > 0$ and $d(X_2, X_3) > 0$. If in contrast it is believed that, instead of the trend, the oscillation component plays an important role in explaining $Y$ it is more reasonable to consider a semi-metric based on Fourrier transform of the curves in order to have $d(X_2, X_3) \approx 0$ and $d(X_1, X_2) > 0$.

In conclusion, while choosing a semi-metric we first have to identify the most relevant features in the curves that may explain the variability in the response $Y$. Once identified, we can then build a *projection-type semi-metric* after considering the appropriate bases (e.g. polynomial, B-Spline, Wavelets, Fourrier, Principle Component Analysis) on which we can project our curves.

*Some commonly used semi-metrics*

As discussed above the concept of semi-metric in functional data analysis plays an important role theoretically as well as practically. We present here three semi-metric families, but many others can be built. The first two are well adapted for rough curves, whereas the third one is for smooth data. To start with, the Principle Component Analysis (PCA) is a popular method for showing data in reduced dimensional space in many multivariate analysis. PCA approaches were recently extended to functional data and used for many different statistical purposes. Thus, functional PCA can be used to build semi-metrics in order to compute similarity between curves in a reduced dimensional space. Indeed, this type of semi-metrics are usable only if the curves are rough and the datasets are balanced, when the curves are observed at the same points and the grid of measurements sufficiently fine. Let $\tilde{X}^q = \sum_{k=1}^{q} \big( \int X(t) v_k(t) dt \big) v_k$ be a truncated version of the expanansion of $X = \sum_{k=1}^{\infty} (\int X(t) v_k(t) dt) v_k$, where $v_1, v_2, \cdots$ denotes the orthogonal eigen functions of the covariance operator $\Gamma(s,t) = \mathbb{E}(X(s)X(t))$ associated with the eigen values $\lambda_1 \geq \lambda_2 \geq \cdots$. Therefore, a class of semi-norm can be introduced from the classical $L^2$-norm as follows:

$$\|X\|_q^{PCA} = \sqrt{\sum_{k=1}^{q} \left( \int X(t) v_k(t) dt \right)^2}.$$

Then, we define the semi-metric based on FPCA truncated component at $q$ as follows:

$$d_q^{FPCA}(X_1, X_2) = \sqrt{\sum_{k=1}^{q} \left( \int \{X_1(t) - X_2(t)\} V_k(t) dt \right)^2}.$$

Another way to build a new family of semi-metrics based on Partial Least

Square (PLS) when we observe an additional response by adapting multivariate partial least squares regression (MPLSR) method. The MPLSR can be used to predict a multivariate response from independent variables when there is a high degree of collinearity between the predictors and the number of predictors is very large compared to number of observations. The MPLSR method computes a simultaneous decomposition of the set of predictors and the set of responses such that the components are taken to maximize the covariance between the two sets of variables. The MPLSR, in particular, gives $p$ components, each corresponding to a response, where the computed components rely on parameter known as the the number of factors, the larger this parameter, the better data fitting. Taking many factors, on the other hand, can result in components with high variability. Based on this, the number of factors is equivalent to the number of dimension retained in PCA. The main difference with PCA is that the components performed in PCA only explain the predictors, whereas in the PLS approach the components are also relevant to the multivariate response. let $v_1^q, \cdots, v_p^q$ be the vectors of $\mathbb{R}^j$ performed by MPLSR where $q$ denotes the number of factors and $p$ the number of scalar responses. Thus, the semi-metric based on the MPLSR is defined as follows:

$$d_q^{PLS}(X_1, X_2) = \sqrt{\sum_{k=1}^{p} \left( \sum_{j=2}^{J} w_j (X_1(t_j) - X_2(t_j)[v_k^q]_j \right)^2},$$

where $w_1, \cdots, w_J$ are the weights.

One more method to build a family of semi-metrics between curves is consider a distance between one of their derivatives. For quite smooth curves, this type, namely, semi-metric based on higher order derivatives, could be appropriate. We consider the

following semi-metric given two observed curves $X_1$ and $X_2$:

$$d_q^{deriv}(X_1, X_2) = \sqrt{\int \left(X_1^{(q)}(t) - X_2^{(q)}(t)\right)^2 dt},$$

where $X^{(q)}$ being the $q^{th}$ order derivative of $X$.

**Nonparametric regression estimation: infinite dimensional case**

In this section we focus on explaining the extension of the concept of kernel local smoothing in the case of infinite-dimensional random variables. The expression of the regression operator $m(\cdot)$ can be then easily defined and will take similar form as the one in (2.6) after making the necessary adjustments.

Let $(X_t)_{1,\cdots,n}$ be $n$ functional random variables valued in $\mathcal{E}$ and let $x$ be a fixed element of $\mathcal{E}$. Then, an extension of kernel local smoothing would be to transform $n$ functional random variables $(X_t)_{1,\cdots,n}$ into $n$ quantities as follows:

$$\frac{1}{V(h)} K\left(\frac{d(x, X_t)}{h}\right)$$

Where $d(\cdot,\cdot)$ is a semi-metric defining the topology of the functional space $\mathcal{E}$, $K$ is a kernel, and $V(h)$ is the volume of $\mathcal{B}(x,h) = \{x' \in \mathcal{E}, d(x,x') \leq h\}$, which is the ball of center $x$ and radius $h$.

This method requires $V(h)$ to be defined. In other words, a measure on $\mathcal{E}$ must be available to quantify the volume of the ball $\mathcal{B}(x,h)$. When $\mathcal{E}$ is a finite dimensional space, the Lebesgue measure is used to quantify $V(h)$. However, when $\mathcal{E}$ is an infinite-dimensional space there no universal measure that could be used to calculate $V(h)$. As a result, in order to avoid having to choose a specific measure, we construct the

normalization directly using the probability distribution of functional random variable. Therefore a functional kernel local weighted variables are defined as follows:

$$\Delta_t = \frac{K\left(\frac{d(x,X_t)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(x,X_1)}{h}\right)\right)}, \qquad t = 1, \ldots, n.$$

By considering such normalization we have to technically be able to quantify $\mathbb{E}(K(h^{-1}d(x,X_1))$.

Observe that, since $X_1 \in \mathcal{B}(x,h)$ then $d(x,X_1) \leq h$. Therefore, one gets

$$\mathbb{E}\left(K\left(\frac{d(x,X_1)}{h}\right)\right) = \int_0^h K\left(\frac{u}{h}\right) d\mathbb{P}\left(d(x,X_1) \leq u\right).$$

Note that at this stage, in order to be able to quantify the above integral, one needs to evaluate the so-called *small ball probability*

$$F(u) := \mathbb{P}\left(d(x,X_1) \leq u\right).$$

Several authors were interested in this problem and found that it is possible to evaluate the small ball probability for some specific functional processes. Below we give some examples:

*Case 1:* When $X$ is a standard Ornstein-Uhlenbeck process. In other words the functional space $\mathcal{E} = \mathcal{C}\left([0,1],\mathbb{R}\right)$ and the semi-metric is the supremum norm. Then one can show that $F(u) \sim C_x e^{-\pi^2/8hu^2}$.

*Case 2:* When $X$ is a standard general diffusion process and $\mathcal{E} = \mathcal{C}\left([0,1],\mathbb{R}^d\right)$ and the semi-metric is the supremum norm. Then $F(u) \sim C_x e^{-Ch^{-2}}$.

*Case 3:* When $X$ is a fractal-type process of order $\tau$, then $F(u) \sim Cu^{\tau}$.

*Case 4:* When $X$ is of exponential-type with order $(\tau_1, \tau_2)$ then $F(u) \sim Ce^{-\log(1/h^{\tau_2})h^{-\tau_1}}$.

In general we work with any functional space endowed with a certain abstract semi-metric $d(\cdot, \cdot)$. [36] suggested to consider the following decomposition of the small ball probability:

$$F(u) = f_1(x)\phi(u) + o(\phi(u)), \quad \text{as } u \to 0, \tag{2.10}$$

where $0 < f_1(x) < \infty$ is a deterministic functional and $\phi(u)$, called the *concentration function*, is a real function tending to zero as its argument goes to zero. The function $\phi(\cdot)$ measures how densely packed are the considered elements of $\mathcal{E}$ in an infinite-dimensional ball of radius $u$. Observe that the small ball probability of processes given in Case 1 to case 4 satisfy the decomposition (2.10). [37] gave several further example of processes (such as the Hilbert autoregressive process of order 1) for which the decomposition in (2.10) is also satisfied.

**Remark 2.** *Note that decomposition* (2.10) *helps in quantifying* $\mathbb{E}\left(K\left(h^{-1}d(x, X_1)\right)\right)$ *and is also valid when* $\mathcal{E} = \mathbb{R}^d$. *Indeed, in such case* $f_1(x)$ *will be the density function of* $X$ *and* $\phi(u) = u^d 2\pi^{d/2}/\Gamma(d/2)$, *where* $\Gamma(\cdot)$ *is the Gamma function. Note that* $u^d 2\pi^{d/2}/\Gamma(d/2)$ *represents the volume of the hypersphere of radius* $u$ *in* $\mathbb{R}^d$.

Now, given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, copies of $(X, Y) \in \mathcal{E} \times \mathbb{R}$, one defines a kernel-type estimator of the regression operator $m : \mathcal{E} \to \mathbb{R}$ as follows:

$$m_n(x) = \frac{\sum_{t=1}^{n} Y_t K\left(\frac{d(x, X_t)}{h}\right)}{\sum_{t=1}^{n} K\left(\frac{d(x, X_t)}{h}\right)}, \qquad \forall x \in \mathcal{E}. \tag{2.11}$$

*The curse of the infinite dimension*

Consider nonparametric regression estimation where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$, if the dimension is greater than 3 (e.g. $d \geq 3$), then the estimator of the regression function converges more slowly to the true regression function. In other words, as the dimension of the predictor increases, the rate of uniform convergence $(\log n/n)^{r/(2r+d)}$ ($r$ is the degree of smoothness of the regression function with respect to the Euclidean norm) of $m_n(x)$ toward $m(\cdot)$ decreases. Note that the rate $(\log n/n)^{r/(2r+d)}$ obtained in the finite dimensional case is acceptable since by the rule of l'Hospital the limit in infinity of $\log n/n$ is equal to the limit of $1/n$. In other words the optimal convergence rate is of order $n^{-r/(2r+d)}$.

Since in functional data setting the predictor $X$ is supposed to be infinite dimensional then we expect that $m_n(x)$ in (2.11) would have poor theoretical properties. Indeed, if $X$ is an exponential-type process and $d(\cdot, \cdot)$ is the $L^2$-norm $\|\cdot\|_2$ as defined in (2.8), one can show that, in such case, the convergence rate of $m_n(x)$ towards $m(x)$ will be of order $(\log n)^{-\nu}$, for some $\nu > 0$. Compared to the rate $(\log n/n)^{r/(2r+d)}$ obtained in the finite dimensional case, a rate dependent on the logarithm of the sample size is statistically unacceptable (since it cannot be written as a power of $n$).

*How the choice of the semi-metric helps in solving the curse of infinite dimension?*

To be able to answer such question let us first introduce some definitions and results.

**Definition 1.** *[Fractal-type Process] A variable $X$ is said to be of fractal order $\tau$, with respect to the semi-metric $d(\cdot, \cdot)$, if there exists some finite constant $C > 0$ such that the*

*associated concentration function $\phi(\cdot)$ is of the form*

$$\phi(u) \sim Cu^{\tau} \quad \textit{as} \quad u \to 0.$$

The following theorem established the convergence rate of the regression operator estimator $m_n(x)$ when the predictor $X$ is a fractal-type process.

**Theorem 1.** *[see [34], p. 208]*

*Assume that $X$ is of fractal order $\tau$. Then, for $x$ fixed, we have almost surely*

$$m_n(x) - m(x) = \mathcal{O}\left(\left(\frac{\log n}{n}\right)^{r/(2r+\tau)}\right),$$

*where $r$ is the degree of smoothness of $m(\cdot)$ with respect to the semi-metric $d(\cdot,\cdot)$.*

**Remark 3.** *Observe that Theorem 1 allows to obtain the finite-dimensional convergence rate only when the predictor $X$ is a fractal-type process. This result is very important since if $X$ is any functional random variable, if we find a way to transform it into a fractal-type process then Theorem 1 becomes applicable and the optimal rate could be reached.*

The following lemma represents the key point to transform any functional process $X$ into a fractal-type process.

**Lemma 1** (see [34], p. 213)**.** *Let $H$ be a separable Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and let $\{e_j, j = 1,\ldots,\infty\}$ an orthogonal basis. Let $k \in \mathbb{N}^{\star}$ be fixed. Let $X = \sum_{j=1}^{\infty} x^j e_j$ be a fixed element in $H$.*

$(i)$ *$\forall(X_1, X_2) \in H \times H$, the function defined by*

$$d_k(X_1, X_2) = \left( \sum_{j=1}^{k} \langle X_1 - X_2, e_j \rangle^2 \right)^{1/2}$$

*is a semi-metric on the space $H$.*

$(ii)$ *Let $X = \sum_{j=1}^{\infty} x^j e_j$ be a squared integrable random element of $H$. If the random variable $\mathbf{x} = (x^1, \ldots, x^k)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^k$ with a continuous density function $f$, then the process $X$ is fractal order $k$ with respect to the semi-metric $d_k$ in the sense of Definition 1.*

Note that a given functional random variable $X$ can be projected on some orthonormal bases such Fourrier, Wavelets, functional PCA to be transformed into a fractal-type process. Then, one can use any of the corresponding semi-metrics described in Subsection 2.2.2.

### Missing data and imputation techniques

The problem of missing data is another factor to take into account in this thesis. In practice, despite the modern technology, which allows to collect data at a very fine time scale, financial data can still be missing. For instance, there are some regular holidays, such as Thanksgiving Day and Christmas, for which stock price data are missing. There are many other technical reasons (such as breakdown in devises recording data, computers' sudden shutdowns, ...) that make stretches of data missing. In the literature of financial data analysis, it is commonly assumed that the data are completely observed which is not realistic. Therefore, the problem of missing data arises whenever there is a disturbance in the sequence of the series in terms of observations, hence it

is necessary to address this such problem. The statistical inference with missing data can be found in statistical literature, hence we will refer to [23] work in this section. Missing data are not simply unobserved values that must be filled with imputed data or removed from the analysis. The missing data pattern may reveal valuable information. Understanding the nature of missing data is important to make accurate statistical inference. Researchers often consider missing data mechanism in choosing data imputation method. The missingness pattern defines which values are missing and observed in the data matrix, while mechanism deal with the relationship between missingness and the values of the variables in data matrix, where rows and columns represent observations and variables, respectively. Assumptions are needed to characterize the missingness process because the missing data mechanism is unknown in practice. In [23], they classified the missingness mechanism into three types: (i) missing completely at random (MCAR), (ii) missing at random (MAR) and (iii) missing not at random (MNAR). MCAR has probability of missingness do not rely on the values of the data, neither missing or observed. For example, lost data due to technical errors, e.g. miscalibration of MRI machine. MAR relaxes MCAR assumption and requires that the probability of missingness depends only on observed values. For example, unable to tolerate MRI sequences, predictable from participant's daily living activities. If the missingness mechanism is neither MCAR or MAR, hence the missingness assumption is MNAR, which depends on the value the outcome would have taken had it been observed. If the assumptions about missingness mechanism do not fit the situation with the data, then the results of the imputation approach may not reflect the actual situation. Several statistical methods were developed for handling missing data problem. Single imputation methods such as regression imputation was introduced in [23], single imputation methods can be

used to impute one value for each incomplete variable. Regression imputation is done using regression model that estimates the relationship between the observed values of the response and the predictor totally observed by applying ordinary least squares. Then, replace the missing observations of the response itself by their predictive values from regression. To illustrate the core idea of this method, consider an example, suppose the oil price contains missing data that has the value of another variable gold price, which is highly correlated with oil price. First, predict the missing value of oil price from gold price and then to fill in these missing values. Hence, missing at random mechanism holds since the probability of missingness is related to other variable not to value of the variable with missing data itself. Regression imputation method may result unbiased imputed values under MAR assumption.

# CHAPTER 3: VOLATILITY ESTIMATION UNDER MAR ASSUMPTION

In this chapter, we focus on the estimation of the regression and the volatility functions in heteroscedastic regression model, when the response variable $Y$ is a real-valued random variable and the predictor $X$ is infinite-dimensional random variable. In Section 3.1, we first consider the estimation of the regression and conditional variance operators when a complete observed sample is available at hand. Then, in Section 3.2 the estimation of the same parameters is considered when the predictor is completely observed whereas the response variable is subject to a missing at random mechanism.

## Case of completely observed data

Let $Y$ be a real-valued random variable and $X$ an infinite dimensional predictor taking values in a certain functional space $\mathcal{E}$ endowed with a certain semi-metric $d(\cdot,\cdot)$. One of the most common problems in statistics is to to understand how $Y$ and $X$ are concomitant. Regression models represent one of the most used tools to answer such question. That is, we assume that the relationship between the response variable $Y$ and its predictor $X$ is described by the following model:

$$Y = m(X) + \eta, \tag{3.1}$$

where $m : \mathcal{E} \to \mathbb{R}$ is called the regression operator and $\eta$ is the error term which is usually supposed to be independent of $X$ and with zero mean and constant variance. Notice that (3.1) is called a homoscedastic functional regression model since the error term $\eta$ has a constant variance. Several authors in nonparametric functional data analysis were interested in the estimation of the operator $m(\cdot)$. For instance, an extension of the Nadaraya-Waston estimator in the finite dimensional case to the model (3.1) was

proposed in [38] for (i.i.d) functional data, where the mean squared convergence rate and the asymptotic normality were established under strong mixing condition for the functional kernel regression estimator of the operator $m(\cdot)$. However, there are a number of well known processes where the mixing properties do not satisfy them, such as the linear AR(1) process, which is not strong mixing. Therefore, studying the asymptotic properties of the nonparametric regression estimator under more general assumptions was considered in the literature. In [37], the Nadaraya-Watson estimator was used to estimate $m(\cdot)$ and under stationarity and ergodicity assumption, a uniform almost sure consistency rate and the asymptotic normality were obtained.

Assuming that the errors have a constant variance and are independent of $X$ is a strong assumption which is usually not fulfilled when we analyze real data. In this thesis we relax such strong condition and we assume the following heteroscedastic functional regression model:

$$Y = m(X) + U(X)\varepsilon, \tag{3.2}$$

where $U : \mathcal{E} \to \mathbb{R}^+$ is the conditional standard deviation operator.

Model (3.2) allows the errors to be dependent on the predictor $X$ and to have a non constant conditional variance. Therefore, this model is more realistic and more appropriate to consider especially when we deal with financial time stochastic processes. Despite its interest from a theoretical, as well as practical point of view, less attention was given to the estimation of the operators $m(\cdot)$ and $U(\cdot)$ in the literature of nonparametric functional data analysis. To the best of our knowledge, the only article that considered the estimation of the regression and the conditional variance operator for ergodic processes is [22].

Let us now focus on the identification of the parameters $m(\cdot)$ and $U(\cdot)$.

**Proposition 1** (*Identification of the operator* $m(\cdot)$)**.**

*Suppose that* $\mathbb{E}(\varepsilon|X) = 0$*, almost surely (a.s.), then* $m(X) = \mathbb{E}(Y|X)$ *a.s.*

**Proof**. Observe that, by applying the conditional expectation on equation (3.2), one gets,

$$\begin{aligned}
\mathbb{E}(Y|X) &= \mathbb{E}\{m(X) + U(X)\varepsilon|X\} \\
&= \mathbb{E}(m(X)|X = x) + \mathbb{E}(U(X)\varepsilon|X = x) \\
&= m(x) + U(x)\underbrace{\mathbb{E}(\varepsilon|X = x)}_{=0} \\
&= m(X).
\end{aligned}$$

■

**Proposition 2** (*Identification of the operator* $U^2(\cdot)$)**.**

*If* $\mathbb{E}(\varepsilon|X) = 0$*, and* $\mathbb{V}(\varepsilon|X) = 1$*, a.s., then* $U^2(X) = \mathbb{E}\left\{(Y - m(X))^2|X\right\}$ *a.s.*

**Proof.** From equation (3.2) one easily obtains $(Y - m(X))^2 = U^2(X)\varepsilon^2$. Then, by taking the conditional expectation, with respect to $X$, at both sides we obtain:

$$\begin{aligned}
\mathbb{E}\{(Y - m(X))^2|X)\} &= \mathbb{E}\{(U^2(X)(\varepsilon)^2|X)\} \\
&= U^2(X)\underbrace{\mathbb{E}((\varepsilon)^2|X)}_{=1} \\
&= U^2(X).
\end{aligned}$$

■

**Remark 4.** *Observe that* $U^2(x) := \mathbb{V}(Y|X) = \mathbb{E}\{(Y - m(X))^2|X)\}$ *a.s., which will be called* **residual-based conditional variance**. *Notice that the residual-based variance*

*$U^2(X)$ can be seen as a regression function of the squared errors, obtained after fitting a regression model to the data, on the predictor $X$. Alternatively, and following simple calculation, one can also show that*

$$\begin{aligned} U^2(X) &= \mathbb{E}\{Y^2 - 2Y \times m(X) + m^2(X)|X)\} \\ &= \mathbb{E}(Y^2|X) + m^2(X) - 2m(X)\underbrace{\mathbb{E}(Y|X)}_{=m(X)} \\ &= \mathbb{E}(Y^2|X) - m^2(X). \end{aligned}$$

*From now on, we denote $\widetilde{U}^2(x) := \mathbb{E}(Y^2|X) - m^2(X)$, which is called a* **difference-based conditional variance**.

Let us now focus on the nonparametric estimation of $U^2(x)$ and $\widetilde{U}^2(x)$. For this let us consider $(X_t, Y_t)_{t=1,\ldots,n}$ n-copies of strictly stationary process $(X, Y) \in \mathcal{E} \times \mathbb{R}$. Suppose that the observations are generated according to the following nonlinear heteroscedastic functional regression model:

$$Y_t = m(X_t) + U(X_t)\varepsilon_t, \qquad t = 1, \ldots, n, \tag{3.3}$$

where $\mathbb{E}(\varepsilon_t|X_t) = 0$ and var$(\varepsilon_t|X_t) = 1$. Here, it is assumed that the errors $(\varepsilon_t)_{t=1,\ldots,n}$ are dependent on the predictor $(X_t)_{t=1,\ldots,n}$.

**(a) Nonparametric difference-based conditional variance estimator**

As discussed above, remember that the difference-based conditional variance of $Y$ given $X = x$ is defined as follows:

$$\begin{aligned} \widetilde{U}^2(x) &= \mathbb{E}(Y^2|X = x) - (\mathbb{E}(Y|X = x))^2 \\ &:= \widetilde{m}(x) - (m(x))^2. \end{aligned} \tag{3.4}$$

Therefore, a plug-in estimator of $\widetilde{U}^2(x)$ can be obtained by replacing $\widetilde{m}(x)$ and $m(x)$ by their nonparametric estimators. That is

$$\widetilde{m}_n^c(x) = \frac{\sum_{t=1}^{n} Y_t^2 K\Big(\frac{d(X_t - x)}{h_m}\Big)}{\sum_{t=1}^{n} K\Big(\frac{d(X_t - x)}{h_m}\Big)} \quad \text{and} \quad m_n^c(x) = \frac{\sum_{t=1}^{n} Y_t K\Big(\frac{d(X_t - x)}{h_m}\Big)}{\sum_{t=1}^{n} K\Big(\frac{d(X_t - x)}{h_m}\Big)}, \tag{3.5}$$

where $K$ is a kernel function and $h_m$ is a sequence of positive real numbers decreasing to zero as $n$ goes to infinity. It worth noting that one can consider different bandwidth and kernel for $\widetilde{m}_n$ and $m_n$. For simplicity reason, we consider here the same tuning parameters for both regression functions.

Consequently a nonparametric difference-based conditional variance estimator of $\widetilde{U}^2(x)$ is defined as follows:

$$\widetilde{U}_n^{2,c}(x) = \widetilde{m}_n^c(x) - (m_n^c(x))^2. \tag{3.6}$$

**Remark 5.** *Even though it is easy to calculate, the main drawback of the difference-based conditional variance estimator is that in practice it may lead to negative values of volatility which is absurd.*

**(b) Nonparametric residual-based conditional variance estimator**

To overcome the drawback of the difference-based conditional variance estimator stated in Remark 5, [15] suggested a nonparametric estimator of the so-called residual-based conditional variance which is defined, for any fixed $x \in \mathcal{E}$ as

$$U^2(x) = \mathbb{E}\Big((Y - m(X))^2 | X = x\Big). \tag{3.7}$$

A Kernel-type estimator of $U^2(x)$ can be obtained following the algorithm described below:

**Step 1:** Estimate the regression operator at the sample curves $X_1, \ldots, X_n$ using the formula of $m_n^c(x)$ in (3.5).

**Step 2:** Compute the squared residuals $r_t := (Y_t - m_n^c(X_t))^2$ for all $t \in \{1, \ldots, n\}$.

**Step 3:** Use the sample $(X_1, r_1), \ldots, (X_n, r_n)$ to estimate $U^2(x)$ at any fixed $x \in \mathcal{E}$ as follows:

$$U_n^{2,c}(x) = \frac{\sum_{t=1}^{n} r_t K\Big(\frac{d(X_t - x)}{h_u}\Big)}{\sum_{t=1}^{n} K\Big(\frac{d(X_t - x)}{h_u}\Big)}, \tag{3.8}$$

where $h_u$ is a sequence of positive real numbers decreasing to zero as $n \to \infty$.

Note that, for simplicity reason, we consider the same kernel used to estimate the regression function. In practice two different kernels might be considered.

**Remark 6.** *In contrast to the difference-based conditional variance estimator defined in* (3.6)*, the residual-based estimator always takes positive values since it represents the regression of the squared residuals* $(r_t)_{t=1,\ldots,n}$ *on the predictor* $(X_t)_{t=1,\ldots,n}$.

**Theorem 2.** *(Uniform consistency, see [22])*[1] *Under some regularity conditions, we have, for* $\alpha, \beta > 0$,

$$\sup_{x \in \mathcal{C}} \Big| U_n^{2,c}(x) - U^2(x) \Big| = \mathcal{O}_{a.s.}(h_m^{2\beta} + h_u^{\alpha}) + \mathcal{O}_{a.s.}(\lambda_n' + \lambda_n^2),$$

*where* $(\lambda_n')_n$ *and* $(\lambda_n)_n$ *are two sequences of positive numbers tending to zero as* $n \to \infty$.

[1]Denote by $\mathcal{O}_{a.s.}(u)$ a real random function $g$ such that $g(u)/u$ is almost surely bounded. Here $\mathcal{C}$ denotes a "compact set" of the functional space $\mathcal{E}$.

The uniform consistency result, displayed in Theorem 2, is fundamental to establish the convergence of the predictor, at a horizon $n+H$, of the volatility operator at any out-of-sample curve $X_{n+H}$ say $U_n(X_{n+H})$.

**Theorem 3.** *(Asymptotic distribution, see [22]) Under some regularity conditions, we have,*

$$\sqrt{\alpha_n}\left(U_n^{2,c}(x) - U^2(x)\right) \rightsquigarrow N(0, \sigma^2(x)),$$

*where* $\rightsquigarrow$ *denotes convergence in distribution,* $\sigma^2(x) = M_2 U^2(x)\omega(x)/M_1^2$, $M_j = K^j(1) - \int_0^1 (K^j)'(u)\tau_0(u)du$ *for* $j \in \{1,2\}$, $\omega(x) := \mathbb{E}\left((\varepsilon_1^2 - 1)^2 \,|X = x\right)$ *and* $\alpha_n$ *is the convergence rate (for more details see [22]).*

Theorem 3 states the asymptotic distribution of the residual-based estimator of the conditional variance. This result plays an important role in building asymptotic confidence intervals for $U^2(x)$. It is worth noting here that the asymptotic variance depends on several unknown quantities such as $\omega(x)$ and the conditional variance $U^2(x)$. By replacing the unknown quantities by their empirical version, [22] established asymptotic confidence intervals for $U^2(x)$.

## Case of Missing At Random data

The principle objective of this section is to update the local constant estimator to incomplete data case. An extension of the work in [11] who applied Nadaraya-Watson estimator in complete data to missing at random mechanism is investigated. With the objective of estimating the regression and variance functions, we take the heteroscedastic regression model given in (1.1), where the error terms $\varepsilon_t$ depend on $X_t$. In our case, the response variable $Y_t$ is not completely observed and subject to MAR at

any discrete time $t$, but the predictor variable $X_t$ is totally observed. There are mainly two strategies that can be followed in the context of missing data. The first one only uses observes complete data, resulting simplified estimation. The second one is based on simple imputation techniques, resulting imputed estimation, which consists in applying the simplified estimator to estimate the incomplete observations of the response variable and then applying the estimator for complete data to complete sample. To complete the missing data, two imputation techniques will be applied including nonparametric regression imputation estimation and inverse probability weighted imputation. In order to check whether an observation is complete or missing, a new variable $\delta$ is introduced into the model as an indicator of the missing observations, where $\delta$ is assumed to be a Bernoulli random variable. Thus $\delta_t = 1$ if $Y_t$ is observed, and zero if $Y_t$ is missing, for any $0 \leq t \leq n$. We suppose that the Bernoulli random variable $\delta$ satisfies the following assumption:

**(H0)** $\mathbb{P}(\delta = 1|X, Y) = \mathbb{P}(\delta = 1|X) = \pi(X)$,

where $\pi(X)$ is called the conditional probability of observing $Y$ conditionally on $X$ and is often unknown. Assumption **(H0)** allows to conclude that $\delta$ and $Y$ are conditionally independent given $X$.

Motivated by [23] work on statistical methods (such as nonparametric regression and inverse weight probability) for handling missing data, let us recall the heteroscedastic regression model defined as:

$$Y = m(X) + U(X)\varepsilon. \tag{3.9}$$

Let us now focus on the identification and estimation of the parameters $m(\cdot)$ and $U(\cdot)$

under the MAR assumption.

*Simplified Estimator*

**Proposition 3** (*Identification of the operator* $m(\cdot)$ *under MAR assumption)***.**

*Assume that the data generating model is given by* (3.9) *and suppose that* **(H0)** *holds true. Then one gets, for any fixed* $x \in \mathcal{E}$*,*

$$m(x) = \frac{\mathbb{E}(\delta Y|X=x)}{\mathbb{E}(\delta|X=x)}.$$

**Proof.** Premultiplying equation (3.9) by $\delta$ and taking the conditional expectation, with respect to $X = x$, at both sides we obtain:

$$\begin{aligned}
\mathbb{E}(\delta Y|X=x) &= \mathbb{E}(\delta m(X) + \delta U(X)\varepsilon|X=x) \\
&= \mathbb{E}(\delta m(X)|X=x) + \mathbb{E}(\delta U(X)\varepsilon|X=x) \\
&= m(x)\mathbb{E}(\delta|X=x) + U(x)\mathbb{E}(\delta\varepsilon|X=x) \\
&= m(x)\mathbb{E}(\delta|X=x) + U(x)\mathbb{E}(\delta|X=x)\mathbb{E}(\varepsilon|X=x) \quad \text{(by } \textbf{(H0)}\text{)} \\
&= m(x)\mathbb{E}(\delta|X=x) + U(x)\pi(x)\underbrace{\mathbb{E}(\varepsilon|X=x)}_{=0}.
\end{aligned}$$

Finally, we have

$$m(x) = \frac{\mathbb{E}(\delta Y|X=x)}{\mathbb{E}(\delta|X=x)}.$$

■

**Proposition 4.** *[Identification of the operator* $U^2(\cdot)$ *under MAR assumption]*

*Under model* (3.9) *and assuming that* **(H0)** *is satisfied, then one gets,*

$$U^2(x) = \frac{\mathbb{E}(\delta(Y - m(X))^2 | X = x)}{\mathbb{E}(\delta | X = x)}, \qquad \textit{for any } x \in \mathcal{E}.$$

**Proof.** Notice that making use of (3.9) one easily obtains $\delta(Y - m(X))^2 = \delta U^2(X)\varepsilon^2$. Then, by taking the conditional expectation, with respect to $X = x$, at both sides one gets:

$$\begin{aligned}
\mathbb{E}\{\delta(Y - m(X))^2 | X = x\} &= \mathbb{E}\{\delta U^2(X)\varepsilon^2 | X = x\} \\
&= U^2(x)\mathbb{E}(\delta | X = x)\underbrace{\mathbb{E}(\varepsilon^2 | X = x)}_{=1} \quad \text{(by } \textbf{(H0)}\text{)} \\
&= U^2(x)\mathbb{E}(\delta | X = x).
\end{aligned}$$

Finally, we have

$$U^2(x) = \frac{\mathbb{E}(\delta(Y - m(X))^2 | X = x)}{\mathbb{E}(\delta | X = x)}.$$

■

Let us consider $(X_t, Y_t, \delta_t)_{t=1,\cdots,n}$ be $n$-copies of $(X, Y, \delta)$ not necessarily i.i.d. In the following we discuss the estimation of the operator $m(\cdot)$, the difference based variance $\widetilde{U}^2(\cdot)$ as well as the residual-based variance $U^2(\cdot)$.

**(a) Simplified Residual-Based Estimator**

Given the shape of the parameter $U^2(x)$ obtained in Proposition 4, we replace each conditional expectation by its nonparametric estimator to obtain a simplified

residual-based estimator of the conditional variance, say $U_n^{2,s}(x)$, defined as:

$$U_n^{2,s}(x) = \frac{\sum_{t=1}^{n} \delta_t r_t^s K\left(\frac{d(X_t - x)}{h_u}\right)}{\sum_{t=1}^{n} \delta_t K\left(\frac{d(X_t - x)}{h_u}\right)}, \qquad \text{for any } t = 1, \cdots, n \tag{3.10}$$

where $r_t^s = (Y_t - m_n^s(X_t))^2$ are the squared residuals obtained after fitting a regression model and

$$m_n^s(x) = \frac{\sum_{t=1}^{n} \delta_t Y_t K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} \delta_t K\left(\frac{d(X_t - x)}{h_m}\right)}. \tag{3.11}$$

**(b) Simplified Difference-Based Estimator**

Let us first recall that the difference-based variance is defined as

$$\widetilde{U}^2(x) = \widetilde{m}(x) - m(x).$$

Therefore, a plug-in estimator of $\widetilde{U}^2(x)$, at a fixed point $x \in \mathcal{E}$, is obtained by replacing $\widetilde{m}(x)$ and $m(x)$ by their nonparametric estimator adapted to the MAR framework. That is:

$$\widetilde{m}_n^s(x) := \frac{\sum_{t=1}^{n} \delta_t Y_t^2 K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} \delta_t K\left(\frac{d(X_t - x)}{h_m}\right)}, \tag{3.12}$$

and $m_n^s(x)$ is as defined in (3.11). Therefore, a simplified difference-based estimator of

the conditional variance is defined as follows:

$$\widetilde{U}_n^{2,s}(x) = \widetilde{m}_n^s(x) - (m_n^s(x))^2. \tag{3.13}$$

*Inverse Probability Weighted Estimator*

The IPW estimator is based on the complete cases but now weighting them with the inverse of the probability that a case is observed as introduced in [39] and [40].

In this way cases with low probability to be observed gain more influence in the analysis and thus represent the probable missing values in the neighborhood. One can look at this approach as an implicit imputation of missing values.

**Proposition 5** (*Identification of the operator* $m(\cdot)$ *under MAR assumption)***.**

*Assuming* **(H0)** *holds true, then one gets, for any fixed* $x \in \mathcal{E}$*,*

$$m(x) = \frac{\mathbb{E}\left\{ \dfrac{\delta}{\pi(X)} Y | X = x \right\}}{\mathbb{E}\left\{ \dfrac{\delta}{\pi(X)} | X = x \right\}}.$$

**Proof.**

The regression operator $m(\cdot)$ can be seen as the minimizer of the loss function $\mathbb{E}\left( (Y - m(X))^2 \, | X \right)$. We can show as detailed below that, almost surely,

$$\mathbb{E}\left\{ \frac{\delta}{\pi(X)} \left(Y - m(X)\right)^2 | X \right\} = \mathbb{E}\big\{ \left(Y - m(X)\right)^2 | X \big\}. \tag{3.14}$$

Indeed, using a double-conditioning, we obtain:

$$\begin{aligned}\mathbb{E}\Big\{\frac{\delta}{\pi(X)}(Y-m(X))^2|X\Big\} &= \mathbb{E}\Big\{\mathbb{E}\Big(\frac{\delta}{\pi(X)}(Y-m(X))^2|X,Y\Big)|X\Big\}\\ &= \mathbb{E}\Big\{\frac{(Y-m(X))^2}{\pi(X)}\underbrace{\mathbb{E}(\delta|X,Y))}_{\pi(X)}|X\}\\ &= \mathbb{E}\{(Y-m(X))^2|X\}.\end{aligned}$$

By taking the first derivative, with respect to $m$, at the left hand side of equation (3.14), and taking $X=x$, the regression function is a zero of the following equation:

$$\mathbb{E}\Big\{\frac{\delta}{\pi(X)}(-2)(Y-m(X))|X=x\Big\}=0.$$

Therefore, a simple calculation allows to find that

$$\mathbb{E}\Big\{\frac{\delta}{\pi(X)}Y|X=x\Big\}=m(x)\mathbb{E}\Big\{\frac{\delta}{\pi(X)}|X=x\Big\}.$$

Finally, one gets

$$m(x)=\frac{\mathbb{E}\Big\{\frac{\delta}{\pi(X)}Y|X=x\Big\}}{\mathbb{E}\Big\{\frac{\delta}{\pi(X)}|X=x\Big\}}.$$

■

**Proposition 6.** *[Identification of the operator $U(\cdot)$ under MAR assumption]*

*Assuming* **(H0)** *holds true, then one gets, for any fixed $x\in\mathcal{E}$,*

$$U^2(x)=\frac{\mathbb{E}\Big\{\frac{\delta}{\pi(X)}(Y-m(X))^2|X=x\Big\}}{\mathbb{E}\Big\{\frac{\delta}{\pi(X)}|X=x\Big\}}.$$

**Proof.**

The conditional variance operator $U^2(\cdot)$ can be seen as the minimizer of the loss function $\mathbb{E}\{((Y - m(X))^2 - U^2(X))^2|X)\}$.

We can show as detailed below that, almost surely,

$$\mathbb{E}\left\{ \frac{\delta}{\pi(X)} \left((Y - m(X))^2 - U^2(X)\right)^2 | X \right\} = \mathbb{E}\left\{ \left((Y - m(X))^2 - U^2(X)\right)^2 | X \right\}. \tag{3.15}$$

Indeed, using double-conditioning, we obtain:

$$\begin{aligned}
\mathbb{E}\left\{ \frac{\delta}{\pi(X)}((Y - m(X))^2 - U^2(X))^2 | X \right\} &= \mathbb{E}\left\{ \mathbb{E}\left( \frac{\delta}{\pi(X)}((Y - m(X))^2 - U^2(X))^2 | X, Y \right) | X \right\} \\
&= \mathbb{E}\left\{ \frac{((Y - m(X))^2 - U^2(X))^2}{\pi(X)} \underbrace{\mathbb{E}(\delta | X, Y))}_{\pi(X)} | X \right\} \\
&= \mathbb{E}\{((Y - m(X))^2 - U^2(X))^2 | X\}.
\end{aligned}$$

By taking the first derivative, with respect to $U^2$, at the left hand side of equation (3.15), and taking $X = x$, the conditional variance function is a zero of the following equation:

$$\mathbb{E}\left\{ \frac{\delta}{\pi(X)}(-2)((Y - m(X))^2 - U^2(X)) | X = x \right\} = 0.$$

Therefore, a simple calculation allows to find that

$$\mathbb{E}\left\{ \frac{\delta}{\pi(X)}(Y - m(X))^2 | X = x \right\} = U^2(x)\mathbb{E}\left\{ \frac{\delta}{\pi(X)} | X = x \right\}.$$

Finally, one gets

$$U^2(x) = \frac{\left\{ \mathbb{E}\frac{\delta}{\pi(X)}(Y - m(X))^2 | X = x \right\}}{\mathbb{E}\left\{ \frac{\delta}{\pi(X)} | X = x \right\}}.$$

■

Let us consider $(X_t, Y_t, \delta_t)_{t=1,\cdots,n}$ be $n$-copies not necessarily i.i.d. In the following we discuss the estimation of the operator $m(\cdot)$, the difference based variance $\widetilde{U}^2(\cdot)$ as well as the residual-based variance $U^2(\cdot)$

**(a) IPW Residual-Based Estimator**

Given the shape of the parameter $U^2(x)$ obtained in Proposition 6, we replace each conditional expectation by its nonparametric estimator to obtain an inverse probability weighted residual-based estimator of the conditional variance, say $U_n^{2,IPW}(x)$, defined as:

$$U_n^{2,IPW}(x) = \frac{\sum_{t=1}^{n} \frac{\delta_t}{\pi_n(X_t)} r_t^{IPW} K\left(\frac{d(X_t - x)}{h_u}\right)}{\sum_{t=1}^{n} \frac{\delta_t}{\pi_n(X_t)} K\left(\frac{d(X_t - x)}{h_u}\right)}, \qquad \text{for any } t = 1, \cdots, n \quad (3.16)$$

Where $r_t^{IPW} = (Y_t - m_n^{IPW}(X_t))^2$ are the residuals squared obtained after fitting regression model and

$$m_n^{IPW}(x) = \frac{\sum_{t=1}^{n} \frac{\delta_t}{\pi_n(X_t)} Y_t K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} \frac{\delta_t}{\pi_n(X_t)} K\left(\frac{d(X_t - x)}{h_m}\right)}. \quad (3.17)$$

The unknown probability function $\pi(X)$ is estimated nonparametrically as follows:

$$\pi_n(x) = \frac{\sum_{t=1}^{n} \delta_t K\left(\frac{d(X_t - x)}{h_\pi}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_\pi}\right)}. \quad (3.18)$$

**(b) IPW Difference-Based Estimator**

Let us first recall the difference-based variance is

$$\widetilde{U}^2(x) = \widetilde{m}(x) - m(x).$$

Therefore, a plug-in estimator of $\widetilde{U}^2(x)$, at a fixed point $x \in \mathcal{E}$, is obtained by replacing $\widetilde{m}(x)$ and $m(x)$ by their nonparametric estimator adapted to the MAR framework. That is

$$\widetilde{m}_n^{IPW}(x) = \frac{\sum_{t=1}^{n} \frac{\delta_t}{\pi_n(X_t)} Y_t^2 K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} \frac{\delta_t}{\pi_n(X_t)} K\left(\frac{d(X_t - x)}{h_m}\right)}, \tag{3.19}$$

$m_n^{IPW}(x)$ and $\pi_n(x)$ as defined in (3.17) and (3.18), respectively. Therefore, an inverse probability weighted difference-based estimator of the conditional variance is defined as follows:

$$\widetilde{U}_n^{2,IPW}(x) = \widehat{m}^{IPW}(x) - (m_n^{IPW}(x))^2. \tag{3.20}$$

**Data-driven smoothing parameters selection**

Cross-validation approach in choosing the smoothing parameters of the above estimators is used. As a result, we select the smoothing parameters as shows below.

$$h_m^{opt,\star} = \arg\min_h \sum_{t=1}^{n} \frac{\delta_t}{P_n(X_t)} \{Y_t - m_{n,-t}^{\star}(X_t; h)\}^2 \tag{3.21}$$

and

$$h_u^{opt,\star} = \arg\min_h \sum_{t=1}^{n} \frac{\delta_t}{P_n(X_t)} \{r_t - U_{n,-t}^{2,\star}(X_t; h)\}^2, \tag{3.22}$$

where $m^{\star}_{n,-t}(\cdot)$ (resp. $U^{2,\star}_{n,-t}(\cdot)$) is a "leave-one-out" version of $m^{\star}_{n}(\cdot)$ (resp. $U^{2,\star}_{n}(\cdot)$), that is

$$m^{\star}_{n,-t}(X_t) = \frac{\sum\limits_{i=1,i\neq t}^{n} \frac{\delta_i}{P_n(X_t)} Y_i K\left(\frac{d(X_i - x)}{h}\right)}{\sum\limits_{i=1,i\neq t}^{n} \frac{\delta_i}{P_n(X_t)} K\left(\frac{d(X_i - x)}{h}\right)}.$$

$$U^{2,\star}_{n,-t}(X_t) = \frac{\sum\limits_{i=1,i\neq t}^{n} \frac{\delta_i}{P_n(X_t)} r_i K\left(\frac{d(X_i - x)}{h}\right)}{\sum\limits_{i=1,i\neq t}^{n} \frac{\delta_i}{P_n(X_t)} K\left(\frac{d(X_i - x)}{h}\right)}.$$

$\star \in \{s, IPW\}$ and $P_n(\cdot) = 1$ if $\star = s$ and equals $\pi_n(\cdot)$ if $\star = IPW$.

The optimal bandwidth for $\pi_n$ is also obtained cross-validation approach, that is

$$h^{opt}_{\pi} = \arg\min_{h} \sum_{t=1}^{n} \{\delta_t - \pi_{n,-t}(X_t; h)\}^2, \tag{3.23}$$

where

$$\pi_{n,-t}(X_t) = \frac{\sum\limits_{i=1,i\neq t}^{n} \delta_i K\left(\frac{d(X_i - x)}{h}\right)}{\sum\limits_{i=1,i\neq t}^{n} K\left(\frac{d(X_i - x)}{h}\right)}.$$

**Remark 7.** *Despite the lack of literature on functional data analysis framework when dealing with incomplete data, special attention was devoted to investigate the statistical properties of the regression estimator for missing data. Recently, [25] considered the inverse probability weighted estimation for the functional regression operator based on i.i.d functional sample in which the response is missing at random, where the asymptotic properties of the proposed estimator were obtained under the mild conditions. The obtained results are as follows:*

**Theorem 4.** *(See [25]) Under regularity conditions, If* $h^{\beta}(n\phi(h))^{1/2} \to 0$ *as* $n \to \infty$*, then*

$$\sqrt{\alpha_n}\,(m_n(x) - m(x)) \rightsquigarrow N(0, \sigma^2(x)),$$

*where* $\rightsquigarrow$ *denotes convergence in distribution,* $\sigma^2(x) = \dfrac{M_2 U^2(x)}{M_1^2 \pi(x) f(x)}$*, the constants* $(M_j)_{j=1,2}$ *are defined in Theorem 3,* $f(x)$ *is a nonnegative bounded function and* $\alpha_n$ *is the convergence rate (see [25] for more details).*

CHAPTER 4: VOLATILITY ESTIMATION WITH IMPUTED DATA

In this chapter, we focus on the estimation of the regression and volatility operators, when the response variable $Y$ is a real valued random variable and the predictor $X$ is infinite-dimensional random variable. Our purpose is to correct the simplified and the inverse probability estimators by filling in the incomplete data using the imputation techniques. Therefore, general classes of imputed estimators of the regression and volatility functions are considered.

**A class of imputed volatility estimators**

A general class of imputed estimator for the regression function is defined as follows:

$$m_n(x) = \frac{\sum_{t=1}^{n} \left\{ \frac{\delta_t}{P_n(X_t)} Y_t + \left(1 - \frac{\delta_t}{P_n(X_t)}\right) m_n^{\star}(X_t) \right\} K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_m}\right)}. \tag{4.1}$$

In contrast, a general class of imputed estimator for the volatility function is defined as follows:

$$U_n^2(x) = \frac{\sum_{t=1}^{n} \left\{ \frac{\delta_t}{P_n(X_t)} r_t^{\star} + \left(1 - \frac{\delta_t}{P_n(X_t)}\right) U_n^{2,\star}(X_t) \right\} K\left(\frac{d(X_t - x)}{h_u}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_u}\right)}, \tag{4.2}$$

where $P_n(x)$ is some sequence of quantities with probability limits $P(x)$, $r_t^{\star} = (Y_t - m_n^{\star}(X_t))^2$ for any $t \in \{1, \cdots, n\}$ and $\star \in \{s, IPW\}$.

We have a particular interest in some special cases are described as follows:

**Case 1:** *Nonparametric Imputed Estimator*

When $P_n(x) = 1$, then $\star = s$, therefore, a nonparametric (NP) imputed regression and volatility estimators of $m(x)$ and $U^2(x)$ are obtained.

**(a) NP imputed Residual-Based Estimator**

$$U_n^{2,NPI}(x) = \frac{\sum_{t=1}^{n} r_t^{NPI} K\left(\frac{d(X_t - x)}{h_u}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_u}\right)} \tag{4.3}$$

and

$$m_n^{NPI}(x) = \frac{\sum_{t=1}^{n} Y_t^{NPI} K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_m}\right)}, \tag{4.4}$$

where $Y_t^{NPI} = \delta_t Y_t + (1 - \delta_t) m_n^s(X_t)$ and $r_t^{NPI} = \delta_t r_t^s + (1 - \delta_t) U^{2,s}(X_t)$.

**(b) NP imputed Difference-Based Estimator**

$$\widetilde{U}_n^{2,NPI}(x) = \widetilde{m}_n^{NPI}(x) - (m_n^{NPI}(x))^2, \tag{4.5}$$

where

$$\widetilde{m}_n^{NPI}(x) = \frac{\sum_{t=1}^{n} Y_t^{2,NPI} K\left(\frac{d(X_t - x)}{h_m}\right)}{K\left(\frac{d(X_t - x)}{h_m}\right)}, \tag{4.6}$$

$m_n^{NPI}(x)$ is as defined in (4.4) and $Y_t^{NPI} = \delta_t Y_t + (1 - \delta_t) m_n^s(X_t)$.

**Case 2:** *IPW Imputed Estimator*

When $P_n(x) = \pi_n(x)$, then $\star = IPW$, therefore, inverse probability weight (IPW) imputed regression and volatility estimators of $m(x)$ and $U^2(x)$ are obtained.

**(a) IPW imputed residual-based estimator**

$$U_n^{2,IPWI}(x) = \frac{\sum_{t=1}^{n} r_t^{IPWI} K\left(\frac{d(X_t - x)}{h_u}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_u}\right)} \tag{4.7}$$

and

$$m_n^{IPWI}(x) = \frac{\sum_{t=1}^{n} Y_t^{IPWI} K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_m}\right)}, \tag{4.8}$$

where $Y_t^{IPWI} = \frac{\delta_t}{\pi_n(X_t)} Y_t + \left(1 - \frac{\delta_t}{\pi_n(X_t)}\right) m_n^{IPW}(X_t)$ and
$r_t^{IPWI} = \frac{\delta_t}{\pi_n(X_t)} r_t^{IPW} + \left(1 - \frac{\delta_t}{\pi_n(X_t)}\right) U_n^{2,IPW}(X_t)$.

**(b) IPW imputed difference-based estimator**

$$\widetilde{U}_n^{2,IPWI}(x) = \widetilde{m}_n^{IPWI}(x) - (m_n^{IPWI}(x))^2, \tag{4.9}$$

where

$$\widetilde{m}_n^{IPWI}(x) = \frac{\sum_{t=1}^{n} Y_t^{2,IPWI} K\left(\frac{d(X_t - x)}{h_m}\right)}{\sum_{t=1}^{n} K\left(\frac{d(X_t - x)}{h_m}\right)}, \tag{4.10}$$

$m_n^{IPWI}(x)$ is as defined in (4.8) and $Y_t^{IPWI} = \frac{\delta_t}{\pi_n(X_t)} Y_t + \left(1 - \frac{\delta_t}{\pi_n(X_t)}\right) m_n^{IPW}(X_t)$.

**Smoothing parameters selection**

Cross-validation approach in choosing the smoothing parameters of the above estimators is used. As a result, we select the smoothing parameters as shows below.

$$h_m^{opt,\bullet} = \arg\min_h \sum_{t=1}^{n} \{Y_t^\bullet - m_{n,-t}^\bullet(X_t)\}^2 \tag{4.11}$$

and

$$h_u^{opt,\bullet} = \arg\min_h \sum_{t=1}^{n} \{r_t^\bullet - U_{n,-t}^{2,\bullet}(X_t)\}^2, \tag{4.12}$$

where $\bullet = \{c, NPI, IPWI\}$. Notice that when the data are completely observed, then $Y_t^\bullet = Y_t$ and $r_t^\bullet = (Y_t - m_n^c(X_t))^2$.

## CHAPTER 5: NUMERICAL ANALYSIS THROUGH SIMULATED DATA

In this section, we carry out simulation study to assess the quality of the proposed estimation methods. Let us consider $(X_t, Y_t, \delta_t)_{t=1,\cdots,n}$ be a strict stationary process valued in $\mathcal{E} \times \mathbb{R} \times \{0, 1\}$. The functional covariate variables $X_1(\lambda), \cdots, X_t(\lambda)$, where $t$ takes $100$ equally spaced values in $[-1, 1]$, are generated by

$$X_t(\lambda) = A(2 - \cos(\pi\lambda\omega)) + (1 - A)\cos(\pi\lambda\omega),$$

where $\omega \sim N(0, 1)$, $A \sim$ Bernoulli$\left(\frac{1}{2}\right)$ and $\lambda \in [-1, 1]$. A sample of $100$ simulated curves is displayed in Figure 5.1.

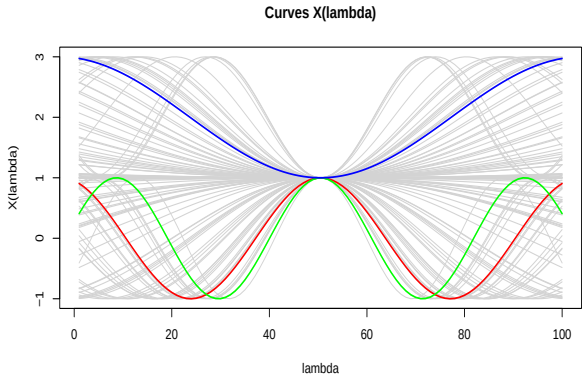



Figure 5.1. A sample of simulated curves $X_t(\lambda)$.

To generate the response variable observation, we consider the following heteroscedastic regression model:

$$Y_t = m(X_t) + U(X_t)\varepsilon_t,$$

where

$$m(x) = \int_{-1}^{1} \lambda x(\lambda) d\lambda, \quad U(x) = \int_{-1}^{1} |\lambda| x^2 d\lambda \tag{5.1}$$

Regarding the errors, four models of generation are considered for $\varepsilon_t$:

**Model 1:** The $\varepsilon_t$'s are i.i.d, distributed according to $N(0,1)$.

**Model 2:** $\varepsilon_t = \frac{1}{2}\varepsilon_{t-1} + \xi_t$, where $\xi_t \sim N(0,1)$.

**Model 3:** $\varepsilon_t = -\frac{1}{2}\varepsilon_{t-1} + \xi_t$, where $\xi_t \sim N(0,1)$.

**Model 4:** $\varepsilon_t = \frac{1}{2}\varepsilon_{t-1} + \xi_t$, where $\xi_t \sim \texttt{Bernoulli}\left(\frac{1}{2}\right)$.

Figure 5.2 shows the generated response variable for each of the four scenarios.

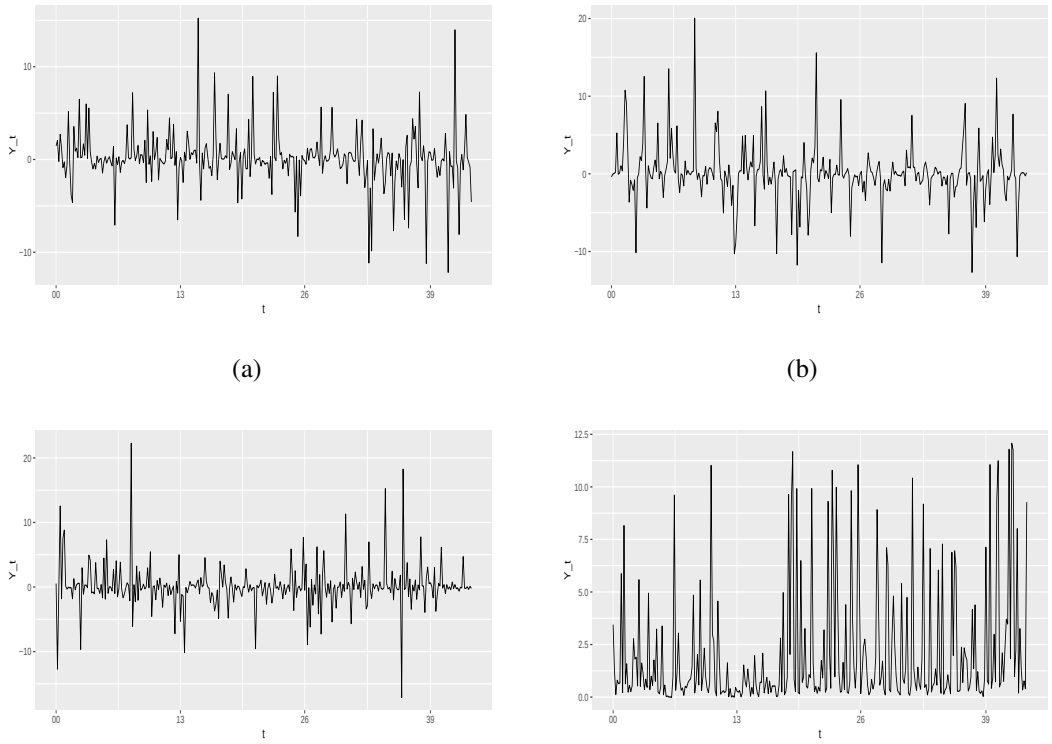


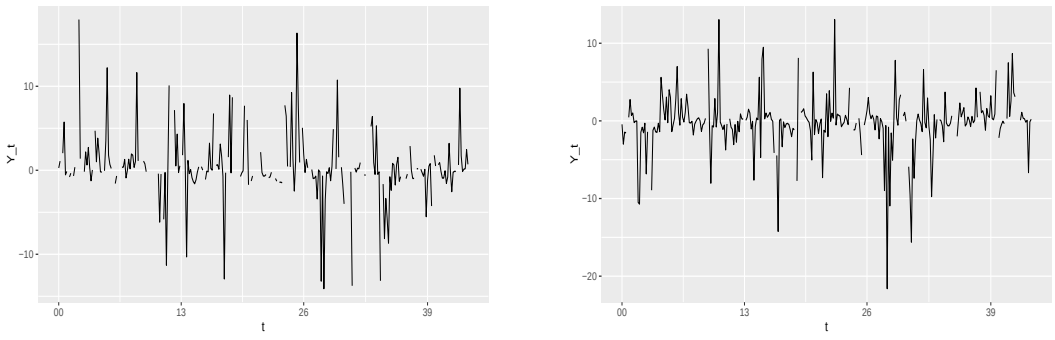

Figure 5.2. The generated process $Y_t$ for Model 1 (a), Model 2 (b), Model 3 (c) and Model 4 (d).

We suppose that missing at random observations in the response variable $Y$ are generated according to the following probability:

$$\pi(x) = \mathbb{P}(\delta = 1 | X = x) = \text{expit}\left(2\alpha \int_{-1}^{1} x^2(\lambda) d\lambda\right)$$

where expit$(u) = \dfrac{e^u}{1+e^u}$ and $\alpha \in \{0.2, 0.8\}$.

**Remark 8.** *Observe that according to the value of $\alpha$ one gets different MAR rate. Higher is the value of $\alpha$, higher will be $\pi(x)$. Therefore, smaller will be the missing data rate for $Y$. Indeed, when $\alpha = 0.8$ the MAR rate will be $20\%$ and $60\%$ for $\alpha = 0.2$ . Figure 5.3 shows an example of the process $Y_t$ that is affected by MAR mechanism for alpha values of $0.2$ and $0.8$, respectively.*




(a) (b)

Figure 5.3. Missing at random in the generated process $Y_t$ when $\alpha = 0.2$ (a) and $0.8$ (b).

In this simulation, we consider the four models mentioned above using the regression and the variance operators as defined in (5.1). Since the choice of the kernel is not determinant factor of the quality of estimation, then the Gaussian kernel is used as defined in Chapter 2. The choice of the bandwidth is based on the cross-validation criterion. Based on the smoothness of the curves $X_t(\lambda)$, we consider a semi-metric for the regression and the conditional variance functions estimation the usual $L_2$-norm of the first derivatives of the curves, is defined as follows:

$$d(X_t, X_s) = \left[ \int_{-1}^{1} \left\{ X_t^{(1)}(\lambda) - X_s^{(1)}(\lambda) \right\}^2 d\lambda \right]^{1/2}, \quad \forall t \neq s.$$

Our purpose is to estimation the conditional variance at a fixed curve

$$x_0(\lambda) = \cos(\pi\lambda/4) \qquad \text{for} \lambda \in [-1, 1].$$

To assess the consistency of the estimation, we generated $B = 500$ samples and from each sample we estimate the conditional variance and evaluate the square error. That is, at iteration $b$ we have

$$\text{SE}_b = \left(\mathcal{U}^2_{n,b}(x_0) - U^2(x_0)\right)^2,$$

where $\mathcal{U}^2_{n,b}(x_0)$ represents either the complete, simplified, IPW, NP imputation or IPW imputation estimator.

Tables 5.1 and 5.2 display some summary statistics of the square errors obtained for each model with missing at random rate equals $60\%$ and $20\%$, respectively. We can see that the estimators obtained after missing data imputation provide better results. Moreover, one can observe that the higher the MAR rate, the lower the quality of estimation. Finally, one can notice that the dependence structure in the data plays an important role in the quality of estimation. Indeed, small errors are obtained in model 1 (corresponding to the i.i.d case), then higher is the dependence structure higher will be the estimation errors.

Table 5.1. Quartiles of the SE obtained for each model when MAR = 60%.

| Model | Comp($\times 10^{-2}$) | | | Simp.($\times 10^{-2}$) | | | NPI.($\times 10^{-2}$) | | | IPW.($\times 10^{-2}$) | | | IPWI.($\times 10^{-2}$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ |
| Model1 | 1.25 | 4.03 | 21.97 | 1.41 | 4.62 | 24.73 | 1.25 | 4.29 | 21.67 | 1.40 | 4.59 | 23.65 | 1.29 | 4.53 | 22.38 |
| Model2 | 1.78 | 7.50 | 35.46 | 2.27 | 8.42 | 39.84 | 2.04 | 7.42 | 35.9 | 2.23 | 8.19 | 39.01 | 2.10 | 7.57 | 36.56 |
| Model3 | 2.10 | 5.07 | 39.10 | 2.63 | 6.93 | 44.81 | 2.12 | 6.08 | 39.56 | 2.59 | 6.92 | 43.01 | 2.30 | 6.81 | 40.49 |
| Model4 | 2.21 | 7.38 | 46.04 | 2.28 | 7.85 | 46.59 | 2.24 | 7.55 | 46.38 | 2.25 | 7.43 | 46.38 | 2.22 | 7.43 | 46.18 |

Table 5.2. Quartiles of the SE obtained for each model when MAR = 20%

| Model | Comp($\times 10^{-2}$) | | | Simp.($\times 10^{-2}$) | | | NPI.($\times 10^{-2}$) | | | IPW.($\times 10^{-2}$) | | | IPWI.($\times 10^{-2}$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ |
| Model1 | 0.4 | 3.57 | 11.48 | 0.58 | 5.06 | 14.06 | 0.51 | 4.54 | 11.02 | 0.58 | 5.09 | 14.36 | 0.51 | 4.94 | 12.66 |
| Model2 | 0.85 | 3.88 | 19.79 | 0.97 | 5.88 | 23.77 | 0.94 | 4.46 | 19.42 | 0.97 | 6.33 | 23.77 | 0.96 | 5.98 | 20.82 |
| Model3 | 0.91 | 4.00 | 22.32 | 0.97 | 8.24 | 26.87 | 0.99 | 5.56 | 20.92 | 0.94 | 8.37 | 26.61 | 0.89 | 7.9 | 21.54 |
| Model4 | 0.36 | 6.79 | 19.22 | 0.38 | 7.03 | 21.10 | 0.34 | 6.95 | 19.67 | 0.38 | 6.85 | 20.95 | 0.36 | 6.81 | 20.30 |

# CHAPTER 6: APPLICATION TO HIGH-FREQUENCY FINANCIAL DATA

In this chapter, we are interested in estimating and forecasting the volatility of the log returns of Brent crude oil closing price in US dollars per barrel given the natural gas closing price in US dollars per MMBtu.

The relationship between oil and natural gas prices has been a topic of interest to researchers and market practitioners for many years. In particular, the volatility of oil prices has been a major concern for investors and market participants. A number of studies have investigated the impact of natural gas prices on the volatility of oil prices, with the aim of developing models that can better capture this relationship.

[41] used a VAR model to examine the relationship between crude oil and natural gas prices in North America. Another study by [42] used a BEKK-GARCH model to investigate the volatility spillovers between crude oil and natural gas prices in the United States. Moreover, [43] used a copula-GARCH model to examine the relationship between Brent oil and natural gas prices. The authors found that the volatility of Brent oil prices was indeed affected by natural gas prices, and that their model outperformed other traditional GARCH models in terms of forecasting accuracy. The authors found that the relationship between the two prices varied depending on the market regime, and suggested that this could be important for risk management purposes.

Overall, these studies suggest that incorporating the relationship between natural gas prices and the volatility of Brent oil prices can lead to improved forecasting accuracy and risk management. However, the precise nature of this relationship may vary depending on the specific market conditions and modeling approach used.

*Data preliminary analysis*

The data covers the trading days from February 14, 2020 to February 14, 2023. Figure 6.1 displays a 1-day frequency time series of Brent crude oil and natural gas closing prices. One can see from Figure 6.1 that there is a correlation between the two prices.

We consider the Brent crude oil closing price, which observed at a daily frequency from February 14, 2020 to February 14, 2023, while the natural gas closing price is observed every minute over the same time period. The returns of Brent crude oil is calculated as follows: $r_t^o := \log\left(\frac{P_t^o}{P_{t-1}^o}\right)$, where $P_t^o$ is the daily closing price at day $t$ of the Brent crude oil. Similarly, the 1-minute frequency of return of natural gas is obtained according to the following formula: $r_m^g := \log\left(\frac{P_m^g}{P_{m-1}^g}\right)$, where $P_m^g$ is the price of natural gas at a minute $m$.

Table 6.1 contains the descriptive statistics as well as the results of the statistical tests on the price and returns for Brent and gas. The maximum natural gas and prices are six times relative to minimum values, while the maximum Brent crude oil and prices are six times relative to the minimum values. The average of the natural gas is slightly greater than the median, while the mean of Brent crude oil is less than the median. The observation of their skewness is consistent with this. The distribution of natural gas prices has a positive skewness, on the other hand, the skewness of Brent crude oil price is negative. The kurtosis of all price series is around $2$, indicating that the distribution natural gas and Brent crude oil prices are platykurtic with flat tails. The table also shows a test for normality, both the two prices and returns distributions are not expected to be normal at $5\%$ level, according to Jarque-Bera test. Furthermore, the Box-Pierce test for the presence of an autocorrelation up to $10$ lags reveals that all the returns squared of

daily frequency of natural gas and Brent crude oil have autocorrelation. These findings support the use of the heteroscedastic regression model to estimate and forecast the return series of natural gas and Brent crude oil. Last but not least, the ADF test shows that all the return series are stationary.

Table 6.1. Descriptive Statistics for Brent Crude Oil and Natural Gas Closing Price.
Note: The symbol*denotes the statistical significance at $5\%$ level; Jarque-Bera and Box-Pierece refer to the empirical statistics of the test for normality and autocorrelation, respectively.

| Descriptive Statistics | Price | | Return | |
|---|---|---|---|---|
| | Gas (Days) | Brent (Days) | Gas (Days) | Brent (Days) |
| Mean | $4.168$ | $71.79$ | $0$ | $0$ |
| Median | $3.673$ | $73.09$ | $0$ | $0$ |
| Std.dev | $2.095$ | $24.788$ | $0.038$ | $0.028$ |
| Minimum | $1.528$ | $19.61$ | $-0.177$ | $-0.308$ |
| Maximum | $9.757$ | $129.03$ | $0.431$ | $0.166$ |
| Skewness | $0.762$ | $-0.009$ | $1.064$ | $-1.989$ |
| Kurtosis | $2.543$ | $2.100$ | $18.516$ | $29.799$ |
| Jarque-Bera | $115.81^*$ | $36.989^*$ | $11201^*$ | $33521^*$ |
| Box-Pierce$(10)$ | $10428^*$ | $10640^*$ | $11.162$ | $32.12^*$ |
| Box-Pierce$^2(10)$ | $10197^*$ | $10484^*$ | $18.277^*$ | $112.51^*$ |
| ADF test | $-0.9524$ | $-2.003$ | $-10.798^*$ | $-9.5171^*$ |
| Observations | $1097$ | $1097$ | $1096$ | $1096$ |

Figure 6.2 shows that the most likely prices for the Brent oil and the natural gas do not in general exceed \$80 and \$5 respectively.

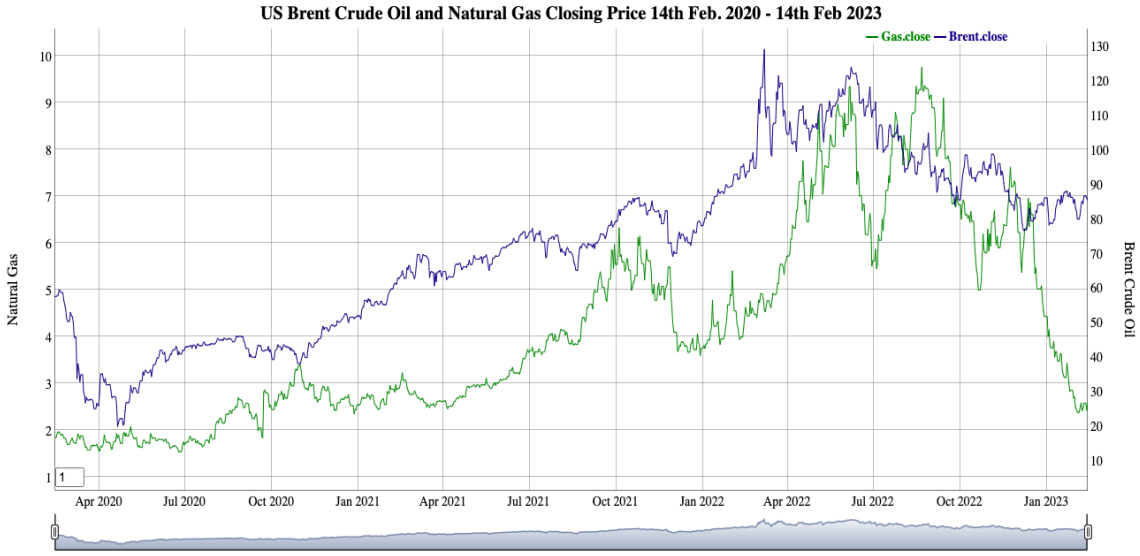



Figure 6.1. Closing Price for Brent Crude Oil and Natural Gas.

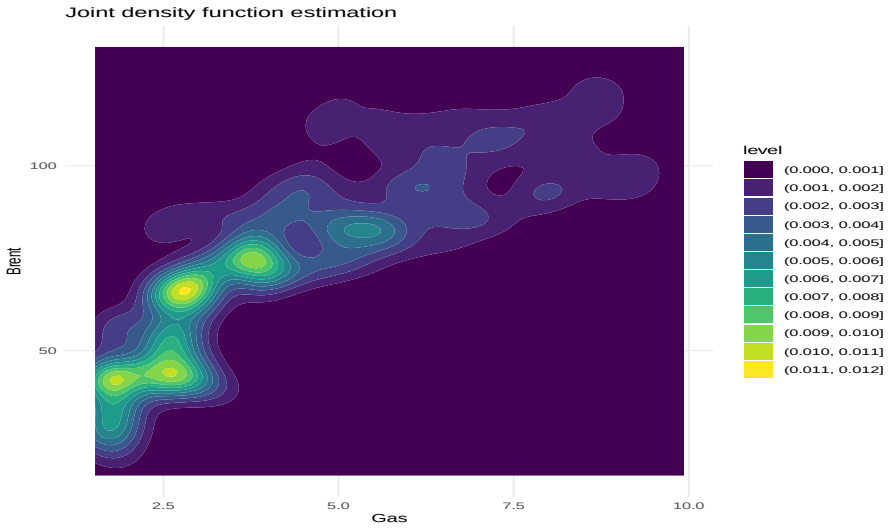



Figure 6.2. Joint density estimation of daily Brent oil and Natural Gas prices.

*The random sample construction*

Our sample here can be denoted as follows: $(X_t, Y_t)_{t=1,\cdots,1096}$, where the sample size $n = 1096$ is the total number of trading days from February 14, 2020 to February 14, 2023. Note that $Y_t = r_t^o$ and the functional-valued discrete-time process is defined

as follows:

$$X_t(m) = r_t^g\left(m + (t-1) \times 1439\right) \qquad t = 1, \ldots, n, \quad \forall m \in [0, 1439).$$

Figure 6.3(a) displays a sample of three 1-minute frequency curve of the natural gas and Figure 6.3(b) shows all intraday (1-minute frequency) curves from February 14, 2020 to February 14, 2023.

**Definition 2.** *Let $Z \in L^2$ space be a functional random variable with mean $\mu(t) = \mathbb{E}(Z(t))$, then the covariance function is defined as follows:*

$$\begin{aligned} \Sigma(t,s) &= cov(Z(t), Z(s)) \\ &= \mathbb{E}\left(\left(Z(t) - \mu(t)\right)\left(Z(s) - \mu(s)\right)\right). \end{aligned}$$

*Given a sample $Z_1(t), \ldots, Z_n(t)$, an estimator of $\Sigma(t,s)$ is defined as:*

$$\widehat{\Sigma}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i(t) - \widehat{\mu}(t)\right)\left(Z_i(s) - \widehat{\mu}(s)\right),$$

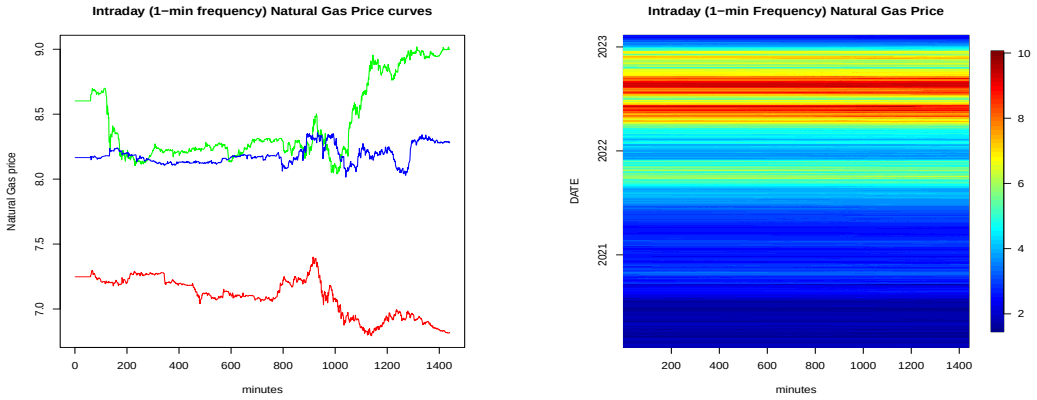
*where $\widehat{\mu}(t) = n^{-1}\sum_{i=1}^{n} Z_i(t)$ an estimator of mean function $\mu(t)$.*

Figure 6.4 displays the estimated covariance operator of the price of natural gas and shows that prices are highly correlated in the morning and until 03:30pm. One can also observe a high correlation in the evening around 08:00pm.

Figure 6.5(a) displays a sample of three 1-minute frequency curve of natural gas return and Figure 6.5(b) shows the stochastic process of the daily Brent Oil return, with the dots representing the three preselected days.

Observe that the data are initially completely observed. Therefore, we artificially generate missing observations in order to validate our methodology.

We assume that the missing at random mechanism is generated according to the




(a) (b)

Figure 6.3. (a) Sample of three intraday (1-minute frequency) Natural Gas price curves. (b) All historical intraday (1-minute frequency) Natural Gas price curves.

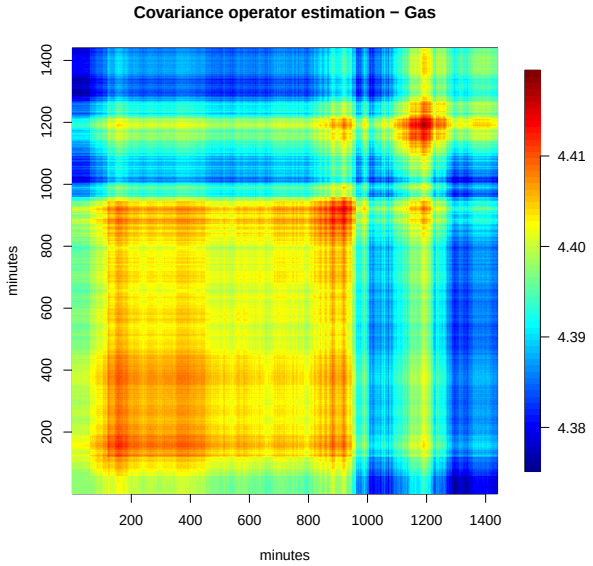


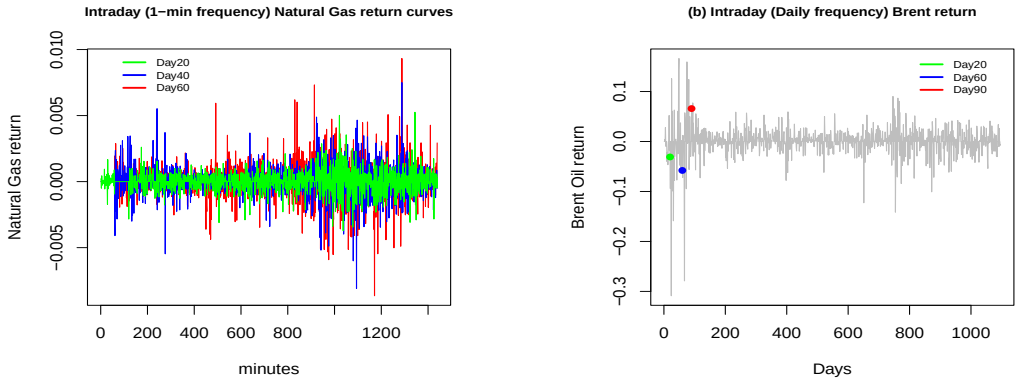

Figure 6.4. Estimated covariance operator of the intraday (1-minute frequency) Natural Gas prices.

following probability distribution:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \langle \alpha_0, X\rangle + c,$$

where $\alpha_0(t) = \sin(2\pi t), \quad \forall t \in [0, 1440]$ and $c = 2$.

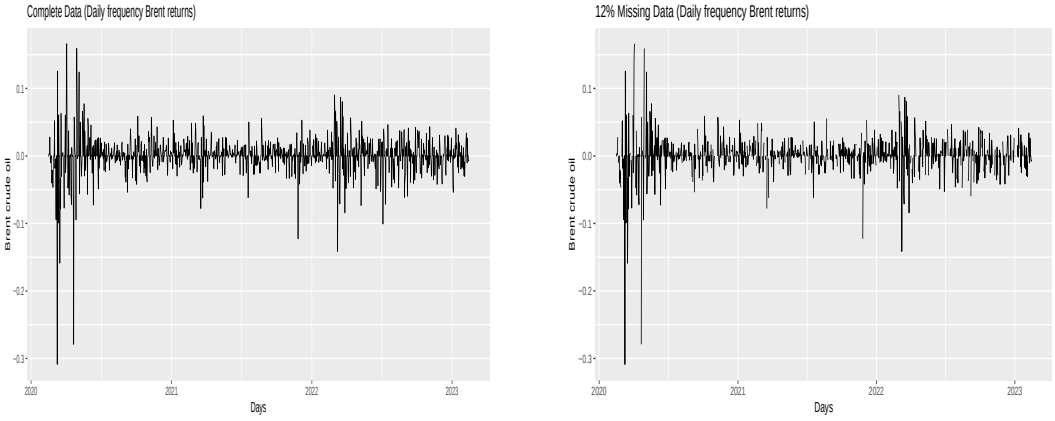
Figure 6.6(b) displays daily Brent Oil return where $12\%$ of the observations are




(a) (b)

Figure 6.5. (a) Sample of three intraday (1-minute frequency) Natural Gas return curves. (b) The stochastic process of daily Brent Oil return and the dots represent the corresponding three preselected days.

missing at random.




(a) (b)

Figure 6.6. (a) Daily Brent crude oil returns for complete data. (b) Daily Brent crude oil returns at %12 MAR.

*Daily Brent oil return volatility estimation and forecasting*

Our purpose is to estimate and forecast the daily volatility of the Brent Oil price log-return using, as a predictor, the intraday (one-minute) frequency log-return of natural

gas. The original sample is split into training and testing subsamples. The training sample is selected from February 14, 2020 to June 30, 2022. However, the remaining period from July 1, 2022 to February 13, 2023 will be for evaluation of forecast of the daily volatility of the Brent Oil price log-return. For the tuning parameters, we considered here the quadratic kernel, the bandwith is chosen using cross-validation approach and the PCA semi-metric is considered to estimate the volatility.

Note that the term "Volatility" refers to a latent variable, which cannot be observed directly but is approximated from other variables that can be observed. Therefore, the concept of the realized volatility was first introduced by [44], who defined it as a non-parametric estimator that is independent of the parameter distribution. In this context, to evaluate the estimation and forecast of the volatility, one considers the so-called *realized volatility* computed based on the 1-hour frequency of Brent oil over same period. Thus, the realized volatility is considered as the true value of volatility that we can refer to in order to assess the performance of our estimators. Given 1-hour frequency one calculates the realized volatility on a specific day $d$ as follows:

$$RV_d = \sum_{h=1}^{24} r_{d,h}^2,$$

where $r_{d,h}$ is the value of the log return of the Brent observed at hour $h$ on the day $d$.

Figure 6.7 shows the In-Sample and Out-Of-Sample (the green-shaded area) sets for the realized and estimated volatility of the Brent returns. Figure 6.7(a) shows that the most volatile period for the Brent is the 2020 year. This is due to an important event named "the 2020 Russia-Saudi Arabia conflict, COVID-19", which is a combination of two factors that led to a significant drop in oil prices, including a supply-demand oil conflict between Russia and Saudi Arabia, and another supply-demand disruption from

the COVID-19 pandemic, which has clearly impacted oil supply-demand because of the lockdowns around the world. The confinements and shutdowns of economic activity lowered demand, resulting in the collapse of oil prices. Besides, Vienna, a major oil producer, were unable to reach an agreement to reduce oil production in response to the COVID-19 pandemic, and immediately after that, Saudi Arabia and Russia began a pricing war that significantly lowered the price of oil. However, the period of lowest volatility is between 2021 and 2022, displayed in Figure 6.7(a). Because in January 2021, oil prices started to rise due to demand outside Europe and reductions in OPEC countries. It can also be noted that the volatility increased steeply during the starting year 2022. Reaching a new peak after 2020, indicating another major high-risk event the war between Russia and Ukraine and its impact on the global economy after the COVID-19 epidemic. However, the Out-Of-Sample set (green-shaded area) shows that the forecast of the realized volatility of the Brent return is getting stable, indicating that the volatility levels are low starting from July 1, 2022 to February 13, 2023. Figure 6.7(b) seems to indicate that the estimated volatility of the Brent return fits well with the true value of volatility. Figure 6.7(c), (d), (e), and (f) present the estimated volatility of the Brent return at a $12\%$ MAR rate. It is clear that the four graphs retain the same pattern of the true value of volatility, even with incomplete observations or after imputation. As a criterion for measuring estimators accuracy in estimating and forecasting in In-Sample and Out-Of-Sample, respectively, we consider the absolute error, defined as

$$AE_d := |\mathcal{U}_d^2(X_d) - RV_d|,$$

where $\mathcal{U}_d(X_d)$ denotes the volatility estimation/forecast obtained either with complete data, missing data or imputed data.

Table 6.2 provides summarized statistics of absolute error for each estimator at a $12\%$ MAR rate. The results for the estimators in In-Sample subset based on median absolute error shows that the nonparametric imputed and inverse probability weighted imputed estimators are more effective than simplified and inverse probability weight estimators, respectively. As a result, the inverse probability weighted imputed method is the best choice for estimating the volatility of the daily Brent oil return when there are $12\%$ of the observations are missing at random. Regarding the estimators in Out-Of-Sample subset based on lower quantile absolute error shows that the nonparametric imputed and inverse probability weighted imputed estimators perform better than simplified and inverse probability weighted estimators, respectively. Therefore, the best method for forecasting the volatility of the daily Brent oil return when there are $12\%$ of the observations are missing at random is the nonparametric imputed approach. However, the complete case is the only one that yields efficient estimators among others of the estimated and forecast volatility of the daily Brent oil return in both In-Sample and Out-Of-Sample subsets, according to the absolute error measure based on lower quantile.

Table 6.2. Summary Statistics of the AE obtained for each estimator when MAR=$\%12$.

| Estimators | In-Sample (IS)$\times 10^{-4}$ | | | | Out-of-Sample (OoS)$\times 10^{-4}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | Mean | $Q_{25\%}$ | $Q_{50\%}$ | $Q_{75\%}$ | Mean |
| Complete | 0.0496 | 0.4954 | 7.3966 | 7.5302 | 1.41353 | 6.12136 | 7.75438 | 6.29141 |
| Simplified | 0.0401 | 0.8101 | 7.9516 | 7.5359 | 1.59297 | 6.24067 | 7.87776 | 6.36422 |
| NP Imp. | 0.0416 | 0.7807 | 7.6034 | 7.5697 | 1.27913 | 6.13609 | 7.68484 | 6.30530 |
| IPW | 0.0388 | 0.8554 | 8.1461 | 7.5856 | 1.5929 | 6.2215 | 8.0110 | 6.4060 |
| IPW Imp. | 0.0479 | 0.6400 | 7.4681 | 7.4882 | 1.43097 | 6.16744 | 8.30704 | 6.50837 |

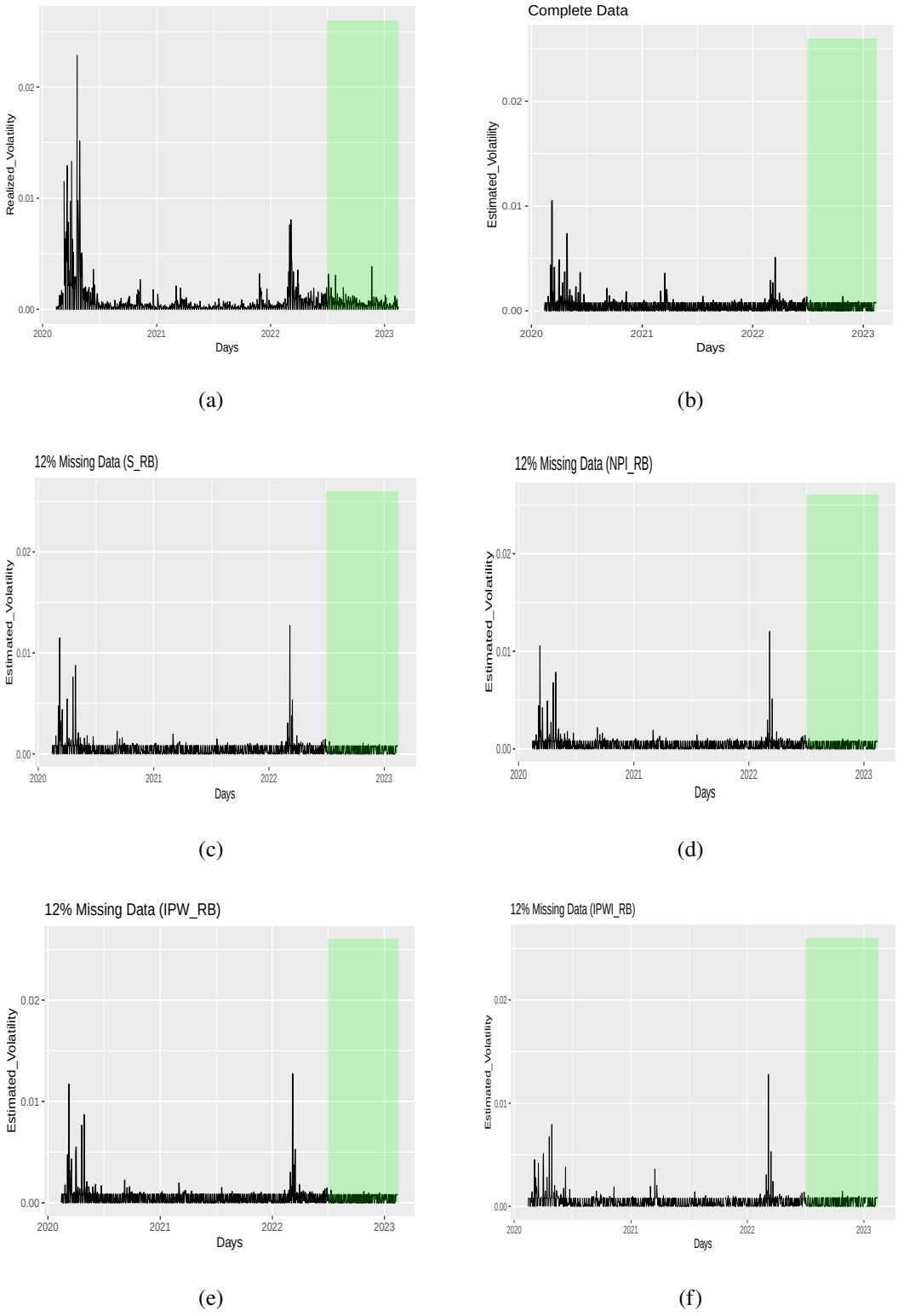



Figure 6.7. (a) Realized Intraday Volatility (hourly frequency) Brent log-returns. (b) Estimated Intraday Volatility (daily frequency) Brent for complete data. Estimated Intraday Volatility (daily frequency) Brent at $12\%$ MAR for Simplified (c), Nonparametric Imputed (d), Inverse Probability Weight (e) and Inverse Probability Weight Imputed (f) estimators.

# CHAPTER 7: CONCLUSION AND PERSPECTIVES

This thesis deals with the nonparametric estimation of the regression and volatility functions in a nonlinear heteroscedastic functional regression model. A Nadaraya-Watson type estimator is used when the response variable is a real-valued random variable and subject to missing at random mechanism, while the predictor is functional in nature taking value in an infinite dimensional space endowed with a certain semi-metric and completely observed. We start by introducing the simplified and inverse probability weighted estimators based on the complete observed data. Then, these initial estimators were used to impute missing values and to define the estimators of the regression and volatility operators after data imputation. A simulation study was performed to assess the performance of the proposed estimators and revealed that imputation techniques improve the quality of estimation. An application to high-frequency financial data is also considered and the estimation and the forecast of volatility of 1-day frequency Brent returns given the 1-minute frequency natural gas intraday returns was investigated. Results show that the initial estimators provide a reasonably good results. Moreover, after imputation of missing data, the quality of estimation/forecasts was improved significantly.

Note that several predictors could be involved in modeling the high-frequency Brent returns such as historical data of the volatility, the exchange rate, geopolitical events indicators, macroeconomic factors (inflation, interest rate, economic growth) and weather conditions. The integration of such predictor requires a more sophisticated models. That may include general additive models and partially linear models among many others. One can also consider the case that the predictor itself is subject to some missing mechanism. These idea might be implemented in future research projects.

REFERENCES


[1] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, pp. 77–91, 1952.

[2] J. Danielsson, *Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk, with Implementation in R and Matlab*, ser. Wiley Finance Series. John Wiley & Sons, Ltd, United Kingdom, 2011.

[3] J. Fan and Q. Yao, "Nonlinear time series: Nonparametric and parametric methods," *Springer-Verlag*, 2003.

[4] R. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, pp. 987–1007, 1982.

[5] M. Borkovec, "Asymptotic behavior of the sample autocovariance and auto correlation function of the ar(1) process with arch(1) errors," *Bernoulli*, vol. 6, pp. 847–872, 2001.

[6] S. Ling, "Estimation and testing of stationarity for double autoregressive models," *Journal of the Royal Statistical Society: Series B*, vol. 66, pp. 63–78, 2004.

[7] N. H. Chan and L. Peng, "Weighted least absolute deviations estimation for an ar(1) process with arch(1) errors," *Biometrika*, vol. 92, pp. 477–484, 2005.

[8] C. Francq and J.-M. Zakoian, *GARCH models. Structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

[9] J. C. Escanciano, "Asymptotic distribution-free diagnostic tests for heteroskedastic time series models," *Econometric Theory*, vol. 26, no. 3, pp. 744–773, 2010.

[10] K. Ghoudi, N. Laïb, and M. Chaouch, "Joint parametric specification checking of conditional mean and volatility in time series models with martingale difference innovations," *Preprint*, 2022.

[11] N. Laïb, "Kernel estimates of the mean and the volatility functions in a nonlinear autoregressive model with ARCH errors," *J. Statist. Plann. Inference*, vol. 134, pp. 116–139, 2005.

[12] G. Collomb, "Nonparametric regression: An up-to-date bibliography," *Statistics*, vol. 16, no. 2, pp. 309–324, 1985.

[13] G. Collomb and W. Härdle, "Strong convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation for dependent observations," *Stochastic Process. Appl.*, vol. 23, pp. 77–89, 1986.

[14] G. G. Roussas, "Nonparametric regression estimation under mixing conditions," *Stochastic Process. Appl.*, vol. 36, no. 1, pp. 107–116, 1990.

[15] J. Fan and Q. Yao, "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, vol. 85, no. 3, pp. 645–660, 1998.

[16] L.-H. Chen, M.-Y. Cheng, and L. Peng, "Conditional variance estimation in heteroskedastic regression models," *J. Statist. Plann. Inference*, vol. 139, pp. 236–245, 2009.

[17] Y. Aït-Sahalia, "Testing continuous-time models of the spot interest rate," *The review of financial studies*, vol. 9, no. 2, pp. 385–426, 1996.

[18] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *J. Polit. Econ.*, vol. 81, no. 3, pp. 637–654, 1973.

[19] K. C. Chan, G. A. Karolyi, F. A. Longstaff, and A. B. Sanders, "An empirical comparison of alternative models of the short-term interest rate," *The journal of finance*, vol. 47, pp. 1209–1227, 1992.

[20] O. Vasicek, "An equilibrium characterization of the term structure [reprint of J. Financ. Econ. **5** (1977), no. 2, 177–188]," in *Financial risk measurement and management*, ser. Internat. Lib. Crit. Writ. Econ. Vol. 267, Edward Elgar, Cheltenham, 2012, pp. 724–735.

[21] J. Fan, J. Jiang, C. Zhang, and Z. Zhou, "Time-dependent diffusion models for term structure dynamics," in 4, vol. 13, Statistical applications in financial econometrics, 2003, pp. 965–992.

[22] M. Chaouch, "Volatility estimation in a nonlinear heteroscedastic functional regression model with martingale difference errors," *J. Multivariate Anal.*, vol. 170, pp. 129–148, 2019.

[23] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, ser. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1987, pp. xvi+278.

[24] A. Pérez-González, J. Vilar-Fernández, and W. González-Manteiga, "Nonparametric variance function estimation with missing data," *Journal of Multivariate Analysis*, vol. 101, pp. 1123–1142, 2010.

[25] L. Wang, R. Cao, J. Du, and Z. Zhang, "A nonparametric inverse probability weighted estimation for functional data with missing response data at random," *J. Korean Statist. Soc.*, vol. 48, no. 4, pp. 537–546, 2019.

[26] N. Ling, L. Liang, and P. Vieu, "Nonparametric regression estimation for functional stationary ergodic data with missing at random," *J. Statist. Plann. Inference*, vol. 162, pp. 75–87, 2015.

[27] P. M. Robinson, "On the consistency and finite-sample properties of nonparametric kernel time series regression, autoregression and density estimators," *Ann. Inst. Statist. Math.*, vol. 38, no. 3, pp. 539–549, 1986.

[28] R. S. Singh and A. Ullah, "Nonparametric time-series estimation of joint dgp, conditional dgp, and vector autoregression," *Econometric Theory*, vol. 1, no. 1, pp. 27–52, 1985.

[29] H. Bierens, "Uniform consistency of kernel estimators of a regression function under generalized conditions," *Journal of the American Statistical Association*, vol. 78, pp. 699–707, 1983.

[30] ——, "Kernel estimators of regression functions," *In Advances in econometrics: Fifth world congress*, vol. 1, pp. 99–144, 1987.

[31] U. Grenander, "Stochastic processes and statistical inferences," *Arkiv för Matematika*, vol. 1, 1:195–277, 1950.

[32] C. R. Rao, "Some statistical methods for comparison of growth curves," *Biometrics*, vol. 14, no. 1, 1:14–17, 1958.

[33] J. O. Ramsay and B. W. Silverman, *Functional data analysis*, Second, ser. Springer Series in Statistics. Springer, New York, 2005, pp. xx+426.

[34] F. Ferraty and P. Vieu, *Nonparametric functional data analysis*, ser. Springer Series in Statistics. Springer, New York, 2006, pp. xx+258, Theory and practice.

[35] F. Ferraty and P. Vieu, "Functional nonparametric statistics: A double infinite dimensional framework," in *Recent advances and trends in nonparametric statistics*, Elsevier B. V., Amsterdam, 2003, pp. 61–76.

[36] E. Masry, "Nonparametric regression estimation for dependent functional data: Asymptotic normality," *Stochastic Processes and their Applications*, vol. 115, pp. 155–177, 2005.

[37] N. Laïb and D. Louani, "Nonparametric kernel regression estimation for functional stationary ergodic data: Asymptotic properties," *J. Multivariate Anal.*, vol. 101, no. 10, pp. 2266–2281, 2010.

[38] F. Ferraty, A. Mas, and P. Vieu, "Nonparametric regression on functional data: Inference and practical aspects," *Aust. N. Z. J. Stat.*, vol. 49, no. 3, pp. 267–286, 2007.

[39] W. D. Flanders and S. Greenland, "Analytic methods for two-stage case-control studies and other stratified designs," *Statistics in Medicine*, vol. 10, no. 5, pp. 739–747, 1991.

[40] L. Zhao and S. Lipsitz, "Designs and analysis of two-stage studies," *Statistics in Medicine*, vol. 11, no. 6, pp. 769–782, 1992.

[41] M.-L. Liu, Q. Ji, and Y. fan, "How does oil market uncertainty interact with other markets? an empirical analysis of implied volatility index," *55*, vol. Energy, pp. 860–868, 2013.

[42] Y. Chen, F. Qu, W. Li, and M. Chen, "Volatility spillover and dynamic correlation between the carbon market and energy markets," *Journal of Business Economics and management*, vol. 20, no. 5, pp. 979–999, 2019.

[43] R. Aloui, S. Hammoudeh, and K. Nguyen, “A time-varying copula approach to oil and stock market dependence: The case of transition economies,” *Energy Economics*, vol. 39, pp. 208–221, 2013.

[44] R. Merton, “On estimating the expected return on the market: An exploratory investigation,” *Journal of Financial Economics*, vol. 8, no. 4, pp. 323–361, 1980.